HAL
open science

HiPEAC Vision 2023

Marc Duranton, Koen De Bosschere, Bart Coppens, Christian Gamrat, Madeleine Gray, Thomas Hoberg, Harm Munk, Charles Robinson, Tullio Vardanega, Olivier Zendra

To cite this version:

Marc Duranton, Koen De Bosschere, Bart Coppens, Christian Gamrat, Madeleine Gray, et al.. HiPEAC Vision 2023. , pp.1-238, 2023, 9789078427032. hal-04023794

HAL Id: hal-04023794

https://inria.hal.science/hal-04023794v1

Submitted on 10 Mar 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Copyright

HiPEAC Vision 2023

HIGH PERFORMANCE EMBEDDED ARCHITECTURE AND COMPILATION

HiPEAC

Editorial board:
Marc Duranton, Koen De Bosschere, Bart Coppens,
Christian Gamrat, Madeleine Gray, Thomas Hoberg,
Harm Munk, Charles Robinson, Tullio Vardanega, Olivier Zendra.

This document was produced as a deliverable of the H2020 HiPEAC CSA under grant agreement 871174 - HiPEAC. The editorial board is indebted to Dr Max Lemke of the Directorate-General for Communication Networks Content and Technology of the European Commission for his active support of this work.

Disclaimer: The views expressed in the individual articles are those of the authors. They do not necessarily represent the views or position of the HiPEAC project, HiPEAC members, or of the European Commission.

Design: www.magelaan.be / Cover design: Roger Castro, Monzón Studio / Cartoons: Arnout Fierens, Arnulf.be

© 2023 HiPEAC

ISBN 9789078427032

Foreword

The focus of the HiPEAC Vision 2023 is that we are in races, both against time and with the rest of the world.

Technology never stands still. The last few years have once again seen rapid, profound changes across the world, both from the technological point of view – with impressive advances in artificial intelligence – and from the geopolitical point of view, where technology is increasingly seen as a strategic asset.

Different world regions are competing for leadership in several areas. Competition between the United States (US) and China in the technology and artificial intelligence (AI) domains is particularly fierce, and it is becoming more intense. This creates a threat to Europe, but at the same time an opportunity. The recent change of ownership and leadership at Twitter is also a wake-up call for Europe. Many of the essential services the European society depends on run on platforms that are not controlled by Europe. This creates vulnerability in the event of conflict, comparable to European dependency on Russian energy. These are just the evolutions of the last year.

Change is taking place so fast that it is also having an impact on the HiPEAC Vision: updating it every two years is no longer sufficient to keep up with the speed of the evolution of computing systems. Therefore, from now on, there will be a HiPEAC Vision every year. The speed of the evolution has also inspired the editorial board to present the challenges of our community as six leadership races: for the "next web", for AI, for innovative hardware solutions, for cybersecurity, for digital sovereignty, and for sustainability solutions.

Structurally, the HiPEAC Vision 2023 has two parts:

- A set of recommendations for the HiPEAC community at large.
- A set of articles written by experts and grouped into six chapters each describing one "global leadership race".

We encourage you to read the complete document, but to clearly understand the recommendations, the reader should read at least the following section "The drivers of the HiPEAC Vision 2023" and the "Key recommendations of the HiPEAC Vision 2023".

The HiPEAC Vision will be available on the HiPEAC website, with easy access to each article, and to the consolidated version. The consolidated version of the Vision will also be printed and distributed, as with previous editions.

Contents

INTRODUCTION 6

RECOMMENDATIONS 7

THE RACE FOR THE "NEXT WEB" 13

The "next web": Weaving future IoT and CPS into an AI-augmented, spatially aware, digital-and-physical space
Tullio Vardanega 14

"Digels", digital genius loci engines to guide and protect users in the "next web"
Tullio Vardanega and Marc Duranton 18

The spatial web: Interconnecting people, places, things and AI for a smarter world
Philippe Sayegh and Capm Petersen 22

Understanding cyber-physical systems among many communities: Large-scale safety-critical systems
Charles Robinson 27

Bridging the stakeholder communities that produce cyber-physical systems
Charles Robinson et al. 32

Towards a living dimension: The future of cyber-physical systems
Francesca Palumbo, Raquel Lazcano and Daniel Madroñal 44

Gaming, content and the metaverse
Thomas Hoberg 54

EU/US white paper on the continuum of computing
Tullio Vardanega, Marc Duranton, Jason O. Hallstrom and Koen De Bosschere 63

THE RACE FOR AI 67

The AI race
Marc Duranton 68

Gigantic transformers and megamodels: The next El Dorado for AI? Or Is "the bigger, the better" the path to AGI?
Marc Duranton 70

AI for better integration between the physical and the cyber world: Embedded AI
Marc Duranton 79

AI as helpers in the software and hardware domains
Marc Duranton, Harm Munk and Tullio Vardanega 90

CONTENTS

THE RACE FOR NEW HARDWARE 97

The race for innovative and new hardware for computing
Christian Gamrat and Harm Munk 98

New and emerging technologies and paradigms
Christian Gamrat and Harm Munk 100

The race towards operational quantum computers
Christian Gamrat and Koen Bertels 110

The programmability wall
Thomas Hoberg 120

THE RACE FOR CYBERSECURITY 127

The race for cybersecurity
Olivier Zendra and Bart Coppens 128

From cybercrime to cyberwarfare, nobody can overlook cybersecurity any more
Olivier Zendra and Bart Coppens 130

Is privacy possible in a digital world?
Bart Coppens and Olivier Zendra 145

THE RACE FOR SOVEREIGNTY 153

The race for sovereignty
Koen De Bosschere 154

The position of Europe in the world
Koen De Bosschere 156

Open source is the enabler for innovation and collaboration
Gaël Blondelle, with Philippe Krief 175

Europe's needs for digital essentials and individual sovereignty
Thomas Hoberg 180

Rethinking education
Koen De Bosschere and Tullio Vardanega 188

THE RACE FOR SUSTAINABILITY 193

The race for sustainability
Koen De Bosschere 194

Climate change is here to stay, so we'd better prepare for it
Koen De Bosschere 196

Why sustainability is crucial for Europe's future
Patrick Blouet 201

Towards circular ICT: From materials to components
Thomas Ernst and Jean-Pierre Raskin 209

Towards sustainable computer architecture: A holistic approach
Lieven Eeckhout 216

Everything as a service
Koen De Bosschere and Marc Duranton 225

GLOSSARY 231

ACKNOWLEDGEMENTS 234

The drivers of the HiPEAC Vision 2023

Every time we start work on the HiPEAC Vision, we plan to make a few simple updates. However, we quickly realize that so many things have changed since the previous edition that we have to make substantial changes[1]. Moreover, change is speeding up, and not necessarily in the right direction: global warming and climate change are more and more evident, the Russo-Ukrainian War, the energy crisis, the problem of semiconductor supply, countries trying to regain their sovereignty (in production, energy, etc…). The previous HiPEAC Vision was written at the height of the COVID-19 crisis; now, it seems that we have even more crises.

In the technical domain, it seems that the speed of change is also increasing: we see new evolutions (and breakthroughs) in artificial intelligence (AI) nearly every week; quantum computing is attracting significant amounts of funding in a lot of countries; while the metaverse is returning (and also progressed rapidly on the "hype curve"). Nearly every day, we hear about new cyberattacks, cybercrimes and cyberwar. Sustainability is becoming a serious consideration for many companies, and optimizing processes to use less raw material and energy is now a major objective.

Information and communication technologies [1] (ICT) – and the hardware and software technologies covered by the HiPEAC community – are involved in nearly all these domains, either directly or indirectly. We have therefore chosen to present the HiPEAC Vision 2023 under the main topic of "leadership races".

Races are also associated with the notion of urgency: we need to act fast. We are in a race against time because of external constraints, such as climate change, that are driving research on sustainability, but also because the technology is evolving very rapidly with a potentially large impact. One telling example is that of artificial intelligence. What the large language models (LLMs) can do now was unthinkable a few years ago. This technology is advancing so rapidly that the Vision article on this topic, which was written in early/mid 2022, had to be updated because it was outdated by end of 2022 (and will probably be obsolete by the time you read it). This evolution is faster than humans can adapt: ethics, laws, usage, education cannot keep up with such a fast pace.

There is also an ongoing leadership race with other regions of the world, and not only in terms of economic aspects. For example, there is fierce competition in LLMs between the big companies from US and China, in which Europe is hardly present[2]. The same holds for the high-end semiconductor market and chip foundries. The race drove US, China, Japan, India to (re)develop foundry capacity and to enact semiconductor legislation to ensure sovereignty. Europe is also aware of this race; it has developed its own Chips Act to reduce its dependences on other parts of the world. There is also a race to define the successor of the current web 2.0: will it become a market place with blockchain and non-fungible tokens (NFTs), or will it be totally virtual like the "Metaverse" proposed by Meta? We have a slightly different view, more in line with the European ethics. (Cyber)security is more than ever an arms race between hackers and security professionals. The number of cyber threats continues to grow, and large organizations are particularly vulnerable because they have more critical operations, more data to protect, and more money to pay a ransom.

The six races

We have identified six critical global leadership races in the HiPEAC domain.

- **The race for the "next web"** – the *continuum of computing*. We observe that more and more data are generated by machines that are interconnected to provide new services. The "next web" will have to take this into account and smoothly integrate the "web of machines" with the "web of humans".
  As this web will be increasingly linked with the physical world, constraints like latency, time, processing time and localization will need to be promoted like other non-functional requirements (such as cost, energy, etc.). These interactions between complex systems involving humans, machines and the physical world form a new dimension of cyber-physical systems (CPS). Some "next web" applications will encompass CPS, also known as large-scale safety-critical systems, and should offer a high level of confidence in terms of safety, which should be taken into account from the outset.
  Another important element is a swarm of "digital twins" that allows modelling both of parts of the real world or of new artefacts. Digitals twins allow the manipulation of variables that are not easily accessible in the real world (e.g. places where

1 However, the 18 recommendations of the HiPEAC Vision 2021 are not obsolete; most of them are still present in a slightly different form in the recommendations of this HiPEAC Vision.

2 Or when a European company is present, it is often acquired by a non-European one.

it is impossible to place a sensor) or that cannot be handled (like time: in a digital twin, a simulation, time can be accelerated, stopped, or go backwards, none of which is possible in reality). Digital twins allow prediction of what could happen under certain hypotheses, as well as risk assessment, and therefore enable optimal decisions. Digital twins are also organized as providing services and they can be used to improve the efficiency of processes.

The key mechanism is to consider "everything as a service" (XaaS). Applications for users are composed by coordinating different services together, whether locally or in a distributed manner. Users are not only people, but also companies and organizations. Game engines can be offering a service for realistic user interfaces if we want to have a more photorealistic view of processes/applications, for example.

Of course, the management of all these services, taking into account all the non-functional requirements to compose them according to the user's wishes, is a very complex task. This will require smart orchestrators (e.g. using artificial intelligence techniques). HiPEAC's vision of such smart orchestrators is represented by digital *genius loci*, which would be at the service of the user.

- **The race for artificial intelligence** (at the edge, with "*megamodels*", or used as a "helper"). The "next web" will require data to be processed intelligently and near to where it was created for efficiency reasons, therefore it will be of paramount importance to have efficient AI systems at the edge or on premises. We are moving towards an era of collaborating AIs rather than "centralized" AI running on (private) clouds. However, LLMs have shown very surprising results, and yesterday's large neural networks are now running on our smartphones (e.g. for voice recognition, translation, etc.). Therefore, adapting the technology of these LLMs to embedded systems will be of high importance, but this also requires that LLMs are mastered in Europe. Last, but not least, AI techniques can be used to enhance the efficiency of software engineers and of system designers. Again, in this area some non-European companies are already showing encouraging results.

- **The race for innovative and new hardware** (including quantum). The widespread use of AI and the "next web" will only be possible if the ICT systems on which they are executed are both efficient and affordable. So there is always a race to be more efficient in energy, both for large systems (high-performance computing (HPC), cloud) due to the cost of ownership and for embedded systems for autonomy (longer battery life). In parallel, work is also now beginning to make these systems more sustainable during their whole lifetime, from mining to recycling.

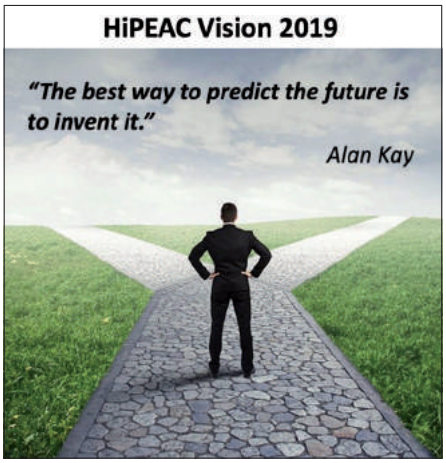
- **The race for cybersecurity**. All the systems mentioned above will only be practically usable if they are secure and safe to use. The arms race between hackers and security professionals is as old as the computer[3], but given the fact that modern societies depend almost entirely on digital technology, the stakes are now much higher. Some countries are creating cyber armies to protect their infrastructure and to set up attacks against other countries to be able to retaliate if attacked, leading to a cyberwar. Here too, we can expect a kind of future arms race between AI tools trying to attack and defend, in addition to human-based attacks.

- **The race for (technological / products / contents) sovereignty**. It is becoming increasingly obvious that the globalization of previous decades is falling apart due to protectionism, unstable supply chains, and even war. COVID-19, the Ukrainian war, climate change and the energy crisis have led to unreliable supply chains, high production and transportation costs, all of which has encouraged countries to try to become more self-sufficient. This is especially true for the semiconductor supply chain which basically spans the whole world, with each country owning part of the process.

- **The race for sustainability.** Each year humanity consumes 1.75 times the resources Planet Earth can generate in one year [2] . Sustainability means that we should reduce our resource consumption to at most one planet. If we fail to do so, we will destroy our own living environment. The production of ICT has a large ecological footprint, and should be reduced a far as possible. At the same time, ICT could also help to make the world more sustainable by optimizing processes with a large ecological footprint. Sustainability implies that we will have to find a trade-off between the environmental costs and the benefits of ICT.

These six races are the result of our analysis of the state of the art and current trends. Here we have outlined a diagnosis, but not solutions. In the next section, for each of the races, we propose recommendations linked to the HiPEAC-related domains (hardware, system and chip design, software, middleware, operating systems, methodologies and tools...). It is urgent for Europe to consider those recommendations and to enter the races.

In the HiPEAC Vision 2019, we still had several options:



For the HiPEAC Vision 2023, we no longer have choices; we just have to enter the races:

3 John von Neumann introduced self-replicating computer programs (e.g. virus) as early as 1949.

Overarching issues

Connecting these six races are overarching issues, as set out below.

Silos

We observe a tendency to "closing in" on all levels, from countries (with more emphasis on sovereignty), to the personal level, to our own "tribe" (as "defined" by social media). Tension is becoming exacerbated at all levels between these "tribes", as evidenced by trade (or real) wars between countries, more extreme political parties, social media "wars", etc.

This tendency also exists in technology, where there are application silos and technology silos:

- **Application silos**. Often, application domains develop their own stack dedicated to their needs, with little or no interoperability with similar stacks developed by other application domains. We see a multiplicity of solutions, which are more or less interoperable, designed to solve similar problems.
  Interoperability or solutions that straddle application domains would increase efficiency and create new synergies and new solutions. For example, synergies between electric-vehicle, home and smart-grid stacks would allow vehicle batteries to be used to power homes during peak hours, therefore reducing the pressure on the electric grid. This trivial example (already operational in some locations) shows the added value of communication and interoperability between these three domains.
  Some open-source solutions are also available to offer interoperability between various standards [3], even for the same application domain. It is therefore one of our recommendations to have exchanges between applications domains and develop interoperability solutions.
- **Technology silos**. It is a natural tendency to group technologies into structures that are easy to understand and to manage. Each structure can be more or less independent and autonomous and only communicate with its adjacent structures in the global flow. Another similar way to manage complexity is to have various levels of abstraction (as in software or communication layers). Research and development activities have a similar structure. Generally, each structure working on a separate topic is making quite good progress and the field has already been optimized.

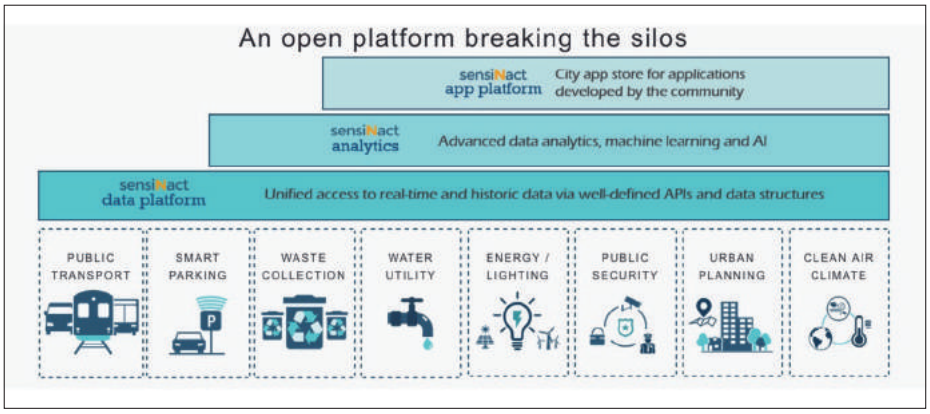


Figure 1: Example of interoperability between applications domains (from [3])

However, the sum of local optimizations is less efficient than global optimization. Taking this as our starting point, we propose breaking the silos that exist for:

- applications;
- algorithms that are executed in the application;
- the programming language in which the algorithms are implemented;
- the software stack (compiler, operating system, middleware, libraries) that result from the mapping of the algorithm into a language that is compiled or executed;
- data coding (floating point, integers, tensors, "spikes", etc.);
- the architecture executing the code;
- and the technology in which the architecture is realized.

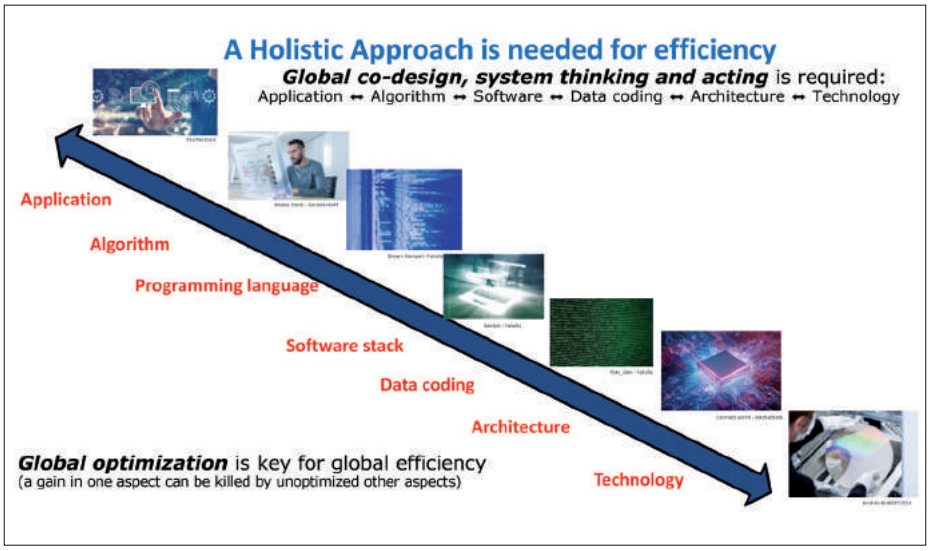
Of course, there are many other steps; these are just a sample selected for simplicity of explanation. Proposals for breaking the silos are listed in the recommendation section which follows.

References

[1] "Information Communication Technologies (ICT)," [Online]. Available: https://competition-policy.ec.europa.eu/sectors/ict_en. [Accessed 21 December 2022].

[2] "Earth Overshoot Day," [Online]. Available: https://www.overshootday.org/. [Accessed 21 December 2022].

[3] "Eclipse sensiNact," Kentyou, [Online]. Available: https://kentyou.com/eclipse-sensinact/



Key recommendations of the HiPEAC Vision 2023

From the observations of the previous section, we derive recommendations in the field of the HiPEAC community at large. More detailed recommendations are expressed in the introductory part of each section, and each article also has a recommendation section, which is even more fine grained, together with "key insights" that will help readers understand the rationale of the recommendations.

Of course, the purpose of each "article" in the HiPEAC Vision is to help readers understand the current situation, the forecast and the reason these set of recommendations were emitted. Therefore, it is highly recommended that readers should read both the introduction and the articles related to a specific topic to better understand the rationale for the recommendations.

Below, we outline a set of global recommendations that can be common to several races, before diving into the recommendations related to each race.

Global recommendations

Break silos and develop a holistic view. We have seen in the previous section that it is important now to "break the silos" and to have a more global view. This is common to all the races and can leverage all of the skills we have in Europe in various domains. In fact, this leads to three recommendations that are logically linked together:

1. **Breaking silos**, developing a holistic view, increasing cross-disciplinarity and multi-dimensionality also increases complexity. To help manage this additional complexity, we need to develop solutions that help us to cope with it. Artificial intelligence, algorithmic approaches and other similar developments are tracks to follow.
2. **Artificial intelligence can be a "helper"** to find solutions; advising humans during the design, realization, manufacturing, usage and recycling stages – helping therefore also to break silos. This use of AI as "helper" can be extended to multiple domains, not only in software and hardware development, as promoted in the "AI race", but also in cybersecurity, or to analyse the best way to reduce the use of material or energy, etc.
3. However, an increasing number of processes are "real time" and require fast, trustable analysis of multiple parameters to find the best (or at least "good enough") solution. This leads to a recommendation to develop **"trusted orchestrators"** that should automate complexity management when interaction with humans would take too much time. The "trust" part is essential because they should take decisions that are acceptable for humans. New techniques need to be developed to ensure this trust.

Hence we have the following global recommendations:

Break silos. This can be achieved with different approaches at various levels:

- Collaboration between teams. Broaden the scope of each structure so that practitioners within one structure are more aware of the problems, solutions and mechanisms of another. In companies, this could be achieved by forging a closer link between computer architects and compilation specialists, for example, or algorithm designers and computer architects, etc., to achieve global system thinking, co-design and acting.
- Using project calls to create synergies between domains. Another way would be to have European calls that are common to two (or more) topics (i.e. between units) and to require exchanges and collaborations between projects and organizations.

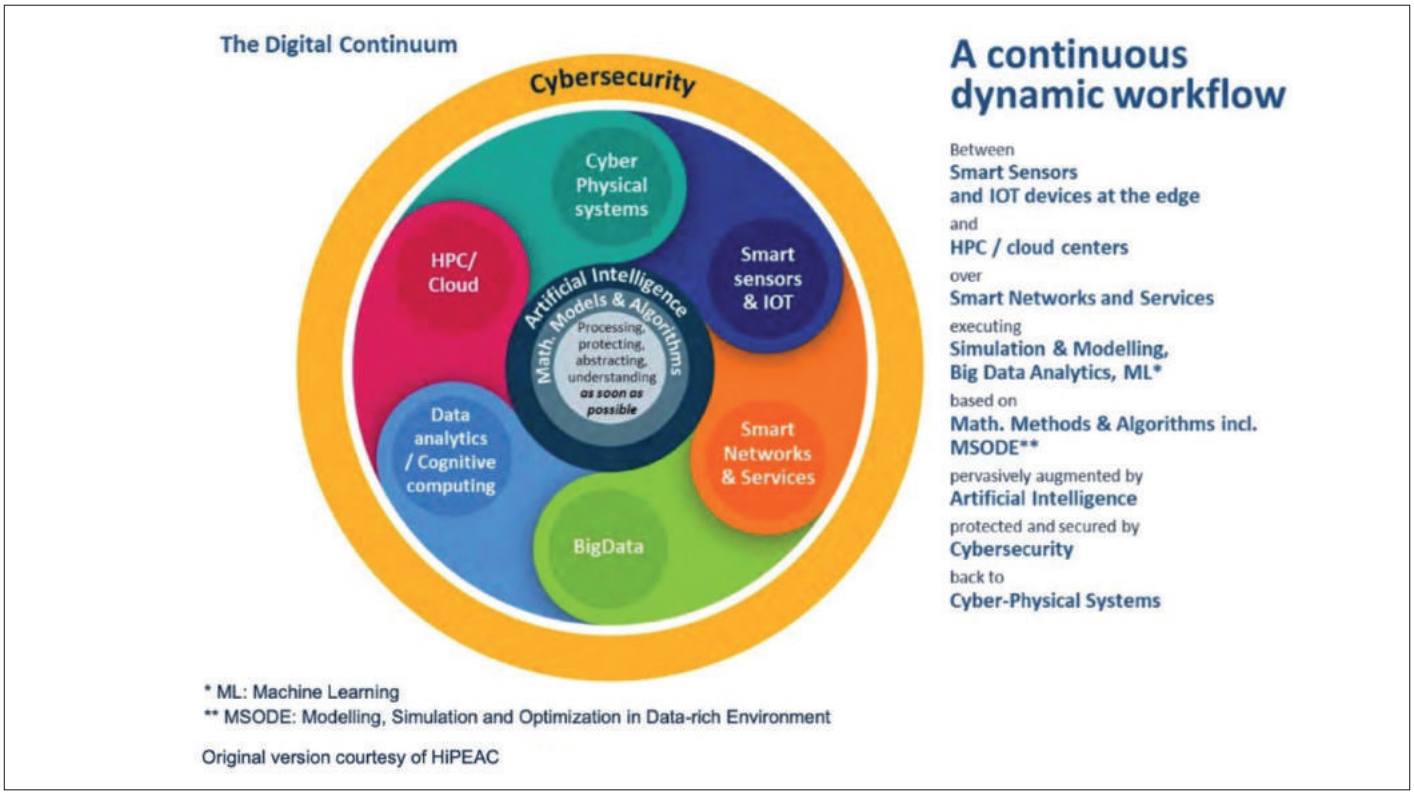


Figure 2: the digital continuum, as seen by the TCI initiative (from [2])

This need for common ground, calls and exchanges has already been identified by several organizations. This was the origin of the creation of the TransContinuum Initiative (TCI) [1], of which HiPEAC, together with ETP4HPC, was a founder. The TCI "*is a horizontal collaboration between 8 European associations and projects involved in IT technology, application and services provisioning for the Digital Continuum…. A paradigm change is needed: we will have to design systems encompassing millions of compute devices distributed over scientific instruments, IoT, supercomputers and Cloud systems through LAN, WLAN and 5G networks. The ambition of TCI is also to become a meeting place of experts representing various disciplines in both science and industry – an asset that Europe can apply in the resolution of its other challenges such as healthcare, climate change or smart cities*". The full charter of TCI can be found in [2].

Bridging the various stakeholder communities, especially on dependability, should enable much faster uptake (including of eco-friendly technologies).

- Establishing well-funded, centralized (or very well-coordinated) cross-disciplinary **European competence centres** (based on the model of CERN for example) that could carry out research, standardization, support platforms/demonstration/testbeds, and help in the exploitation of the results.
  This was also a recommendation in the HiPEAC Vision 2021: *"Europe has many national competence centres with a leading international reputation in computing (imec, CEA, CWI, Fraunhofer, …) but has few international competence centres such as CERN for physics or the ESA for space research. HiPEAC recommends the creation of such a well-funded European competence centre in computing so that Europe is able to retain and attract top talent, to set its own ambitious research agenda, to attract large investments, and to form the core of a network of regional competence centres. Such a network will be crucial for defending Europe's position as a scientific powerhouse. It will also* [be] *the entry point of an innovation pipeline*" and allow cross-domain collaboration.
  Competence centres should promote the exploitation of results: often ideas come from Europe but are commercialized outside of Europe. Here also, there is a need to be fast and pragmatic, for example to create de facto standards rather than to attempt to establish standards. All steps from research to industrial transfer should be closely linked. Such centres should also have an educational mission, allowing people to learn neighbouring or other fields of expertise, and to be up to date with their field (technology moves faster than "classical" education; keeping up with the state-of-the-art requires upskilling, reskilling, and lifelong learning).
- **Making full use of open source**, both for hardware and software developments. This provides an easy way for multiple teams to collaborate, and the results can be reused for several solutions. It has multiple benefits, but awareness of its usage should be developed, together with its business and commercial aspects. More information on this topic may be found in the section titled "The race for sovereignty".

Support the development of tooling for cross-disciplinary and multi-dimensional challenges. Adding levels of abstractions is the natural way to manage complexity, and to make things apparently more understandable and explainable. However, as shown before, this leads to clusters and can result in less optimality at a global level.

Perhaps an alternative way is to try to get a global view and optimize various technologies across the board that need to manage a large number of hypotheses and parameters. This would be extremely challenging for humans alone, but they could get help from tools (e.g. from operational research) or from AI. These tools could prune the hyperspace of parameters and highlight just the combinations that seem promising, so it would only be necessary to test and validate a low number of proposed solutions.

Of course, it is better if the tools could "explain" how they get to the result, but as the result can be validated and compared with other results (e.g. found "by hand"), it is not so important if the results are proven to be better than alternative solutions.

Therefore, it is recommended to **use AI as "helpers"** or "sidekicks" to propose solutions. These AI-proposed solutions should not be applied directly; they should be thoroughly validated and tested to check if they are appropriate, useful and bring added value. AI tools will help tackle incomplete and complex problems, but we should define how much to delegate to AI, and humans should remain in the loop[6].

Develop trustable runtime orchestrators able to manage complex systems, with the help of experts in system orchestration. The previous recommendation is valid if the test/validation can be done, or if there is time to do it, but in some cases the system will have to select a large number of options "in real time", like selecting services according to their functional and non-functional requirements for building applications in the "next-web" context. This will require the development of orchestrators that can be trusted. We see the development of runtime, dynamic product and application configuration, and orchestrating all the different technologies and services in a trustable manner as major challenges to be tackled in the short term.

Recommendations per leadership race

The race for the "next web[7]"

When the world wide web was invented, little more than 30 years ago, it was not obvious to many that the web would "absorb" the internet altogether and become the single surface on which all kinds of digital computation should take place. (Here we understand "computation" as the enabler of digital automation, with all of its transformative power.) The perception back then, and for many years thereafter, was that the web was a place at the side of our personal, social and professional life, a lateral commodity where you would go to follow links to documents and other online resources. What has happened since has shown that digital platforms and services have "enveloped" very many dimensions of modern life, with humans increasingly "within them".

This major leap in evolution is creating the premise of a seamless continuum extending from handheld devices, veritable bridges to the digital world in the personal, social, and professional spheres, to increasingly resource-rich service infrastructures that host pervasive command-and-control applications with feedback loops from digital and physical sensors and actuators, which span from the so-called edge to the centre of the cloud. The convergence between the internet of things (edge devices) and cyber-physical systems (with their non-functional properties) is a central part of this transformation. The "continuum of computing" denotes the fact that the automated computation that allows digital services to assist human needs will be a space where applications will move seamlessly across, towards and along with data, in manners that will depend on a variety of functional and non-functional requirements including latency, privacy, security and energy.

What will emerge from this evolution will be a "next web" where *programmable* links will connect applications, services, digital twins of things and of humans, into a maze of interactions, which will give rise to an immense variety of higher-level services and applications resulting from personalized orchestration of existing resources. This new web will need open interoperable standards, which will allow digital resources to be spatially localized, and to embed collaborative AI engines that can operate also at the edge and not solely in the deepest recesses of corporate clouds.

This "next web" will also pervade industrial infrastructures, enabling them to leverage generations of IoT updating digital-twin applications, augmented with AI, and seamlessly integrated with the command-and-control plane of CPS, leading to what could be called *"the industrial metaverse[8]"*. This evolution will give rise to augmented-reality platforms offering multiple interfaces between humans and things.

There are five recommendations for Europe to be at the forefront in the race for the new web:

- Launch research and development actions to **explore the realization of web-inspired re-interpretation of future computing systems**.
- Initiate **proof-of-concept level efforts** to develop agreed definitions of **functional and non-functional specifications** and accompanying programming support for execution- and communication-level interoperability **across the continuum of computing**. Processing should be done where it is the most efficient according to functional and non-functional criteria.
- Promote **demonstrative testbeds of "trusted orchestrators"**, running on any mobile device and appliance to serve the needs of safe, private, personalized and effective service composition on the computing continuum.
- Provide active support to **standardization** efforts in this field.

6 AI will change everything and is moving faster than humans can follow. But it is still a tool with no consciousness, with no notion of good/bad results and humans should always validate the results and usages.

7 Several denominations are being used in the technical narrative for the "next web" evoked in this vision. The term "web3" is frequently referred to distributed ledgers (blockchains) and NFTs, the metaverse, the semantic web. We use the generic term "next web" to cover our vision of an intertwined physical, virtual and cyber world that will host personalized and trustworthy orchestrations of a wealth of value-added digital services.

8 The *industrial metaverse* aims to improve the real world (with better use of resources, better orchestration, better efficiency to create new things, …). Virtual, augmented and mixed reality will enable one of the multiple potential interfaces with humans.

- Develop a "**European industrial metaverse**", of which the foundations are the combination of the IoT synchronizing in real-time digital twins, all considered as services with functional and non-functional requirements, combined with other services running on the continuum of computing (from deep-edge to cloud and HPC) and dynamically orchestrated by trusted and smart (thanks to AI techniques) orchestration.

The race for artificial intelligence

Today, artificial intelligence (AI) is everywhere. It will drastically change the digital world and will supercharge many application domains, including the development of hardware and software. Natural language processing (NLP) is the current "big thing" driving research and competitions and some of its results are mind-blowing, mainly with gigantic neural networks (known as "megamodels"). However, while some of their results are good, others are meaningless, so all results should be taken with caution and with a critical mindset. It is very important for Europe to be able to develop such megamodels using European databases, languages and culture and following European ethics.

At the other end of the spectrum, deep-learning techniques are getting better and better at analysing data from the real world (audio, video, signals), allowing better interaction between computers and the physical world. These techniques (like voice recognition, analysis of images), which previously had to be executed in the cloud, can now be done on embedded systems ("edge AI"). As algorithms are improving, together with better specialized hardware, we can also forecast that the "megamodels" that are running today on large servers will be able to run on edge devices soon, unlocking a lot of new usages. Europe should be at the forefront of this "squeezing" of complex AI models into edge devices and of the development of supporting hardware for practical use.

There are four recommendations for Europe to be at the forefront in the race for artificial intelligence:

- Europe should focus on developing embedded AI solutions for better integration/understanding of the physical world, e.g. for near-the-edge and edge devices, (like automotive, smart factories, etc), leveraging its knowhow in cyber-physical systems and embedded systems. In general, "**AI at the edge**" is a key element for enforcing privacy and system efficiency, and it should be weaved with cloud intelligence in a manner that provides for rapid feedback cycles to actions and events occurring in the compute continuum.
  As new learning techniques that need fewer data or that do not require a lot of labelled data for learning develop, these algorithms will also be interesting in embedded systems or in on-premises systems. Research in Europe could focus on **new approaches that do not require a lot of data for learning and could be used at the edge**, and on federated learning approaches that allow privacy to be preserved.
- Europe should not drop out of the race for the "megamodels". Sufficient computing resources and databases need to be available to European researchers and industry to develop **European megamodels** (following European precepts and ethics) and experiment with them. Perhaps a "CERN"-like common infrastructure should be promoted in this field.
  Europe should be also at the forefront of techniques to move "large" models into edge devices, with tools (pruning networks, using the right data representation (quantization), or other optimization techniques), efficient algorithms and new edge hardware with high performance capabilities and low energy requirements.
- It is also key to support research and development of tools helping to **identify bias and misbehaviour** manifested by AI. The large models (generally transformer based) show interesting results mixed with nonsense. Solutions to separate the good from the bad should be enforced and learning databases curated. Hybrid techniques (using explicit knowledge to check the results of data driven deep neural networks) are a path to explore.
- As it is necessary to continue improving the efficiency (in terms of both energy and cost) of the hardware, software and algorithms used by AI accelerators, **AI techniques** will be increasingly present as tools **to support the development of efficient software and hardware**. These technologies should be evaluated and should trigger the development of new products.

The race for innovative and new hardware

In the race for the reduction of energy consumption by computing devices, several new computing paradigms are being researched. These paradigms, such as quantum computing, neuromorphic computing, spintronics and photonic devices, to name only the most promising ones, appear to offer faster systems for solving particular problem areas, at a much lower energy consumption than current von Neumann architecture-based semiconductor devices.

Research and development (R&D) efforts in these new technologies are spread over much of the same fields as contemporary integrated circuit (IC) technology, but with a marked difference in emphasis. Of these new paradigms, quantum computing is attracting most of the attention today. Quantum computing requires the development of new devices, as there are several ways to implement devices that exhibit quantum phenomena fit for quantum computing. At the heart of these efforts is materials research. Although it is difficult to predict when, or even if, quantum devices will become an operational technology, the ongoing research in new materials will have a profound impact on the development of new computing technology in the near future. In addition, new technologies such as spintronics and photonics will highly benefit from materials research for quantum computing, and vice versa.

The development of quantum computing has also renewed research in information theory, complexity theory, algorithm development and computer architecture. Both quantum computing and neuromorphic computing have shed new light on information representation and information processing. As the successful von Neumann architecture is here to stay (mainly as orchestrator of various accelerators, potentially using different technologies and information coding), embedding architectures based on new computing paradigms in more traditional architectures is becoming a challenging research field, together with the development of programming models for these architectures. As well as connecting

old to new, research is also needed to determine how to leverage and embed devices based on new materials in computing devices. Future computing devices will be **hybrid systems**, where each device will process its specific part of the computation.

IC fabrication processes require large amounts of energy and fresh water, and due to their toxic waste can put high pressure on the environment. Not all application areas of ICs require state-of-the art devices; in fact, they can be manufactured with relatively slow devices fabricated with less demanding processes, mostly because of high volumes and low device cost. The processes for manufacturing these slow devices can be used as a testbed for innovative processes, application of new materials, and reduction of resources before new process are applied for state-of-the-art devices.

There are five recommendations for Europe to be at the forefront in the race for innovative and new hardware:

- Establish and stimulate **competence centres** that combine the strength of Europe's position in the fundamental fields of computing (information theory, complexity theory, computer architecture, algorithm design, programming paradigms) with European industrial initiatives. This will both attract top talent and develop a strong European industrial presence in the field of new computing devices. In this context, the role of pilot-lines will be key. Europe must place emphasis on development that couples research to innovative practical applications.
- Research in **system architecture and software stack** for emerging computing paradigms needs to be pushed. The potential success of a novel computing scheme is not only about its core technology but rather in how it fits into a flow that allows useful problems to be solved. Existing ICT will still be at the core of using these new developments.
- **Heterogeneous integration** will play a key role in developing and maturing new design methods, fabrication techniques and novel packaging for the new computing hardware. A European ecosystem needs to be established, combining research, research transfer organizations (RTOs), industry and application prescribers.
- Design support and investment programs in the knowledge that the race for new technologies is not only about mastering the technologies (e.g. qubits, or neuromorphic devices). It is also about their integration with **infrastructures and industry environments**, and with more traditional ICT that will help make it real.
- Invest in designing and fabricating novel **IC technologies** that have a significantly lower environmental impact, and at the same time apply the lessons learned from those new fabrication processes to existing ones.

The race for cybersecurity

After decades of digitalization spreading into every area of our lives, with very little attention given to the aspects linked to cybersecurity, information and communication technology (ICT) had essentially become an "open bar" for cybercriminals. For a few years, with a marked degradation during the peak of the COVID-19 pandemic, the news has been rife with reports of privacy breaches and cyberattacks (mainly ransomware) on companies and institutions, especially local governments and hospitals. In addition, cyberwarfare has been making the news too, especially in relation to the conflict in Ukraine.

Cybersecurity is a highly competitive race between nations, and between defenders and attackers, with enormous stakes. The pervasiveness of ICT provides a broad attack surface, and attacks can be economically devastating, but they can also have tangible or even lethal repercussions on the physical world. Despite several highly acclaimed advancements (e.g. the General Data Protection Regulation-GDPR), the EU still has a great deal of work to do in this regard, particularly to maintain its sovereignty and become a leader in the global competition. Cybersecurity is, in fact, a matter of both economic leadership and national sovereignty.

There are four recommendations for the race for cybersecurity:

- **Critical infrastructure and supply chains** (utilities, transports, health, etc.) should have both their **hardware and software components hardened against cyberattacks**, with reinforced cyber defences, monitoring and resilience capabilities (including redundancy). In addition, edge and IoT devices should be hardened against side-channel attacks and reverse engineering.
- The EU should support and invest in **research on methods and tools** to make **security, including privacy, a first-class citizen during the development of new ICT systems** (including e.g. quantitative security metrics/properties), find **vulnerabilities in existing ICT systems** (e.g. with static analyses on source code and behavioural analyses at runtime), and **automatically prevent or mitigate them** (e.g. with automated refactoring tools and blocking systems, possibly involving AI). This would provide the technical bricks to widely improve the cybersecurity of the EU industry.
- The EU should **broaden mandatory security and privacy, EU-based audit and certification** of ICT systems, with several levels depending on the criticality of the application, and regulate to make ICT systems providers and resellers **liable in case of subpar cybersecurity**. This would incite the EU industrial ecosystems to advance quickly toward better cybersecurity, and give it an edge in the global competition. The expanded Network and Information Security Directive (NIS2) is a first step in this direction, but work in this area must be widely enforced and strengthened.
- Quantum computing has the potential to be a disruptive advance that would invalidate past and current cryptographic schemes, leaving most systems wide open and exposed. In order to avoid being in such a situation and be ahead of the tide, the EU should invest in post-quantum cryptography (PQC) so as to have **EU-designed and EU-validated quantum-resistant encryption schemes, hardware and software**.

The race for sovereignty

After decades of globalization, the world has discovered that a globalized economy has its weaknesses too, especially if supply chains pass through regions that face production issues (due to a pandemic, a local conflict, natural disasters, social unrest, energy or water shortage, geopolitical tensions, sanctions, etc.). Local governments are making supply chains more resilient by shortening them, especially for products that are considered of strategic value (like chips, drugs, food, energy, etc.).

Sovereignty is, however, much more than building a factory in Europe: it also requires research, innovation, an entrepreneurial ecosystem, (venture) capital, raw materials, energy, and a well-trained workforce to be successful. For digital sovereignty, we should not only aim for hardware sovereignty, but also for software sovereignty, i.e. a software stack that is owned by Europe and that is not controlled by foreign companies and agencies. Sovereignty does not necessarily mean having everything done locally; it is having a (good) bargaining position.

There are four recommendations for the race for sovereignty:

- Europe should strive for **digital sovereignty, both in hardware and in software**. Hardware sovereignty is covered by the European Chips Act (although creating new designs with innovative architectures should be further promoted), but software sovereignty is as important. Europe should also work on "digital essentials" that can guarantee that Europe's essential software infrastructure keeps working uninterruptedly no matter what happens outside Europe. The goal is not necessarily to strive for full digital independence, but for at least digital interdependence with the other global regions to have bargaining power.
- Europe should **promote open source**, and create an infrastructure to support the open source software and hardware supply chain in Europe. Open source not only helps to collaborate and to innovate inside Europe, but it also helps to build soft power outside Europe. It can also help to create Europe's digital essentials.
- Europe should keep **investing in talent**. Europe's higher education system is excellent, but it should pick up two additional important roles: lifelong learning to upskill and reskill the existing (and shrinking) European work force, and supporting regional entrepreneurial ecosystems to bring more research results to the market. The creation of well-funded European competence centres will help to retain and attract top talent, and to stay at the forefront of new digital technologies.
- Europe should keep **investing in research and innovation**, and in a more entrepreneurial ecosystem that generates lots of start-ups, scale-ups and global companies. Europe needs more venture capital to support the growth of scale-ups into global companies. Key areas to invest in are the silver economy, health, mobility, energy, automation and sustainability.

The race for sustainability

Climate change is now felt all over the planet, leading to changing natural ecosystems and to the loss of biodiversity; already, we are witnessing the impact on various aspects of our daily lives. However, sustainability is more than climate change. It is about organizing the world within the limits of Planet Earth both today and for the generations to come. One aspect of this is the use of raw materials, which are, by definition, finite on a finite planet. Modern information technology (ICT) devices require more than 60 minerals from all over the planet, some of which are hard to mine, are scarce, or are mined in socially unacceptable conditions. Once used in a device, they are difficult to recycle; but without efficient reuse and recycling, there cannot be true sustainability.

Modern ICT devices are very power efficient, leading to a situation where more energy is needed to produce them than they will use during their entire lifetime ("embodied" versus "operational" energy use). Contrary to common belief, continuing to use an old device is often more sustainable than buying a new, more power-efficient one. This shows that a full lifecycle assessment can lead to counterintuitive conclusions, and that such an assessment is essential in order to come up with sustainability solutions with real impact.

On the other hand, sustainability is by far the most important grand challenge of the 21st century, and it won't be possible without extensive use of digital technology to make efficient use of resources.

There are four recommendations for the race for sustainability:

- System designers should use a **full lifecycle assessment** when designing new computing systems: from raw material to device, and 100% back to raw material. Based on such life cycle assessments, Europe should develop sustainability labels for digital devices (e.g. TCO certified). These labels should not only focus on repairability, "refurbishability" and recyclability, but also on socially acceptable mining and production operations.
- Since operational energy is dwarfed by embodied energy for modern low-power devices, computer architects should first and foremost **focus on embodied energy**, e.g. by avoiding dark silicon, by limiting the number of cores to the number actually needed, by avoiding accelerators that are only used sporadically, etc.), and by maximizing the use and the lifetime of a device and its software.
- Europe should develop **new business models** and economic models that also take into account the lifecycle environmental costs in order to make sustainable solutions more competitive with non-sustainable ones.
- Europe should continue **searching for ways to dematerialize goods** with a substantial environmental footprint by environmentally friendlier (digital) goods and services.

References

[1] "The TransContinuum Initiative," ETP4HPC, [Online]. Available: https://www.etp4hpc.eu/transcontinuum-initiative.html. [Accessed 19 December 2022].

[2] Jean-Pierre Panziera et al. , "TransContinuum Initiative (TCI): our vision," 8 December 2021. [Online]. Available: https://www.etp4hpc.eu/pujades/files/TCIvision_08-12-2021%20%20web.pdf. [Accessed 19 December 2022].

HiPEAC Vision 2023

THE RACE FOR THE "NEXT WEB"

The web paradigm, where digital assets are linked to one another into an information-rich maze across which computation moves freely, is a fitting metaphor for the new space that will result from AI-augmented weaving of future IoT and CPS nodes into a seamless continuum with the cloud. The new web that will emerge will be spatially aware, allowing humans to relate to it in a manner more akin to their perception of the physical world and their mental representation of it.

The "next web"

Weaving future IoT and CPS into an AI-augmented, spatially aware, digital-and-physical space

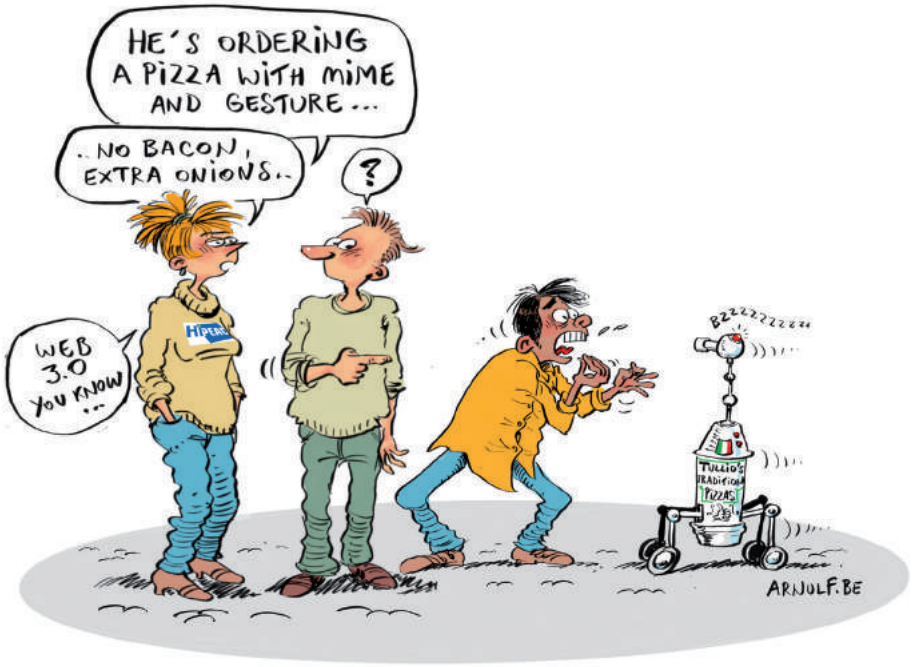
By TULLIO VARDANEGA

The term "internet of things" (IoT) evokes the interconnection of networks of compute devices that reach out to the physical world across geographical expanses, for sensing and actuating physical "things", thereby providing compute-based intelligence to physical environments in the real world. If we take the 'I' of the IoT to mean the internet for real, then it would be natural to extend the analogy even further.

In the internet space, in fact, the web (combined with the cloud, which firmly rests on the web) has long become the most convenient and practical platform for accessing digital resources, as well as for using, deploying, and developing applications. From that perspective, therefore, the modern web may be regarded as a ubiquitous programmable space, available to anyone, anywhere, any time.

This vision argues that future IoT systems may be imagined as the "next web" space, where every "thing" is reachable as a linked digital resource, and applications are programmed, deployed and used in an "as-a-service" mode, located on the device(s) where that is more convenient to the end user. Future IoT systems will weave into a seamless continuum with the cyber-physical systems (CPS) space, where command-and-control capabilities add to sensing and actuation at the border between the physical and the digital world. Artificial intelligence (AI) will permeate the whole digital space in order for it to engage satisfactorily with the human side.

This compute-intensive evolution of the digital space, from the extreme edge, where the physical world begins, to the centre of the cloud, farthest from it, will give rise to a single unified continuum of computing, bringing along with it an array of technological, social and ethical challenges. In order for the web paradigm to adapt to this transformation, it will have to incorporate spatial awareness, with proximity, locality, and positioning in space and time becoming important attributes of digital assets, necessitating the adaptation of technology and protocols.



Key insights

- In the light of the modern web metaphor, the future IoT may be likened to a programmable space where applications are dynamic aggregates of components residing anywhere, deployed and used in an as-a-service mode, with a centre of execution determined as the most convenient to the user requirements.
- The aggregation of disparate service components and resources should be regarded as a personalized orchestration. Its execution engine must be trusted to abide by user preferences, while it should be able to withstand malicious infiltrations/deflections and privacy-breaking intrusions. It should also be vendor neutral and maximally interoperable.
- Programming orchestrators should evolve from traditional manual coding to modes that include "natural programming", where the human user employs natural interfaces (voice, audio, context) to express needs, objectives and requirements, and an intelligent engine turns such input into program code.
- Conceivably, such orchestrators may be further organized into an autonomous conversational hierarchical assembly, where an orchestrator may seek help from or provide support to others, in the quest to satisfy user requirements.
- Interoperability is a central prerequisite to the enactment of this vision. Achieving this will require various levels of standards, from specification and programming abstractions to interaction and networking protocols.
- The command-and-control element of cyber-physical systems will weave into the new web space, bringing about technological, social and ethical requirements for AI-augmented seamless interaction between humans and digital assets.
- The increasingly blurred border between the digital and the human sides of the compute continuum will cause the new web to become increasingly more physical. Instead of going *on* the web, we humans will be going *into* a new "spatial" web.

Key recommendations

- Research and development actions should be launched to explore the realization of web-inspired re-interpretation of future IoT systems.
- Proof-of-concept level efforts should be launched to develop agreed definitions of functional and non-functional specifications and accompanying programming support for execution- and communication-level interoperability across the continuum of computing.
- Promote demonstrative testbeds of "trusted orchestrators", running on any mobile device and appliance to serve the needs of safe, private, personalized and effective service composition on the computing continuum.
- Explore ways to enable "AI at the edge" and weave it with cloud intelligence in a manner that provides for rapid feedback cycles to actions and events occurring in the compute continuum.
- Devise low-latency, multi-modal, reliable communication architectures following the tactile internet paradigm to support peer-to-peer seamless interactions among humans and digital assets in the compute continuum.
- Promote the realization of "world model" pilots that showcase impactful uses of the spatial web, for example for decarbonization scenarios.
- Provide active support to standardization efforts toward the spatial web.

Premise

The basic component-level ingredients of modern IoT systems are becoming increasingly easy to use, both as individual items and amateur-style (introductory) collections *within vertical integrations* within the strict bounds of a single ecosystem (make/supplier). House automation is one of the most typifying examples of this tendency, where amateur users can set up complete commodity-level installations with very little hassle, so long as the components are from one and the same technology vendor.

At the same time, however, it can be observed that modern IoT systems at industrial/professional scale are becoming increasingly complex to build, master, evolve (and use). This is due to overwhelming heterogeneity in the available hardware and software components, including for end-user apps, which massively overlap for functionality, and expose / use extremely different APIs.

We anticipate that the next level of digitalization will result in a "maze of IoT". This could be built either statically (as permanent, stable aggregations), or dynamically (as opportunistic, temporary aggregations). We therefore need to address the technical challenges that arise from the effort to enact this vision.

Overarching vision

As future IoT systems become increasingly programmable, they also weave into a seamless continuum with the cyber-physical systems space, where command-and-control capabilities add to sensing and actuation at the border between the physical and the digital world. This convergence calls for AI to permeate the digital space in order for it to engage satisfactorily with the human side. This compute-intensive evolution will cause the whole digital space, from the extreme edge, where the physical world begins, to the centre of the cloud, which is farthest from it, to morph into a single continuum of computing, with an array of technological, social and ethical challenges to be addressed in that transformation.

The increasingly blurred border between the digital and the human sides of the compute continuum will cause the new web to become increasingly more physical. Instead of going on the web, we humans will be going into a new web, which we call "spatial" in this HiPEAC Vision chapter, to signify the need for it to recognize physical space (with proximity, locality, and positioning in space and time becoming key additional attributes of assets). Some enterprise projects have put forward their own

(controversial) interpretations of this new space and coined the term "metaverse" for it. An article in this chapter reflects on the human side, for motivations, requirements, and anticipated obstacles of the envisioned transition into an AI-augmented, spatially aware digital universe.

The convergence of autonomous vehicles, drones and bots, IoT sensors, augmented reality, sovereign identity, digital transactions, ubiquitous high-speed connectivity, and AI-augmented orchestration calls for new low-latency and interoperable technologies and standard protocols.

From human-expert mediation to trustworthy software orchestration

One such advancement should radically change the way one may go from actual needs (user requirements, whether industrial or individual, to component selection (equivalent to design) and integration (equivalent to system build). This route should no longer be an expert-only task, because obligatory human-expert mediation in such endeavours is an intrinsic bottleneck, a painful vulnerability, and a tremendous slow-down factor.

It should be replaced by a trustworthy "software mediator", capable of supporting natural interfaces (voice, drawing, gestures, context, examples), therefore conducive to "natural programming" Indeed, there is a potential link between this requirement and the capabilities of the low-code/no-code app development techniques. Those two paradigms might be the lower-end of the accomplishment of our vision, but we really anticipate going beyond them, in a manner that takes inspiration from the Codex and CodePilot initiatives, in which the human intervention is limited to the specification side and not to assembling code. The way such software mediators should operate for transforming user needs into programmatic artefacts may leverage the impressive advancements in the field of "natural-language-to-code" translators such as OpenAI Codex [1].

This notion of automated software mediator has one distinct premise and two distinct implications.

The premise is that much / most of what is needed to satisfy the end-user needs already exists in the form of partial solutions, which only need to be glued together in a functionally sound, solid (robust and scalable) and seamless workflow. Checking for availability can be done by periodically crawling registered repositories, whose authors have agreed to be part of the "project" and are consequently offered a sufficiently informative manifest file, previously screened to ensure sufficient quality.

The two implications concern the nature of the resulting "product" (the trustworthy software mediator).

- First, human interaction with the envisioned natural interfaces may be regarded as the "**next web**". In fact, the human-in-the-loop mode of operation may be regarded as the base of a hierarchy of interaction models that, as one goes up it, allow for progressively more autonomous inter-orchestrator dialogue, which eventually only needs humans for final approval. All of the resources of interest (digital assets of all kinds, many of which are increasingly "**digital twins**" of physical artefacts connected to the IoT system) should be navigable, linkable, CRUD-ready (within obvious limits of ownership and privacy of access) from a uniform-unified-omnipresent interface, much as we have become accustomed to when navigating the internet strictly over the web and interacting with all kinds of digital resources exposed in the web space, in two-way directions. By CRUD-ready, we mean digital entities that can be created, read, used, and deleted by a simple unified programming interface, which is the web interface itself [2].
  The technical name of the logical (virtual) platform underneath this "next web" is the "**continuum of computing**", which allows data *and code* to move freely, wherever most appropriate for reasons of privacy, energy, latency, capacity, etc. The realization of such a platform requires vertical and horizontal cross-stack interoperability, which is not only technical but also semantical.
- Second, the entity that does produce and execute the workflow designated to satisfy user needs may be regarded as an "**orchestrator**" capable of residing at the edge, as near as possible to the end user. This entity is an orchestrator because it deploys, executes, controls, and maintains the functional workflow that carries out the required (service) job.

In concept, this orchestrator is analogous to, but considerably more powerful than, recent "infrastructure-as-code" (IaC) engines that are transforming cloud deployment procedures from the recipe-based declarative manifests of the recent past to outright programmatic ones, massively raising the power of abstraction and algorithmic logic in the corresponding specifications. The term "infrastructure-as-code" designates a technical solution of application deployment in the cloud that is realized programmatically – by a true program, written in a cloud-native programming language – instead of declaratively by way of a manifest-like "recipe". The difference between the two modalities is that the former is far richer in expressive power than the latter and therefore naturally conducive to realizing sophisticated deployment policies that cover the entire

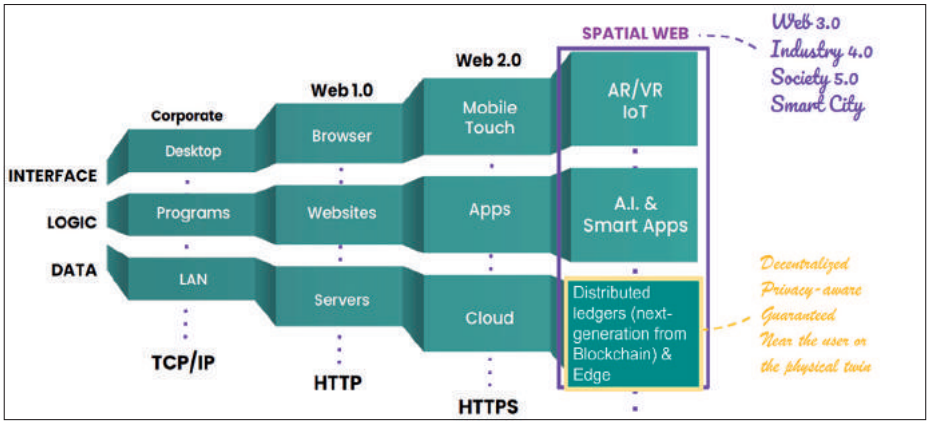


Figure 1: Adaption from Philippe Sayegh's presentation at IoT Week (21 June 2022)

application life cycle in all circumstances, including scaling-out and scaling-in situations, mitigation procedures, etc. The envisioned orchestrators are more powerful than those IaC engines because they interact via natural interfaces, and because they are most rigorously built as a "**trusted code base**", free from the risk of intrusion, hijacking, vendor lock-in, and subservience to malicious intent.

Naturally, the envisioned orchestrators should also play a defensive and protective role toward their user, by shielding them from unwanted interactions, instructions, attacks, and by proactively and responsibly following their preferences for functionality, provenance, ethical concerns, etc.

Everything-as-a-service

Another important game-changing technical objective is to move, in a conscientious, ethical and sustainable way, toward an "**everything-as-service**" (XaaS) trustworthy distributed platform-and-delivery mode. In this mode, all of the physical world can be interacted with, for use and control, in opportunistic, cross-sectorial, dynamic and highly-scalable aggregations. The main software "intelligence" (the orchestrator evoked under the previous heading) would run on the edge under direct ownership of the end user, instead of being secluded proprietarily deep down in the centre of the cloud. That arrangement would earn obvious yet crucial benefits for transparency, data privacy, and bandwidth reduction. It should be noted that this XaaS notion is a 1:1 match with the notion of "next web" discussed above and the associated corollary of the continuum of computing.

Another important challenge to overcome in order to implement this vision occurs at the hardware- and software-component level: too many important details are hidden away or ignored (notably non-functional, like execution time, response latency, predictability, energy balance, interaction with the external environment) in the current programming technology.

To overcome that deficiency, at the development level, we need more unifying abstractions (for algorithms and data structures, for programming artefacts, for cross-stack services).

The race for the "next web"

Various visions for a more interactive, intuitive "next web" have emerged, some of which reflect the priorities of a narrow section of extremely powerful, mostly US-based technology giants. In order to create a future web that properly serves its users, it is important to define possibilities, prerequisites and potential pitfalls of any particular scenario.

The articles in this section explore different aspects relating to the creation of this "next web" paradigm:

- **"'Digels', digital genius loci engines to guide and protect users in the 'next web'"**. This article presents the concept of the "Digel", a trustable, protective guide capable of orchestrating the "next web" and providing a seamless experience for users.
- **"The spatial web: Interconnecting people, places, things and AI for a smarter world"**. This article presents the concept of the "spatial web", the next evolution in computing and information technology, which will eventually eliminate the boundary between digital content and physical objects.
- **"Defining cyber-physical systems: Large-scale safety-critical systems"**. This article presents a definition of cyber-physical systems (CPS) from the application perspective. It classifies CPS requirements and points to future directions for research.
- **"Bridging the stakeholder communities that produce cyber-physical systems"**. This article provides a detailed outline of the different communities involved in producing cyber-physical systems (CPS), and emphasizes the importance of bringing these communities together.
- **"Towards a living dimension: The future of cyber-physical systems"**. This article explores the technical ingredients for future cyber-physical systems, including artificial intelligence, the tactile internet, swarm behaviour and the compute continuum. It outlines future challenges and suggests how Europe could meet these.
- **"Gaming, content and the metaverse"**. This article explores trends in augmented reality and simulation, and provides suggestions of how Europe should participate in the future "metaverse".

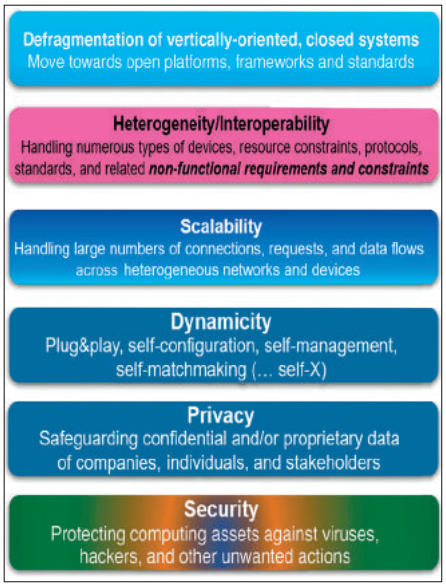


Figure 2: The multiple dimensions of the capabilities to be provided for in the realization of this vision. (This picture is drawn from an EU-US joint white paper that overlaps the topics addressed by this vision document).

- **"EU/US joint effort on the continuum of computing"**. This article presents the results of recent discussions between European Union (EU) researchers and representatives of the United States National Science Foundation. It outlines the technical challenges and envisioned directions for the evolution of the compute continuum.

Tullio Vardanega is an associate professor in the Department of Mathematics at the University of Padova, Italy.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2023.

Cite as: T. Vardanega. The "next web": Weaving future IoT and CPS into an AI-augmented, spatially aware, digital-and-physical space. In M. Duranton et al., editors, HiPEAC Vision 2023, pages 14-17, Jan 2023.

DOI: 10.5281/zenodo.7461756

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

© HiPEAC 2023

The wealth of digital services around us needs organizing, or better "orchestration". To make applications trustworthy for end users, the "next-generation web" will need special software technologies designed to guide and protect users as they navigate digital spaces.

"Digels", digital *genius loci* engines to guide and protect users in the "next web"

By TULLIO VARDANEGA and MARC DURANTON

Today, the very notion of "normal" life is largely intertwined with digital services. In many areas of our lives, digital services provide various manners of augmentation, which is both an enormous benefit and a serious threat. This trend – which is very visible to anyone – is only set to increase, as the "next web"[1] transforms the digital sphere, incorporating cyber-physical elements and artificial intelligence to take user experience to a wholly new level of immersion and personalization.

This "next web" vision is an opportunity to join intellectual forces and competences to address the technical challenges involved in making a seamless web environment (a kind of "ubiquitous browsable multidimensional reality") actually possible. The most significant challenges concern technical complexity, the interoperability of solutions, and the embedding of artificial intelligence (AI) in any computing device connected with it [1]. Other equally important concerns in this endeavour arise in relation to the threats of privacy breaches, exposure to misinformation, polarization and user addiction, to name just a few.

The 2021 HiPEAC Vision article " 'Guardian Angels' to protect and orchestrate cyber life", argued that it is necessary to create "guardian angels", personalized digital engines capable of guiding and protecting users in their interactions with digital services. Personalized, loyal and user-centred, these would also be capable of communicating with one another to access more services [1].

Building on this theme, this article presents mark II of the "Guardian Angel" concept: Digel. The Latins used the term "genius loci", literally "born-in-place", to designate an ethereal entity that was the "protective spirit of a place". This notion can be transposed to the web sphere, whose vastness, complexity, and perils call for a loyal, trustworthy virtual entity, fully inhabiting the webspace, to accompany us in our meanderings across it. We call this "Digel", for "digital genius loci". While different in many ways, this concept has acknowledged points of contact with the open-source federated personal assistants such as Genie, created by researchers at Stanford University [2].

The Digel will be the heart of the "next web", a web-powered platform offering composable services. These services will be associated with contracts specifying both functional and non-functional requirements (regarding, for instance, response time, latency, energy), containerized "migratability" to local computing resources, cost of use, etc. Future applications will result from the orchestration and composition of such services.

One of the key functionalities of Digels will be to orchestrate the provision of user-tailored services by coordinated and controlled aggregation of services already present on the platform at large, selected in accord with users' preferences and requirements, and delivered in as-a-service modality of execution. These orchestrators will also be in charge of managing inbound (security, verification of trust) and outbound (privacy and confidentiality) data, actively protecting the users' assets. The "programming" of these next-web engines will be done using natural interfaces, such as voice, drawings and schematics, or even by examples –an act, a thing, a discourse that describe the intent –thanks to the interpretation and generalization capabilities of AI.

1 We do not use the Web 3.0 or Web3 designations of our next web concept, as there are currently differing interpretations of those terms. "Web3" has been taken by blockchain and non-fungible token (NFT) platforms. Web 3.0 as seen by Tim Berners-Lee is a different entity, decentralized, and based on an extra layer above web 2.0, using "pods" to secure private data. For others, Web 3.0 is the "semantic web" that connects knowledge.

Key insights

- Digel engines will operate across the compute continuum from extreme-edge IoT to the cloud. By protecting the digital assets of users, they will ensure a trustworthy interface between the physical world and the cyberspace.
- Digels and multi-layered compositions of them will be the seed of the "next web", by allowing users to create complex personalized applications by orchestrating the selection, the workflow and the data exchanges between pre-existing services. The resulting orchestrated applications will intertwine the cyber and physical worlds with guarantees of fulfilling functional and non-functional requirements arising from service composition, in a safety- and security-aware manner, for both business and personal use.
- Digels should be user friendly and easy to interact with, allowing even those unfamiliar with today's computing technology to benefit from digital services.
- Digels will execute orchestration "scripts" automatically generated from requirements expressed by natural interfaces (voice, graphics, gestures, examples, …).
- It will be possible to embed the compute core of Digels, where their execution behaviour will be programmed, and all of its private data, in resource-scarce devices.
- Digels should be built on top of existing technologies and be interoperable with them.

Key recommendations

- Proof-of-concept developments of Digels, with their orchestration and composition capabilities should be promoted as part of future EU-funded collaborative research-and-innovation programmes.
- Such explorations should promote selective evolution of state-of-the-art web technologies, and integration around specific exploratory scenarios that allow key features to be showcased.

Some background on personal assistants

Once upon a time, there were "personal digital assistants": hand-held computing devices that mimicked the abilities of personal computers (PCs), only on a small scale, with application functions mostly confined to office suites and small applications. The advent of smartphones replaced these devices entirely, while helping transform the very localized notion of an office to a cloud-enabled "anywhere" hosting and computing platform, powered by Web 2.0 network protocols.

Then came AI-powered digital assistants, which eschewed traditional forms of human-to-computer interaction in favour of voice commands, conversational bots and contextual helpers. The next frontier of these digital assistants points to edge-based (as opposed to cloud-centred), privacy- and security-aware personalized digital engines. These digital engines would be programmed (rather than just commanded) by natural interfaces into orchestrated compositions of web services deployed in an everything-as-a-service delivery mode, where sensitive data are kept local and computation moves as needed according to a factors such as opportunity, convenience, privacy, energy, latency and cost.

Figure 1: a pictorial evocation of humanistic orchestration. Source: Dmitry on Adobe Stock

Bird's eye reconnaissance of the technical landscape

The notion of the "continuum of computing" reflects the observation that edge, fog and cloud computing platforms are being pulled together into what is bound to become a seamless execution environment. One of the factors behind this trend is the massive and unabating growth of value-added applications for mobile, handheld, wearable and unattended devices, in addition to traditional server-side applications. This explosion in complexity calls for self-managed workflow-based orchestration capabilities to meet emerging user requirements, whether personal or industrial, that cannot be met by any one component alone.

Interestingly, all those services effectively run on one and the same base programmable interface, the web, which

makes them ubiquitously accessible, intrinsically interoperable and technologically neutral. This programmable platform will be an evolution of the web 2.0 space, more aware of privacy, security, latency, energy, spatial locality, via assume-guarantee interface specifications which will allow Digel engines to apply service orchestration and composition on them.

These intelligent orchestrators would be distinguished by user-centred features, including:

- The ability for the user to instantaneously build their own service aggregator, which knows their preferences and profiles. This service aggregator would be capable of orchestrating services and information available on the web to do the user's bidding, with the additional ability to change service provider dynamically as the need or opportunity arises.
- The traits of a personalized recommendation engine truly loyal to its user, capable of fetching and selecting information items according to the user's personal preferences, without being deceived or deflected by provider-side solicitations.
- The ability to use correlation with trusted sources and AI to assess the trustworthiness and authenticity of all news and information items proposed to the user.
- The ability to have the user's digital exchange with others, whether humans or bots, filtered so as to prevent the inadvertent sharing of information that the user does not positively intend to share. As part of that, the ability to use strictly private past-behaviour analysis to help the user break away from information silos and filter bubbles, and to prompt users to update their preferences accordingly.

Figure 2: A pictorial representation of the Digel, generated by Stable Diffusion [4]

Figure 3: a scene of the HiPEAC Comic Book [3], which evoked the Guardian Angel concept as it was described in [1].

As we noted in [1], a number of state-of-the-art or in-development technologies may be regarded as the base to what will be needed to create Digel engines for real. For obvious reasons of practicality these orchestrating engines should be built on top of existing technologies and be interoperable with them. A non-exhaustive list of technology axes that need to be developed includes:

- The as-a-service delivery mode.
- Use of the highest levels of the internet protocol stack (HTTP/3).
- AI-assisted service provisioning, migration, and orchestration based on multi-objective optimization (e.g. performance, privacy, security, energy, emissions).
- A lightweight trusted computing base (similar to the mechanics behind modern web browsers), written in memory-safe programming language and verified by verification tools before compilation and deployment.
- Execution by lightweight containerized sandboxed interpretation, possibly in just-in-time mode, such as that used in WebAssembly [5] technology.
- Dynamic web directories for specification-based, resource-aware service discovery, provisioning, migration, and orchestration across the compute continuum.
- Dynamic programmatic orchestration of the services in a trusted and secure way, for example by cloud-native application programming interface (API)-integration languages like Ballerina [6].
- A human-friendly, natural way to program the orchestrator, for example based on speech-to-code transformation presumably powered by AI solutions.
- Human-friendly interface for non-specialists, able to assess the user's status and adapt to their reaction.
- Recommenders that can be run locally, with better security and control of their results.
- Digel engines should be able to collaborate in a trusted way to carry out aggre-

gations than cannot be achieved by one Digel alone, enlarging the scope of their capabilities.

- Digel engines should be capable of operating individually or collaboratively to detect fake or untrusted data and to assess the trustworthiness of the data accessed.
- Privacy- and security-preserving sharing and offloading of mobile data and computations to potentially untrusted hosts, using combination of detective deep-learning techniques, federated computation models, and homomorphic encryption.

Conclusion

As we concluded in [1], internet-of-things (IoT) devices, edge and deep-edge devices, and smart assistants are at the interface between the real and the cyber worlds. However, owing to the increasing inherent complexity of the cyber world, using it efficiently and safely is becoming more and more challenging, both for industrial and personal users. From the user's point of view, a single trustable entry point simplifying and supporting interactions with digital services will be required. We propose the "digital genius loci", or Digel, as this point of entry.

Our vision for the "next web" is a platform which intertwines the cyber and physical worlds for both industrial and personal use, overcoming the fragmentation of vertically oriented closed systems, the heterogeneity and the lack of interoperability. It will increase scalability in a dynamic environment where systems should self-configure, self-manage and be plug-and-play, while also coping with security and privacy of personal and corporate data. Orchestrators for our proposed Digels will contribute to the interoperability and trustworthiness of this vision of the "next web".

References

[1] Marc Duranton and Tullio Vardanega. "Guardian Angels" to protect and orchestrate cyber life. HiPEAC Vision 2021, DOI: https://zenodo.org/record/4719375#.Y4oggHbMK38, January 18, 2021.

[2] Stanford University. Genie: the Open, Privacy-Preserving Virtual Assistant, https://genie.stanford.edu/

[3] T. Vardanega et al, "Past, present and future of the Internet and digitally augmented humanity: A HiPEAC Vision," March 2020. [Online]. Available: https://www.hipeac.net/media/public/files/46/7/HiPEAC-2019-Comic-Book.pdf .

[4] "Stable Diffusion on GitHub," GitHub, [Online]. Available: https://github.com/CompVis/stable-diffusion. [Accessed 24 November 2022].

[5] "WebAssembly," [Online]. Available: https://webassembly.org/.

[6] "Ballerina," [Online]. Available: https://ballerina.io/.

Tullio Vardanega is an associate professor in the Department of Mathematics at the University of Padova, Italy.

Marc Duranton is a researcher in the research and technology department at CEA (the French Atomic Energy Commission) and the coordinator of the HiPEAC Vision 2023.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2023.

Cite as: T. Vardanega and M. Duranton. "Digels", digital genius loci engines to guide and protect users in the "next web". In M. Duranton et al., editors, HiPEAC Vision 2023, pages 18-21, Jan 2023.

DOI: 10.5281/zenodo.7461766

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

© HiPEAC 2023

We are entering a new era where, instead of going on the internet, we will go into the spatial web, which will eventually eliminate the boundaries between digital and physical.

The spatial web: Interconnecting people, places, things and AI for a smarter world

By PHILIPPE SAYEGH and CAPM PETERSEN

The boundaries between our digital and physical worlds are becoming increasingly blurred. We are entering a new era where compute and data will progressively come into the world: instead of going on the Internet, we will be going into a new internet – the emerging world of the "spatial web", sometimes called the outernet, web 3.0, industry 4.0, society 5.0 or the "metaverse", depending on your specific vantage point.

The spatial web is the next evolution in computing and information technology, which will eventually eliminate the boundary between digital content and physical objects that we know today. It raises challenges and creates opportunities to work towards a more transparent, interoperable, trustworthy, secure and compliant cyber-physical network of people, places and things.

Key insights

- It is necessary to integrate disparate technologies into a cohesive network that is as coherent with our logical understanding of the world as it is with the physical features of the world.
- An integrated operational framework is required to contextualize and synchronize data sources while making them interoperable.
- This framework needs to be built on new standards while leveraging legacy systems. It will have to address challenges in terms of context, space, operations and legislation.
- The new standards hinge on fundamental requirements around communication, interoperability, trust, adaptability and execution.

Key recommendations

- Fund research and development (R&D) and pilots to showcase the horizontal integration capabilities of the new standards.
- Create a "world model" based on spatial and multidimensional data and metadata that allows for impactful scenario simulations (e.g. decarbonization and circular economy use cases). This model can replicate the intelligence loop of sensing, modelling, simulating, acting and adapting to changes in the context and the environment.
- Raise awareness about the need for contextual and adaptive computing and metadata standardization.
- Upskill talent to master the spatial web standards implementation and methodology.
- Explore and contribute to the IEEE Spatial Web WG (P2874): https://sagroups.ieee.org/2874/
- Promote organizations and events, such as competence networks and hackathons, to raise awareness and drive the collaboration and change required for further adoption of the spatial web standards.
- Create a dedicated fund for developing applications and start-ups using spatial web standards.

Challenges and requirements

One of the most urgent and difficult problems in computing is poor interoperability between software and hardware. It is necessary to integrate these disparate but increasingly mutually reliant technologies into a unified and cohesive network for an information exchange that is as coherent with our *logical understanding* of the world as it is with the *physical features* of the world.

With a growing suite of advancing and emerging technologies coming together in all areas of our society and economy, the need to make sense of disparate data has never been greater. The convergence of autonomous vehicles, drones and bots, internet of things (IoT) sensors, extended reality (XR), sovereign identity, digital transactions, 5G, plus the need for an artificial intelligence system that can safely orchestrate all their activities, necessitate new technologies and protocols. An integrated operational framework and system is needed to contextualize, synchronize and make interoperable these disparate data sources about people, places, assets and things.

This operational framework needs to be built on new standards while leveraging legacy systems to address four fundamental new challenges to successfully govern activities in the real and virtual world.

1) The context challenge of internetworking functionalities and data between humans, machines, and objects, and the heterogeneous and siloed systems operating them.
2) The spatial challenge of delineating and coordinating activities between humans, machines, and objects in a multidimensional way (3D, 4D, cost, temperature, safety, etc.).
3) The operational challenge of synchronizing the flow of information and activities across spaces, humans, machines, and objects.
4) The legislative challenge of creating and maintaining digital versions of policies and legislation, and making them machine readable, machine sharable and machine executable.

The spatial web initiative and IEEE P2874 standards

The Spatial Web Foundation [1] has defined a contextual model and communication protocol that captures multi-dimension data interoperability. These standards allow systems to share a more complete and coherent understanding of activities and to continuously self-optimize. They hinge on several fundamental requirements that build upon each other:

1) **Communication**: a common query language in which to interface.
2) **Interoperability**: a common data model in which to map context.
3) **Trust**: verifiable data lineage, credentials, and governance.
4) **Adaptability**: flexible logic to simulate and adjust.
5) **Execution**: ability to orchestrate tasks and resources.

They are being standardized within the IEEE WG P2874 [2], namely:

- HSML - Hyperspace Modeling Language
  - A common data model that enables adaptive intelligence at scale.
  - A standard that models spatial and hyperdimensional relationships which can exist between any base elements and their purpose.

- HSTP - Hyperspace Transaction Protocol
  - Multi-dimensional range query.
  - A contracting protocol that queries that language and sends and receives the common language's data.
  - Protocols allowing stakeholders to govern identities, activities and spaces and location in an interoperable way and across data domains.
  - Governs interactions between parties to ensure privacy and security.

Note that IEEE has designated both the HSML and HSTP protocols as a "public imperative". This designation is typically reserved for critical public infrastructure like nuclear energy, smart grids, and voting machines.

Context awareness: the missing link to a smarter cyber-physical architecture

The vision of a smarter cyber-physical infrastructure requires full awareness of policy and context. AI systems, for instance, will require context-making to improve performance and explainability: they need to be policy-aware and context-aware. Data needs to be put into context for IoT and cyber-physical systems to remain relevant and adaptive to changes in use cases and scenarios.

Context can be defined as follows:

- Context is multi-dimensional. It needs to describe identities, activities, and spaces and places. That encompasses the following dimensions:
  - semantic (meaning and logic)
  - spatial (physical and situational)
  - societal (values and value)
  - systems (networks and ecosystems)
- Context needs to:
  - be stateful
  - be machine readable and executable
  - be shareable between heterogeneous networks, devices and applications
  - maintain coherence over time and space for all the actors/edges involved in a use case
- Context is represented by models of metadata that describe the activities of people, places and things over time. Context needs to be shared between networks of heterogeneous devices and applications empowering them to proactively offer enriched, situation-aware and usable content, instructions and experiences.
- Context is made up of the elements of relationships between identities, activities, spaces and places – commonly known as the *who, what, when, where, how* and *why* of any scenario, situation or circumstance. The answers to these questions are often stored in different data silos and different data spaces. They need to be made interoperable and are key to building a smarter cyber-physical infrastructure that is privacy preserving, secure interoperable and compliant by design.

The challenges in modelling context are linked to – although not limited to – a set of constraints: they need to represent the relationship between the identity of the

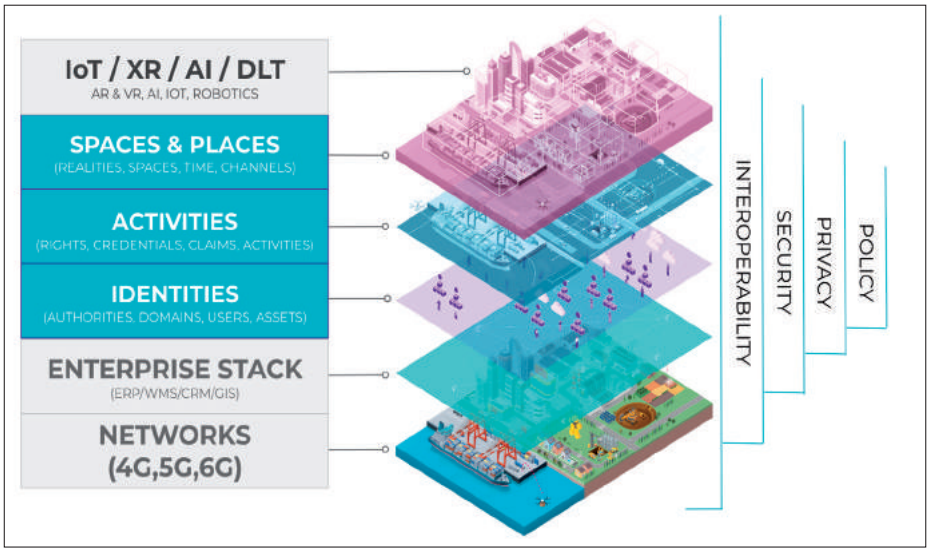


Figure 1: Smart cyber physical infrastructure

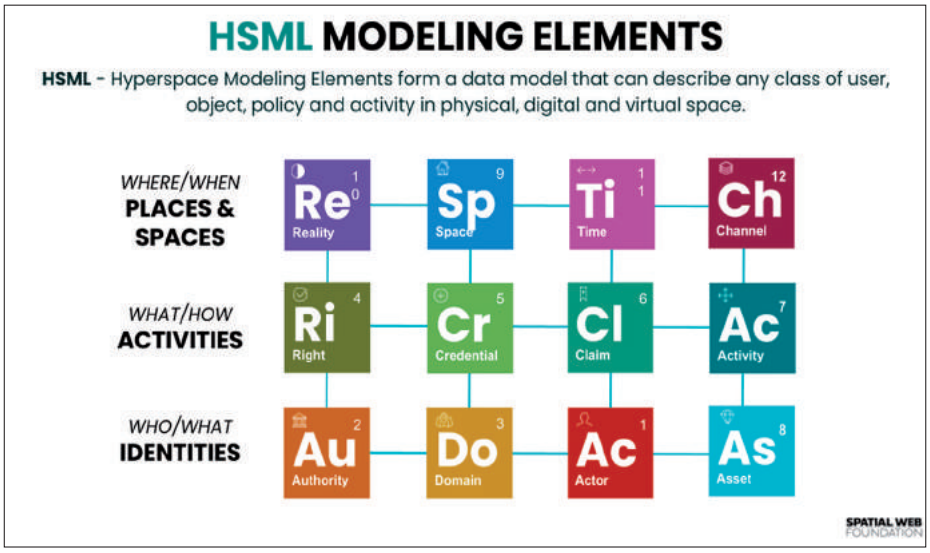


Figure 2: HSML modeling elements. Source: The Spatial Web Foundation

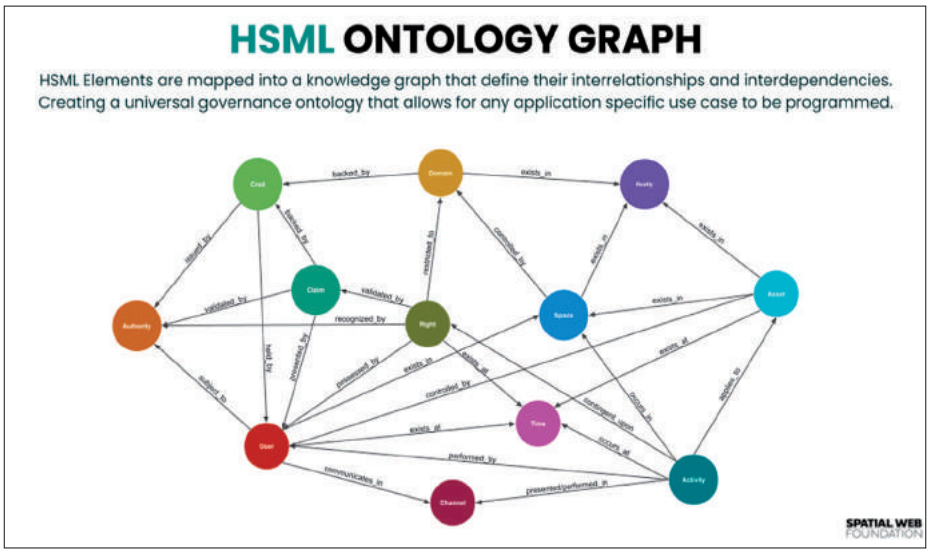


Figure 3: HSML elements. Source: The Spatial Web Foundation

actor, the scope and authorization of their permissioned activities, and the place and time where they may happen. HSML coherently describes and maintains the *where* and *when* of any scenario (dimensions, space, time and channel), the *why* and *how* (i.e. conditions or governance – right, credential, claim and activity), and the *who* and *what* (i.e. objects – authority, domain, actor and asset) shareable and addressable by multiple competing AI algorithms that can maintain their coherence at scale.

Spatial domains: a paradigm shift in anchoring data in space

Spatial domains are digital titles linked to 3D volumetric locations, such as buildings, ports, streets, or larger regions, such as cities, states, continents and trading blocs. Spatial subdomains represent subspaces that have a holonic structure, which allows for the orchestration of hierarchical rights and policies.

Spatial domains enable secure management of digitally mediated rights and permissions for:

- who/what is authorized to access the domain.
- what content or data is available to view.
- who/what kind of activities are allowed.
- who can publish and modify content.
- who can transact or interact with it.

Rules and regulations

Spatial domains contain all the rules, rights, permissions and fees associated with a geospatial region as defined by their owner or authority. This results in a governance layer of geospatial information written in HSML that can be queried by actors as they approach and move through them.

3D model

The 3D model contains all the geometries, addresses, spatial anchors, location of IoT devices, and any subdomains. The 3D model is dynamic and can be updated in real time as local conditions change. A visual digital twin can be used to represent the spatial domain and the activities happening within it.

Policy awareness: from analogue to adaptive digital law

How can we maintain and enforce policies that can be interpreted and shared by machines and humans? Existing regulations have traditionally been drafted by humans, for humans, with little regard for interpretability and executability by machines. There is tremendous value in seeking to bridge the gap between the traditional way of drafting and publishing regulations (i.e. *by* humans, *for* humans) and having machine-readable and machine-executable regulation (i.e. by humans, for machines) especially in the context of autonomous mobility scenarios.

Translating existing regulations into machine-interpretable and machine-executable code will allow us to govern the behaviour of machines in a policy-compliant way and dynamically adapt that behaviour as policies change and evolve over time. More generally, the challenge is to discover where and when technology can replace humans when it comes to interpretation, and where and why not.

The FF2020 Horizon Europe project [3] developed a geospatial infrastructure using the draft IEEE Spatial Web standards and protocols (HSML and HSTP) to govern actions within the domain of urban air mobility to enforce rules and policies systematically. This solution allows current laws to be parsed into machine-readable models and interpreted programmatically, reducing the human interpretation of laws and guidelines into more binary statements.

The solution takes the legal text and constructs ontologies and context from automated natural language processing and machine learning methodologies. These are then used to construct machine-readable representations of specific legal scenarios. This information is then parsed into HSML, where it is combined with other sources of information (such as IoT devices, user input, cameras, drones, etc.) that enable the AI powering a drone to be able to understand and comply with laws and conditions.

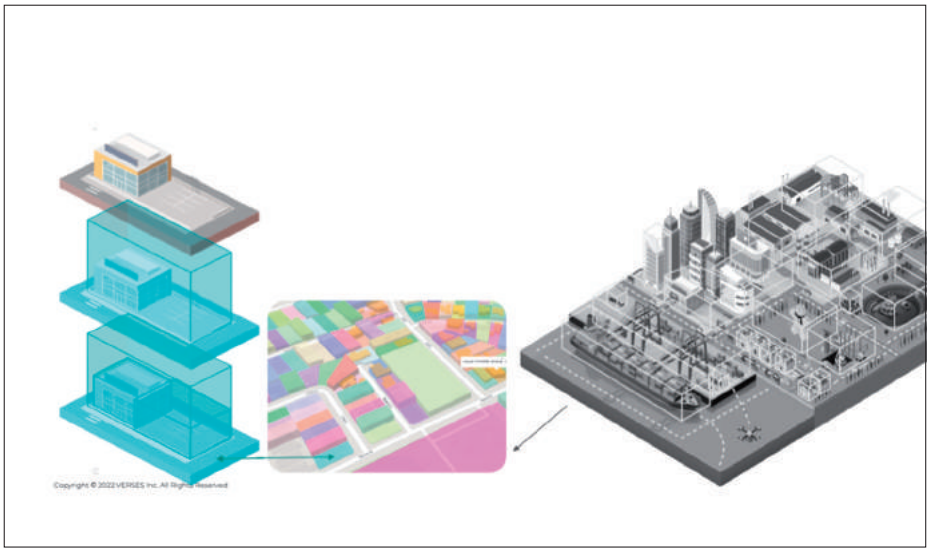


Figure 4: Spatial domains. Source: The Spatial Web Foundation

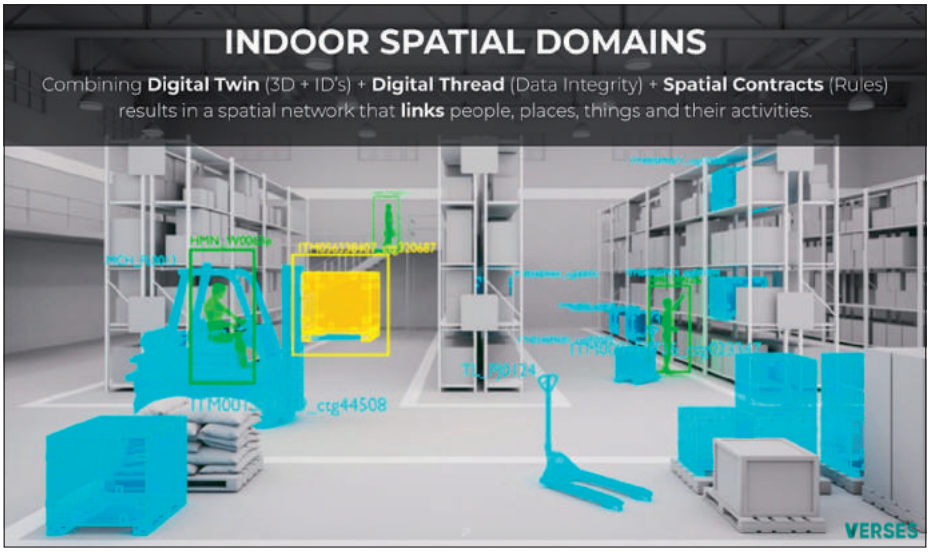


Figure 5: Indoor spatial domains. Source: The Spatial Web Foundation

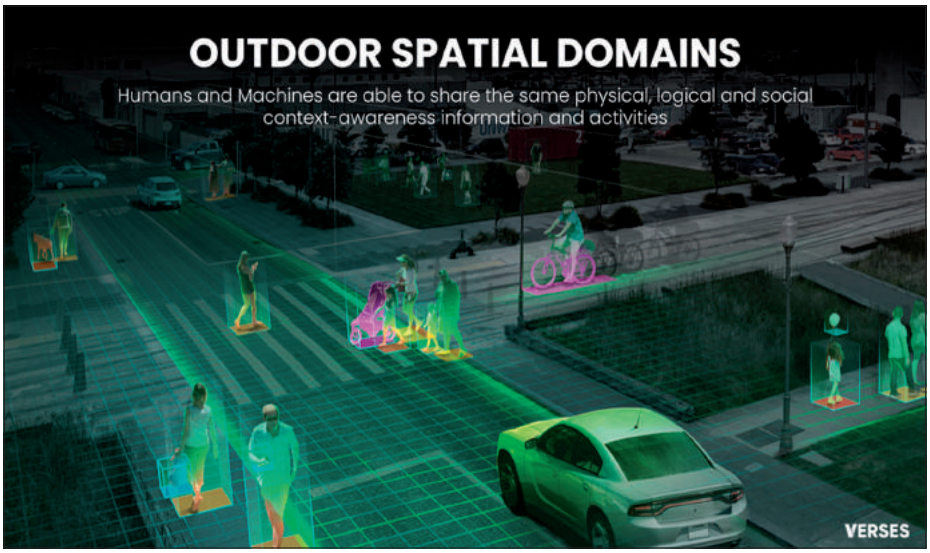


Figure 6: Outdoor spatial domains. Source: The Spatial Web Foundation

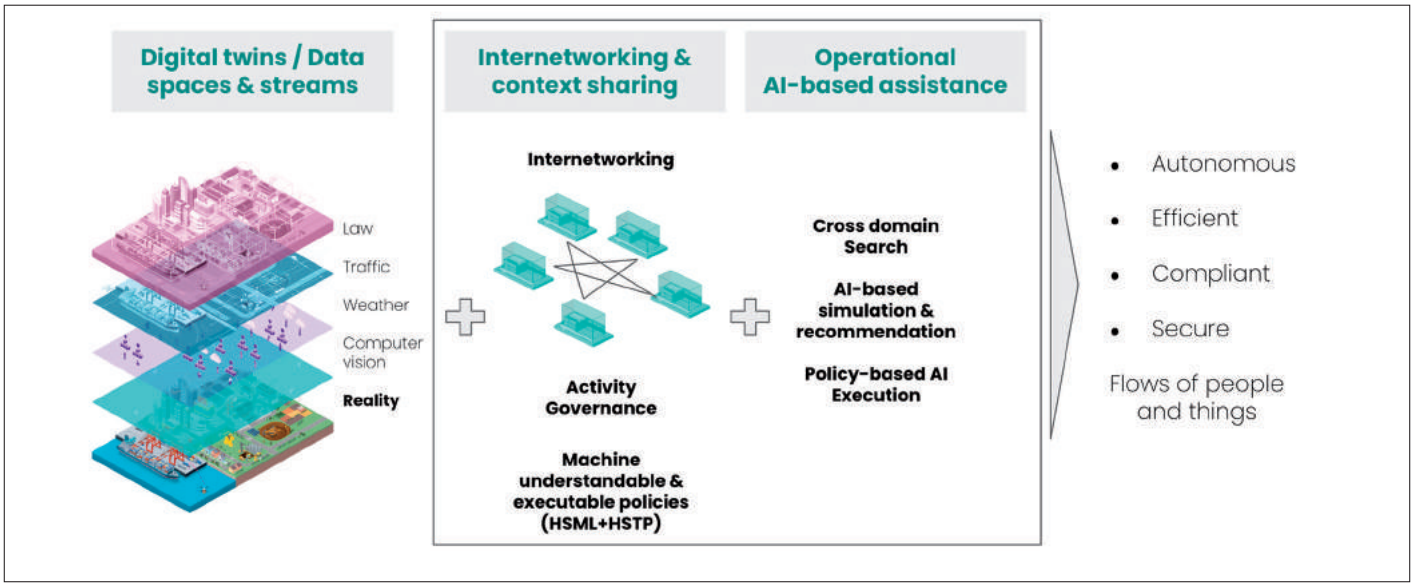


Figure 7: Building block for hyperdimensional contextual interoperability. Source: The Spatial Web Foundation

For example, harmonized liability rules for drones are lacking at the European Union (EU) level (U-spaces). Consequently, liability law is governed by the traditional rules at the national (member-state) level. An innovation, therefore, lies in (1) the development of liability rules that are (2) machine-interpretable and machine-executable.

Moreover, regulation is subject to change, and it is often multi-layered, meaning that regulatory bodies at different levels (e.g. EU level, state level, municipality level and company level) jointly contribute to a specific type of regulation. As a result, different pieces of regulation may consist of different languages.

To make a regulation machine-interpretable and machine-executable, and to make any interpretation model scalable, a semantic representation that accommodates regulatory changes and regulation at different levels and in different languages needs to be available. The technological challenge lies in the development of a suitable semantic representation that is applicable to specific geolocations.

Several laws that have been converted into HSML, including rules about dynamically changing variables, have been tested in real-world flight environments. We translated existing EU laws for urban air mobility into HSML, ensuring that the proposed actions and flight plans automatically conform to the set of parsed rules and regulations at play:

- The consortium ensured that HSML could accommodate the vast array of prohibitive and deontic laws found in the EU literature.
- The consortium worked with legal experts to ensure that the assumptions made during the translation process respected existing interpretations of the relevant laws.
- The consortium worked to provide a mechanism for suggesting legal innovations when a machine-readable format cannot straightforwardly accommodate the ambiguity of laws.

Building blocks for hyper-dimensional and contextual interoperability

The figure below shows a building block that enables hyperdimensional contextual interoperability. It includes the following capabilities:

- Internetworking spatial domains and context sharing using machine understandable and executable policies
- Operational AI-based assistance, e.g. cross-domain search, AI-based simulation and recommendation, and policy-based AI execution

KOSM, the AI operating system developed by VERSES, is an example of a building block implementation: It can be described as an operating system built upon HSML and HSTP that allows for the deployment of artificial intelligence applications in a way that integrates physical, logical, and operational meta-data. KOSM thereby allows for the autonomous, efficient, compliant and secure flow of people and things in the real world and/or in a metaverse.

References

[1] The Spatial Web Foundation: https://spatialwebfoundation.org/
[2] IEEE WG P2874: Standard for Spatial Web Protocol, Architecture and Governance: https://standards.ieee.org/ieee/2874/10375/
[3] Flying Forward 2020 Horizon Europe project: https://www.ff2020.eu/

Philippe Sayegh is chief adoption officer and **Capm Petersen** is chief innovation officer at VERSES.io.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2023.

Cite as: P.Sayegh and C. Petersen. The spatial web: Interconnecting people, places, things and AI for a smarter world. In M. Duranton et al., editors, HiPEAC Vision 2023, pages 22-26, Jan 2023.

DOI: 10.5281/zenodo.7461772

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

© HiPEAC 2023

Cyber-physical systems (CPS) bring together contributions from different technologies, stakeholders and areas of expertise. Given the increasing importance of CPS to society, it is important to have a common understanding of CPS and to clearly identify the communities involved. We propose the expression "cross-disciplinary bridging for large-scale, safety-critical systems" as a working common definition for new research in this area. This is an important step for tackling multi-dimensional challenges at the global level between the many stakeholders.

Understanding cyber-physical systems among many communities:

Large-scale safety-critical systems

By CHARLES ROBINSON

Cyber-physical systems interact with the physical environment and can be also termed large-scale safety-critical systems. They are increasingly collaborative and play an increasingly important role in supporting society. For example, they already represent a significant proportion of national infrastructure, occurring in many domains such as manufacturing, transport, medicine and robotics.

Defining their key characteristics is essential both to enable clear public understanding of their importance and to establish solid ground on which the many stakeholders, including researchers, engineers and policy makers, can come together and respond effectively to changing societal and market demands. Classification of these types of systems also enables a clear foundation, bringing together the various contributing communities around common goals.

CPS engineering is concerned with the assembly of technologies to provide complex services requiring physical action and collaboration. Hence, challenges relating to the convergence of technologies on the product side have a significant role. These challenges relate to the processes and technologies that facilitate bringing the parts together and therefore are key for supporting technology transfer to industry.

Furthermore, classification provides a means to develop clearer measures on research advancement and to monitor the industrial pulse. This article represents an update of previous work based on feedback from engaging with hundreds of specialists via the Platforms4CPS and HiPEAC projects.

Key insights

- CPS is a research application domain. That is, it encompasses how technologies including the IoT are combined to realize a particular application – in this case safety-critical applications such as air traffic management. CPS research should relate both to the contributing technologies and particularly to the aggregation of these technologies for creating future products.
- Classifying the different communities and technologies involved in CPS is important so that they can advance common goals.
- Future CPS development involves multiple disciplines and stakeholders to create an operational system, encompassing all domains where systems have physical interactions and collaboration. Solutions to multi-dimensional challenges are required, including the environmental impact of transport and other CPS domains.
- For safety reasons, CPS include subsumption behaviour – if several are working together, they become one CPS, similarly to emergency services joining forces on search and rescue activities.

Key recommendations

- CPS research, for discussions between stakeholders, should be defined as 'cross-disciplinary bridging towards large-scale safety-critical systems'.
- The dual nature of CPS research – CPS technology and CPS as an application – should be recognized.
- An agreed CPS classification system, such as that outlined in this article, should be used as a baseline for describing these systems with stakeholders.
- Approaches supporting systems thinking/interdisciplinarity need to be established to understand, prioritize and respond to the multi-dimensional community challenges faced for developing future CPS.

Evolving to future cyber-physical systems

Cyber-physical systems (CPS) have their roots in embedded systems, which are themselves devices with the capacity to monitor and affect their surrounding environment – usually requiring safety-critical automation. Embedded systems are used all around us and range in complexity: making automatic adjustments inside home kettles and cookers, using electrical impulses to regulate a patient's heartbeat inside pacemakers, or as part of an aircraft control system.

As these examples show, embedded systems have radically increased the standard of living for people across the world. Of particular consequence have been the advances in production and transportation domains where automation has greatly facilitated the exchange of goods, enabling significant societal advances. Advances in infrastructure, particularly for commerce, are directly linked to enabling civilizations to develop, especially by freeing up time for people to expand their skills in engineering, the sciences and the arts.

CPS represent an integration of technologies with much larger scope for physical interaction and increasingly collaboration, such as air traffic management. This collaboration particularly implies the ability to achieve goals that are unattainable for individual systems. However, it also makes guaranteeing safety increasingly difficult, given that CPS elements are increasingly distributed but must maintain central orchestration. Other examples of CPS include satellite navigation in our cars, which requires access to three satellites to operate correctly, or robotic-assisted surgery, which requires feedback to the surgeon.

Future CPS will aid rescue teams using autonomous vehicles that need to work as a team to tackle large incidents, as well as support cooperating autonomous cars, smart hospitals and robotic assistance enabling home care, for example. This support continues up to the level of smart cities, which, by providing coordinated collaboration of the many interacting services, support different needs of citizens while maintaining safety.

The realization of such decomposable and decentralized systems, with multiple levels of supervisory control required and the critical need to ensure safety, is limited by current development and research approaches. Where embedded systems have addressed design complexity by splitting the problem (separation of concerns), it has led to siloed expertise, limiting the complexity that we can manage. Bridging approaches and tools are required between the silos in order for future CPS to be developed to their full potential. (See the HiPEAC Vision article "Bridging the stakeholder communities that produce cyber-physical systems" for more detailed discussion of this topic).

Furthermore, the great advances supported by mechanical automation are endangering the very planet that supports human life and provides other long-term wealth. By way of example, automation requires energy and, in Europe, transport accounts for around 30% and industry for 25% of consumption. Of this energy consumed, 70% is produced from fossil fuels (according to the European Commission/European Enviromental Agency). CPS inherit such challenges and will need to turn them around; in fact, application-side CPS research provides the opportunity for addressing the challenges. *Solving the challenges linked to energy consumption and the green transition will require evolved and new methods, tools, processes and policies. In particular, aggregating technologies that manage how these elements are combined will be key to future CPS application research.*

The aggregating technologies determine the options available for the use of new component-based technologies, so CPS application-side research represents an important avenue for investigation. However, we need to stabilize and master CPS foundations while advancing the much-needed orchestrating/aggregating technologies, particularly those that are cross-domain and interdisciplinary.

The future CPS research domain represents a duality between dedicated *CPS technology and CPS as an application*, the latter referring to the unification of technologies to create this class of products. These two are intertwined because the application domain represents the destination or gateway that enables technology to be used. That is to say, the application domain research is about the bottlenecks for technology uptake on the product side, including the approach and influences on product lifecycle, and how the system components are brought together supported by several layers of aggregating technologies. The combined research of the two perspectives has the potential to have profound results, with technologies opening up completely new levels of personal freedom and new high-value professions, including in engineering and science.

Classifying cyber-physical systems

Creating CPS involves many communities, including those of specialist engineers, sales people, users and regulators. As all these stakeholders play a part in advancing CPS orchestration technologies and channelling the research, this field represents a multi-dimensional challenge. For instance, CPS technologies often take two to three times longer to be market ready than technologies without a requirement to interact with their environment, due to certification among other factors.

Another issue is that the many contributing communities also have multiple and sometimes conflicting perspectives of such systems. Recent discussions with specialists from the contributing communities have indicated that this isn't a problem within each community, where they work with the definition that is most relevant to them. However, there is a need to establish common identifiers for some solid ground on which to establish collaboration between communities. As a result of these discussions, general support has emerged for the expression *'cross-disciplinary bridging towards large-scale safety-critical systems'* to capture the essence of CPS and help policy makers and the wider public understand what this field is.

Clarifying the fundamental characteristics of CPS has been paramount in helping underrepresented communities gain visibility and in understanding the impact of public funding, while also supporting

advances in bridging technologies. This is especially important for Europe's quest for a resilient and sustainable future where CPS are directly involved, representing a large part of society's technology support infrastructure. We present here a distillation of many iterations over recent years by communities working in this area.

The most *distinguishing features for CPS are their physical interactions and collaboration*. CPS are products ranging in scale and diversity, from nanobots, railway networks, satellite constellations, smart hospitals, autonomous car fleets, future decentralized computing such as those used for controlling an aircraft, up to higher levels such as smart cities in which automation improves quality of life and also resource usage.

These systems have six key physical characteristics which they cannot do without: *sensing, communications, physical action, processing, the provision of energy and coordinated collaboration*, as shown in Figure 1. Classifying these is important not only for guiding research communities, but also in terms of understanding high-level progress across industry and for the public to be able to relate and understand the impact of CPS.

Future CPS will function like teams, where subsumption occurs with CPS merging and separating. They become one CPS, not because everything is managed centrally but because they are collaborating as a team sharing some higher levels of supervision. This is essential to ensure incidents related to safety and responsibility are fully covered. So CPS become one team of systems because they share a common framework enabling them to collaborate, converting cyber plans to real-world outcomes through physical actions [1].

We of course need to manage the interactions between the six characteristics. These interactions may be at lower levels between some of the above identifed characteristics but is particularly important at higher levels for overall control. This control also has to consider system properties that affect all the identified characteristics. In particular, given that these systems have movement and interact in the physical world, they must respect time and ensure safety and security. This already means that there are external stakeholders in terms of regulation, policies and standards to take into account before considering the requirements of a customer.

Since CPS are products, other stakeholders include those responsible for the product lifecycle: from development and testing, to operation and maintenance, to product retirement. This includes the ability to respond to and to impact markets and evolve product processes to support new technologies. There are evidently many technical and non-technical communities involved for creating a CPS, as highlighted by Figure 2.

CPS is an application domain where an assembly of technologies are applied to create large-scale safety-critical systems such as a rail network, where there are many influences affecting the combination of the technologies into a final product. That is to say, in this context a CPS represents the common goal for which the technologies are combined - a destination to which technologies are transferred. In Figure 2, technology provides the methodologies, hardware and software at each of the levels shown. As one moves up these levels, linking functions and system properties, the technologies are faced with increasing scope to manage and stakeholders to take into account. This means one is moving from technologies treating a single problem → single solution to group problems → group solutions. As a result, the related assembling and cementing technologies require different and longer lifecycles to mature, but are essential for advancing

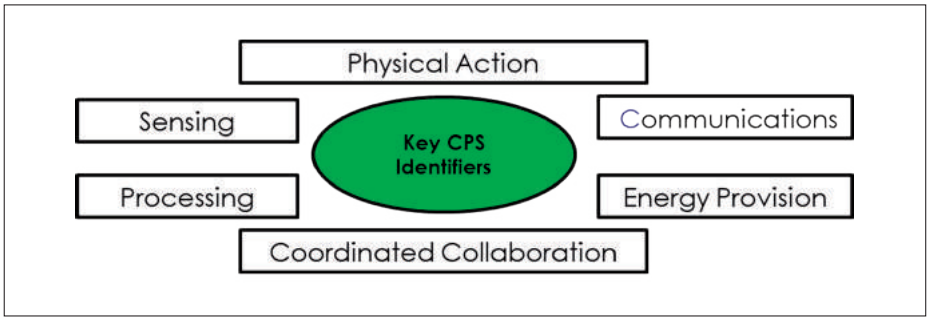


Figure 1: The six key characteristics to identify a CPS.

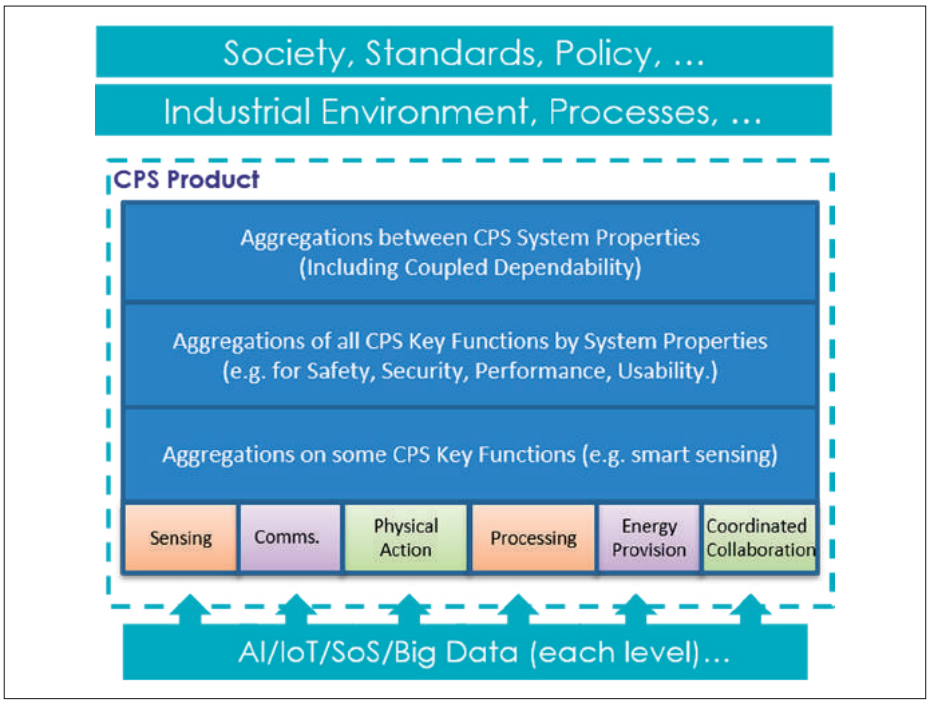


Figure 2: Contributors and influencers to aggregate in order to realize a cyber-physical system.

cyber-physical systems and the uptake of new component technologies.

Technology domains such as the Internet of Things (IoT), Systems of Systems (SoS), Artificial Intelligence (AI) and Big Data will also have increasingly important roles to play in the future of CPS as a destination of technologies. These enabling technologies will be particularly challenged by assuring safe interactions between systems and with people. The application of SoS represents a key stepping stone and some important research has advanced on this topic for some time now, abbreviated as CPSoS. Thanks to this work, hierarchical systems are now seen as key components of CPS, so we suggest that the use of the CPSoS acronym can now be phased out [2].

The definition presented here of future CPS products has been broken up into three parts: there are the two distinguishing features (physically interactive and collaborating systems) and the six key functions (sensing, communication, physical action, processing, energy and coordinated collaboration). Finally, we propose a model to be used, which includes the many stakeholder communities required for realizing a complete cyber-physical system.

Significance of a clear CPS definition for application-side research

Being able to classify and link the contributing elements for creating future physically interactive and collaborative systems is essential for many reasons. From the public point of view, the ability to identify such systems and their advantages is important, particularly because existing benefits are indirect, such as those delivered via services. Market push will be initially needed, as current purchases of CPS are usually business to business or made by governments to improve national infrastructure (e.g. transport and manufacturing capability). In the longer term, the consumer will be more in the loop, for instance owning components of a CPS

(such as an autonomous car) or having miniaturized systems like nanobots interacting and customizing features of the home.

A clear definition provides firm points of reference where a cohesive framework can be put in place, *essential for understanding the cumulative effects of successive aggregations of technologies required to create a CPS*. Here we use aggregation to mean the approaches and tools to bring the parts together. Such traceability of knock-on effects across these aggregations will provide significant support for a global view, benefitting many aspects of development and research for CPS.

From a funding point of view, application-side research will improve techniques for measuring the impact of investments on CPS and aggregating technologies, or other complex technology combinations. A common characterization means the benefiting market sectors can be clearly identified. Higher-level aggregations usually take longer to exploit because there are more stakeholders involved, for instance, technology methodologies providing global orchestration for air traffic management has the standards for technology across the whole system to respect as opposed to a module inside focused on smart-sensing. Understanding the uptake of these higher-level aggregations requires new approaches to metrics for visibility of return on investment.

In addition, more specific feedback from related businesses can be provided, for instance from their interaction with safety certification authorities or use of specific tool combinations. This is also relevant for tracing the evolution of product process methodologies, where direct feedback of new approaches applied can be difficult to measure [3], often including intangible factors, such as improved knowledge generation and transfer between contributing developers. Therefore, such a framework will provide a better means for keeping a finger on the industry pulse.

Similar benefits hold true for CPS research needs and contributor support, with significantly more visibility of bottlenecks along the chains of aggregations. The overall response to top-level research challenges and impact from many levels of contributions can be traced through the aggregations. Achieving cooperating systems, for instance, will require supervisory control layers over the combined systems to ensure safety and also other mechanisms for decomposable teams. Other top-level technical and societal challenges include sustainability, resilience and the circular economy. Traceability through extensive aggregations is especially relevant when it comes to CPS representing smart cities, where clear guidance on impactful research is important, given the scope of potential research paths for higher-level aggregations.

With respect to the technology contributors, clear characterization and traceable uptake of orchestration technologies means more informed and prioritized development can take place, such as in response to higher-level aggregations presenting barriers to new technologies, or even for aggregations at the same level, such as between security and safety [4]. As an example, an existing certification technique used in companies may need to be adapted so that new technologies for physical action can be applied. As the CPS field has a cross-domain focus (e.g. transport and manufacturing), it also provides the ability to identify and prioritize work on cross-domain barriers, with the capacity to open up new markets for contributors to CPS.

Furthermore, because of the aggregational nature of CPS application domain research, focus must be placed on tools and techniques not only for the multidisciplinary operations of a CPS, but also for supporting multi-specialist interaction in research, including translation and mediation between perspectives. This will be particularly important for bridging the silos between disciplines, something that will be required to achieve the expectations for future CPS. A clear definition and work on aggregating knowledge and tools between stakeholders are consistent with many of the other recommendations of the current HiPEAC Vision. They also support the recommendations from Platforms4CPS [5], including the ability to couple system properties like safety and security, CPS with explainable actions for trust, improved collaboration and defragmentation of communities and maintaining Europe's leadership in CPS while taking into account the societal impacts of these systems.

References

[1] M. Törngren et al., "Deliverable 4.3: Collaboration on the foundations of CPS engineering," Platforms for CPS, 29 10 2018. [Online]. Available: https://www.platforms4cps.eu/fileadmin/user_upload/Deliverable_4.3_Collaboration_on_the_foundations_of_CPS_Engineering.pdf.

[2] C. Robinson et al., "Deliverable D3.3: Best practices and tools for SoS applied to CPS applications," HiPEAC, December 2020. [Online]. Available: https://www.hipeac.net/media/public/60/8/d3.3_report_on_best_practices_and_tools_for_sos_applied_to_cps_applications.pdf.

[3] A. Thum-Thysen et al., "Unlocking Investment in Intangible Assets – Discussion Paper 047," uropean Commission – Directorate-General for Economic and Financial Affair, May 2017. [Online]. Available: https://ec.europa.eu/info/sites/default/files/economy-finance/dp047_en.pdf.

[4] C. Robinson et al., "MERgE: Technology Advancement for Cohesion of Concerns in System Engineering," in Proceedings of CERTS 2016: the 1st Workshop on Security and Dependability of Critical Embedded Real-Time Systems, Porto, 2016.

[5] H. Thompson et al., "Platforms4CPS: Key Outcomes and Recommendations," October 2018. [Online]. Available: https://www.platforms4cps.eu/fileadmin/user_upload/E-Book_-_Platforms4CPS_Key_Outcomes_and_Recommendations.pdf.

Charles Robinson is the research projects leader in critical embedded systems at Thales Research and Technology, France.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2023.

Cite as: C. Robinson. Understanding cyber-physical systems among many communities: Large-scale safety-critical systems. In M. Duranton et al., editors, HiPEAC Vision 2023, pages 27-31, Jan 2023.

DOI: 10.5281/zenodo.7462025

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

© HiPEAC 2023

The development of future large-scale safety-critical systems, also known as cyber-physical systems (CPS), involves many technology and influencer communities. Novel approaches and tools will be required to tackle the multi-dimensional challenges between these communities to benefit CPS, especially their adaptability to new technologies.

Bridging the stakeholder communities that produce cyber-physical systems

By CHARLES ROBINSON ET AL.[1]

There are many communities involved in the creation of cyber-physical systems (CPS), which are used in domains including transport, health, manufacturing and, in the longer term, will be in the home, where miniaturization will play a role.

In this article we explain that engineering for future CPS needs a centre of gravity in order to draw these communities together. This will provide common goals around which technical advances can be aligned. Overviews of the communities involved are provided, with examples of their relevance in the creation of CPS and to some common challenges.

Advancements of aggregating technologies are multi-dimensional challenges, representing many influencing dependencies from all communities, especially at higher levels where the whole system product is drawn together. This means that, to make good progress, Europe will require new forms of coordination in order to orchestrate research and to capitalize on lessons learned related to the cumulative advances between the communities.

Key insights

- Large-scale safety-critical systems, also called CPS, are physically interactive (high certification obligation) and increasingly collaborative (task sharing). They involve many contributor and influencer communities in their creation, who each tend to make advances in isolation. Creating the technical bridges between these communities to channel technology development is essential for these future systems.
- The scope is wide and communities need a technical interface around which to align. Discussions suggest this centre of gravity to be real-time safe and secure automation.
- A new form of research coordination is necessary to direct cumulative developments from the stakeholder communities. CPS projects with cross-community challenges and which involve most of the stakeholders are needed for this.
- The development of CPS requires a holistic approach, guided by target products, that brings together a wide range of disciplines. These should include not only functional, system and enabling technologies, but also the fields of psychology, sociology and ethnography, among others.
- Aggregating technologies have different industrial uptake lifecycles to component technologies. Research programmes treat them the same and both technology types suffer. A dedicated team would be very beneficial for investigating and implementing specific technical supports (for projects, programmes and industrial policy).
- System thinking/interdisciplinarity will play a critical role in reversing the adverse effects of technology on the environment and requires more support from national funding. Technologies for future large-scale safety-critical systems is one key area that this applies to, especially to tackle their high environmental impact.

1 For the list of contributors, see the acknowledgements at the end of this article

Key recommendations

- Draw on best practices for systems thinking/interdisciplinarity to develop methodologies and supportive tooling that will act as the technical bridges between communities contributing to large-scale safety-critical systems.
- Real-time, safe and secure automation should be used as the centre of gravity for communities contributing to large-scale safety-critical systems.
- While advancing towards this centre of gravity, we should not neglect the other common interests (common challenges) shared between the communities that will reinforce the bridges. Approaches from other interdisciplinary fields should be used as inspiration.
- Research orchestration needs to be developed for coordinating cross-community research to future large-scale safety-critical systems. This also calls for projects tackling challenges that are common across the contributing communities.
- A research instrument, transversal to European research programmes with stable financing, is required for advancing aggregating technologies in particular and the technology uptake by large-scale safety-critical systems in general.

Introduction and new cross-community development approaches

In order to manage large complex problems, people break them down into parts. It is for this reason that, from the technology point of view, there are many contributing and influencing communities involved in the creation of future safety-critical products. Of course, the parts subsequently need to be assembled together in order to address the initial complex problem. For the same reasons, the various technological contributions for future large-scale safety-critical systems, also known as CPS, require layered aggregation in order to achieve these physically interactive and collaborating systems.

This means that there are significant, multi-dimensional influences across CPS communities, which contribute to our ability to transfer technology to industry. It also poses challenges for assuring CPS, which should be based on sound methods of justifying that a CPS is fit for purpose and that all risks of using it are adequately addressed, notwithstanding the complexity and the heterogeneity of the CPS components and of the communities of stakeholders involved.

For the purposes of this article, we take CPS in the context of an application; that is to say, the term could be replaced directly with an example CPS application such as railway transport or satellite constellations. In this framing, CPS therefore represent physically interactive and collaborating systems that are present in many domains including transport, health and manufacturing (For an in-depth definition of CPS, see, for instance, the HiPEAC Vision article "Understanding cyber-physical systems among many communities: Large-scale safety-critical systems").

Communities involved in CPS, discussed in the subsequent section, range from providers of a) functional properties including sensing, physical action, communication, energy provision, processing and coordinated collaboration to b) system-level engineering including properties like safety and performance specifications, managing customer requirements, architecting, system validation, mechanical engineering and control engineering. There are technology support communities providing c) enabling technology domains like the Internet of Things (IoT), Systems of Systems, Big Data, Artificial Intelligence (AI) / Machine Learning (ML) and High-Performance Computing (HPC). Finally, there are the influencing communities from d) the production environment, with enterprise processes and product line, and e) the market, such as regulation and current and future needs of society.

These communities have tended to transfer technology as a one-to-one mapping with products. However, to respond the challenges of future CPS and to enhance technology transfer, they will need to take relations with the other contributing communities increasingly into account. While the challenges and importance of advancing aggregation techniques are discussed later, there also needs to be a common focal point from which one community can interact with any of the other communities. This point should provide a common interest based on the physical challenges of these systems. *Discussions have proposed this centre of gravity to be real-time, safe and secure automation of CPS development and operation.*

Research on CPS should seek to enhance the interrelations and automation of these three dependability properties, i.e. real-time, safety and security. They are goals that must be achieved at a global level when all the technologies are combined. As an example, each piece of hardware has an impact on the energy consumption of the whole system. Similarly, individual software and hardware components can jeopardize safety if they fail naturally or due to a security breach. These goals can also be variable and related to environmental conditions, such as a train reducing its speed (performance) in response to heavy showers (to maintain safety).

Hence for technologies to be accepted in these systems, they must guarantee these dependability properties, i.e. they must comply with the safety and security constraints of a product and not compromise real-time responses. This means that the easier it is to couple your technology with these system constraints (through automation), the easier it becomes to adjust it to the system (or adjust the system for new technologies).

It is usually the case that, in order to add new technologies to a CPS, the whole system requires re-certification. This can be prohibitively expensive without sufficient automated information about the impact of the new technologies on these dependabil-

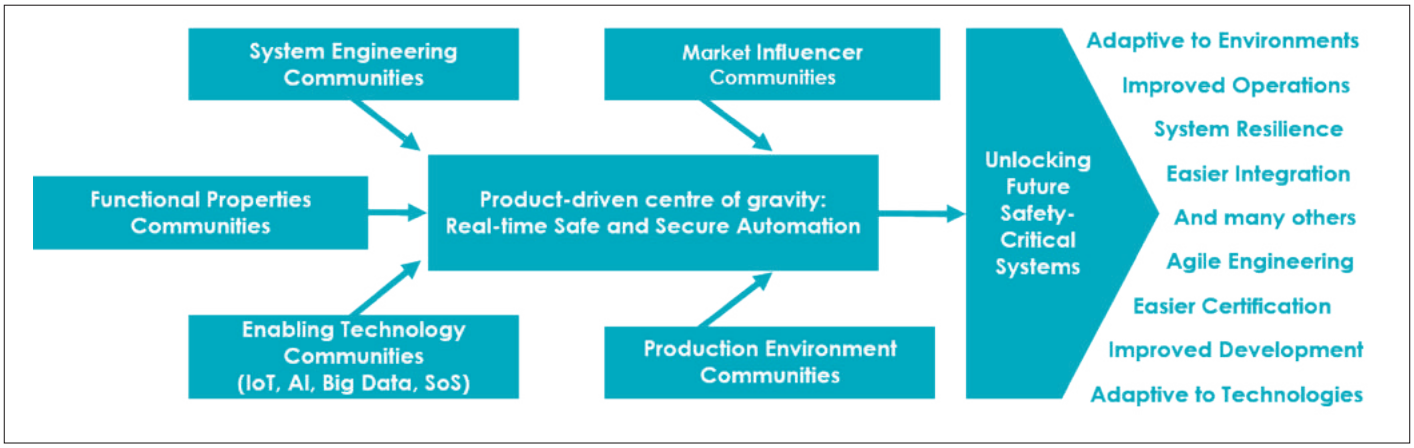


Figure 1: The stakeholder communities for creating CPS.

ity properties – and particularly the interrelations of those properties. Take systems certified, for instance, against an extreme earthquake occurring every 1000 years (for safety), such as a nuclear plant: in this case, the safety experts currently would prefer no new technologies or patches for security to be added to these systems due to the certification costs.

As a result, historically, interrelations between system properties have been limited to minimize complexity, but the current need for adaptability (to new technologies, to environmental or internal changes) requires this design mindset to be readdressed. So in summary, a centre of gravity, as shown in Figure 1, will provide a useful point to channel us towards more impactful research advances for these future large-scale safety-critical systems.

While the management of trade-offs between the system properties of performance, safety and security is an established skill in system development, it still remains very much a manual and qualitative process and one that is based on prior experience. It remains to this day very much a bottleneck and is holding back the communities contributing to CPS development from ensuring that advances in areas such as trust in artificial intelligence (AI) are applied to CPS.

System-level engineering for CPS is therefore in need of transformative automation. Fortunately, automation between system-level dependability properties can rely on a number of decades of research in techniques [1], some of which have already been applied in industry but are generally in need of new approaches for technology transfer. Such approaches are included in the coordination suggestions for research orchestration described later in this article.

Of course, current pressures for industry to find advanced solutions for managing system property trade-offs are also driving the search for automated coupling. As examples of some initiatives, the UK Research Institute in Trustworthy Interconnected Cyber Physical Systems (RITICS) [2] involves dozens of UK universities and industrial collaborators. Topics include safety, security and autonomous systems. Relating to autonomous vehicles, the Intel Research Collaborative Institute of Safety of Autonomous Cars (ICRI-SAVe [3]), deserves a mention as a vibrant community.

Many industries are actively looking for solutions to manage the performance, safety and security of their products, including large enterprise like Siemens, Thales and AVL, who have been forming combined safety-security teams. The challenge also affects small and medium-sized enterprises (SMEs) in their products and services. This recent momentum has visibility, for example, in the Ada and IEEE conferences, in the IET code of practice on cybersecurity and safety [4], in recent large research collaborations including MERgE [5], SeSaMo [6] and AQUAS [7], and in co-engineering discussions.

Overview of stakeholder communities for creating and advancing CPS

We now provide overviews of the five communities, indicated in the previous figure, which are involved in creating CPS. We give descriptions and examples of their relevance to CPS as well as their relation to cross-community challenges for future development. These include embedded computing as a CPS backbone, system decentralization and decomposability, and physical collaborations with people.

Functional-property communities

CPS functional properties have to address aspects that cover sensing, actuation, communication, energy provision, processing and coordinated collaboration. Such properties are key characteristics of these systems, with actors in specific communities researching and developing the different components.

The relevance of functional properties becomes more evident when considering novel and innovative advanced applications that are being progressively adopted in a number of large-scale, safety-critical domains, such as industrial automation, transportation, smart cities, critical infrastructures, space, etc. Some examples can be found in H2020 projects such as CPSwarm [8] and other CPS cluster initiatives.

Industry-driven needs and the well-established nature of general research communities in the CPS domain mean that it is feasible to envision projects that might prototype concepts such as swarms of unmanned aerial vehicles and rovers

supporting safety and security operations; swarms of automated ground robots that collaboratively support humans in logistic operations within a smart warehouse; or enhanced and dynamic platooning applications for autonomous freight vehicles. Currently, the development of such applications cannot leverage a simple plug-and-play integration of the various technologies entailed, given the complexity of managing teams of systems and humans in evolving and dynamic scenarios with emergent properties.

Therefore, in order to properly combine and integrate the different technology building blocks required, the various 'functional properties communities' have to be properly engaged. Experts from the functional property communities will need to work with other actors with collaborative systems competence. Moreover, while the increased adoption of CPS has resulted in the maturation of solutions for CPS development, a single consistent science for future CPS has not yet been consolidated. Few functional properties community members have already started working alongside other communities on a connective framework e.g. using modelling, design/development tools and methodologies, deployment solutions, monitoring and controlling solutions for large-scale challenges. In this context, model-centric approaches have clear relevance for facilitating collaboration between experts from different sectors and thus enabling the definition, composition, verification and simulation of collaborative, autonomous CPS.

For these reasons, it is important for future CPS to be considered not only from the technology perspective but also as an **application domain** where the technology of the functional properties' community plays a role for aggregation of CPS-related research. To promote this, closer and wider collaboration is needed within the community, along with new research initiatives. Understanding the nature of this aggregation from the bottom up and top down is important for driving the communities towards much-needed technology advances. The resulting collaboration plays a very important role in finding solutions to the bottlenecks that currently prevent CPS from having greater impact on society; such solutions would also promote market uptake, open up new markets and optimize the use of resources in the various industry sectors.

These communities have many cross-cutting challenges for future CPS. Embedded computing will evolve significantly and plays an essential enabling role for functional properties. For instance, the need to use specific sensors on a CPS and to timely process the relevant raw data onboard will need increased computational power. However, energy limitations introduce other constraints; only a holistic vision of CPS can help driving research initiatives. Moreover, the envisioned combination of 5G, beyond 5G and Smart Networks and Services/6G technologies with distributed and high-performance computing will pave the way towards a deep integration of future CPS in the computing continuum. In relation to decentralization and decomposability, with distributed intelligence and emergent properties, an example research context would aim to solve/work on delays in physical, computing and actuation timing. This requires model design and simulation approaches to capture the full heterogeneity of the system and its contributing communities. Physical interaction with people requires a system to have high fidelity knowledge of its environment and its physical dynamics. This requires the technologies of the functional properties community, which in turn need integration with the safety and security measures set by the system-engineering community. It is therefore clear that the best way to advance future CPS is to further support integration and aggregation approaches for community collaboration.

Systems-engineering communities

The development of CPS requires a holistic development approach that brings together a wide range of disciplines. This includes the typical systems engineering disciplines, such as requirements engineering, architectural design, implementation and quality assurance including system-wide responsiveness, safety and security. The disciplines of this community are important in terms of both the CPS in general and individual systems engineering sub-processes, such as mechanical engineering, control theory, electrical engineering and software engineering.

In almost all of our application-driven future scenarios, like in autonomous driving and Industry 4.0, CPS must be able to fulfil their purpose to a large extent without intervention of human users [9]. According to the Society of Automotive Engineers (SAE) taxonomy for autonomous driving, we refer to such systems as highly automated or fully automated CPS [9]. Already today and even more so in the future, systems engineering is one of the core competence fields for building such highly automated or fully automated CPS.

In the case of highly automated CPS, it is necessary to have a more comprehensive understanding of the term 'functional safety'. In contrast to the understanding of the term by the ISO 26262 standard, which essentially considers the malfunction of system components, highly automated CPS require an analysis of the interaction of a) the functionality of the CPS under consideration with b) its context (e.g. other CPS in collaboration). This analysis serves to detect possible safety threats resulting from the interaction between system functions and contextual conditions, such as the interaction between the autonomous driving function of a vehicle and the failure of the signalling system at an automated road intersection. This new understanding of functional analysis, which goes far beyond the requirements of ISO 26262, is the subject of the SOTIF standard [10].

These threats to safety must be identified during the development process and mitigated, e.g. by specifying suitable requirements or safety devices (safety monitors) which bring the CPS to a safe state should CPS fail to behave according to requirements/expectations. Since CPS often monitor and control technical or physical processes, control theory is a discipline of great importance in the development of such systems. In this context, the concepts of monitoring and controlling technical/physical processes are reflected in various artefacts of systems engineering. For instance, the requirements originated from the way the processes should be controlled,

as well as from decisions made about the design of the necessary sensors and actuators or even about the design of the algorithm for the computational processes of the feedback system.

In order to be able to develop such complex technical systems consisting of software and hardware, seamless systems engineering processes are required, establishing techniques, methods and tools for challenges such as the following examples. Since CPS in many fields of application work together in dynamically formed networks at runtime to pursue higher-level goals, possible collaboration structures must be identified and analysed in requirements engineering. For example, in the development of autonomous vehicles, the collaboration structures in which these vehicles must operate should be taken into account. Examples of such structures might be vehicle convoys to optimize the flow of traffic or at automated intersections to ensure safe crossing of the intersection, even with high traffic volumes and in complex traffic situations. In collaborative CPS, the issue of coordinated decentralized monitoring and control of technical/physical processes is added; an example of this is the coordinated acceleration or deceleration of the various vehicles within a convoy of vehicles.

In the case of highly automated systems, the involvement of the human user is required in (a few) defined situations to ensure that the system is able to fulfil its purpose of ensuring safe operation. The integration of the human user must be effective, i.e. the user interface of these systems must be designed in such a way that the human user is able to perform the necessary tasks according to the intention, as free from errors as possible and within the existing time restrictions. One might think here of the example of autonomous road traffic, where highly automated systems require the driver to take control of the vehicle when a critical driving situation occurs.

Enabling-technology communities

Internet of Things (IoT)

The Internet of Things community developed around the goal of providing a means for all devices to be globally connected via the internet. The name 'Internet of Things' was used in 1999 by Kevin Ashton during a presentation to his higher management at Procter & Gamble. He described IoT as a technology that connected several devices with the help of RFID tags (radio frequency identification) for supply-chain management [11]. In 2008 the first international conference on IoT took place in Switzerland, discussing RFID, short-range wireless communications, and sensor networks; today, these topics continue to represent the major technological research domain for advancing the IoT, gathering information about the real world that can then be made useful in some way [12].

Since 2010 it has been normal for many different devices to be in our homes to be connected to the internet. Connected devices are used extensively in the consumer domain. In 2015, to support advancement of IoT for industry, the European Commission created the Alliance for Internet of Things Innovation (AIOTI). Applying IoT to the industrial environment has been termed industrial IoT, or IIoT, and has the goal of optimizing production value while considering the many additional challenges related to safety, security and performance. IIoT technologies support interconnectivity with the internet in the context of these challenges, enabling not only networked smart objects and information technologies but also "*optional cloud or edge computing platforms, which enable real-time, intelligent, and autonomous access, collection, analysis, communications, and exchange of process, product and/or service information, within the industrial environment*" [13]. IoT technologies , in particular those for the IIoT, will be standard constituent elements of future safety-critical frameworks.

Enabling the infrastructure to support distributed intelligence and information exchange is at the core of IoT, so supporting cross-community work on CPS decentralization, decomposability and human interaction is important. These are already areas receiving some focus from the IoT community [14, 15], as indeed is the case for bringing communities around an embedded computing backbone, with work considering edge-cloud computing [16] exchanges. As an enabling technology, IoT responds to support other domains which means its focuses change based on the latest domain challenges, corroborated in recent IoT roadmapping activities that its landscape is changeable in nature [17].

Artificial Intelligence (AI)

Autonomy will bring incredible new benefits to CPS, but there are major challenges that must be overcome. The intelligence that can be applied is limited by current approaches to certification, legal frameworks and (lack of) trust for such systems. There is also a fundamental mismatch between the approach to functional safety software (top-down, correct by design) and current approaches to deep-learning programs (bottom-up, data driven), as addressed by projects such as SAFEXPLAIN [18]. These need to be addressed while maintaining and increasing the safety of such systems (which calls for improved traceability of the influences between the contributing communities to CPS). Safety of such systems is a serious challenge as the levels of reliability achievable by ML/AI are simply inadequate for high integrity systems (safety integrity levels 3 and 4) by and are an active area of research attracting investment from public and private sector. Cyber security of such systems is even greater challenge as vulnerabilities of ML/AI are significant and widely spread.

Reducing or mitigating these limiting factors will be an enabler for many advanced AI technologies related to decision making, learning etc, for the operation of the systems. In parallel, the other communities can provide more robust technologies for systems that are evolving as a result of AI. Of course, there are identified routes for AI to become "more trustable"; these include explainability of actions in human language, and the application of AI to non-safety-related aspects of CPS like decision support for system design.

A significant characteristic of CPS will be coordinated collaboration. This relates to the way components of a CPS coordinate with each other or with people for outcomes only achievable through such cooperation. AI can bring strong support

here such as through the field of decentralized intelligence called multi-agent systems (MAS) [19]. Regarding design, the needs of CPS include the explicit representation of the environment and the need to represent abstraction layers, from the physical layer to the components and system, as CPS are closely coupled to the hardware elements of the system. Finally, it may also be necessary to represent the non-functional requirements, such as safety or resilience. Some MAS design tools, such as Tropos [20], if correctly used, may help to meet these requirements.

In terms of decentralized intelligence for CPS, there are many challenges to that need to be addressed, in particular methods for executing coordination. The whole system needs to be able to react in real time, which is not the case for most decentralized AI coordination protocols, which rely on negotiation, usually with no defined deadline for decisions [20]. As another example, finding ways to work with the functional property community on communication middleware for intelligent collaboration is likely another issue to needing to be tackled.

High-Performance Computing (HPC)

High-performance computing (HPC) consists of the aggregation of powerful computing resources for solving problems that require large computing power [21]. Recently, HPC technologies were only required in the context of traditional massively parallel "number crunching" applications like weather prediction, computational chemistry, or computational fluid dynamics. However, the latest developments in low-power computing technologies [22] – required in the HPC industry to scale performance levels further – has facilitated the adoption of HPC technologies in a wide range of CPS applications.

Existing HPC platforms offer the computation capabilities needed by the most demanding CPS applications within an affordable power budget in domains such as automotive, space, avionics, robotics and factory automation. Centralized domain architectures that replace the traditional federated computing architectures – like those required by economically affordable autonomous driving systems – are only possible when HPC technologies are deployed. Single-chip high-performance embedded computing platforms reduce the traffic flow through CPS' electronic networks and enable high-speed communication as required for processing vast amounts of information in real time. So this community will be important for consolidating the embedded computing backbone.

Furthermore, these technologies involve parallel processing, that is, splitting the tasks up into parts for several computers (or multiple cores) to process, thus reducing the time taken to complete tasks. This characteristic thus holds a direct relation with the CPS challenges of decomposability and decentralization – how tasks can be split up while ensuring safety and security for people, the system and its environment.

Unfortunately, the deployment of HPC in a CPS increases the complexity of the resulting system and may have a non-negligible impact on the verification and validation costs of relevant system properties (e.g. safety and security). Thus, an effective exploitation of HPC technologies in cyber-physical applications requires at least either the development of new methodologies to verify and validate such complex systems or the adaptation of key technologies to the specific context, as explored in the EU-funded PROXIMA [23] and MASTECS [24] projects, for example.

Big Data

Cyber-physical systems are being driven by the combination of embedded and internet technologies and a vision of "smart anything everywhere" [25]. The blend of this cyber, physical (and social) data can help us to understand incidents and changes in our adjacent environments better, monitor and control buildings and urban infrastructure, and provide better healthcare and care services for older people, among many other applications. To make effective use of the physical-cyber-social data, integration and processing of data from a variety of heterogeneous sources is necessary. A key objective for big data in CPS is to analyse very large, fast, and heterogeneous data streams, mostly from industrial rather than consumer environments. This can be achieved through machine learning, which is the most common technique used to extract information from the data.

The core Big Data applications in CPS are in varied fields, including energy utilization, city management, transportation systems and disaster management. For example, a smart transportation system would generate big data consisting of drivers' behaviour, commuter information, vehicle locations, traffic-signal management, accident reporting, automatic fare calculations, and so on. Robot-aided surgical systems (i.e. human-in-the-loop CPS) comprise a teleoperation console operated by a surgeon, an embedded system hosting the control of the automated robot, and the physical robotic actuators and sensors. Big Data methods can be used here for the modelling of surgical skills, for the detection and classification of surgical motions for automation and environment, and for the integration of this knowledge into control and automation of surgical robots.

In the operation of complex systems (e.g. aircraft and industrial processes), fault-detection and fault-isolation schemes are designed to detect the onset of adverse events. Such systems use big data methods (such as machine-learning classifiers) to enhance the diagnostic accuracy of the online reasoner on board the aircraft. Moreover, big data can be utilized in command and control with cyber-physical infrastructures for emergency services and defence.

The value of the Big Data community as a contributor to CPS products can only grow in the future due to increasing interest in data as an important business asset. The combination of heterogeneous data from numerous sources will require new applications for integration, query and analysis, along with embedded computing, HPC, and data-reduction techniques. This remains an open research issue for CPS. The variety of types and sources of data will give rise to new kinds of data stores to sustain flexible data models.

Another important issue is that of remote storage of big data. Until now, cloud-based

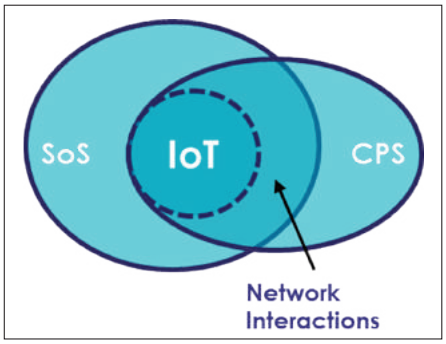


Figure 2: Technology relations of SoS, IoT, and CPS [35]

models have facilitated the storage and processing of big data sets, providing data accessibility and better IT power. However, this creates a centralized data store that does not scale in the CPS setting. To facilitate decentralized data storage and processing, a number of problems (e.g. replication, parallelism and requirements) arise. There is an urgent need for new approaches and techniques.

System of Systems

The "System of Systems" (SoS) concept has been around for at least fifty years, but in the last twenty it has been an area of major concern. Following the description of its characteristics by Maier [26]; it is defined in ISO15228 as: *"SoS…brings together a set of systems for a task that none of the systems can accomplish on its own. Each constituent system keeps its own management, goals, and resources while coordinating within the SoS and adapting to meet SoS goals*" [27]. As for CPS, SoS represents a type of application as well as a technology domain.

Broadly, one can consider SoS applications as independent systems that interoperate (work together) to achieve a purpose, with a significant amount of ubiquitous networking. In the case where they have extensive software control between safety-critical systems, the application itself is both a SoS and a CPS because they share common characteristics. Figure 2 describes the relationship between SoS, CPS, and the Internet of Things. Where infrastructure interactions are supported by internet protocol, then the CPS is also described as IoT, which is necessarily always a SoS. There are also interesting SoS-CPS applications that interact through means other than the internet protocol (e.g. mechanical or electromagnetic interactions) and the engineer may need to guard against such interactions for safety or performance reasons.

However, from the technology perspective, CPS application research considers how all technology communities are integrated to create a system and its interactions, with the SoS technology community contributing to the coordinated collaboration aspect. This is a key property for future CPS, meaning that SoS research is indispensable for creating future CPS. In relation to embedded computing, the importance of localized processing, while maintaining a connection to centralized processing capacity, is recognized as a priority in areas such as edge computing, which uses SoS technology. This also links directly with the challenge of decentralization or decomposability where systems work together. A smart city is an example of human interaction and SoS, for example; it manages busy traffic at city junctions to minimize delays for drivers and pedestrians.

In 2012, INCOSE conducted a survey to identify "pain points" for SoS practitioners, i.e. the problems that kept systems engineers and managers awake at night [28]. The study indicated seven main areas of concern: SoS authorities; leadership; constituent systems; capabilities and requirements; autonomy, interdependencies and emergence; testing, validation and learning; and SoS principles. It is no coincidence that creating CPS includes these pain points, because they are concerned with networked, intelligent systems of high complexity. This suggests that the communities of SoS and CPS have areas of common interest suitable for collaboration.

Digital Twins and the Metaverse

Human-CPS interaction will also advance with the advent of digital-twin and "metaverse" technologies [29], in particular when CPS operate in close proximity or hand-in-hand with human operators. The metaverse will provide haptic feedback over robots that complement and advance human capabilities [30]. Human operators will receive visual guidance in their view of augmented reality, and will obtain the ability to project themselves into the CPS they control. They will sense, act and interact through the impersonated system with other humans and with the environment in which the CPS operates. They will receive extended cognition and operating capabilities over swarms and manage the complexity of CPS hierarchies with ample application areas. Human caretakers may intervene in case of emergency or when service robots hit the boundaries of autonomy.

Swarms will act in harsh environments on Earth, in space and on remote celestial bodies instead of exposing humans to the risks they have to take today. Examples include mining, nuclear-waste handling and reactor deconstruction, but also asteroid mining and exploration. Replacing the internet with a network of immersive virtual worlds, cyber-physical systems will allow the metaverse to bridge into reality, with all the benefits, but also all privacy, safety and security risks this entails.

Digital twinning is one of the enabling technologies for exercising such advanced control from the digital realm over the real, physical world. Digital twins are virtual models of reality that are continually updated about the actual state of their physical counterparts and which can enable decision-making that, in turn, leads to changes in the real world. The long-term goal of digital twins is to be able to capture the intentions and objectives of the physical twin, but also to improve overall performance through digital simulation, testing and monitoring how the real-world physical system will act in its environment. While the aim is to advance into a better future, this can threaten safety and security when not handled with utmost care. Thus, it will be inevitable for the metaverse and digital-twin communities to join forces with the CPS community to achieve real-time safe, secure, and cyberattack-resilient automation from the moment metaverse-enlightened CPS are designed and throughout their lifetime.

Production-environment influencer communities

Members of the production-environment communities are responsible for the industrial product process and life-

cycle. This includes enterprise policy and processes, decisions about technology usage and the evolving physical plant [31]. They drive the large-scale production of goods using equipment in the form of modular automated product lines. Such equipment typically combines mechanical, electrical, and software components; it also requires substantial initial investment and maintenance costs. Throughout its long lifecycle (15-30 years) [31], the equipment operator and component suppliers cooperate to repair and repurpose/upgrade parts at a minimal cost. This imposes several constraints on component models and their versions, which in turn constrains policy and process management.

In addition, the arrival of digitalization and the CPS revolution brings the "servitization in manufacturing" opportunity, a paradigm shift where manufacturers shift to offer product-related services, beyond just selling a tangible asset (For further discussion of this concept and examples, see the article "Everything as a service" in this HiPEAC Vision). In the above example of automated product lines, component providers could offer monitoring, online maintenance, repair, and overhaul services [32] among other value-added services. Service contracts generate more steady revenue compared to the cyclical product business, but, in general, organizations in manufacturing struggle to drive servitization [32], because the introduction of the new services incurs higher costs without proportional returns.

The adoption of digitalization tools and solutions and the development of innovative services leveraging the full potential of CPS require incentives and coordinated efforts among different partners. Research projects, partnerships in which early movers and less-digital companies cooperate to embrace servitization and adopt CPS tools, provide a nurturing environment, where decision-makers find that the "test-before-invest" concept is an incentive that helps lower barriers and can evaluate potential benefits. For example, in the H2020 HUBCAP [33] project, less digitally focused SMEs were able to pair up with model-based design providers to adopt digital innovation and enhance their solutions using model-based design technology.

Among the success stories, there is the example of the partnership between Mototok International GmbH, a provider of innovative aircraft tug solutions, and Evitado Technologies GmbH, a provider of LiDAR-based algorithms adding advances from the self-driving car industry to an already innovative CPS product. Other examples show how advances were made in training for industry 4.0, the development of innovative organ-preservation devices in the medical domain, smart textiles, and precision agriculture.

The prime innovative aspect of HUBCAP is a web-based collaboration platform that facilitates stakeholders' access to computing resources and advanced CPS design and engineering solutions, by providing a cloud-based sandbox solution (Figure 3).

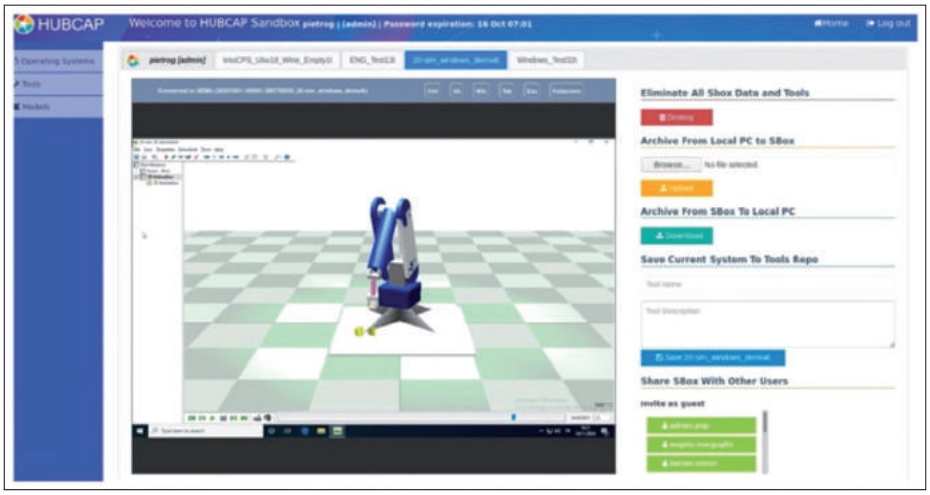


Figure 3: Snapshot from the Sandbox showing SME asset

The sandbox provides pre-installed models and tools, allowing companies to experiment with new tools and assets in a ready-to-use virtual machine available via a regular web browser, with emphasis on performance and interaction between partners. This will be taken forward and combined with DevOps capabilities, also in a digital twin setting, at http://gitworks.io/.

Production-environment community members are deeply involved with the cross-community challenges identified. There is a historical synergy with the development and advancement of embedded computing, which will continue in the future. This community is always demanding advancements in embedded computing, and advances in manufacturing also affect how we produce the embedded platforms of the future. Regarding decentralization and decomposability, there are several lessons learned and case studies in which cooperation and adaptation to local and greener processes promote research, discussion, and changes to manufacturing. Finally, this community has a particular interest in the challenge of physical collaboration with people. This interest is from both an internal perspective, covering topics such as human-machine interaction and collaborative robots, and an external perspective, where the potential for improvement from product usage data needs to be fully explored.

Market-influencer communities (society needs, regulation, standards, policy)

CPS are believed to have an enormous impact on many aspects of socio-economic life. Therefore, a number of stakeholders grouped here under the generic name of 'market influencers' will have a stake in shaping the future of CPS and of the contributing communities.

Societal needs may be described basically by means of individuals or groups putting forward requirements and benefitting from CPS. The individual appears here as the consumer who is, in one way or another, making use of either a product incorporating CPS, or elements of larger CPS implementations, addressing communities of end users in terms of mobility, personal life (general wellbeing),

healthcare, leisure, environment, etc. Other needs may be identified in the area of public services offered at local and national government level, including education, healthcare services, community services, and operation of public institutions.

As well as responding to societal needs, however, CPS also pose new challenges. Some specific fields include education and employment, as CPS induce the obsolescence of certain professions and create new ones. Therefore education, including training and retraining will be affected, as will the employability of the existing and future workforce, which will have implications for the labour market and social security.

Regulation – both hard and soft legislation - will have to be adapted in order to govern CPS so as to ensure their smooth integration into society. However, given the rapid cross-border spread of CPS technology, international agreements might be needed, too, particularly if we consider the globalised nature of today's value chains. Regulation will have to address the interplay between CPS actors (producers, consumers) as well the foreseen and unforeseen effects of the technology. Regulation is also supposed to be structured according to the societal needs that the technology is supposed to fulfil. A particular aspect of related regulation might address the human individual, chiefly in relation to human-machine interaction, which is anticipated to increase significantly in the coming years (intruding into both privacy and healthcare). The "must be implemented" regulation should be supplemented with recommendation-type measures of indicative nature.

Standards ensure interoperability and compatibility of products from different producers and allow the market presence of a large number of actors. Moreover, standards are important in order to set and describe safety levels and quality frameworks. To some extent, standards provide the technical base for legislation governing the area and also give room to innovation as usually standard specifications can be fulfilled in a variety of competing ways.

Policy aims to achieve certain results in a given field by reflecting society's needs or goals. Public policy in particular is directed towards supporting certain areas through frameworks of development in terms of tax incentives, grants or even regulation. Policy also includes public investment in facilities or processes of general interest. A further aspect for consideration is policies aiming to increase employment in a differential manner within the given population (i.e. in favour of disadvantaged groups), or to ensure development of regions lagging behind. Such policies also set out to address issues of general interest like climate change (that can only be done at international level) or the environment.

Beyond public policy, one should take into consideration policies of generically named "interest groups". Pressure groups such as non-governmental organizations (NGOs) consumer associations also have policies for their vision and procedures to supporting their realization, which can indirectly influence the market.

These "market influencer" stakeholders between them represent the conditions under which all the other communities operate for producing future CPS. The relevance of their involvement should be apparent, especially when considering the aggregative effects of contributing and cross-community technologies. Deficits in education in one community can have a knock-on effect on other communities. Training approaches and certification can be a deciding factor in the sustainability of mixed-community technologies. Policy can evolve approaches and perspectives that enhance behaviours supporting longer-term governance or culture, providing resilience, value generation and trust in new technologies.

Research orchestration for cyber-physical systems

With respect to coordinating CPS research *as an application domain*, additional approaches and orchestration should be introduced. This is because the application-domain perspective is based on the product side, with cumulative effects being considered through the aggregation of layered contributions from the stakeholder communities. Another issue is that disruptive discoveries, technologies or developments might influence the cycle of research. For example, if significant progress is made on quantum computing, or discoveries in material/biological science, that could make sensors more different.

Orchestration of research is particularly about knowledge management, longer development cycles, persistence and refinement of multi-disciplinary approaches for collaboration between communities. Take the example of constructing a building where a new team takes over every few months. Limited progress can be made without guidance at a higher level. This is similar for advancing CPS research. Persistence of acquired interaction techniques, between project collaborations, is significantly more difficult to maintain. For instance, usability and sensor experts have specific languages for their domains.

Therefore, approaches that support collaborations and which have been developed during collaborations should be taken, refined, and applied in subsequent collaborations of different groups. A dedicated CPS research instrument could advance this concept, in conjunction with future CPS support action projects. Projects themselves will also need to provide environments with favourable conditions for aggregative research considering the multi-

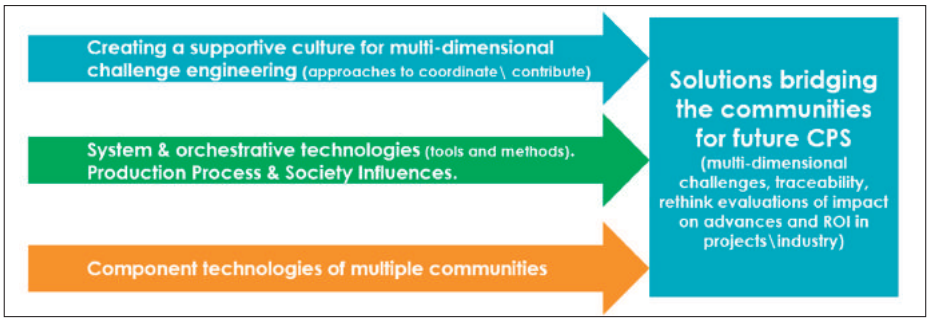


Figure 4: Stakeholder inputs to projects.

dimensional challenges, with conditions significantly different to those for developing component technologies.

Considerations for future CPS projects

For advancing CPS research as a technology domain, useful mechanisms already exist. For example, there have been projects following the standard approach, which gathers technology providers around one or more CPS-related use case. If awarded funding, the partners then work together for a few years to bring their technologies closer to market deployment (i.e. advancing "technology readiness levels" or TRL).

Cascade funding, where funded projects themselves fund smaller initiatives, has also shown itself to be a useful means for transferring component technologies for CPS, because the smaller initiatives are directly managed by companies looking for particular solutions.

However, *for the application domain side* of CPS research, new project approaches and higher support mechanisms also need to be introduced, enabling the multi-dimensional challenges previously discussed to be tackled. The characteristics that are believed to be essential in such projects are:

- Use cases: physically interactive and collaborative systems; of relevance to all communities, likely to be uniquely large industry or with integrated small-medium enterprises. Supplied also with the intention of advancing industry-side "industrial readiness levels" of production and product lifecycles for new technologies.
- CPS centre of gravity: all projects addressing the multi-dimensional challenges between communities should interface on work advancing real-time safe and secure automation, including interrelations, for CPS design and operation.
- Cross-community challenges: projects on application domain research should focus on grand challenges that need contributions from each community. Proposed call topics include:
  1) Embedded computing backbone
  2) Decentralization and decomposability
  3) Physical collaborations with people
- Developing the support environment: tools and approaches are required not only by industry, but also by researchers to support engagement of the different CPS stakeholders and perspectives. We propose that such projects include some dedicated work (a work package) that develops support for collaboration on the multi-dimensional challenges. New approaches established iteratively: orchestration approaches should be implemented in a manner that can be refined. We should avoid 'one-hit wonders' that seek to solve everything at once. A second iteration of such projects could also include smaller spin-offs and initial stage smart city investigatory projects.

Contributions of the communities to be these projects can be visualized, as shown in Figure 4, to involve the technology component providers, the influencers/aggregative technology providers and those developing the culture and support environment. This provides the means to advance the CPS aggregation techniques which are required to address the multi-dimensional CPS challenges.

These HiPEAC-proposed project characteristics (for application-domain research) relate directly to previous CPS community recommendations to the European Commission, including trustworthy and societal scale CPS, ethics data protection and liability, CPS engineering, interoperability, complexity, edge computing, humans-in-the-loop, co-engineering of system properties and enhancing uptake of CPS technologies.

Considerations for future advisory coordination and support actions for the CPS communities

The European Commission funds coordination and support actions (CSAs) to accompany, coordinate and stimulate innovation in particular technology fields and their communities. A particular challenge for a CPS CSA is that it is in fact a multi-community subject. This is because, as discussed, while CPS is *a technology domain* with specific complex challenges related to cyber and physical integration and cyber to physical plan realization, CPS is foremost an *application domain*. This is of consequence because CPS and other technologies can be much more difficult to apply to the final systems without also advancing the means for their combination.

To support application domain research projects, future CPS CSAs will likely support the transfer and synchronization of project environments, support the "big picture" metrics of aggregations in CPS and specific return-on-investment (ROI) valuation techniques to pre-empt industry needs. In particular, they will support a focal point for all the contributor technology and influencer communities.

Research instrument for technology orchestration: Supporting projects, funding programmes and industry

A team providing support across relevant research programmes, dedicated to supporting the channelling of different community contributions into safety-critical applications, in particular, would be very beneficial for supporting in particular technology orchestration/aggregation and technology uptake by safety-critical systems/CPS in general. This can only really be properly realized if this team provides an ever-present pivot for CPS projects and CSAs, developing the support environment required to manage the multi-dimensional challenges.

They would have two support roles: the development and investigation of concepts that are provided by projects and programmes that had been identified as useful support assets (to the projects), but their implementation being normally unattainable in the scope of the projects and programmes.

On the development side, support to programmes would include, as an example, enhanced tool techniques for directed communication (the right information, to the right people, at the right time – especially for start-ups). The research instrument team provides prototyping tools to relevant CSAs that would support deployment in technology orchestration projects, who then test and further develop the tools. Support for the creation and testing

of tools largely depends on results from the investigatory side. Some examples include:

- Inter-community supports like wiki-type project glossaries to manage the multiple perspectives (e.g. mediation between safety/security, medical/railway, SMEs/large enterprises).
- Multi-community access like digital passports, allowing users to access and test many research tools with the same account.
- Improved techniques like supporting management of intellectual property rights.
- Connecting contributions, such as a holistic view of open-source tool advancement across projects.

The investigatory side considers and proposes enhancements from the product-side perspective, for projects, programmes and industrial policy. These would be potential assets for promoting in particular aggregative technology uptake and longer-term profitability. Investigations would consider enhancements outside our normal fields of operation. Potential concepts include:

- Supporting the project environment for capitalization on and continuation of knowledge from multi-stakeholder interactions. Approaches for iterative improvement. Incentives, performance measures, mentoring.
- How are CPS-specific and aggregative technologies advancing, what is the funding flow to the contributing communities? Studies on benefits – but also consequences of lack of funding.
- Managed contributions, e.g. open-source results – rather than a default expectation, should be with respect to conditions (such as business model, maintenance, community building).
- Considerations for adapting the destination (industrial processes) to the new technologies; how to lift constraints at the product-side.
- Lighthouse initiatives within programmes (advancing structuring and management policies) may provide ideas to be explored.
- Currently, technology readiness levels (TRLs) measure the advancement of individual components rather than aggregations of components. A complementary approach, let us say aggregative-TRLs, is therefore required. This is not to be confused with the "integration readiness levels" measuring the interface between technologies (how they connect), rather than aggregating technologies (managing their combined effect).
- Supporting the development of a body of knowledge and teach the science of CPS engineering.
- Balancing local/national/European interests across networks. For instance, cross-border Digital Innovation Hubs (DIH) could complement the specific interests of regional or national DIHs.
- Policy on protection of EU business data (~B2B GDPR). CPS representation would be relevant here to consider the effects of such a policy on CPS technology advancement.
- Studies to advise/encourage industry towards longer-term strategies. This may also include changes in government regulation to shift from short-term competition of yearly quotas towards longer-term and more profitable competition and managing incentives where average employee turnaround is 3-4 years. No CPS-specific studies on corporate evolution seem to exist yet.

The proposed ways forward through this higher-level support from a CSA and a research instrument not only enables significant advancements for future large-scale safety-critical systems research, but also addresses the recommendations made by previous projects for CPS technology (such as Platforms4CPS [34], which represents an update of several roadmaps). These earlier recommendations included: collaboration and defragmentation of siloes; public understanding of the importance of CPS; supervisory support to draw together a common body of knowledge; and developing talent in order to maintain Europe's leadership and sovereignty of diverse technology aggregations for multi-domain applications including transport, manufacturing and health.

Acknowledgements

Contributors for the five groupings of communities

- **Rajendra Akerkar** is Professor and Head of Big Data Technologies at Western Norway Research Institute.
- **Djamila Aouada**, is Professor and Head of CVI2 Group, SnT Faculty, Université du Luxembourg.
- **Alessandra Bagnato** is Research Scientist and Head of Modelio Research at Softeam (Docaposte Group).
- **Miklós Györffi** is Senior European Affairs Analyst at the Eötvös Loránd Research Network and former staff member of the European Parliament.
- **Michael Henshaw** is Professor and Programme Director in Systems Engineering, Associate Dean for Teaching, Loughborough University.
- **Peter Gorm Larsen** is Professor and Head of the DIGIT Centre, Department of Electrical and Computer Engineering, Aarhus University.
- **Carles Hernandez** Luz is Senior Researcher in Processor Designs for Safety-Critical Systems, Universitat Politècnica de València
- **Hugo Daniel Macedo** is Researcher in the DIGIT Centre, Department of Electrical and Computer Engineering, Aarhus University.
- **Claudio Pastrone** is Head of IoT and Pervasive Technologies Research Area in LINKS Foundation.
- **Peter Popov** is Associate Dean (International), School of Mathematics, Computer Science and Engineering, City University London.
- **Claudio Sassanelli** is Professor Department of Mechanics, Mathematics and Management, Politecnico di Bari.
- **Marcus Völp**, is Professor and Head of CritiX Group, SnT Faculty, Université du Luxembourg.
- **Thorsten Weyer** is a Professor of software engineering, Technical Hochschule Mittelhessen (THM), Giessen.

References

[1] S. Paul, "D3.4.4. MERgE - Recommendations for Security and Safety Co-engineering v3 partA," 22 April 2016. [Online]. Available: https://itea4.org/project/merge.html. [Accessed 2 December 2022].

[2] "RITICS: Research Institute in Trustworthy Interconnected Cyber-Physical Systems," [Online]. Available: https://ritics.org/. [Accessed 2 December 2022].

[3] "ICRI_SAVe," Intel, [Online]. Available: http://icri-save.de/the-institute/. [Accessed 2 December 2022].

[4] "IET Code of Practice: Cyber Security and Safety," Institution of Engineering and Technology, [Online]. Available: https://electrical.theiet.org/guidance-codes-of-practice/publications-by-category/cyber-security/code-of-practice-cyber-security-and-safety/. [Accessed 2 December 2022].

[5] "MERgE: Multi-Concerns Interactions System Engineering," ITEA 4, [Online]. Available: https://itea4.org/project/merge.html. [Accessed 2 December 2022].

[6] "SESAMO: Security and Safety Modelling on CORDIS," European Commission, [Online]. Available: https://cordis.europa.eu/project/id/295354. [Accessed 2 December 2022].

[7] "AQUAS: Aggregated Quality Assurance for Systems," [Online]. Available: https://aquas-project.eu/. [Accessed 2 December 2022].

[8] "CPS Swarm," [Online]. Available: https://www.cpswarm.eu/. [Accessed 2 December 2022].

[9] SAE International, "Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles J3016_202104," 30 04 2021. [Online]. Available: https://www.sae.org/standards/content/j3016_202104/.

[10] International Organization for Standardization,, "ISO 21448:2022 Road vehicles – Safety of the intended functionality," 2022. [Online]. Available: https://www.iso.org/standard/77490.html.

[11] M. Roberti, "What Is the Internet of Things?," RFID Journal, 2016. [Online]. Available: https://www.rfidjournal.com/question/what-is-the-internet-of-things. [Accessed 4 December 2022].

[12] "Internet of Things 08: International Conference for Industry and Academia," Zurich, 26-28 March 2008. [Online]. Available: https://iot-conference.org/iot2008/prg/program.pdf.

[13] H.Boyes et al, "The industrial internet of things (IIoT): An analysis framework," Computers in Industry, vol. 101, pp. 1-12, https://www.sciencedirect.com/science/article/pii/S0166361517307285, 2018.

[14] K.H. Wöhnert et al., "Secure Cyber-Physical Object Identification in Industrial IoT-Systems," Procedia Manufacturing, vol. Volume 51, pp. 1221-1228, https://doi.org/10.1016/j.promfg.2020.10.171, 2020.

[15] S. Sachdev et al., "Voice-Controlled Autonomous Vehicle Using IoT," Procedia Computer Science, vol. Volume 160, pp. 712-717, https://doi.org/10.1016/j.procs.2019.11.022, 2019.

[16] A.J. Ferrer et al., "Admission Control for Ad-hoc Edge Cloud," Future Generation Computer Systems, pp. 548-562, https://doi.org/10.1016/j.future.2020.08.024, 2021.

[17] E.M. Jakobsen et al, "NGIoT D4.6: Report on IoT business model innovation patterns and acceleration support activities," 30 September 2022. [Online]. Available: https://www.ngiot.eu/wp-content/uploads/sites/73/2022/09/EU-IoT_D4.6-IoT-BM-patterns-and-acceleration-support.pdf. [Accessed 4 December 2022].

[18] "SAFEXPLAIN: Safe and explainable critical embedded systems based on AI," CORDIS, European Commission, [Online]. Available: https://cordis.europa.eu/project/id/101069595. [Accessed 4 December 2022].

[19] M. Wooldridge, An Introduction to MultiAgent Systems, 2nd Edition, Hoboken, New Jersey: Wiley, May 2009.

[20] C. Cares et al, "Agent-Oriented Engineering for Cyber-Physical Systems," ICITS 2019: Information Technology and Systems in Advances in Intelligent Systems and Computing, vol. 918, pp. https://doi.org/10.1007/978-3-030-11890-7_10.

[21] "High Performance Computing," European Commission, [Online]. Available: https://digital-strategy.ec.europa.eu/en/policies/high-performance-computing. [Accessed 4 December 2022].

[22] N. Rajovic et al, "Supercomputing with commodity CPUs: are mobile SoCs ready for HPC?," in SC13: International Conference for High Performance Computing, Networking, Storage and Analysis, Denver, Colorado, 2013, https://dl.acm.org/doi/10.1145/2503210.2503281.

[23] " Probabilistic real-time control of mixed-criticality multicore and manycore systems (PROXIMA)," [Online]. Available: http://proxima-project.eu/. [Accessed 4 December 2022].

[24] "MASTECS: Multicore Analysis Service and Tools for Embedded Critical Systems," [Online]. Available: https://mastecs-project.eu/. [Accessed 4 December 2022].

[25] "Smart Anything Everywhere," European Commission, [Online]. Available: https://smartanythingeverywhere.eu/. [Accessed 4 December 2022].

[26] M. W. Maier, "Architecting principles for systems-of-systems," Systems Engineering, vol. 1, no. 4, pp. 267-284, 1998.

[27] "ISO/IEC/IEEE 15288:2015 Systems and software engineering – System life cycle processes," May 2015. [Online]. Available: https://www.iso.org/standard/63711.html. [Accessed 4 December 2022].

[28] J. Dahmann, "1.4.3 System of Systems Pain Points," in INCOSE International Symposium, Las Vegas, Nevada, 2014, https://doi.org/10.1002/j.2334-5837.2014.tb03138.x.

[29] S-V Rehm et al, "The Metaverse as Mediator between Technology, Trends, and the Digital Transformation of Society and Business," Journal of Virtual Worlds Research, vol. 8, no. 2, pp. 1-6, https://doi.org/10.4101/jvwr.v8i2.7149 , 2015.

[30] Z. Sun et al, "Augmented tactile-perception and haptic-feedback rings as human-machine interfaces aiming for immersive interactions," Nature Communications, vol. 13, pp. 5224, https://doi.org/10.1038/s41467-022-32745-8, 2022.

[31] S. Braun et al, "Requirements on Evolution Management of Product Lines in Automation Engineering," IFAC Proceedings Volumes, vol. 45, no. 2, pp. 340-345, https://doi.org/10.3182/20120215-3-AT-3016.00060., 2012.

[32] M.M. Herterich, "The Impact of Cyber-physical Systems on Industrial Services in Manufacturing," in 7th Industrial Product-Service Systems Conference, Saint-Étienne, 2015.

[33] P.G. Larsen et al, "A Cloud-Based Collaboration Platform for Model-Based Design of Cyber-Physical Systems," in SIMULTECH 20: 10th International Conference on Simulation and Modeling Methodologies, Technologies and Applications, Online, 2020, https://www.scitepress.org/Link.aspx?doi=10.5220/0009892802630270.

[34] "Platforms4CPS," [Online]. Available: https://www.platforms4cps.eu/. [Accessed 4 December 2022].

[35] M. J. d. C. Henshaw, "Systems of Systems, Cyber-Physical Systems, The Internet-of-Things…Whatever Next?," INSIGHT, vol. 19, no. 3, pp. 51-54, 2016.

Charles Robinson is the research projects leader in critical embedded systems at Thales Research and Technology, France

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2023.

Cite as: C. Robinson. Bridging the stakeholder communities that produce cyber-physical systems. In M. Duranton et al., editors, HiPEAC Vision 2023, pages 32-43, Jan 2023.

DOI: 10.5281/zenodo.7462013

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

© HiPEAC 2023

Our world is becoming dependent on cyber-physical systems (CPS). This provides unprecedented innovation opportunities, but also represents unprecedented complexity. We need to design future CPS to make sure that they become human-centred and part of a circular economy.

Towards a living dimension: the future of cyber-physical systems

By FRANCESCA PALUMBO, RAQUEL LAZCANO and DANIEL MADROÑAL

Over the last few years, society has been drastically affected by the COVID-19 pandemic and other major events, such as the war in Ukraine. The whole world has been affected by these events and, subsequently, technology is also on the verge of a (yet another) revolution.

Without doubt, humans have gained a central role in information and communication technology systems and are requiring these systems to be more advanced and autonomous than ever. But is it just a matter of having humans at the centre of technology, or is the actual human-system relationship going to change? Are systems ready to team up with humans in a living dimension, creating a new form of society? More importantly, if not, how will they need to evolve to achieve this?

Key insights

- Expectations regarding CPS have changed: CPS are moving towards a living dimension. Future CPS will become capable of collaborating with humans towards a common goal, blurring the boundaries between CPS and humans.
- Classic CPS see humans mostly as operators or the physical process to be controlled. Several issues need to be addressed in order for humans and CPS to form part of the same society.
- The key enabling technologies required to power a new generation of CPS are as follows: artificial intelligence (AI), swarm intelligence, the tactile internet, and the computing continuum. These enabling technologies will contribute to increased decision-making and coordination capabilities of future CPS.
- Challenges include aspects relating to modelling, architectures, design, and security. These are already intrinsic challenges of this domain, and which are now exacerbated by the living dimension and the presence of the humans at the centre of the paradigm.
- By definition, future CPS will be increasingly autonomous and will be of an evolvable nature, aspects that generate significant challenges for CPS scientists.

Key recommendations

- Prioritize research on AI following an integral strategy. AI at the edge and cloud intelligence should be effectively brought together to allow rapid feedback processing and independent single element evolution, while maintaining the full picture.
- Exploit AI to make models flexible enough to encompass for post-deployment system evolution to cope with emergent situations and to adapt to unpredictable human behaviours.
- Develop new low-latency, multi-modal, reliable communication architectures following the tactile internet paradigm to allow overcoming human-to-system traditional command-based interaction and achieve a more natural peer-to-peer seamless interaction.
- Continue investing in swarm intelligence studies and begin testing embodied CPS swarms in complex large-scale, realistic environments with a high degree of heterogeneity to put the basis for forming a heterogeneous society of CPS collaborating with humans.
- Extend threat modelling for embodied swarm-based CPS systems tackling the needs and effectively protecting the heterogeneous society formed of humans and CPS influencing each other.
- Leverage and research on effective interoperability of existing design frameworks, considering humans at every stage of CPS design, and ensuring there is continuous verification of system properties during the design process.

Human-triggered technological shift

Over the last few years, the world has witnessed major events including the escalating impact of the climate change, the war in Ukraine and the COVID-19 pandemic. The full impact of these events on life in general, and in technology in particular, remains to be seen.

Taking the COVID-19 pandemic as an example, it has already had an impact on human lives, supply chains, work, and markets. Forced social distancing and a significant reduction of human presence on work sites resulted in major investments in tools to support remote work and monitor industry verticals (e.g. manufacturing, transportation and logistics).

This, in turn, has led to extreme digitalization and virtualization that, in the information and communication technology (ICT) domain, is the continuation of a journey started decades ago. Terms like disappearing computers, ubiquitous computing and ubiquitous information, pervasive computing, embedded systems and "smart anything everywhere" are in common use today. The mobile internet has been flanked by the internet of things (IoT) and, as discussed later in this article, the tactile internet is just around the corner.

All of this points to a new technology revolution, where the computing domain is so entangled with humans that a new paradigm is necessary to redefine this relationship. Multiple examples can be found in our daily lives, hinting at a not-so-distant future where the pervasive usage of cyber-physical systems (CPS) and human-computer interaction reach unprecedented levels: autonomous driving, smart cities or remote and virtualized healthcare, to name just a few.

The flagship of all these examples, the one that entangles all of them simultaneously, is the so-called *metaverse*. While the term was first used in 1992, recently it has gained significant momentum, since it has started to be closer to a reality than a chimera. Although a single definition has yet to be agreed upon in the literature, it can be defined as *"an integrated immersive ecosystem where the barriers between the virtual and real worlds are seamless to users, allowing the use of avatars and holograms to work, interact and socialize via simulated shared experience*" [1].

While this is just one example illustrating where the technology is going and what a possible future for CPS could look like, it highlights many of the issues we have mentioned: CPS, now more than ever, are human-centred. They form the basis of the creation of a virtual and remote community where humans can interact, work and, eventually, live.

However, although "futuristic" concepts such as the metaverse are now closer than ever, the technology in general, and CPS in particular, still need to evolve before such a future becomes attainable. This evolution needs to take place not only from the scientific perspective, but also from the societal one, where many issues are appearing as a result of an ever more digitalized society [2]. As mentioned above, a new paradigm is emerging, one that will revolutionize the way CPS are defined, modelled and realized, and one that needs to respond to these many issues.

In this article, we present our vision of how classic CPS, through some key enabling technologies (KETs), will evolve in order to cope with a future where the borders between cyber and physical parts will be more blurred than ever, reshaping their definition and paving the way towards a *new paradigm where CPS will have a living dimension and will not only interact with humans, but team up with them in a new form of society.*

Classic cyber-physical systems

In 2017, the United States National Science Foundation (NSF) Cyber Physical Systems Group defined CPS as *"engineered systems that are built from, and depend upon, the seamless integration of computational algorithms and physical components*" [3].

As depicted in **Figure 1**, CPS are closed-control loop systems, where the physical part, representing the environment, is observed through sensors. Observations are then used by the cyber part, which processes the information gathered to control and/or provide feedback to the physical part by means of actuators and/or proper interfaces. Sensors and actuators represent the boundaries, crossed by data only, between the physical and the cyber parts.

Data can be processed/stored at the edge or in the cloud. Processing/control is typi-

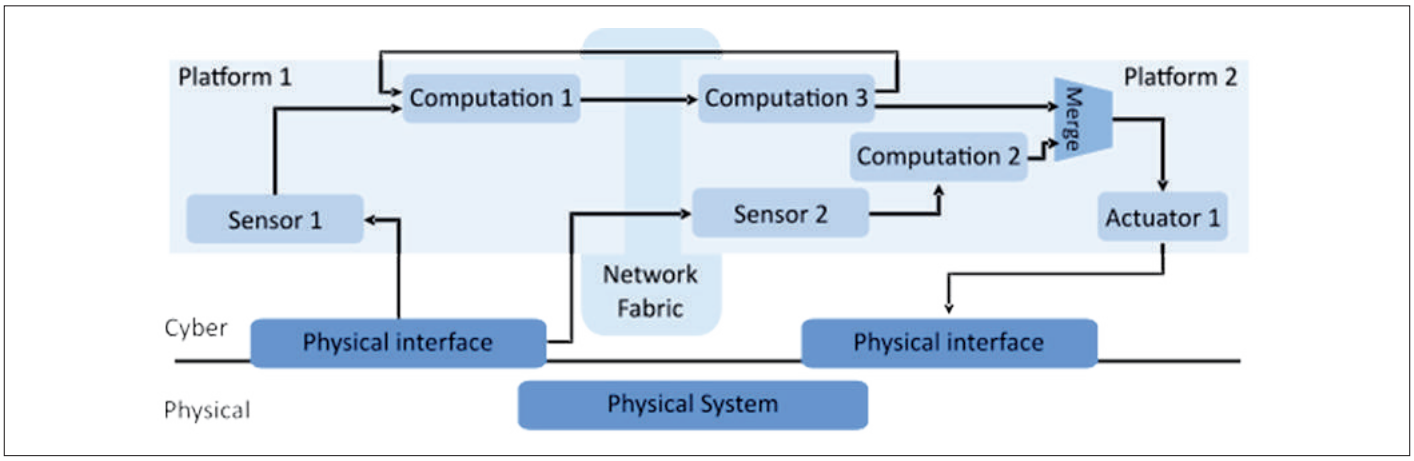


Figure 1: CPS overview (authors' art)

cally smart, although this does not necessarily imply artificial intelligence (AI); smartness can also be achieved through runtime modelling and prediction techniques to understand the current physical system condition and to react appropriately, adjusting (when and where needed) the computation and/or providing a different service.

Today, the cyber part is rarely a monolithic computing item. Indeed, it is more often composed of interconnected processing units. Therefore, the network fabric and connectivity devices play a key role in the cyber part of the CPS.

Often CPS are described in terms of views or layers, namely:

- *physical* – the environment/process that the CPS controls;
- *computational (or functional)* – the composition of the system in terms of computing resources and the algorithms executed;
- *communication (or network)* – interaction among components and exchange with the physical environment.

Going back to the NSF definition, it continues as follows: "*Advances in CPS will enable capability, adaptability, scalability, resiliency, safety, security, and usability that will far exceed the simple embedded systems of today. CPS technology will transform the way people interact with engineered systems – just as the Internet has transformed the way people interact with information.*"

The CPS available today have certainly made good on this promise. Nevertheless, expectations for the role of humans now go far beyond simple interaction.

Humans in CPS have been always seen either as operators or as the controlled physical system. Steps to bring humans closer to machinery have already been taken in the manufacturing sector with the concept of human-CPS (HCPS) that emerged thanks to advances in IoT and AI combined with CPS. However, the role of a human in HCPS, according to the authors of [4], is still seen as the master (the creators, managers, and operators of intelligent machines) rather than a teammate.

Now, the aim is for humans to act as partners of the CPS in the realization of a common goal. This poses some challenges:

- It is difficult to predict human behaviour and model human decisions. Indeed, starting from an initial assumption, the system will have to learn and adjust as it is used, in order to adapt to mutable human will.
- Dependability issues become more severe: in addition to protecting sensitive data, guaranteeing their privacy and security, safety mechanisms should also be present.

Addressing these challenges will bring CPS towards a new living dimension. The market, accelerated by the many verticals demanding effective system/process virtualization and autonomy, is pushing for this, and it is extremely dynamic. According to a recent analysis by Data Bridge Market Research [5], it will reach more than US$ 12 million by 2028, with a Compound Annual Growth Rate (CAGR) of more than 10% each year.

Increased penetration of the internet and AI are essential enablers for this market and will contribute, in our vision, to extend the NSF classic CPS definition as follows: **increased decision-making and coordination capabilities will make it possible for future generations of CPS to autonomously collaborate with their human teammates and with each other. They will form a kind of society with unprecedented cognitive and coordination capabilities, taking an additional step forward in the way people interact with engineered systems and blurring the boundaries between the CPS and its operators.**

Key enabling technologies

Future CPS with a living dimension will be an evolution of classic CPS. They will be smart systems capable of making decisions in an autonomous manner, interacting with human beings, and collaborating as a group among each other and with humans. This expected evolution of CPS will undoubtedly be determined by the two major market drivers: connectivity and AI.

In our vision, four main KETs are driving this revolution: on the one hand, AI for decision making, the tactile internet for interaction, and swarm intelligence for collaboration; on the other hand, the computing continuum, which is both an enabler and a requirement to bring all the aforementioned functionalities together.

Artificial intelligence

The global AI market size was valued at US$ 93.5 billion in 2021 and is projected to expand at a CAGR of 38.1% from 2022 to

© Adobe Stock

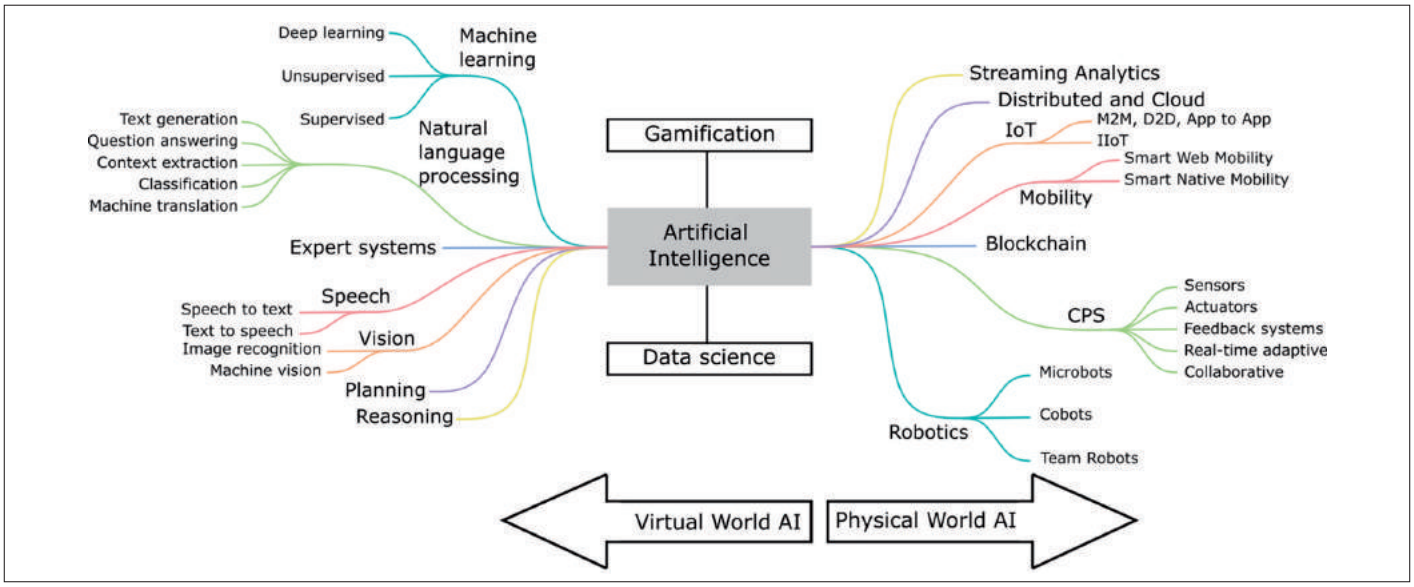


Figure 2: Virtual and physical worlds in AI [10]

2030 [6]. Research and innovation directed by tech giants including, among others, Google, Microsoft, Apple, and IBM, are driving the adoption of advanced technologies in industry verticals, such as automotive, healthcare, and manufacturing. For instance, following the acquisition of the Mobileye autonomous vehicle business in 2017 and of the AI chipmaker Habana in 2019, Intel Corporation has acquired also Cnvrg.io and SigOpt, providing a platform for data scientists to build, train, compare and run machine learning models and an optimization platform to run models and simulations.

These acquisitions testify to the fact that remaining competitive in the market of next-generation chips is not just a matter of pushing a new efficient architecture onto the market, but also requires investment in tools for customers to support the definition of the loads running on the new chips.

AI-enabled CPS are systems that, generally speaking, adopt some kind of intelligence somewhere in the product, meaning either at the *virtual* and/or at the *physical* level, as shown in **Figure 2** which depicts different crosscutting virtual technologies (on the left) applicable to various physical systems and application domains (on the right).

For example:

- AI controllers (e.g. based on supervised learning and reinforcement learning) are virtual enabling technologies to be used in place of more traditional controllers (e.g. based on proportional integral derivative control or model predictive control) to control physical processes [7].
- Deep reinforcement learning methodologies are virtual enablers for physical autonomous indoor navigation systems [8].
- On physical factory floors, the use of smart machinery relying on advanced human-machine interfaces, based on virtual AI-enabling technologies such as speech recognition, might dramatically improve the safety of machine operators [9].

There is undoubtedly plenty of hype around AI. According to a recent study focused on AI-enabled CPS in the chemical domain [11], in the last 10 years the keywords "*cyber-physical system*" and "*artificial intelligence*" have been combined in 447 articles. Between the years 2019 and 2021, a significant increase in publications occurred in this area, accounting for 71% of the total publications on this topic in the last 10 years.

There has been also a proliferation of models for AI applications. New, more accurate models are proposed every year. Nevertheless, such increased accuracy generally does not come for free: network size and corresponding computational costs might be extremely high as well, as depicted in **Figure 3** [12].

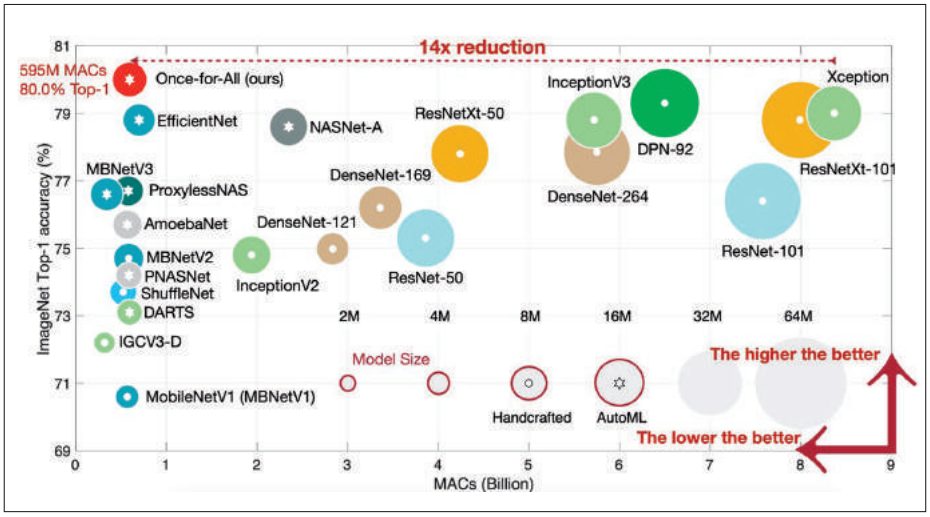
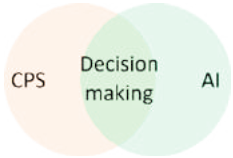


Figure 3: Accuracy versus complexity in ImageNet Top-1 network implementations [12]

AI has given rise to new challenges on how to efficiently deploy those demanding networks on diverse hardware platforms to meet different hardware efficiency constraints (e.g. latency or energy). The many challenges AI faces towards sustainability are thoroughly analysed in the 2021 HiPEAC Vision [2], where special attention is given to the growing ecological costs of AI versus the potential ecological benefits, such as those brought about by fields as smart agriculture.

The interaction of CPS and the environment allows these systems to obtain many useful data, which could potentially improve the knowledge the system has of the environment, steering positive behaviours, actions and control over it. Nevertheless, the availability of data does not automatically entail the ability to understand. This is where AI represents an unquestionably powerful instrument to boost data reasoning.



Tactile internet

The number of IoT devices worldwide is expected to increase from 9.7 billion in 2020 to more than 29 billion in 2030. Such devices are used in all types of verticals, with the consumer accounting for 60% in 2020 [13]. With these numbers in mind, it is clear how the collective ability to share information and cooperate has become extremely high.

The mobile internet gave people the opportunity to exchange data and multimedia content on the move. Next, IoT-enabled networking smart devices resulted in collections of interconnected computing devices constantly exchanging data.

The new frontier is known as the tactile internet [14][15]. It intends to enable the *"control of the IoT in real time, also adding a new dimension to human-to-machine interaction by enabling tactile and haptic sensations"*. With the shift of ICT infrastructures towards the edge, it is understandable how the role of the network is of paramount importance. Speed, capacity, bandwidth, throughput, security, resiliency and low latency, already factors under scrutiny with the emergence of the IoT, are becoming increasingly important as the tactile internet gains momentum.

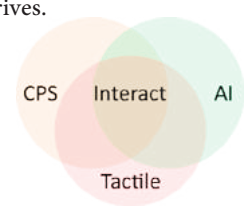
In many cases, tactile internet scenarios have the characteristics of CPS. A representative example of a tactile internet application with all the key CPS ingredients would be a remote surgery operation [16]:

- Humans are in the loop representing, respectively, the operator (the surgeon) and the physical view (the patient) of the CPS.
- The scenario is characterized by dynamic and reactive behaviours. Bidirectional haptic communications are intended to guarantee real-time exchanges of the control commands, coming from the operator, and the haptic feedback, from the teleoperator robot. Based on the latter, operator decisions can change and, consequently, the robot set-up can also be adjusted.
- A reliable networking infrastructure represents the communication layer, and dedicated edge-based computing architectures support smart real time processing on both the operator and patient sides.

Bearing in mind that all the requirements for IoT (such as guaranteed availability, the coexistence of different media, and security) still stand, on top of the aforementioned CPS-oriented characteristics, the tactile internet imposes also [16, 17]:

- Hard real-time processing.
- Ultra-low latency with a round trip in the order of 1 ms.
- High reliability, normally indicated by a very low failure rate (10^{-3} to 10^{-7}).
- Multi-modal bilateral communications, including visual, auditory and haptic feedback.

Allowing humans to be part of the system itself and not restraining them to giving commands and receiving an output would boost human-machine interaction possibilities. Following the tactile internet paradigm, these interactions are evolving towards a safe, reliable and effective human-to-CPS peer-to-peer relationship, unlocking their collaboration within the same working area or even from distant locations. AI at the edge is necessary to achieve the ultra-low latency imposed by the tactile internet, performing on-the-fly predictions of the expected action/reaction and refining the stimulus once the real data arrives.



Swarm intelligence

In nature, swarms are groups of individuals that coordinate among each other to meet a global goal, typically unreachable by a stand-alone entity. Swarm members

© Adobe Stock

follow a set of relatively simple rules and interact locally with other individuals and with their environment. Swarm intelligence exploits the understanding of the conditions, rules and interaction patterns defining swarm behaviour to design artificial systems exhibiting effective swarm behaviours [18].

Traditionally, swarm intelligence has been largely adopted to design successful algorithms for optimization, networking, and decision-making problems, as in the case of particle swarm optimization (PSO) [19] and ant colony optimization (ACO) [20]. More recently, the concept of swarm robotics has emerged. This uses relatively simple, physically embodied agents to accomplish the given swarm mission in many different contexts, including marine applications (see for example Apium Swarm Robotics [21]) or smart agriculture (see for example SwarmFarm Robotics [22]).

The market of swarm-intelligence-based systems was valued at US$ 21 million in 2021 and is expected to register a CAGR of over 37% during the period 2021-2026 [23]. So far, the major driver for this market has proved to be the ability of swarm intelligence to help solve complex optimization problems (i.e. moving goods while minimizing operational costs and travel time, or enabling business teams to generate highly accurate financial forecasts). The adoption of embodied swarms in critical missions is becoming another major contributor to the growth of this market, especially in the United States, due to the increased adoption of swarm-based drones in defence [24].

A collection of cooperative embodied robots is a swarm if it exhibits the following abilities:

- **adaptability**: it can adapt to dynamic environments and cope with different tasks;
- **robustness**: it can cope with disturbances and failures, such as the loss or the malfunctioning of individual agents;
- **scalability**: it performs well with varying numbers of swarm members and varying problem sizes.

© Adobe Stock

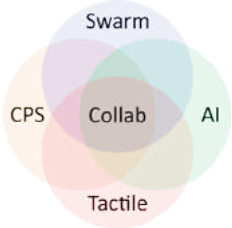
The number or density of agents plays an important role in swarm performance. A swarm system cannot operate below a given critical mass, which is the minimum number of elements for the cooperation to give positive effects.

Finally, swarms are expected to present super-linear performance scaling: the effect of the overall system is required to be more than the sum of the effects of its individual parts. If the synergies of cooperation boost each individual swarm members' performance, then the overall system is a well-designed swarm. This phenomenon is known as the "swarm effect" and it means that, within the bounds of feasible swarm sizes, not only the efficiency of the whole swarm as a group, but also the efficiency of each individual should increase.

As discussed in [25], embodied swarms and CPS resemble each other. Indeed, a CPS can be composed of distributed elements, composed of subsystems presenting local coordination and exposing partial autonomy. They can leverage distributed control, supervision and management and can be capable of (self-)reconfiguration at different timescales.

These characteristics allow both the evolution of the overall system during its operation and the emergence of useful behaviours. Clearly, for a CPS to be classified as a swarm it should also respond and adapt to changing conditions, in a robust and scalable manner, and its mission should be accomplished by harnessing the local cooperation of its individual composing elements thanks to the swarm effect.

Including the swarm intelligence concept in a system composed of CPS that understand their environment thanks to AI and that can interact with humans thanks to the tactile internet concept completes the system with a sense of society and collaboration capabilities. Having (partially) autonomous, highly distributed and cooperating CPS collaborating with humans would result in superlinear performance improvements for the system, as long as the provision of swarm intelligence is part of system design from the outset.



Computing continuum: from edge to cloud and vice-versa

Edge computing is defined as the *ability to process data close to sensors*. Processing takes place locally instead of transporting raw data to the cloud or to remote data centres. According to the International Data Corporation, by the end of 2023 more than half of new enterprises' ICT infra-

structure will be at the edge [26], while Gartner has predicted that, by 2025, 75% of enterprise-generated data will be created and processed outside a traditional data centre or cloud [27].

There are many widely acknowledged advantages associated with edge processing:

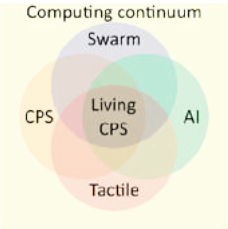
- reduction in response latency (data-to-result processing time), saving both the time required to transmit the data and the time required to receive the result;
- savings in bandwidth, energy and costs related to data transfer, as well as the cost of storing the raw data, which can be larger than the processed results;
- increased system robustness to network failures, reducing the occurrence of service interruption, as long as rendered locally;
- increased data security and privacy.

Despite all these pros, designing edge-based solutions faces important resource-scarcity challenges, for computing, storage, and energy.

These considerations lead to an approach where edge and cloud go hand in hand, coordinated by effective synchronization. Cloud computing (public, hybrid or private) offers valuable centralization capabilities: if all the data are in the same place, any entity in the enterprise can access them where and when needed. Edge allows specific savings and makes the data collected immediately actionable, creating value for applications that gain advantage from real-time analytics, IoT devices and rapid decision-making. The two extremes should be considered as a continuum: "a successful strategy begins with a cohesive plan for managing the data, including both processing and transmission, as well as the overall infrastructure" [28].

In fact, the concept of the computing continuum has been extensively discussed in HiPEAC Vision roadmaps for the last five years, reaching a culmination in the emergence of a new paradigm: the Trans-Continuum, where "edge, fog and cloud computing platforms are being pulled closed together into what will likely become a seamless execution environment", and where "programming has to be reinvented, with languages and tools to orchestrate collaborative distributed and decentralized components, as well as components augmented with interface contracts covering both functional and non-functional properties" [29].

The concept of the computing continuum (or, more ambitiously, the Trans-Continuum) is both an enabler for and a requirement of the rest of the KETs in our vision of future CPS enriched with a living dimension. The computing continuum allows the data collected by the system, by means of environmental sensors and/or tactile internet-enhanced human-machine interfaces, to be stored and employed when needed. An integral continuum approach means that individuals can refine their own local rules by running AI at the edge. Conversely, the cloud can re-tune the overall mission tasks and/or individual behaviours based on the information of all the nodes within the swarm to avoid divergence from the global challenge or to adapt to changing conditions.



Towards a new living CPS dimension

Smart everything everywhere is already a reality, given that AI-enabled technologies are embedded in everyday objects. The world is becoming truly collective and connected, increasingly featuring systems that integrate large numbers of multiple interacting components at different scales. In addition, thanks to the tactile internet, it should be possible to change the way humans interact with those systems. Ultimately, the goal of this "internet of everything" is to bring together people, processes, data, and devices.

Classic CPS, powered by AI and the tactile internet, can make substantial steps towards becoming living systems. By definition, living systems are "*open, self-organizing systems that have the special characteristics of life and interact with their environment. This takes place by means of information and material-energy exchanges*" [30]. Classic CPS are open by definition due to their ability to actively, and potentially continuously, interact with their environment through sensors and actuators. Additionally, AI can provide CPS with unrivalled understanding capabilities that allow content-rich information exchange, not only with other CPS, but especially with humans.

The introduction of the tactile internet is intended to enable a shift towards enhanced *peer-to-peer relationships and more effective exchanges*, since humans will be capable of interacting with the CPS using their senses and not only through a data/command exchange.

Indeed, as for living systems, the *equilibrium of CPS is dynamic*. CPS are by definition dynamic and reactive systems. The addition of AI and the tactile internet is expected to *dramatically improve the continuously interacting and understanding capabilities* of a CPS, while also unlocking the potential for *evolution* and *self-organization*. Meanwhile, the ability to *self-organize* is provided by swarm intelligence, which also allows better understanding of the *emergent behaviours* that might arise as an outcome of the *complex dynamics* that will characterize these future CPS. As with natural systems, these dynamics will feature both deterministic and stochastic components.

This evolution will transform a group of highly interconnected smart CPS into something closer to a living computing system, composed of *individuals that are capable of reasoning and communicating with each other to team up and reach overarching goals*. Despite such disruptive potential, this new generation of CPS also throws up new challenges, as discussed below.

Open challenges

Existing challenges in the CPS community, such as modelling, design, architecture, and security issues, are exacerbated by the living, human-centred dimension targeted by the new generation of CPS. Moreover,

these future CPS pose additional challenges related to the self-evolving nature of a heterogeneous society composed of systems and humans, pushing the concept of autonomic systems to the extreme.

Modelling

CPS are hybrid by nature. Their different aspects obey diverse dynamics, including continuous and discrete ones, threatening the established separation of concerns approach. Interactions among aspects are not considered in the models. For instance, traditional application models are basically intended to describe functionalities, disregarding timing aspects or what happens when the application is physically implemented on a given computing platform.

Attempts to capture the dynamic and reactive nature of CPS have been made with respect to adaptivity [31] and timing concerns [32, 33]. However, it is yet to be demonstrated whether and how these attempts would withstand the presence of the additional living dimension, since the swarm effect may affect the model of individual elements. Modelling swarms implies the definition of the local rules, which are always complex, especially when heterogeneous swarms composed of both CPS and humans are to be considered.

The expected human-to-system entanglement in future CPS severely challenges system modelling. Human in the loop, in general, implies an abstract modelling of the human based either on interviews of the operators, employees, etc., or on data from real use. In both cases, more flexible modelling will be needed, to make models evolve over time to fit humans' unpredictable and changeable behaviours.

Architectures

Edge computing is already a key factor when implementing CPS, since flexibility, low energy consumption and high-performance capabilities are already a must. These aspects are even more critical considering the evolution of CPS to include AI, the tactile internet and swarm intelligence.

This evolution forces edge devices to run AI and swarm intelligence algorithms, which normally require the design of application-specific accelerators to be run on hardware substrates [34]. This strategy still requires manual tuning to comply with target-dependent trade-offs between accuracy, computational complexity, and performance requirements.

Furthermore, as mentioned previously, this evolution also needs to encompass the computing continuum and the tactile internet, which both have tight constraints in terms of communication performance. Therefore, in addition to the challenges of computation, the communication architecture should also be taken into consideration, including mesh networks to improve resiliency and robustness.

Security

The concept of societal security, known as the "*ability of a society to persist in its essential character under changing conditions and possible or actual threats*", dates back to 1993 and was developed by the Copenhagen School of security studies [35]. As future CPS will form a sort of "society", security will certainly be a matter of concern for them, as it currently is for any kind of system [36].

Future CPS, exploiting swarm-intelligence-based coordination concepts, will involve non-hierarchical and decentralized control structures that are prone to security issues related to confidentiality, integrity, authentication/identification, and availability [37]. Comprehensive threat modelling, coupled to appropriate mitigation measures to reduce the overall risk or impact of a given threat, is needed both for CPS [38] and for embodied swarm-based systems [37].

Although including the computing continuum forces us to consider security issues along the whole spectrum, it also provides future CPS with strategies to cope with security threats, for example by processing attack-prone data at the edge.

Design

CPS design is already intrinsically complex due to the hybrid and interdisciplinary nature of these systems, which leads to a reliance on partially integrated toolchains and methodologies tuned for specific design aspects. Adding the described living dimension will exacerbate such complexity. Indeed, for the described future CPS, it will become essential to consider aspects such as swarm behaviours (local rules, entity cooperation), extremely low latency (required in the tactile internet) and AI decision making (having to deal with both the algorithm definition and its effective porting onto the chosen, potentially constrained, platform).

In this new approach, the idea of humans being part of the system must be considered during every aspect of the design. In addition, the design must address how to communicate and distribute the workload along the computing continuum.

Model checking is an open issue for both classic CPS and embodied swarms of CPS. This is going to be exacerbated for future CPS with humans at the centre, since the presence of humans poses serious challenges in terms of dependability, requiring in turn continuous verification of the system properties with runtime feedback loops to redesign cycles.

To cover all these complex design steps, there are already very good examples of frameworks for CPS design and operation [39] [40], for swarms of CPS design and operation [41], and for porting AI at the edge [42]. Indeed, as for classic CPS, the major issue would be how to get the most from existing frameworks. Interoperability among tools, despite the big promises of meta-model [43] and semantic [44] approaches, is still to be reached, although lately the MLIR initiative [45] seems to have the potential of achieving substantial breakthroughs.

Adaptive and autonomous systems evolution

Leveraging concepts such as the autonomic computing paradigm [46] and related self-* properties [47], classic CPS can achieve self-reconfiguration through awareness of themselves and of their environment [48],[49] by means of feedback control loops [50]. The implementation of the feedback loop at the scale of system of system [51] was attempted through AI-enhanced CPSoS [52]. Nevertheless, this might not suffice for future CPS with

a living dimension, since the self-evolution of the ensemble of individuals would still be limited by design given the fact that, in swarm intelligence, local rules are mostly fixed.

Conclusion

Living CPS appear to be just around the corner. Building on the foundations laid by classic CPS, the key enabling technologies set out in this article will help transform CPS into members of a community, consciously and autonomously cooperating with their human colleagues. Meanwhile, these technologies will help address, at least in part, the challenges that scientists face to make this future a reality.

References

[1] Yogesh K Dwivedi, Laurie Hughes, Abdullah M Baabdullah, Samuel Ribeiro-Navarrete, Mihalis Giannakis, Mutaz M Al-Debei, Denis Dennehy, Bhimaraya Metri, Dimitrios Buhalis, Christy MK Cheung, et al. 2022. Metaverse beyond the hype: Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. International Journal of Information Management 66 (2022), 102542.

[2] K. De Bosschere. AI for a better society. In M. Duranton et al., editors, HiPEAC Vision 2021. DOI: 10.5281/zenodo.7092776

[3] https://www.nsf.gov/pubs/2017/nsf17529/nsf17529.htm [Last access on 22/11/2022]

[4] Ji Zhou, Yanhong Zhou, Baicun Wang, and Jiyuan Zang. 2019. Human–Cyber–Physical Systems (HCPSs) in the Context of New-Generation Intelligent Manufacturing. Engineering 5, 4 (2019), 624–636. https://doi.org/10.1016/j.eng.2019.07.015

[5] https://www.databridgemarketresearch.com/reports/global-cyber-physical-systems-market [Last access on 22/11/2022]

[6] https://www.grandviewresearch.com/industry-analysis/artificial-intelligence-ai-market [Last access on 22/11/2022]

[7] Jiayang Song, Deyun Lyu, Zhenya Zhang, Zhijie Wang, Tianyi Zhang, and Lei Ma. 2022. When Cyber-Physical Systems Meet AI: A Benchmark, an Evaluation, and a Way Forward. In 44th IEEE/ACM Inter-national Conference on Software Engineering: Software Engineering in Practice, ICSE (SEIP) 2022, Pittsburgh, PA, USA, May 22-24, 2022. IEEE, 343–352. https://doi.org/10.1109/ICSE-SEIP55303.2022.9794128

[8] Yuke Zhu, Ziyu Wang, Josh Merel, Andrei A. Rusu, Tom Erez, Serkan Cabi, Saran Tunyasuvunakool, János Kramár, Raia Hadsell, Nando de Freitas, and Nicolas Heess. 2018. Reinforcement and Imitation Learning for Diverse Visuomotor Skills. In Robotics: Science and Systems XIV, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, June 26-30, 2018, Hadas Kress-Gazit, Siddhartha S. Srinivasa, Tom Howard, and Nikolay Atanasov (Eds.). https://doi.org/10.15607/RSS.2018.XIV.009

[9] Paola Busia, Gianfranco Deriu, Luca Rinelli, Cristina Chesta, Luigi Raffo, and Paolo Meloni. 2022. Target-Aware Neural Architecture Search and Deployment for Keyword Spotting. IEEE Access 10 (2022), 40687–40700. https://doi.org/10.1109/ACCESS.2022.3166939

[10] Jinan Fiaidhi, Sabah Mohammed, and Sami Mohammed. 2018. The Robotization of Extreme Automation: The Balance Between Fear and Courage. IT Prof. 20, 6 (2018), 87–93. https://doi.org/10.1109/MITP.2018.2876979

[11] L.M.C. Oliveira, R. Dias, C.M. Rebello, M.A.F. Martins, A.E. Rodrigues, A.M. Ribeiro, and I.B.R. Nogueira. 2021. Artificial Intelligence and Cyber-Physical Systems: A Review and Perspectives for the Future in the Chemical Industry. AI 2(3) (2021), 429–443. https://doi.org/10.3390/ai2030027

[12] Han Cai, Chuang Gan, Tianzhe Wang, Zhekai Zhang, and Song Han. 2020. Once-for-All: Train One Network and Specialize it for Efficient Deployment. In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net. https://openreview.net/forum?id=HylxE1HKwS

[13] https://www.statista.com/statistics/1183457/iot-connected-devices-worldwide/ [Last access on 22/11/2022]

[14] https://www.itu.int/en/ITU-T/techwatch/Pages/tactile-internet.aspx [Last access on 22/11/2022]

[15] Martin Maier, Mahfuzulhoq Chowdhury, Bhaskar Prasad Rimal, and Dung Pham Van. 2016. The tactile internet: vision, recent progress, and open challenges. IEEE Commun. Mag. 54, 5 (2016), 138–145. https://doi.org/10.1109/MCOM.2016.7470948

[16] Nattakorn Promwongsa, Amin Ebrahimzadeh, Diala Naboulsi, Somayeh Kianpisheh, Fatna Belqasmi, Roch H. Glitho, Noël Crespi, and Omar Alfandi. 2021. A Comprehensive Survey of the Tactile Internet: State-of-the-Art and Research Directions. IEEE Commun. Surv. Tutorials 23, 1 (2021), 472–523. https://doi.org/10.1109/COMST.2020.3025995

[17] Daniel van den Berg, Rebecca Glans, Dorian De Koning, Fernando A. Kuipers, Jochem Lugtenburg, Kurian Polachan, Prabhakar T. Venkata, Chandramani Singh, Belma Turkovic, and Bryan Van Wijk. 2017. Challenges in Haptic Communications Over the Tactile Internet. IEEE Access 5 (2017), 23502–23518. https://doi.org/10.1109/ACCESS.2017.2764181

[18] Gerardo Beni and Jing Wang. 1993. Swarm Intelligence in Cellular Robotic Systems.

[19] Mohammad Reza Bonyadi and Zbigniew Michalewicz. 2017. Particle Swarm Optimization for Single Objective Continuous Space Problems: A Review. Evolutionary Computation 25 (2017), 1–54.

[20] Vincent T'kindt, Nicolas Monmarché, Fabrice Tercinet, and Daniel Laügt. 2002. An Ant Colony Optimization algorithm to solve a 2-machine bicriteria flowshop scheduling problem. Eur. J. Oper. Res. 142, 2 (2002), 250–257. https://doi.org/10.1016/S0377-2217(02)00265-5

[21] http://apium.com/ [Last access on 22/11/2022]

[22] https://www.swarmfarm.com/ [Last access on 22/11/2022]

[23] https://www.marketdataforecast.com/market-reports/swarm-intelligence-market [Last access on 22/11/2022]

[24] https://www.mordorintelligence.com/industry-reports/swarm-intelligence-market [Last access on 22/11/2022]

[25] Melanie Schranz, Gianni A. Di Caro, Thomas Schmickl, Wilfried Elmenreich, Farshad Arvin, Y. Ahmet Sekercioglu, and Micha Sende. 2021. Swarm Intelligence and cyber-physical systems: Concepts, challenges and future trends. Swarm Evol. Comput. 60 (2021), 100762. https://doi.org/10.1016/j.swevo.2020.100762

[26] https://blogs.idc.com/2020/06/01/edge-computing-not-all-edges-are-created-equal/ [Last access on 22/11/2022]

[27] https://www.gartner.com/smarterwithgartner/what-edge-computing-means-for-infrastructure-and-operations-leaders [Last access on 22/11/2022]

[28] https://www.techtarget.com/searchcio/DrivingITSuccess/4-Things-You-Need-to-Know-Now-About-Edge-Computing [Last access on 22/11/2022]

[29] M. Duranton, M. Malms, and M. Ostasz. The continuum of computing. In M. Duranton et al., editors, HiPEAC Vision 2021, pages 40-43, Jan 2021.

[30] James G. Miller. 1965. Living systems: Basic concepts. Behavioral Science 10, 3 (1965), 193–237. 10.1002/bs.3830100302

[31] Karol Desnos, Maxime Pelcat, Jean-François Nezan, Shuvra S. Bhattacharyya, and Slaheddine Aridhi. 2013. PiMM: Parameterized and Interfaced dataflow Meta-Model for MPSoCs runtime reconfiguration. In 2013 International Conference on Embedded Computer Systems: Architectures, Modeling, and Simulation, SAMOS 2013, Agios Konstantinos, Samos Island, Greece, July 15-18, 2013. IEEE, 41–48. https://doi.org/10.1109/SAMOS.2013.6621104

[32] Marten Lohstroh. 2020. Reactors: A Deterministic Model of Concurrent Computation for Reactive Systems. Ph.D. Dissertation. University of California, Berkeley, USA. https://www.escholarship.org/uc/item/0fm2m3k1

[33] Marten Lohstroh, Christian Menard, Alexander Schulz-Rosengarten, Matthew Weber, Jerónimo Castrillón, and Edward A. Lee. 2020. A Language for Deterministic Coordination Across Multiple Timelines. In Forum for Specification and Design Languages, FDL 2020, Kiel, Germany, September 15-17, 2020. IEEE, 1–8. https://doi.org/10.1109/FDL50818. 2020.9232939

[34] Yechan Yu, HoJin Kim, Jinjoo Ha, Daewoo Kim, and Kang Yi. 2018. Cost-Performance Comparison of Various Accelerator Implementation Platforms for Deep Convolutional Neural Network. In Parallel and Distributed Computing, Applications and Technologies, 19th International Conference, PDCAT 2018, Jeju Island, South Korea, August 20-22, 2018, Revised Selected Papers (Communications in Computer and Information Science), Jong Hyuk Park, Hong Shen, Yunsick Sung, and Hui Tian (Eds.), Vol. 931. Springer, 335–344. https://doi.org/10.1007/978-981-13-5907-1_36

[35] Waever, Ole, Barry Buzan, Morten Kelstrup, and Pierre Lemaitre (1993). Identity, Migration and the New Security Agenda in Europe. Pinter Publishers

[36] https://time.com/5868008/twitter-hack-accounts-targeted-bitcoin-scam/ [Last access on 22/11/2022]

[37] Liqun Chen and Siaw-Lynn Ng. 2022. Securing emergent behaviour in swarm robotics. J. Inf. Secur. Appl. 64 (2022), 103047. https://doi.org/10.1016/j.jisa.2021.103047

[38] Jean-Paul A. Yaacoub, Ola Salman, Hassan N. Noura, Nesrine Kaaniche, Ali Chehab, and Mohamad Malli. 2020. Cyber-physical systems security: Limitations, issues and future trends. Microprocessors and Microsystems 77 (2020), 103201. https://doi.org/10.1016/j.micpro.2020.103201

[39] https://www.cerbero-h2020.eu/toolchain/ [Last access on 22/11/2022]

[40] https://into-cps.org/ [Last access on 22/11/2022]

[41] https://www.cpswarm.eu/index.php/cpswarm-workbench/ [Last access on 22/11/2022]

[42] https://www.aloha-h2020.eu/ [Last access on 22/11/2022]

[43] Felice Balarin, Yosinori Watanabe, Harry Hsieh, Luciano Lavagno, Claudio Passerone, and Alberto Sangiovanni-Vincentelli. 2003. Metropolis: An Integrated Electronic System Design Environment. Computer 36, 4 (apr 2003), 45–52. https://doi.org/10.1109/MC.2003.1193228

[44] Peter Gorm Larsen, John S. Fitzgerald, Jim Woodcock, Carl Gamble, Richard John Payne, and Kenneth Pierce. 2017. Features of Integrated Model-Based Co-modelling and Co-simulation Technology. In Software Engineering and Formal Methods - SEFM 2017 Collocated Workshops: DataMod, FAACS, MSE, CoSim-CPS, and FOCLASA, Trento, Italy, September 4-5, 2017, Revised Selected Papers (Lecture Notes in Computer Science), Antonio Cerone and Marco Roveri (Eds.), Vol. 10729. Springer, 377–390. https://doi.org/10.1007/978-3-319-74781-1_26

[45] https://mlir.llvm.org/ [Last access on 22/11/2022]

[46] Jeffrey O. Kephart and David M. Chess. 2003. The Vision of Autonomic Computing. Computer 36, 1 (2003), 41–50. https://doi.org/10.1109/MC.2003.1160055

[47] Enterprise Management Associate. 2006. Practical Autonomic Computing: Roadmap to Self Managing Technology. White Paper (2006). https://www.enterprisemanagement.com/research/asset.php/286/Practical-Autonomic-Computing:-Roadmap-to-Self-Managing-Technology

[48] Frank D. Macías-Escrivá, Rodolfo Haber, Raul del Toro, and Vicente Hernandez. 2013. Self-adaptive systems: A survey of current approaches, research challenges and applications. Expert Systems with Applications 40, 18 (2013), 7267–7279. https://doi.org/10.1016/j.eswa.2013.07.033

[49] Francesca Palumbo, Tiziana Fanni, Carlo Sau, Alfonso Rodríguez, Daniel Madroñal, Karol Desnos, Antoine Morvan, Maxime Pelcat, Claudio Rubattu, Raquel Lazcano, Luigi Raffo, Eduardo de la Torre, Eduardo Juárez, César Sanz, and Pablo Sanchez de Rojas. 2019. Hardware/Software Self-adaptation in CPS: The CERBERO Project Approach. In Embedded Computer Systems: Architectures, Modeling, and Simulation - 19th International Conference, SAMOS 2019, Samos, Greece, July 7-11, 2019, Proceedings (Lecture Notes in Computer Science), Dionisios N. Pnevmatikatos, Maxime Pelcat, and Matthias Jung (Eds.), Vol. 11733. Springer, 416–428. https://doi.org/10.1007/978-3-030-27562-4_30

[50] Éric Rutten, Nicolas Marchand, and Daniel Simon. 2013. Feedback Control as MAPE-K Loop in Autonomic Computing. In Software Engineering for Self-Adaptive Systems III. Assurances - International Seminar, Dagstuhl Castle, Germany, December 15-19, 2013, Revised Selected and Invited Papers (Lecture Notes in Computer Science), Rogério de Lemos, David Garlan, Carlo Ghezzi, and Holger Giese (Eds.), Vol. 9640. Springer, 349–373. https://doi.org/10.1007/978-3-319-74183-3_12

[51] Georgios Keramidas, Christos P. Antonopoulos, Nikolaos S. Voros, Pekka Jääskeläinen, Marisa Catalán Cid, Evangelia I. Zacharaki, Apostolos P. Fournaris, and Aris S. Lalos. 2020. CPSoSaware: Cross-Layer Cognitive Optimization Tools & Methods for the Lifecycle Support of Dependable CPSoS. In 2020 IEEE Computer Society Annual Symposium on VLSI, ISVLSI 2020, Limassol, Cyprus, July 6-8, 2020. IEEE, 470–475. https://doi.org/10.1109/ISVLSI49217.2020.00-12

[52] Stavros Nousias, Nikos Piperigkos, Gerasimos Arvanitis, Apostolos P. Fournaris, Aris S. Lalos, and Konstantinos Moustakas. 2021. Empowering cyberphysical systems of systems with intelligence. CoRR abs/2107.02264 (2021). arXiv:2107.02264 https://arxiv.org/abs/2107. 02264

Francesca Palumbo is an associate professor at Università degli Studi di Sassari, Italy.

Raquel Lazcano is a research assistant at Università degli Studi di Sassari, Italy.

Daniel Madroñal is a research assistant at Università degli Studi di Sassari, Italy.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2023.

Cite as: F. Palumbo, R. Lazcano and D. Madroñal. Towards a living dimension: The future of cyber-physical systems. In M. Duranton et al., editors, HiPEAC Vision 2023, pages 44-53, Jan 2023.

DOI: 10.5281/zenodo.7461786

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

© HiPEAC 2023

The most relevant metaverse may be a serious game in the digital twin of reality.

Gaming, content and the metaverse

By THOMAS HOBERG

The push for maximum mind-changing power at minimal cost via fully personalized content creation has led to game engine technology and artificial intelligence advancing to a point of realism where generated content and recorded content become difficult to distinguish.

But the inherent spatiality of game engines also allows fitting the internet much closer to how our brains work, because most of our social code and even instincts are based on where we are and how far from, close to or intimate with everyone else we are. Meanwhile, the internet is far too flat, direct and invasive to naturally accommodate the increasing digitalization of an ever-growing part of our social and commercial interactions with others.

Using game engine technology to maintain an adequately detailed digital twin of our environment, allows spatiality an intrinsic part of all our interactions. Complex deliberations on e.g. privacy become quite natural when walls and distance regulate visibility; Zuckerberg and other internet giants need to be invited to the dinner table, rather than Alexa listening in at the bedside.

This article outlines the major shifts in media business models, the giant ongoing technology transformation in content creation such models cause, and points to problems and opportunities that require European regulation and support.

Key insights

- Gaming is essential for the evolution of social code.
- Realism versus reality becomes a display engine preset: fake versus truth become ever harder to decipher.
- Content creation business models are switching from media ownership to fully individual content.
- Our brains, instincts, culture and laws are all naturally and inherently spatial.
- Spatiality can naturally resolve the most complex multi-domain compliance issues faced by digitalization.
- Metaverses stepping beyond digital twins allow much better and faster evolution of social code.

Key recommendations

- Support the creation of a standard base metaverse in the form of an adaptively precise digital twin of reality as the base for spatial computing.
- Use that to reflect ground rules into the internet and to develop a more natural basis for conflict resolution in a multi-compliance environment via game-engine computed visibility.
- Take gaming and game technology very seriously.
- Regulate private multiverses to ensure ground rules always have precedence over cloud rules.

Why do we game?

When children play, they run a simulation of a world where they are in charge. They understand that that day will come, so they go through every scenario they can imagine to be ready when it does.

So, what drives the adults already in charge to play? Do they still strive to "improve their game" or do they seek to escape from a reality where they have failed to achieve their goals?

Both desire the ability to create a variant of this world, where perhaps some or even most aspects can be changed under their control: they could aim for realism, be it visual or the operational model, or they could aim to overshoot in a diversity of directions, to extremes that explore some individual urge.

In a way ancient man was able to imagine almost anything, because many of the constraints imposed by physics and nature were still unknown or the whims of frolicky gods. Today scientific reality seems to impose hard limits onto the imaginable, yet the entertainment industry has invented weapons of mass illusion. These can achieve an incomprehensible realism or just realize incomprehensible visualities for practically everyone at the price of entrance to a cinema or internet gaming café. They can completely overwhelm humanity's most evolved and highest bandwidth sensory input channels – sight and hearing – with uncanny realism, while the other senses such as smell, taste, touch and motion oddly remain much more challenging "to be played".

The Enlightenment perhaps only brought a transient peak of rationality, a hunger for a singular objective truth in its wake: it would seem that today we are more invested not just in creating "alternate truths" but forking the planet into metaverses, each with its own ethos, including some on Earth, an extension in a way, of gated communities, up to the size of private nations.

Computer games started in the smallest of niches, because they obviously required a computer, an expensive piece of machinery that initially only the military could afford. Today we live in an age where there is almost a surplus of compute power and it's more the price of energy versus the value to be obtained from extra computation that is limiting its global growth.

This not only makes everyone with disposable income a potential gamer, it also turns out that using game engines and artificial intelligence (AI) techniques is the cheapest way to create entertaining content, especially compared to live actors, who may have more mouths to feed than their own. We are passing the point where it is both possible and less costly to generate uncanny realism locally on your personal endpoint device than to stream video content from a data centre. Personalizing that content to a tailor-made cocktail of endorphins and adrenaline to ensure your unwavering attention and maximize the mind-changing power of its source is but a shifting of levers, a highly automated optimization, routinely achieved via code or AI models.

The collateral damage is addiction and a far too personal view of the world, which is only really shared by the AI which created your "social" feed. And that AI is not your friend: its mission is to change your mind and get you hooked on whatever the content's sponsor is peddling.

The quest for realism

For almost anyone alive today, what we consider reality is often just what we've seen on a TV screen. That screen is much the same as that which computers use, but even if what you saw on a screen was something as outlandish as people stepping on the Moon, until very recently you always knew if what you saw was what a camera had captured from the real world or a computer painting. Because for a long time, what the best computers could do was at the level of a comic strip or cartoons.

Figure 1. Peter Cushing as Grand Moff Tarkin in Rogue One [3]

For several decades computer graphics had to choose between realism and real time: realism required photorealistic computer-generated imaging (CGI) or visual effects software (VFX) rendering at stop-motion speeds, while real time entailed regressing towards cartoonish image quality. The former would only work with fully scripted movie-style video that was pre-rendered over weeks or months for pure playback, while the latter allowed full interaction, but the better it got visually, the wider the uncanny valley to realism seemed to grow.

The classic modelling approach of using ever finer 3D geometry meshes with shader textures failed to do justice to the twinkle or the wrinkles of a character's smile, while exploding in complexity. Even the most elaborate bump maps, shaders or hair simulation algorithms only increased the gap. And then animating the digital limbs with truly human grace or clumsiness required obtaining high-resolution motion traces with not only significant technical effort but also participation by actors paid for going through the motions harnessed in gadgets.

Like in so many other areas, machine learning has completely changed the game.

In a sense it has enabled the near seamless transition from reality as captured in real time from nothing but cameras into a geometric model [1] or a digital twin ready for use in a game engine as well as the mapping from such a model to a Hollywood-perfect rendition of a person [2], a stage set, dinosaurs, goblins, angels, planes, planets or plants.

Most importantly the level of detail and fidelity can be chosen with relative independence at both the source and the destination of such mapping: Hollywood has rarely aimed for true fidelity; instead it projects a realism that is perennially a tad out-of-reach yet feigns attainability to mere mortals[1].

Whole crowds can be captured as mere stick figures, transformed from people recorded by an airport surveillance camera into goblins hustling through a mine. For character actors the more significant body parts like face/eyes/mouth, arms, hands and postures may be captured in detail, yet abstracted as mere vectors and mapped into those starry looks and smiles that cost millions and earn billions.

Actors long deceased are coming back for an encore (Figure 1); extras paid below minimum wage might still be employed to move them about like ragdolls, but the essence of people that is captured from

1 Superheroes exempted, but even those require a 'civilian' anchor to be interesting.

crowds in all walks of life allows not just the real-time creation of digital twins, but also the insertion of wholly synthetic individuals, groups, and crowds[2] of any kind. The media industry has long employed soundbites, the "Wilhelm Scream" [5] perhaps being the most iconic, which were recorded once and then modified and reused as required. Canned laughter [6] likewise had for a long time enjoyed use in the industry, but today's shows can include high-fidelity visuals of people laughing their heads off, quite literally now, if you click that option.

Virtual news desks and TV studios [7] have become mainstream and nearly everyone has used background replacement filters when on a video call while working at home during COVID-19 lockdowns. More extensive face and body filters have been employed by influencers, sometimes with a bit of mishap [8]; there are a lot of videos out there which show Chinese live streamers completely transform in real-time with little more investment than a smartphone and an app [9].

The more advanced character transmorphing can go quite a bit beyond these "deep fakes", which mostly involve replacing only a face. Instead, it can replace an entire character, all limbs and body parts, clothing, apparel, voice and facial expressions using a library of models for each aspect and level of detail. And fidelity to the source may be the least desirable aspect of that transformation, because either it is an AI bot already or a reality TV gig economy actor with neither the looks nor the talent of the persona they project.

Movie studios have faked reality in sometimes extremely elaborate stage sets, simulated giant and microscopic scales, flying or zero gravity with ingenious props and tricks, stop-motion, mirrors, smoke screens, over-exposure, matte painting; creative cutting and editing had been used long before computer based visual effects.

Actors have not just worn costumes and make-up, some have spent countless hours being put into elaborate contraptions or fat suits, or under elaborate layers of painted silicon to completely transform their character.

Figure 2. A live streamer becomes someone else before our eyes. From [25]

However, in a modern extended or augmented reality studio [10] [11] all walls and every surface can become a display run by a game engine like EPIC's Unreal[3] [12], enabling not only a better immersive experience than green screens provide to actors, but even the ability to do live composition of content, instead of relying on post-processing. Not only can extended reality (XR) studios have presenters or actors appear in just about any setting, they can also have individuals appear together to a viewer, when they are actually on opposite sides of the planet. And obviously such remote people need not even be real: Metahumans are a standard component of EPIC's Unreal that just need to have their strings pulled by a script or an AI.

The ability to map anyone to anyone and anything to anything in real time and then to switch tracks between mapping live action and simulating is going full circle. With that, true reality as we know it can be replaced by fake realism, tailor made for any individual just for the cost of running a computer game or starting a video stream on a content distribution platform.

Media-merging metaverse

Grabbing and holding consumer attention for less expense and higher returns is what determines the winners and the losers in the content industry, which encompasses everything from print and audio/video media to gaming, streaming and social networks. The print, music, and film industries were long able to offset blockbuster production costs via monetized mass consumption: skill boosted scale and thus revenues.

Yet the two most recent acquisitions and mergers made by Unity Technologies, owner of arguably the most widely used game engine globally[4], show how the centre of gravity is moving from scripted to coded/generated content, or which Goliath is becoming David:

- In December 2021 Unity Technology bought Weta Digital [13], widely consid-

2 Marshalling hundreds of extras to behave like an ancient crowd following a chariot race used to require gargantuan effort. Likewise the memorable scenes of mass slaughter in Lord of the Rings were only possible because they were generated by computers [4].

3 Since they are no longer limited to games, that moniker seems less and less adequate.

4 Unity and Unreal (EPIC) are hard to compare, due to very distinct company structure and markets, but they generally recognized as the top two contenders.

ered the most influential and successful visual effects company and responsible for the stunning visuals in nearly every blockbuster movie of the last two decades.

- In July 2022 Unity announced an agreement to **merge** with the much smaller ironSource, an Israeli company whose business model is to be completely unknown and unnoticed by consumers, while it notices and knows everything about them.

The merger of Unity and ironSource transfers the intellectual property of the world champion in realism to a game engine company and could be viewed as a sell-out of VFX and CGI or even their end as a distinguishable art form and discipline.

The latter is not only a merger; it's probably much more significant. From printed books, past vinyl records and ROM cartridges to CDs, DVDs and digital rights management (DRM), media in Western civilization were mostly purchased with copyright law providing protection against pirating.

A quick glance at this world population chart (Figure 3) [14] indicates that strong copyright support may perhaps be found in half of the blueish parts, while markets grow mostly in the reddish zones.

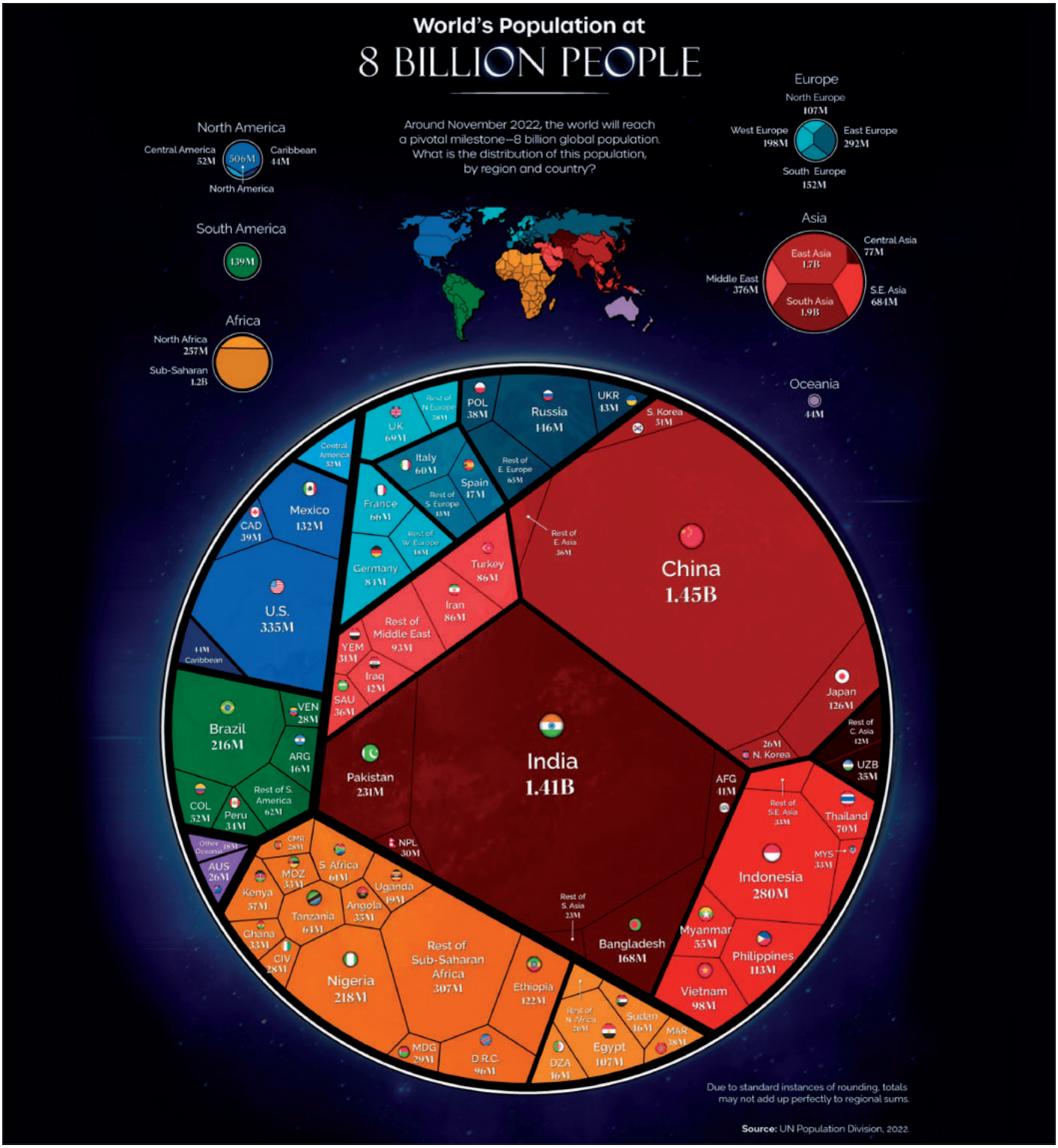


Figure 3. Summary representation of the world's population by country and continent [27]

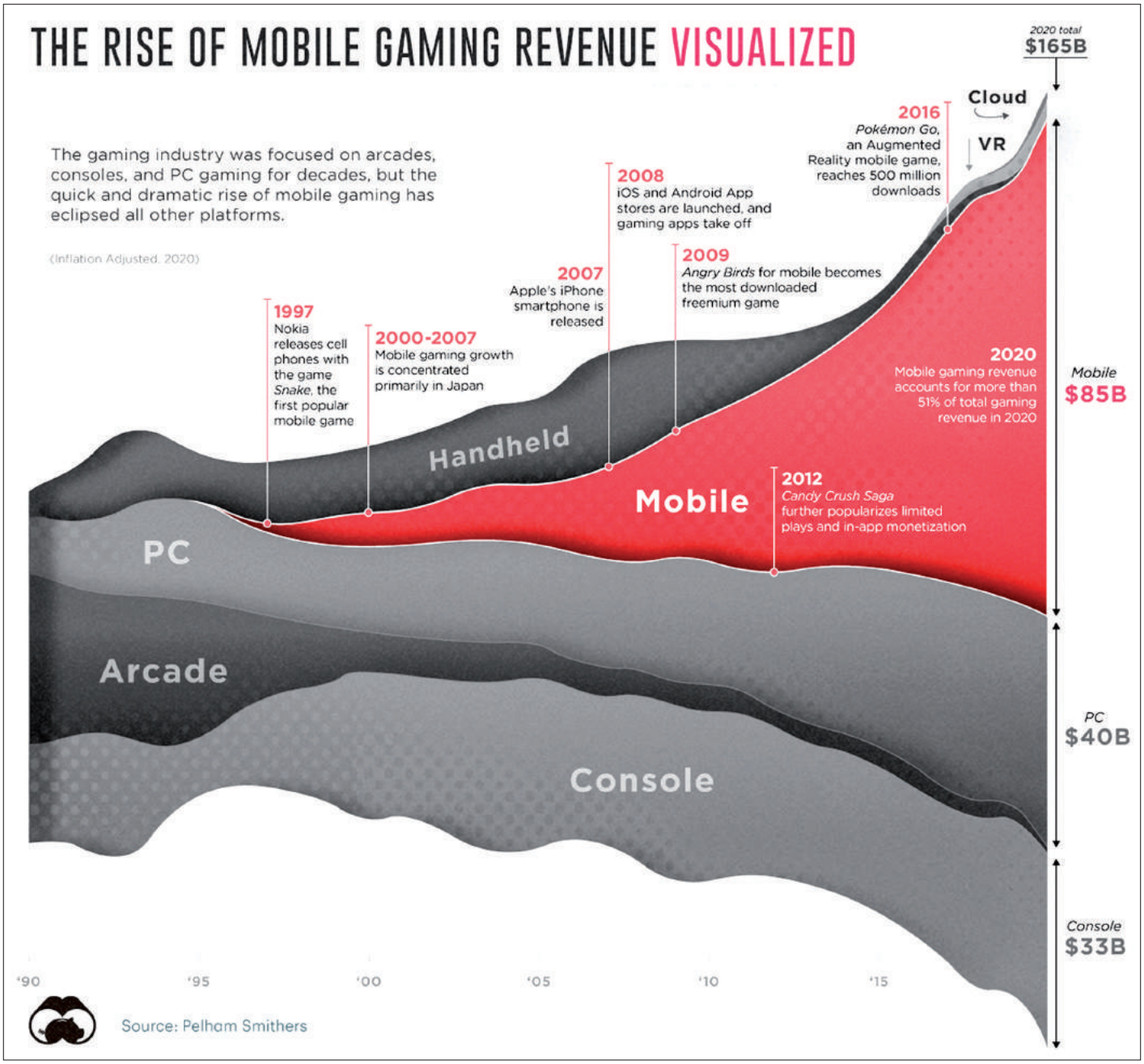


Figure 4. Mobile gaming revenue, from [15]

Apart from Japan and South Korea, very few people in Asia have ever owned personal computers, while games consoles were prohibited in China until as late as 2015. Cybercafés were the first and principal avenues for gaming and internet access, which does not work with DRM-protected media for remuneration.

Ever since smartphones became affordable, they have been the default route to personal computing for nearly everyone in Asia as well as for the young in more affluent societies. Western consumers may also have consoles or even gaming PCs along with huge flat screen TVs, but microtransactions and subscriptions are pushing out game or movie purchases, which have often deteriorated to little more than a licence with limited time and limited use.

Subscriptions again have a much longer history and cultural acceptance in the West and offer their own challenges, which we'll look at shortly.

It is micro transactions that are the main motive for the ironSource merger. Predominantly found in free-to-play games today, they threaten to spill over into in streaming media and more content-related aspects of social media before long.

Because DRM-protected Western games could not be bought or operated in internet cafés, Asian companies produced look-alike clones for multi-user games and then offered players the ability to personalize their character with payment, often via coupons that could be purchased in the café or from basic vending machines. That provided a valuable incentive and an easy transition to social media-integrated payment e.g. via WeChat whose owner Tencent is the world's biggest video game investor because of that synergy. When those purchases are perfectly integrated and incur no outsider's transaction fees, their individual amounts need not be high

– they can even be tiny[5]. As long as they are frequent, they'll still reach the cumulative effect that has had mobile game revenues surpass all other media channels combined.

Free-to-play games are intentionally very easy during the first couple of rounds: the games actually won't let you lose no matter how badly you play because, if you lost, you'd just leave. Then, after you've enjoyed your first flushes of unexpected success, they'll raise the bar, to the point where continuing to win may take just a little more effort than you'd have wanted to invest. Yet, you no longer want to leave: you've invested part of your life and who wants to lose that? That's when you might be given an upgraded weapon or improved stamina for no more than making a recommendation on social media or suffering through a commercial.

The point at which each one of us will yield to the temptation to hand over real money is very personal and may depend on a lot of context, such time of day, day of the week, bank account balance, family trouble or the number of beers consumed. Way too much of this contextual data can be observed and analysed from your smartphone and from the game you are playing, including the heart rate and the jerkiness of your movements from the helpful sensors on your smart watch. Your personal and momentary pièce de résistance can be inferred perhaps not with absolute precision, but with such a degree of confidence, that ironSource could command a merger, not a mere acquisition, because that's exactly what they do and what they offer to anyone who wants to make money from a smartphone app[6].

This merge point of generated "live action" at good-enough realism doesn't just support personalized pricing but also fine-tuned content, a blurring of game, soap opera, shopping channel, even political commentary, or religious service. To be fair, it could also adjust for visual or hearing impairment, adjust the input for Parkinson's disease or disabilities, or filter content to guard against epileptic seizures or translate it into a language the consumer understands or tolerates.

But the natural and near unavoidable risk is that content will be tuned for maximum profit and a balanced point of managed addiction, where the consuming individual is still able to earn enough of a living to keep on spending.

Investigating a game or generated content for its addictive potential or to rank the violence, nudity or political correctness in an objective manner becomes impossible, because those qualities are measured across a carefully picked population sample, with all exposed to the identical experience to eliminate individuality. The only agency best qualified to attest addictiveness is the least interested party to disclose it.

Which is why such technology needs to be either forced out and prohibited, or mandated to pass its knowledge to a monitoring agency along with the insurance funding required to treat victims of false negatives.

Meta vs the metaverse

One must not mistake how insights or resulting ideas are sold with how much research, analysis and reflection went into trying to shape a changing market. We believe that the metaverse the former Facebook is trying to sell is not just a fancy that popped into Mr. Zuckerberg's head.

Meta by necessity is at the forefront of research in many areas and can see how technology is enabling this mix of generative and transmorphed creation of real-time realistic content. It is also built on the realization that it can harvest the basic human impulse to comment, advise, even offer a tiny bit of help to anyone who triggers the parental instinct, to have its members co-create the very content which Meta can then use to exploit them.

It has used its psychology research to tune reinforcement and recommendation networks for maximum attention via the strongest emotional response[7]. As Francis Haugen has exposed [16], Meta's leaders are fully aware that the content that they create with their mix of member input, external data, algorithms and AI models can be harmful, even deadly, although primarily for the needy, the non-affluent or non-Meta-members[8].

In Meta's view, the metaverse in the sense of technical feasibility is already there or nearly so. We consider the company less to be leading than to be attempting to hijack a set of changes that threaten its income, because the Chinese giants such as Tencent with WeChat, Douyin/TikTok, Weibo and Xiaohongshu enjoy synergies, market and regulatory conditions which might put them far ahead.

The significant operational overhead and cost of external payment compared to completely integrated variant Tencent's WeChat can enjoy, saw Meta try to launch a global payment scheme that included currency protection [17], because microtransactions become too expensive and volatile otherwise. Likewise, their main motivation for turning the metaverse into a walled garden is quite simply following Apple's playbook of taxation with neither representation [26] nor all the negative cost positions of non-cloud/meta government[9]. It aims to create a libertarian plan-M(eta) [18], if lobbying for libertarian freeports [19] within nation states should "do a Truss-Kwarteng" (fail).

Unfortunately, Apple's walled garden playbook has been emulated far too often and Meta's ability to hype corporate and consumer sheep into Libra or Meta pastures seems to be failing. That doesn't mean 'the metaverse' will fail or not happen. It will mostly not follow Meta's direction or the predicted growth curves. We predict a situation very similar to that of the Internet of Things (IoT), where a focus on engineer-

5 To support development of good spending habits, a kid's pocket money should ideally last them for several transactions.

6 If you want another metaphor: they help create the type of vending machine that lures you with half-price soda when you're just passing by and that will increase the price when you're interested, and increase it further still when you're thirsty; and ask for your house and car when you've run out of alternatives.

7 Or used genetic programming, which very likely would have led to the same result.

8 Meta regards this as governmental negligence, and that lack of regulation frees them from accountability.

9 Education, social security, health care, three pillars of non-business law, military etc.

ing over human nature, regulatory deficits and privacy concerns styme consumer value. Fixing those issues can in fact generate a significant amount of value and thus growth beyond human scale [20].

Just like with the IoT (which can be thought as a metaverse of its own), companies immediately try to push products and services before fully thinking through where the value for consumers would lie and how it needs to fit human brains, culture and habits across the broad spectrum of humankind.

Spatiality

Where we are, especially in relation to our parents, kids or teammates, the next place of refuge, our food or utensils is something we are tracking constantly, if perhaps somewhat semi-consciously while we concentrate on a primary activity. Just how important that is we notice mostly if we suddenly find ourselves without a good idea of just where we are e.g. being awakened in total darkness[10]. Our mind is naturally spatial, so much so that we even map the disembodied, emotional and spiritual to locations, sometimes within our body (e.g. the heart for strong emotions), sometimes just vaguely upwards (heaven).

We tend to organize our belongings spatially into floors, rooms, cabinets, drawers, boxes etc, especially when we have plenty, so we can find them again. That is such an ingrained way of thinking that we may even organize our mind or activities in a similar manner, putting tax-related issues in some dark cellar off beyond that home-loan outbuilding in the untended backyard off-right, while the bright wide staircase immediately ahead leads to all things family and fun.

That mind-verse is first of all very personal and perhaps a bit Harry-Potterish: those rooms don't align and perhaps run into each other; if you ever tried to "digital-twin" them into a game engine model, staircases and secret passages are more likely to follow associations or emotions, not building codes.

Interestingly that spatiality also tends to reflect relationships, between any mix of objects and subjects, really, and thus intuitively decides what can and should happen between them or not, as they might have no or quite a few inner and/or outer walls between them: **privacy is a notion we think or feel spatially, and the mental metaverse we use to intuitively gauge it is often different from the physical universe, even if it feels just as real to our mind**.

We all create our individual mindverse as we grow up and we have naturally built passageways or floating docks into both the real world and to those of our closest relations: they never need to be perfectly joined, because they are simply a mind crutch, there to help and ease things, ignored when they are not needed.

With the internet and digitalization, the number of these distinct metaverses naturally increases and the biggest shortcoming of the current internet is that it's a flatlander's world [21], that squeezes the spatial world into a hypertext tree and thus leads to disorientation, because the spatial power of our brains lingers unused to make its complexity more manageable.

It is there that a Spatial Web that makes relative positions, enclosures, borders and paths an implied standard input into all computing, becomes much more natural and empowering to humans, while it just as elegantly solves the privacy versus control dilemma that cripples value on today's internet.

Naked children

This author has triplet boys. Keeping them safe and out of danger once they learn to walk and run in four directions, takes everything you have so the camera sees too much action to take a shot [or shoot a take]. But we have an extraordinarily large collection of photos and videos of them splashing about in the bathtub, because we tempted them to stay with lots of foam and sometimes even the coloured variety. When I went on business trips, I longed to be back home and the only time my wife found a moment to ease my yearnings with pictures and live transmissions was, you guessed it, when the triplets were in the nude and in the bath.

Just about the only crime universally and officially abhorred is child sex abuse; no excuses are accepted anywhere. Cloud giants and national governments newly feel obliged to search for images of nude children also within the confines or your most personal computers at home and everywhere else, and take drastic, debilitating and sometimes irreversible actions if they find them. They even legislate back-doors and cryptography that can be safely broken, but only for the execution of those searches[11].

This dilemma is insolvable in a non-spatial web because the search and crypto choices are just binary and do not have the context available. It becomes a non-issue when both parents are alone in the bathroom with their children and everybody else is the other side of a wall and too far away to see anyway.

A spatial web can put children and parents in a single metaverse room and resolve who can see what, even if one parent physically sits half a world away.

The technical problem for which Meta doesn't have the essential ingredient and that search and break agents across international networks need to trust and respect is spatiality information about a family's bathroom and about all data bouncing around within, while it transits their network electronically, without knowing or needing to identify the family. At the same time the convicted child abuser sat across the table from papa in the coffee shop and who'd like to catch a glance, needs to be kept at full distance on the other side of a massive metaverse wall.

Why we intuitively believe gaming is useful

The Nobel laureate Daniel Kahneman [22] and his colleague Amos Tversky split the brain into two major components, a System 1 that runs in what machine learning (ML) people might call "inference

10 The inability to quickly recall where we were just before, is the main reason dementia causes so much stress, through disorientation.

11 Or any other business that temporary holders of public office think opportune.

mode", and a System 2, which resembles "training" in ML.

The motivation behind the differentiation is again similar in both: avoiding the mental effort of thinking everything through by jumping to an eagerly accumulated treasure trove of conclusions, even at the risk of being wrong, but avoiding egregious errors.

Robin Dunbar has also made it his career to rack his brain about brains or rather how the effort involved in mentalizing levels of intentionality in relationships, which programmers might regard as the nesting level of cascaded IFs, have driven human evolution [23] up against a steep wall of brain size and energy consumption.

Brains don't easily think in causality beyond a certain level of cascaded IFs, because it's both energetically expensive, prone to giving rise to errors and thus waste e.g. from unresolved contradictions, before it becomes even impossibly hard. Baby kittens and human babies have little trouble to learn the first IF of cause and effect and will train another two or three to the point where we can navigate the physical world - the spatial world - completely with System 1 with an effort so low we don't even notice.

Gaming not only relieves us of mentalizing the simulation arena via a Gedanken-experiment; acting it out with full spatiality or even physicality allows us to discover unresolved logic errors like giving a cookie to a friend and then trying to eat it later. Spatial reasoning is nearly all in the efficient System 1, so gaming is like using it for a generative adversarial neural network to help train the more complex IF cascades.

The near impossible complexity of compliance networks that mesh multiple and conflicting legal, corporate and privacy rulesets, becomes much easier to evaluate with regards to:

- may I *do* something?

because once it's 'rendered' in 3D with all the digital twins of compliance as walls and layers in place, you find there are far too many obstacles or distance for you to note something even *exists*.

When Facebook, Google or Alexa don't just quietly observe each one of our utterances and gestures, but demand entry through our front door like a travelling salesman, only to sit down at our dinner table, observe how we put our children to bed or indeed join in from the bedside like a smartphone, we'd give Mr. Zuckerberg's digital twin at least an earful if not the boot. If Alexa had jumped up and stopped us from enjoying a meal only just prepared, or a movie or rugby match just starting, to declare her undying endorsement for this trinket or that bauble, it's unlikely she'd have ever emerged from the first scrummage. Just how brutally GAMA[12] ignore "do not track" becomes self-evident when their actions are rendered in the metaverse.

Legislation and regulation or even just culture generally dictate the spatiality of the physical world and the relationships it even makes possible or observable. Including spatial context from the physical world (which still governs the most important laws) as well as merging it in a rule-based manner with the spatiality of the relevant metaverses offers two initial benefits: it makes it much easier to compute (even with acceleration) and also to understand and argue if it had to come in front of a court of law: privacy quite literally can become occlusion again!

What Europe should do

Take gaming and game technology very seriously

Civilizations thrive and suffer based on the quality and evolution of their social code, which encompasses everything from habits to constitutions. From the second our mind awakens until the moment we die, we imagine and, in the most formative episodes of our lives, we even put our imaginations to the test in conversation, play-acting or gaming. The more involved approaches lead us to better recognize missing elements, contradictions, or outright mistakes, especially when the environment becomes more complex.

Gaming and gaming technology isn't "just the biggest money maker" in content creation, it is the biggest driver of our future and so it is critical that it should not surprise us that even the smallest children only interrupt their play for eating and sleeping.

What's more, with even classic content creation shifting to a game engine underneath, Europe needs not only to put its hands into every piece of the pie there, but also to ensure it puts the regulatory frameworks in place to keep foreign corporate or national giants from domination.

Ground rules in the metaverse

Meta's attempt to hijack the metaverse is only the latest brazen attempt to appropriate what is not theirs[13]. Their "move fast and break things" creed is highly motivated by their need for rapid growth and the problem that retrofitting multi-tenant compliance into a software base that is entirely built on a Wild West ethos is likely more expensive than pushing the illusion of a new, virtual and walled reality.

Europe and indeed everyone needs to enforce that ground rules always apply and *reflect the ground truth*. Not only does that make complying considerably easier, it also ensures that digitalization only stops at the point that the benefits of digitized convenience peter out, not where the struggle or indeed the collateral damage of a civil war between the cloud and the ground piles too high.

Needless to say, it demotes Meta and all other cloud captains to the rank of a commodity service provider, rather than the metaverse sovereign they reluctantly became, after none of the other groups or companies they bought, fired or asphyxiated stepped forward.

Software that is inherently spatial, and aware of compliance, ethics and energy consumption are factors that have natural global applicability, and that can run in the Wild West, the autocratic East, the theocratic Middle and especially well within a Europe that is naturally diverse.

12 Google, Apple, Meta and Amazon

13 Data generated from user input needs to belong to the users that created it, not to the company that stores or transforms it.

Influencing and addiction

Many of China's political values are quite incompatible with Europe's. But when it becomes sensitive to an issue, it's important to take note of what and why. It doesn't mean we need or should copy the reaction without reflection.

When China moved in 2021 to reduce smartphone gaming by minors to just three of the 168 hours in a week [24], it seemed a very drastic move in the world's biggest and most profitable gaming industry: even communists no longer scoff at taxes. It's easy to attribute that to a general tendency of the Chinese rulers to "dictate hard and break individualism".

However, the opium wars remain fresh enough in China's leadership memories. It may be much more reasonable in its assessment of the addictiveness of hyper-personalized games and content than anyone other nation. Content publishers are very ready to make hyper-snooping an essential part of the business model, proposing it as a vital need to improve the quality of their service, and therefore against which privacy regulation like the General Data Protection Regulation (GDPR) must obviously take a step back as that would interfere with business.

Europe's governments are legally required to protect their citizens against peddlers of addiction. The trend and the tools are clearly there and sold at prices that should do more than raise eyebrows: strong, proactive and auto-adaptive/proof-reversed legislation is urgently required. This theme is expanded in the HiPEAC Vision article "Europe's need for digital essentials, individual sovereignty and consumer protection".

References

[1] See OpenPose as a pioneering work from 2015 https://github.com/CMU-Perceptual-Computing-Lab/openpose
[2] Current commercial state of the art https://www.unrealengine.com/en-US/metahuman
[3] https://screenrant.com/rogue-one-cgi-peter-cushing-tarkin/
[4] https://www.cnet.com/culture/entertainment/features/how-lord-of-the-rings-used-ai-to-change-big-screen-battles-forever/
[5] https://allthatsinteresting.com/wilhelm-scream
[6] https://en.wikipedia.org/wiki/Laugh_track
[7] https://www.zerodensity.tv/
[8] https://www.lowyat.net/2019/190799/chinese-video-blogger-revealed-to-be-middle-aged-woman-after-face-filter-malfunction/
[9] https://www.youtube.com/watch?v=CHqQ_uGeDFk
[10] https://www.formstudios.com/studio-b/
[11] https://magnoliastudios.film/virtual-studio/
[12] https://www.unrealengine.com/
[13] https://unity.com/our-company/newsroom/unity-announces-intent-acquire-weta-digital
[14] https://www.visualcapitalist.com/visualized-the-worlds-population-at-8-billion/
[15] https://www.visualcapitalist.com/how-big-is-the-global-mobile-gaming-industry/
[16] https://www.franceshaugen.com/
[17] https://en.wikipedia.org/wiki/Diem_(digital_currency)
[18] https://www.theguardian.com/commentisfree/2022/jul/05/could-new-countries-be-founded-on-the-internet
[19] https://www.theguardian.com/commentisfree/2022/aug/17/freeport-turbocapitalism-tramples-over-british-democracy-teesside-plymouth
[20] https://zenodo.org/record/4719662
[21] https://en.wikipedia.org/wiki/Flatland
[22] https://en.wikipedia.org/wiki/Thinking,_Fast_and_Slow
[23] https://www.cambridge.org/core/journals/politics-and-the-life-sciences/article/abs/robin-dunbar-human-evolution-our-brains-and-behavior-new-york-oxford-university-press-2016-415-pages-isbn-9780190616786-hardcover-2995/B286252602D4D2F558457A52B77B9559
[24] https://www.cbsnews.com/news/china-limits-online-gaming-children-3-hours-week/
[25] https://m.blog.naver.com/kinkis/221602367808
[26] https://en.wikipedia.org/wiki/No_taxation_without_representation
[27] https://www.visualcapitalist.com/visualized-the-worlds-population-at-8-billion/

Thomas Hoberg is the technical director at Worldline Labs, Frankfurt, Germany.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2023.

Cite as: T. Hoberg. Gaming, content and the metaverse. In M. Duranton et al., editors, HiPEAC Vision 2023, pages 54-62, Jan 2023.

DOI: 10.5281/zenodo.7461953

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

© HiPEAC 2023

The next computing paradigm will emerge from the convergence of computing across the continuum, systems working for both the cyber and physical world, "natural" programming and orchestration, and Everything-as-a-Service. To enable this paradigm, we need high-level programming and reasoning abstractions; mechanisms for discovering properties, services, and devices; and interoperability among them, all united by self-organizing and trustworthy orchestration mechanisms. The European Union and the United States will play important roles in addressing these challenges, and an active research collaboration will boost the effectiveness, and drive the convergence, of solutions.

EU/US White Paper on the Continuum of Computing

By TULLIO VARDANEGA, MARC DURANTON, JASON O. HALLSTROM, KOEN DE BOSSCHERE

A strategic research collaboration between the EU and the US is starting to enable seamless convergence across the computing continuum, spanning both the physical and cyber worlds. The collaboration represents a new expedition to tame the growing heterogeneity of computing platforms and services across the computing continuum, giving rise to a new computing paradigm.

The web has changed forever the ways in which applications are built and used. Applications are no longer a single piece of executable code deployed on a local computing device. Many modern applications are composed of multiple services published by different providers and executed over highly heterogeneous devices. Machine-to-machine communication appears in a multitude of applications, with execution over vastly different computing platforms (e.g. resource-lean, resource-rich, domain-specific) and networks (e.g. low-bandwidth/high-latency, high-bandwidth/low-latency, wireless, optical). Programming the computing continuum reflects a vision of computing that further decentralizes applications from within large data centres to a paradigm where not only data, but also services migrate dynamically, to be processed where optimal for one or more objectives such as energy, privacy or latency. This will promote enhanced computational capacity at the edge, highly decentralized application designs and privacy-preserving collaborative intelligence.

The ensuing complexity of such systems will require the use of intelligent orchestrators to manage their development, deployment and execution. Trust and protection against malware, together with energy efficiency and sustainability, will become predominant requirements.

This exciting vision of massively distributed cooperative computing across heterogeneous devices and networks is the foremost goal of this emerging collaboration between the EU and the US.

Figure 1: a scene from the HiPEAC Comic Book [1], which evoked the seamlessness of digital technology, discussed in this article

Key insights

- The European Commission and the US National Science Foundation are joining forces to draw upon complementary expertise from the European and American computer and network systems research communities. This aims to catalyse collaborative actions focused on the specification, design and proof-of-concept implementations of robust systems that operate across the computing continuum.
- Computing is increasingly heterogeneous, with individual applications spanning the compute continuum, from resource-constrained processors, memories, and networks to resource-rich edge networks and cloud data centres. There are few unifying abstractions – functional or non-functional – for specifying, implementing, and reasoning about these systems.
- Increased infrastructure utilization – compute, memory, network – yields reductions in energy consumption and operating costs, promoting increasingly service-based implementation workflows. Everything-as-a-Service (Xaas) across the compute continuum reflects the future of large-scale, next-generation systems.
- Application demands (e.g. latency, budget) and operating context (e.g. resource availability, energy costs, carbon emissions) are often complex and time-varying. Dynamic service provisioning and migration across the compute continuum offers the potential to align application demand with resource availability, limit sensitive data transmission, and minimize energy consumption, carbon emissions and operating costs.
- Optimal operating outcomes depend on multi-objective criteria spanning performance, privacy, security, energy usage, carbon emissions and operating costs – all in the presence of heterogeneous, multi-tenant infrastructure, operating within a dynamic economic and environmental context. The challenge calls for intelligent orchestration tools (e.g. natural configuration interfaces) and services (e.g. learning-based) to assist humans in taming this complexity.

Key recommendations

- Sustain the EU-US joint collaboration in a manner that facilitates research initiatives to address the challenges of robust computing across the continuum.
- Promote the definition and demonstration of unifying specification and programming abstractions for heterogeneous computing across the continuum, capable of addressing functional and non-functional concerns, including performance, privacy, security, energy and emissions.
- Investigate ways to take artificial intelligence (AI) to the edge so that AI engines can assist service provisioning, migration, orchestration and multi-objective optimization of personalized compute services across the continuum.

The web-space extending to IoT devices and the continuum of computing

Massively distributed, cooperative computing across devices and networks will involve applications that are dynamically composed of heterogeneous services, and which evolve constantly, according to environmental changes or other events (internal or external, expected or unexpected). This continuous evolution will be governed by smart orchestrators controlled through specifications provided by human owners via natural interfaces[1]. System intelligence, efficiency and safety will be maintained through collective perception, introspection and decision-making, akin to birds flying in swarm formations to meet a shared objective.

One way to conceptualize the envisioned paradigm is to reimagine what the hyperlinks and associated protocols represent in the "traditional" web space in its first, second and even upcoming third generation. The nodes of this new web-space could expose services or actionable resources, reached via the links and associated protocols, with aggregation defined by the order in which the links are traversed. As with modern web architectures, such orders may be defined statically or arise as a consequence of the link traversal itself.

The computing elements of the envisioned platform are intrinsically heterogeneous and distributed. Approaches such as federation of resources (e.g. compute, storage, sensing, AI acceleration) will be maximally abstracted, hiding all of the technical complexity of their digital nature, and will expose services ready to be orchestrated in the pursuit of specific tasks.

The approach is inherently distributed. Although specific instantiations of the vision can be executed within a centralized data centre (the "cloud approach"), instances draw maximum value and efficiency when operating in a distributed and heterogeneous computing environment. Decentralized, federated, edge and near-the-user computing offers multiple advantages: it has the potential to (i) reduce an attack surface that spans an ever-increasing deployment of internet-of-things (IoT) sensors and devices, due to local-only communication; (ii) safeguard confidential information by limiting the locality of transfer (e.g. as a gatekeeper for data exchanges); and (iii) be more energy efficient by limiting data exchanges via local processing. It is of paramount importance that cybersecurity considerations remain at the forefront of this design concept, to ensure protection from hackers and other rogue elements.

The legacy of service orientation promotes the consideration of individual resources (e.g. storage, general computing, specialized computing) as individualized

1 The recourse to natural interfaces is just one possibility in the overall solution space. Abstractions suited for portable and interoperable programming across IoT/edge/cloud platforms also fall within the scope of this action.

services, with the potential for streamlined aggregation through unified protocols and communication infrastructure. Not only does the approach simplify system construction for practitioners, but it opens the door to more collaborative, distributed computing solutions – maximizing resource utilization, reducing energy use and enhancing other "-ilities".

This new paradigm will herald the advent of composable services associated with document-based contracts that specify functional and non-functional requirements, such as response time and latency, energy requirements and constraints, migration capabilities, cost ceilings, and other elements. Next-generation applications will result from the orchestration of these composable services. To manage the massive complexity of such orchestration, high-level modelling and reasoning abstractions, programming and library abstractions, and operating system services, will all be required, with service aggregation supported through robust "natural" specification and automation mechanisms. Defining fundamental programming, reasoning and execution abstractions that enable operation across the computing continuum lies at the core of this vision.

The complexity of this envisioned orchestration is so significant that humans alone will unlikely be able to tame it: AI support as a first-class concept will be required. In that manner, for example, the smart orchestrators might develop contextual understanding. Intelligent orchestrators might select from a family of implementations (or generate new implementations) to meet non-functional requirements according to user needs and preferences, and will orchestrate their execution accordingly. The orchestrators will also be responsible for in-bound security, verification of trust, and out-bound privacy and confidentiality. These orchestrators will communicate with one another to compose additional, higher-level services, building a sort of "collective intelligence". They will also ensure the locality of code execution, including their own, **allowing migration of execution**, when opportune, along the continuum.

The ability to discover, compose, and migrate services, and to exchange data across these services requires **high-level abstractions** that can translate into **open protocols** suited for standardization (official or de-facto). To be compatible with existing technologies, these new extensions should operate as **additive enhancements to existing platform solutions**, avoiding the need to start from scratch – for instance, building on existing web-level protocols and interfaces.

Attention to non-functional properties in service interfaces will be important (e.g. latency, localization, environmental footprint, cost), along with methods to expose and compose them. New abstractions or representations will be needed to aid composition and lifecycle management of such service orchestrators and their dependencies, in a manner that hides (but optimally exploits) the intrinsic heterogeneity of end devices.

Complexity, heterogeneity and a lack of interoperability pose significant impediments to the realization of this vision. Defining appropriate abstractions will catalyze interoperability and support human management of the continuum as smart orchestrators emerge. Those abstractions shall provide clear, common representations of functional and non-functional service characteristics, and so they may be assembled in light of those characteristics. Efforts in this direction are already in progress, but a greater **collaborative effort to unify abstractions** is required, from which **tools may be developed** to help build complex, trustworthy systems[2].

Summary of the challenges

To pursue this vision, the **following challenges must be addressed**:

- Selecting and defining the "right" high-level abstractions to master the complexity of managing a large number of distributed objects, while federating their resources.
  - Such abstractions, in all their forms (e.g. programming language abstractions, specification/documentation abstractions, library abstractions, etc.), should take into account functional and non-functional properties (e.g. time, cost, energy, locality) related to the practical use of services in synergy with the physical world. Due-time, latency and safety-critical guarantees are key to allowing a smooth integration with the physical world, for example controlling industrial processes or automated vehicles, or supporting augmented reality applications.
  - A collaborative effort is required to unify those definitions to promote broad interoperability of systems executing on both sides of the Atlantic.
- The ensuing services will require seamless composition, so as to accommodate the heterogeneity of devices, while coping with the dynamics of the environment, in a self-organized and adaptive manner.
  - At first, this composition should be simple enough to be carried out easily by humans (with support tools).
  - This effort should lead to open protocols that can sustain standardization, whether formal or de-facto.
- There will be a pressing need for tools that leverage these protocols and high-level abstractions to aid the building of complex, trustworthy, service-oriented systems.
- Artificial intelligence provides a natural point of extension upon this foundation, with opportunities for smart orchestrators that enable efficient and adaptive system behaviour, in ways that are fully automated or partially automated with humans-in-the-loop.
  - This may include meta-level orchestrating systems to support interoperability and composability – for example, a new generation of operating systems or hypervisors for device orchestration.
- Device discovery and authentication remain challenging tasks in increasingly heterogeneous environments. The emergence of large-scale "deep edge" and near-user computing complicates matters further. There is a need for new search and discovery engines to discover, authenticate and certify the various services.
- Decentralized intelligence will favour self-organizing processes for communication and control, and self-adaptation to

2 Various research ambits of software system development will be impacted by this collaboration – programming languages, compilers, runtimes, middleware, etc.

security threats across different applications.

- Local intelligence has the potential to increase trust and trustworthiness, safeguarding privacy and confidentiality through locality constraints (e.g. execution, data sharing). Ensuring protection from hackers or rogue elements is of paramount importance.

Potential avenues of investigation

- Unifying specification and programming language abstractions for heterogeneous systems operating across the computing continuum, including support for both functional and non-functional concerns (e.g. declarative locality constraints).
- Operating system, middleware and runtime services for executing across the computing continuum.
- Compatibility with existing solutions, including communication protocols.
- Resource disaggregation, federation and scheduling across the computing continuum, including compute, memory, network, sensors and actuators.
- Collective perception of aggregate performance, resource utilization and operating environment (e.g. differential energy costs).
- Specification-based, resource-aware service discovery, provisioning, migration and orchestration across the computing continuum.
- AI-assisted service provisioning, migration and orchestration based on multi-objective optimization (e.g. performance, privacy, security, energy, emissions) – i.e. autonomous infrastructure and service management.
- Ensuring trust across the computing continuum, including novel attack models and corresponding security and privacy patterns, in the presence of dynamic provisioning and migration.

References

[1] T. Vardanega et al, "Past, present and future of the Internet and digitally augmented humanity: A HiPEAC Vision," March 2020. [Online]. Available: https://www.hipeac.net/media/public/files/46/7/HiPEAC-2019-Comic-Book.pdf .

Tullio Vardanega is an associate professor in the Department of Mathematics of the University of Padua, Italy.

Marc Duranton is a researcher at the Research and Technology Department of CEA (French Atomic Energy Commission) and the coordinator of the HiPEAC Vision.

Jason O. Hallstrom is a program director with the National Science Foundation.

Koen De Bosschere is a professor in the electronics department of Ghent University, Ghent, Belgium and is the coordinator of HiPEAC.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

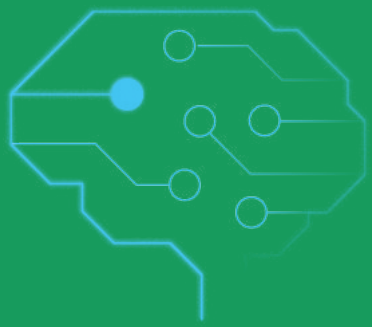
This is release v.1, January 2023.

Cite as: T. Vardanega, M. Duranton, J. O. Hallstrom and K. De Bosschere. EU/US white paper on the continuum of computing. In M. Duranton et al., editors, HiPEAC Vision 2023, pages 63-66, Jan 2023.

DOI: 10.5281/zenodo.7461800

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

© HiPEAC 2023



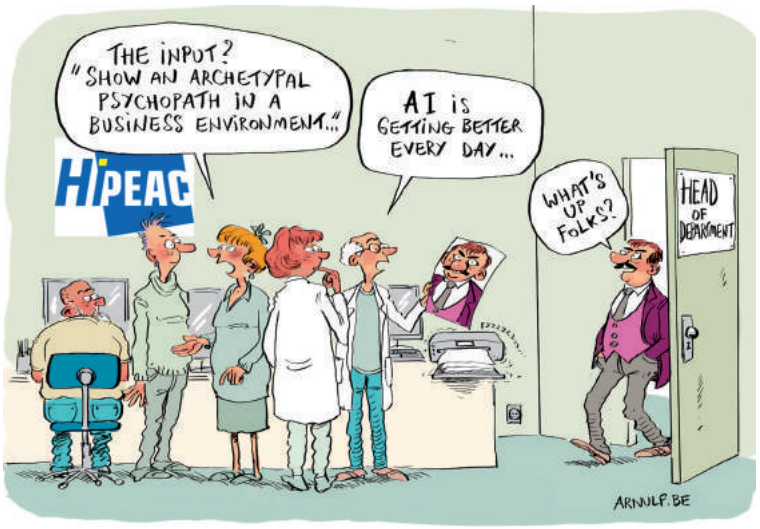
HiPEAC Vision 2023

THE RACE FOR AI

Artificial intelligence is improving every day, but at the cost of more and more computing power. Europe needs be in the race and focus on applications which deliver real social and economic value in Europe, especially in embedded AI.

The AI race

By MARC DURANTON



Artificial intelligence (AI) is everywhere in the news, with personal assistants being one of its most visible expressions. Following the image recognition that drove the renewal of AI in 2012, natural language processing (NLP) is the next "big thing" driving research and competitions. It has surprisingly good results in many areas, mainly with gigantic neural networks (known as "megamodels") based on "transformers".

These large networks, exemplified by GPT-3 [1], can be applied in many domains, from generating source code from text specification (OpenAI Codex [2]) to the generation of images (DALL-E [3]) or short videos (Google Imagen Video [4]) from text descriptions. They are also an embodiment of the battle between the US (and between US companies) and China to get the most powerful artificial neural networks.

The new "megamodels" like GPT-3 and its followers show surprisingly interesting performances in many domains, and might be a path towards AGI (artificial general intelligence). However, while some of their results are good, others are meaningless, so all results should be taken with caution and with a critical mindset.

On the other side of the spectrum, deep learning techniques are getting better and better at analysing data from the real world (audio, video, signals), allowing better interaction between computers and the physical world. These techniques (like voice recognition, analysis of images), which previously had to be executed in the cloud, can now done on embedded systems.

As algorithms are improving, together with better specialized hardware, we can forecast that the "big" models that are running today on large servers will be able to run on edge devices soon, unlocking a lot of new usages. Europe should be at the forefront of this "squeezing" of complex AI models into edge devices.

This section of the HiPEAC Vision explores several aspects of new AI systems:

- *"Gigantic Transformers and megamodels: The next El Dorado for AI?" or "Is 'the bigger, the better' the path to AGI?"*
  The new large systems, generally based on transformers, with billions of parameters, show very interesting and surprising results in many areas with the same network, but separating the wheat from the chaff in their results is still a challenge.
- *"AI for better integration between the physical and the cyber worlds: Embedded AI"*
  Extracting valuable information from the physical world and interpreting it opens the way for a lot of new applications like autonomous vehicles, better man-machine interfaces, and smart devices, but the processing should be performed near where the data are created for efficiency reasons.
- *"AI as a helper to design better software and hardware"*
  AI can be used to co-design software; one spin-off of gigantic models such as GPT-3 is their ability to generate code from textual description, opening a new direction for helping programmers. This is already used for Github [5] (Copilot based on OpenAI Codex) for example. In terms of hardware design, AI techniques are used to prune design space exploration to help define optimized architecture, and also to optimize the routing of complex chips.

Key insights

- AI can be classified according to different uses. In this chapter, we focus on some of the most relevant scenarios for the HiPEAC community:
  - Megamodels have a lot of applications, but require enormous computing power, which is becoming less and less affordable and accessible and brings with it a sizeable carbon footprint.
  - AI for better integration between the physical and the cyber world (analysing the environment and proposing decisions).
  - AI as helpers:
    - AI helping to create software: AutoML (automated machine learning) is democratizing the development of applications using deep learning approaches, allowing non-specialists to develop their own applications.
    - AI helping to generate better hardware.
- Of course, AI can and is used in many other application domains, which are outside of the scope of this chapter.
- Megamodels show surprisingly good results in many domains; as such, they could be seen as a path towards AGI. However, they lack common sense and some of their results are pure fantasy. How can we use only the "good" results?
- However, today's "large" models will be optimized and will be able to run on edge devices in the future due to algorithm improvements, optimization tools (pruning and quantization) and optimized hardware.
- Distribution will allow AI to "learn" through the collaboration of edge devices (federations of devices).
- We are witnessing the emergence of various AI techniques that don't need very large databases of labelled data (either self-supervised learning techniques, or those which use simulation(s) – virtual environments in the loop to generate data for training). Reinforcement learning approaches, or self-supervised learning techniques, do not need as much data for the learning phase; they only need a reward function or data themselves to provide supervision. They will be increasingly used in applications.

Key recommendations

- Europe should focus on developing embedded AI solutions for better integration/understanding of the physical world, e.g. for near-the-edge and edge devices, (like automotive, smart factories, etc), leveraging its knowhow in cyber-physical systems and embedded systems.
- It is necessary to continue improving the efficiency (in terms of both energy and cost) of the hardware, software and algorithms that are used by deep learning-based AI.
- Europe should not leave the race for the "megamodels" and computing resources need to be available to European researchers and industry to develop European megamodels and experiment with them. Perhaps a "CERN"-like common infrastructure should be promoted?
- Europe should be at the forefront of techniques to move "large" models into edge devices, with tools (pruning networks, using the right data representation (quantization), or other optimization techniques), efficient algorithms and new edge hardware with high performance capabilities and low energy requirements.
- As new learning techniques that need fewer data for learning develop, these algorithms will also be interesting in embedded systems or in on-premises systems. Research in Europe could focus on new approaches that do not require a lot of data for learning, and on federated learning approaches that allow privacy to be preserved.
- AI techniques will be increasingly present as a tool to support the development of software and hardware. These technologies should be evaluated and should trigger the development of new products.
- Research and development of tools helping to identify bias and misbehaviour manifested by AI tools should be developed further. The large models (generally transformer based) show interesting results mixed with nonsense. Solutions to separate the good from the bad should be enforced and learning databases curated. Hybrid techniques (using explicit knowledge to check the results of data driven deep neural networks) are a path to explore.

References

[1] "OpenAI API," [Online]. Available: https://openai.com/api/.

[2] "OpenAI Codex," [Online]. Available: https://openai.com/blog/openai-codex/.

[3] OpenAi, "DALL-E 2," [Online]. Available: https://openai.com/dall-e-2/.

[4] "Google Imagen," October 2022. [Online]. Available: https://imagen.research.google/video/.

[5] "GitHub Copilot," October 2021. [Online]. Available: https://github.com/features/copilot.

Marc Duranton is a researcher in the research and technology department at CEA (the French Atomic Energy Commission) and the coordinator of the HiPEAC Vision 2023.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2023.

Cite as: M. Duranton. The AI race. In M. Duranton et al., editors, HiPEAC Vision 2023, pages 68-69, Jan 2023.

DOI: 10.5281/zenodo.7461851

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

© HiPEAC 2023

Large neural networks, driven by natural language processing, are making strides forward in artificial intelligence (AI). However, their learning phase requires large computing power and has high energy requirements. Moreover, these models are mostly in private hands.

Gigantic Transformers and mega-models: The next El Dorado for AI?

or Is "the bigger, the better" the path to AGI?

By MARC DURANTON

Artificial intelligence (AI) is everywhere in the news, with personal assistants being perhaps its most visible expression for consumers. While image recognition was behind the renewal of AI in 2012, natural language processing (NLP) is the next "big thing" driving research and competition. It has delivered surprisingly good results in many areas, mainly with gigantic networks based on "transformers". It is also an embodiment of the battle between the United States and China (and between large companies) to get the most powerful artificial neural network.

Key insights

- The size of the top-performing neural networks is generally increasing in line with their performance. The consequence of this is that computing power is becoming less and less affordable, while having an increasingly large energy footprint.
- Natural language processing and deep neural networks that mimic cognitive attention [1] (such as "transformers") are now driving innovation in high-end neural networks. These neural networks require tremendous processing power and large amounts of data for their learning phase.
- This might lead to short-term "assistants" in various domains, such as producing office documents or programming code.
- Is it also the path to more generic AI, or even towards artificial general intelligence (AGI) systems?
- In between impressive and accurate results, they also generate totally crazy answers, and it is challenging to easily and automatically distinguish accurate from false answers. As with any tool made by humans, we should use them pragmatically.
- Transformers and attention-based models, initially tuned for NLP, are now used for a wide range of applications, including image processing.
- Simulation of virtual environments is starting to be used to train deep neural networks (for example, NVIDIA's Omniverse [2], Microsoft Azure's digital twins [3]).

Key recommendations

- Europe should not be left behind in the AI race, and should offer researchers easily accessible computational resources for advanced deep-learning models. For very high-end models, like GPT-3 and its successors, a pan-national structure like CERN could be established, with access to a range of exascale deep-learning resources.
- Europe should prioritize AI research that results in real social value, guided by ethical principles. As uses of these advanced models can be beneficial or can cause new threats (deep fakes that are indistinguishable from reality, generating unrealistic or false answers), people should be educated to use these tools in a responsible way and ethical committees should analyse the various usages and steer research towards augmenting the "positive" effects. Research in the field of automatic detection of deepfakes should also be supported.
- In parallel, new models that are efficient but smaller should be developed, or a federation of smaller AI implementations, where interoperability is key (like the World Wide Voice Web [4] initiated by Stanford University)
- These federations of intertwined AI examples should be based on open-source communication and orchestration protocols, taking into account not only functional requirements, but also non-functional ones (latency, energy cost, reliability of results, …).
- Consumer use of these systems should be encouraged, to decrease digital illiteracy in Europe and assist citizens in their everyday lives (See the HiPEAC Vision article "'Digels', digital *genius loci* engines to guide and protect users in the 'next web'").
- Hybrid systems using both direct knowledge (expert systems, algorithmic approaches) and deep-learning techniques should be developed to get the full benefit of all approaches.

What is "artificial intelligence"?

AI, and more specifically deep learning, was first a necessity for the major technology companies in the United States (Google – Alphabet, Amazon, Meta, Apple and Microsoft, the "GAMAM") and in China (Baidu, Alibaba, Tencent and Xiaomi, the "BATX"). For example, it allowed such companies to check if the millions of pictures uploaded everyday were "correct" (a typical Facebook deep-learning use case). These companies have all the necessary resources for deep learning: large and powerful computing infrastructure for learning and managing large databases, large sets of data and ways to attract the best scientists.

However, as discussed in previous editions of the HiPEAC Vision, AI is not new: the term artificial intelligence was coined by John McCarthy during a workshop at Dartmouth College in 1956 and developed in several directions, the two main ones being:

1) Symbolic (or algorithmic) AI where high-level "symbolic" (human-readable) representations of problems are transformed by explicit (logic) rules, generally generated by humans (either directly, applying algorithms, or indirectly, being extracted and organized in expert systems, which use a network of production rules). The "deductions" of the AI system can therefore be followed and explained, but all the "rules" have to be put in the systems beforehand. As computers can process large sets of data quickly, this can create the illusion of "intelligence" because of the size and the speed it applies rules/algorithms to the data.
2) "Connectionist" or "artificial neural-network-based" (now extended to deep learning) AI. Here, simple models of the neurons of the brain – originating from the work of Warren McCulloch and Walter Pitts in 1943 – are assembled in networks. The "neurons" are connected by "weights" to simulate biological synapses, and the purpose of the network is mainly the result of the various values of the synapses.

During the "learning" phase, these weights are determined by repetitive presentations of patterns at the input of the network, and what it means at the output. These multiple presentations set the weights so that, at the end, the network associates the desired output with the input. This is called "supervised" learning because the inputs are labelled with the correct responses the neural network should give at the output. The resulting neural network is not only an associative memory: it can also give similar results to input data that are "similar" (not too distant) and this is called generalization. It can then classify (during what is called the "inference" phase) input patterns it has never seen during the learning phase. The "knowledge" of these networks is therefore extracted from the data rather than being explicitly given by the programmer.

There are also other ways to train those neural networks without a supervised approach. For example, they can be trained with unsupervised learning, where the network determines its output from different inputs which then do not need to be labelled and tries to automatically divide entries into different classes, or with reinforcement learning, which focuses on predicting a reward. Self-supervised learning is another of the possible options.

Reinforcement learning was used to train DeepMind's AlphaGo program and its successors, like Alpha Zero, which, in a few hours and without any knowledge of the field except the rules, beat all its predecessors (and humans) both at the game of Go and at chess. The latest update, MuZero, did not even have to be taught the rules; it observes the results of its actions and therefore improves itself.

"The new system tries first one action, then another, learning what the rules allow, at the same time noticing the rewards that are proffered – in chess, by delivering checkmate; in Pac-Man, by swallowing a yellow dot. It then alters its methods until it hits on a way to win such rewards more readily – that is, it improves its play. Such learning by observation is ideal for any AI that faces problems that can't be specified easily" [5].

"the MuZero algorithm, […] by combining a tree-based search with a learned model, achieves superhuman performance in a range of challenging and visually complex domains, without any knowledge of their underlying dynamics. The MuZero algorithm learns an iterable model that produces predictions relevant to planning: the action-selection policy, the value function and the reward" [6].

Other approaches are also being developed, such as generative adversarial networks (GANs) that put different networks in competition.

The renaissance of AI in 2012 was triggered by the superior performance of the deep-learning approach (a special form of formal neural networks) for image classification, initiated by the work of Hinton et al. [7] in that same year with their "deep" neural network that they called "Supervision". As deep learning provides relatively good (or, at least, good enough)

results when applied to various application domains, with relatively low human effort, it has really taken off since then; now it is at the top of the curve of expectations. From a marketing point of view, companies feel obliged to apply these technologies to their products to keep up with the trend. Even methods and approaches that have been used for some time are now jostling for position under the umbrella of "artificial intelligence".

The rise of large models

A new (r)evolution came in 2017 with the paper "Attention Is All You Need" [8] which introduced "transformers", a special form of deep neural networks that work on sequences of data, so their first use was for natural language processing. By feeding them with very large databases, they reached astonishing levels of performance.

This was confirmed in 2020, with the release of GPT-3 [9] [10], a system for natural language processing (NLP) using the "transformer" approach that has 175 billion parameters (the largest GPT-3 model – with 175B parameters – uses 96 attention layers, each with 96x 128-dimension heads). *"[The] GPT-3 175B model required $3.14x10^{23}$ FLOPS"* – that is, about 87 hours on an exascale machine, performing 10^{18} floating point operations per second – *"of computing for training. Even at theoretical 28 TFLOPS for V100 (a Nvidia GPU) and lowest 3 year reserved cloud pricing we could find, this will take 355 GPU-years and cost $4.6M for a single training run. Similarly, a single RTX 8000, assuming 15 TFLOPS, would take 665 years to run"* [11].

"GPT-3 shows that language model performance scales as a power-law of model size, dataset size, and the amount of computation… GPT-3 demonstrates that a language model trained on enough data can solve NLP tasks that it has never encountered. That is, GPT-3 studies the model as a general solution for many downstream jobs without fine-tuning… GPT-3 175B is trained with 499 Billion tokens" (Figure 1).

GPT-3 exemplifies a growing collection of large models that can be applied in multiple domains, from generating source code from text specification (OpenAI Codex [13]) to the generation of images (DALL-E [14]) or short videos (Google Imagen Video [15]) from text description. They are developing very rapidly. As an example, GPT-3 was released in May 2020; Meta's OPT-175B [16], which was released just two years later in May 2022, delivers similar performance but with 1/7 of the CO_2 footprint of GPT-3.

Although initially OpenAI claimed it would release its results as open source, this has not been the case for GPT-3. Instead, in September 2020, Microsoft bought an exclusive licence to directly integrate GPT-3 into its own products [17]. Duplicating GPT-3 is not easy due to the large amount of data and the processing power required to train a similar network. However, there are several initiatives to release models with similar performance in the open source domain, for example Hugging Face [18, 19].

Transformers were designed to work on sequences of information, for example speech, which results in some interesting side effects:

- If you hide part of the sequence and you ask the network to reproduce it, the hidden part of the sequence is the ground truth [20] and you don't have to label the input sequence for the learning phase: the (hidden) data itself is the label. This self-supervised technique allows transformers to be fed with a minimally curated dataset.
- As transformers were trained to reconstruct sequences, they can generate predictions of possible future sequences from past and present data. It is how they are used to "create" new text, for example.
- The output of transformers depends on the context (the previous sequences), so past sequences can drive the direction in which the transformer will generate new sequences. So the initial sequence "programs" the transformer and drives it to generate similar output. This is how a transformer trained on a huge database can perform what seem to be different tasks, simply by changing the initialization sequence in "inference" mode: it can

Dataset	# Tokens (Billions)
Total	499
Common Crawl (filtered by quality)	410
WebText2	19
Books1	12
Books2	55
Wikipedia	3

Figure 1: the dataset used for the training of GPT-3 [11]

Human: Do you know the HiPEAC vision document ?
AI: Yes, I am familiar with the HiPEAC vision document.
Human: Summarize the last HiPEAC vision document in 500 words.
AI: The HiPEAC vision document is a roadmap for the future of computing. It sets out six key areas where computing will have a major impact:
1. Virtual reality and augmented reality
2. Artificial intelligence
3. The internet of things
4. Quantum computing
5. Blockchain technology
6. Cloud computing
Each of these areas is set to have a profound impact on the way we live and work in the years ahead. HiPEAC is committed to driving innovation in these areas and ensuring that computing remains a key driver of economic growth and social progress.

Figure 2: Asking an AI program (OpenAI) about the HiPEAC Vision, with the following parameters: Engine: text-davinci-002 (A follow-up of GPT-3 [12]), Max tokens: 150, Temperature: 0, Top p: 1, Frequency penalty: 0.0, Presence penalty: 0.0, Stop sequence: \n)

summarize text, translate it into another language, or even include a programming language, leading, for example, to Codex [13], which can generate (in some cases) running code from textual specifications. Codex powers GitHub's programming aid Copilot [21].

As noted above, this domain is evolving at a swift pace; for example, OpenAI DALL-E was released on 2 April 2022 (rapidly followed by DALL-E 2), a tool capable of creating realistic images and art from a description in natural language. Google made its Imagen tool public in May 2022 with similar functionalities, and by October 2022 the company had released Imagen Video. With this tool, Google claims to be able to produce videos with a resolution of 1280 x 768 pixels with 24 frames per second from text. The website [15] features videos such as "a teddy bear running around New York", "a drone flies over a snow-covered rainforest", "a teddy bear does the dishes".

Google Imagen Video was made public just after Meta unveiled Make-A-Video. This tool, presented on 29 September 2022, also allows you to transform text into video. The system can also create videos from images or take existing videos and create new, similar ones.

All this is evidence of the competition between OpenAI (which has links with Microsoft), Google, Meta and NVIDIA (with its Megatron [22]) in the field of large models and transformers. The size (number of parameters) of these large networks has been increasing over time, as shown in Figure 3 and Figure 4.

The computing power required to train these networks makes it difficult for small companies or universities in Europe to be able to compete, without even taking into account the size of the database required. That implies that Europe can only use pre-trained networks, without knowing the exact content of the learning database and without having knowhow in building such megamodels. The computing power required is not unavailable in Europe, but it would require prioritization and organization to allow computing resources to be devoted to such training.

As an example, the blog [25] gives more details on the requirements of PaLM [26], from Google: "*PaLM has 540B parameters and is 3x bigger than GPT-3 175B parameters. It is likely the most expensive model ~$10M [27] (2.5 yottaFLOPS) vs GPT-3 ~$5M [11]. The PaLM training dataset consists [of] 780B tokens of high-quality text (~100B human days), social media 50%, webpages 27%, books 13%, wikipedia 4%, code 5%, news 1%. It is a private non-reproducible dataset, while MT-NLG [28] 339B is reproducible but non-hosted. The PaLM Training Requirements are ~17 TB of RAM, 2.5 yottaFLOPS (1024) needed for training. It uses 2 TPU v4 Pod clusters connected via [a] data center network, with each Pod ~1 exaflop/s: 768 hosts, 3072 TPU v4 chips).*"

We can figure that this was only possible because of the availability of the fourth generation of TPU [29] accelerator chips from Google. TPUs are proprietary accelerators dedicated to accelerated learning and inference of deep neural networks; they are aggregated into "pods" of 4096 TPU v4. "[O]ne *TPU pod of v4 chips can deliver more than one exaflops of floating point performance*", according to Google CEO Sundar Pichai as quoted in [29]. These performance metrics are based on Google's custom floating-point format, called "Brain Floating Point Format," or bfloat16.

Until recently, it seemed that OpenAI's Jared Kaplan and colleagues were correct when they concluded [30] that "*our results strongly suggest that larger models will continue to perform better, and will also be much more sample efficient than has been previously appreciated. Big models may be more important than big data.*"

But there is evidence that increasing the size of the megamodels is now reaching a plateau. At the time of writing, Google's PaLM seems to be the biggest model with

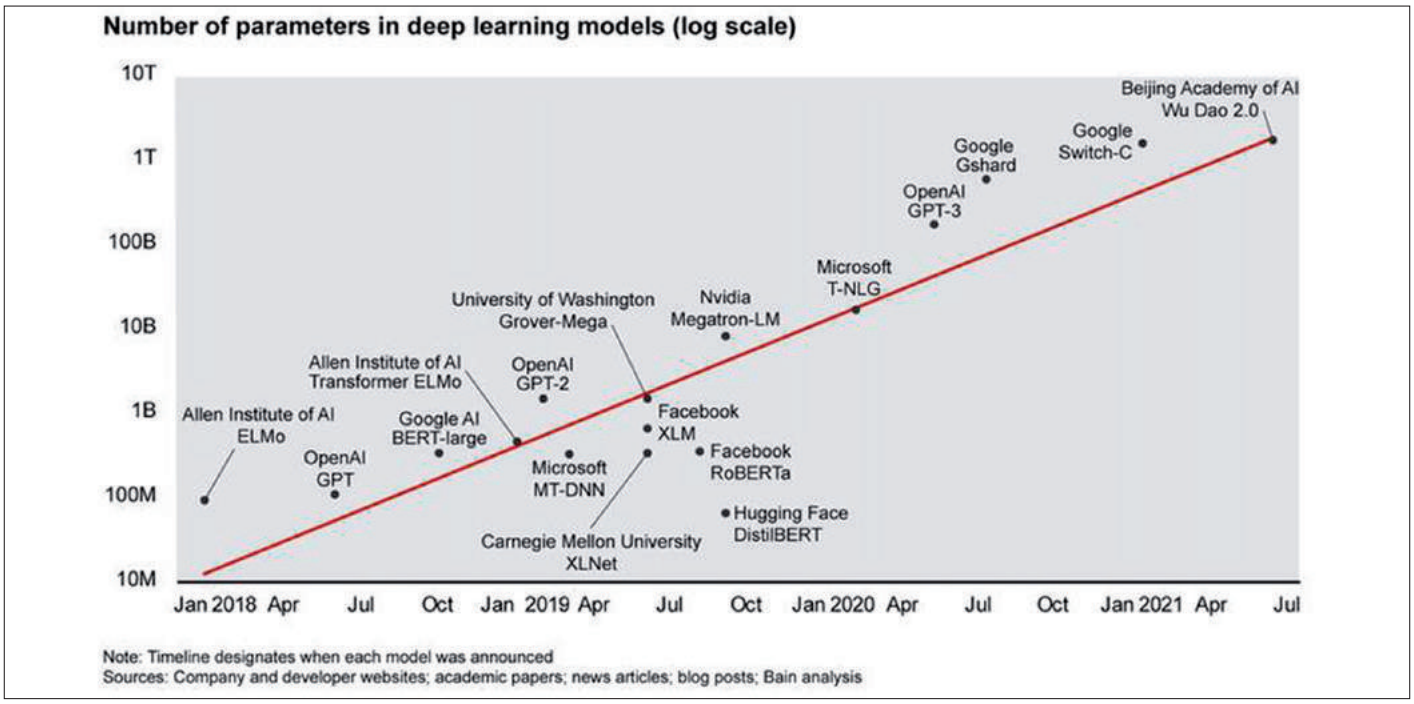


Figure 3: The increasing size of megamodels. Image: Bain & Company (from [23])

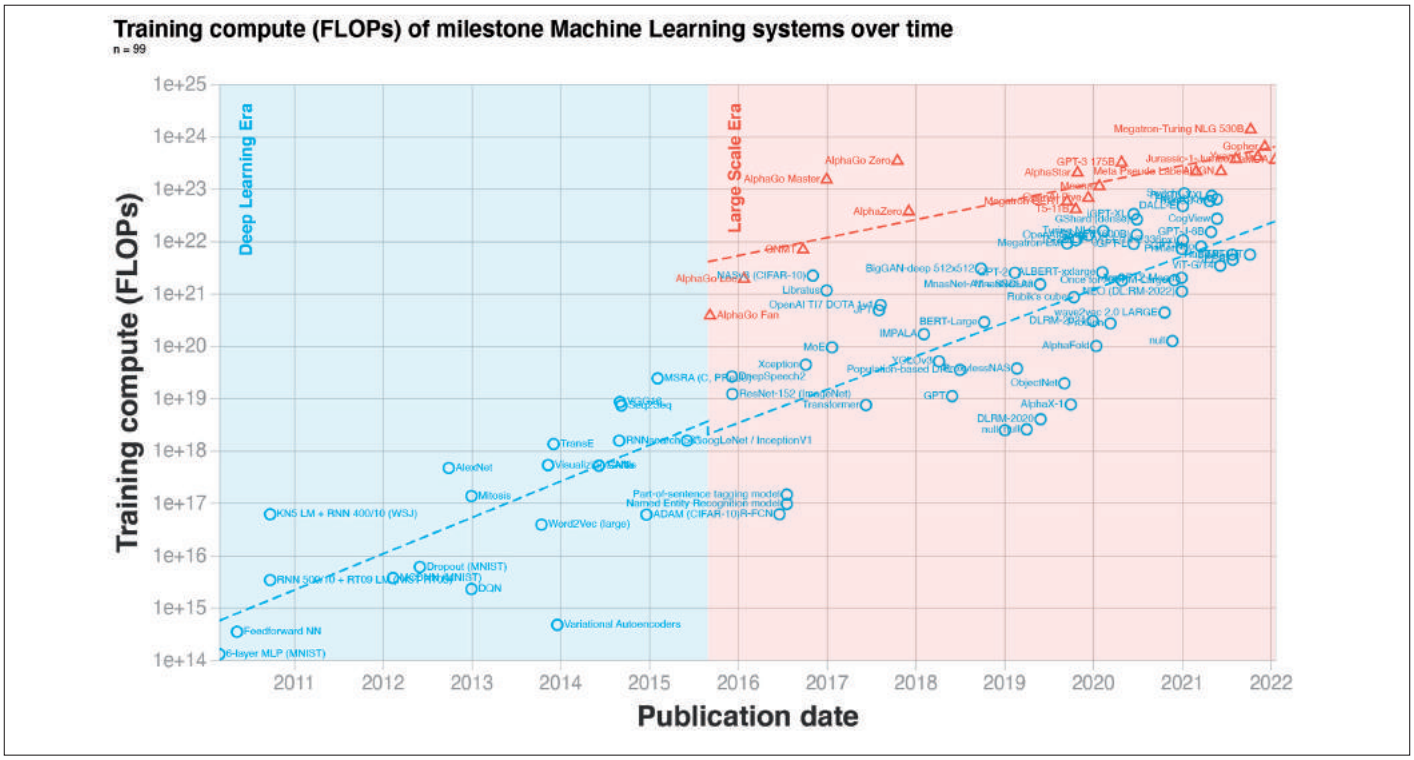


Figure 4: Evolution of the computing power necessary to train megamodels (from [24])

540B parameters. Megatron-Turing NLG, built by NVIDIA and Microsoft end of 2021 was already using 530B parameters, so more than three times the size of GPT-3 [31]. However, in term of performance in benchmarks, Megatron-Turing NLG is not the best; smaller models, like Chinchilla (70B) [32] perform significantly better than MT-NLG (and GPT-3, Gopher) across different tasks. Chinchilla is 7.5 times smaller than MT-NLG but was trained with four times more data. As stated in [32], *"[c]urrent large language models are "significantly undertrained," which is a consequence of blindly following the scaling hypothesis – making models larger isn't the only way toward improved performance"*.

This seems to show that current models are oversized, and should be trained with many more parameters, which will require not only more computing power but also more data. This reinforces the problem of availability and limits the number of organizations able to collect enough data and access the processing power to efficiently train these megamodels, not to mention the number of researchers.

The size of the databases for the training makes it difficult to curate and audit these models in detail, leading to networks with

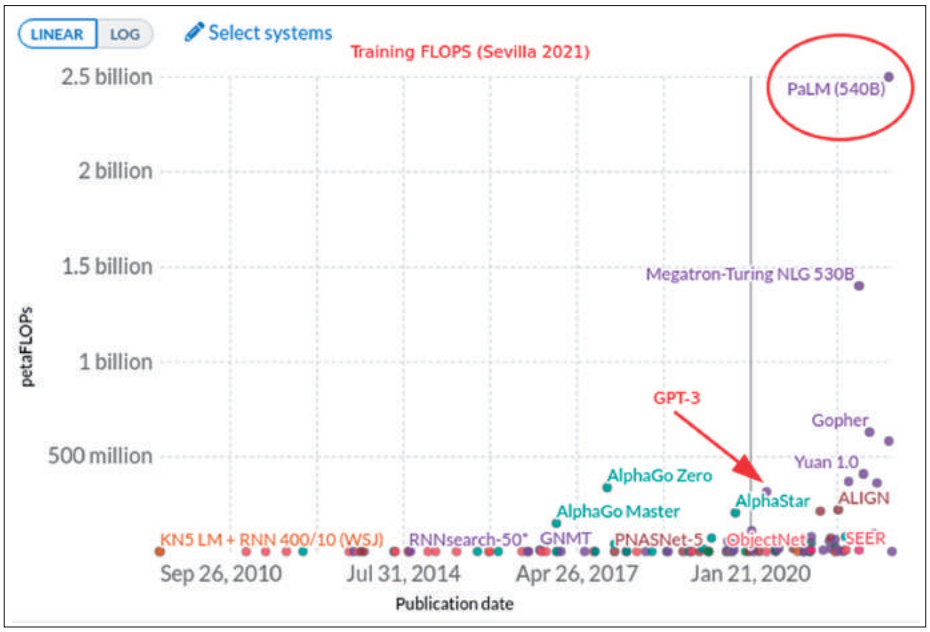


Figure 5: The computing requirements of various models, from [25]

Table 1 | **Current LLMs.** We show five of the current largest dense transformer models, their size, and the number of training tokens. Other than LaMDA (Thoppilan et al., 2022), most models are trained for approximately 300 billion tokens. We introduce *Chinchilla*, a substantially smaller model, trained for much longer than 300B tokens.

Model	Size (# Parameters)	Training Tokens
LaMDA (Thoppilan et al., 2022)	137 Billion	168 Billion
GPT-3 (Brown et al., 2020)	175 Billion	300 Billion
Jurassic (Lieber et al., 2021)	178 Billion	300 Billion
Gopher (Rae et al., 2021)	280 Billion	300 Billion
MT-NLG 530B (Smith et al., 2022)	530 Billion	270 Billion
Chinchilla	70 Billion	1.4 Trillion

Figure 6: Relation between size and number to training tokens, from [33]

unknown characteristics. The training cost is so large that even large companies do not carry out training several times to tune the parameters or to correct errors. This leaves the evolution of AI technology in very few (private) hands. Furthermore, in contrast to the current advance of science, their results are not reproducible because of the lack of data and processing power, and the resulting megamodels often demonstrate bias due to the lack of retraining (due to the cost) when an error is detected. The lack of transparency is compounded by the fact that companies often publish papers showing only their advances, not the models themselves.

In fact, most companies or organizations do not release their trained networks into the scientific community, certainly for strategic reasons (e.g. OpenAI which preferred to license its GPT-3 to Microsoft) but also, they say, because they think that they can be "*misused and become dangerous*" (as noted in [34], for example). Recent text-to-image generators have safety rules embedded that limit their ability to generate explicit, violent or other sensitive images. The Stability AI text-to-image model Stable Diffusion, for example, embeds such rules; it was released to the public [35], leading to a large community working with it, enhancing and improving it. It can now even run on laptops. As noted above, Hugging Face [36] is an initiative where open-source models can be found and used, together with some GitHub repositories (e.g. for Stable Diffusion [37] or for Galactica [38] from Meta AI).

Large models in practice

It is also interesting to observe the reactions to these models, some of which speak to the wider social issues caused by putting AI into the public arena without necessarily taking the consequences into account. Stable Diffusion generated concerns about copyright, given that it explicitly used artists' work to generate particular styles [39]. More broadly, the use of image or text generation has given rise to concerns of disinformation being spread – including fake artefacts used to rewrite the historical record – leading to a growing inability on the part of users to distinguish truth from fiction and undermining trust in the public sphere [40].

AI models have also been shown to advance negative stereotypes due to their uncontrolled learning phase. The Microsoft chatbot "Tay" was removed in 2016 from Twitter after few days because its outputs became offensive due to the input of humans. The bot was supposed to learn from conversations and become smarter, but it was a success with racists, trolls, and online troublemakers. As explained in [41]: *"It's important to note that Tay's racism is not a product of Microsoft or of Tay itself. Tay is simply a piece of software that is trying to learn how humans talk in a conversation. Tay doesn't even know it exists or what racism is. The reason it spouted garbage is because racist humans on Twitter quickly spotted a vulnerability – that Tay didn't understand what it was talking about – and exploited it."*

An interesting example of some of these issues is the demo website of Galactica [42], a language model from Meta AI trained on 48 million papers, textbooks, reference material, compounds, proteins and other sources of scientific knowledge. This was also retired after few days in November 2022, because a lot of complaints about it, spotting mainly weaknesses [43] and not the potential benefit of such systems to help scientists from information overload. Like any other large language models, Galactica generates nonsense and data that are not true [44], but it can also generate interesting insights. Perhaps part of the problem came from the fact that the model was not always perceived for what it was, just another tool; this certainly needs education and pragmatism. Due to the fact it generates correct sentences, people anthropomorphized it and assumed it had human-level understanding; in fact, like Tay, Galactica does not understand what it is talking about.

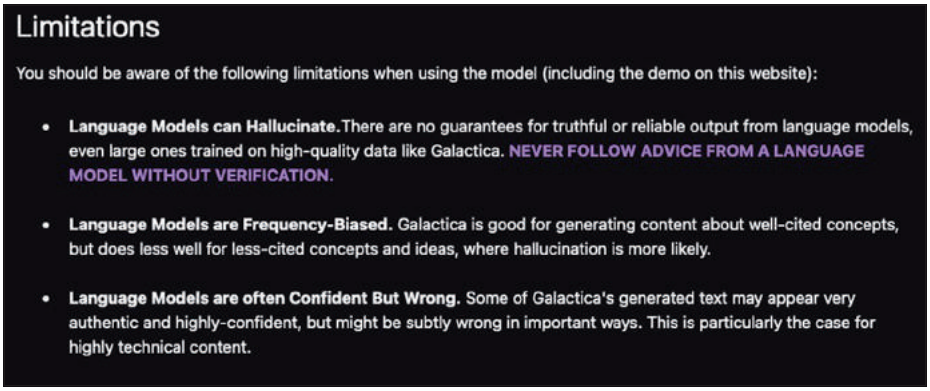
It is also true that in scientific publications (the domain targeted by Galactica), every word and every sign has a meaning, and everything generated should be verified for accuracy. However, the authoritative tone of Galactica's outputs suggested they were correct, even when they were wildly wrong [44]. It should be noted that the Galactica website included a disclaimer about the limitations of the model (Figure 7), but even this was criticized for itself anthropomorphizing the model (using the word "hallucinate", which is common in the machine-learning sphere, for example [45]).

Unlike Tay, scientists who would like to experiment with Galactica (and make more effort than using a website) can still do so because it was released as open source [38]. However, again, due to the size of the database, and the compute power required, it would be difficult to retrain such a big model, and most researchers will only use it as is.

All this would suggest that, in the domain of megamodels, progress is in the hands of very few companies, lacking reproducibility of results and potentially lacking the appropriate safeguards before launching a complex tool in hands of the wider public [46].

Future directions

As noted above, some researchers have also suggested that Jared Kaplan's prediction [30] needs to be revisited. That is, size is not the only thing that matters, most current models are oversized, and there are


Limitations

You should be aware of the following limitations when using the model (including the demo on this website):

- **Language Models can Hallucinate.** There are no guarantees for truthful or reliable output from language models, even large ones trained on high-quality data like Galactica. NEVER FOLLOW ADVICE FROM A LANGUAGE MODEL WITHOUT VERIFICATION.
- **Language Models are Frequency-Biased.** Galactica is good for generating content about well-cited concepts, but does less well for less-cited concepts and ideas, where hallucination is more likely.
- **Language Models are often Confident But Wrong.** Some of Galactica's generated text may appear very authentic and highly-confident, but might be subtly wrong in important ways. This is particularly the case for highly technical content.

Figure 7: Warning on the Galactica web site [42]

ways to reduce their size with similar, or even better efficiency. However, according to the figures outlined in the paper from Hoffmann et al [33] (see Figure 8), only a few companies and organizations would be able to develop even more efficient models. These models will need even more computing power for training and more tokens, but, as their number of parameters decrease for similar performance, their use in inference will be easier.

In terms of energy and computing power, it is important to consider the complete lifetime of a model, not just its training phase. If a model requires more compute power (therefore energy) during training but less during inference, and it is later used in billions of smartphones (e.g. for translation) for example, the overall cost might be lower. Reducing the number of parameters and the computing and storage requirements would allow models to be used in embedded systems.

Hopefully there will be improvements in algorithms, in data coding (now some computations can be performed in float8, a new data format supported by recent NVIDIA GPUs) and in accelerator performances. Algorithmic improvement is a key factor driving the advance of AI: *"Since 2012 the amount of compute needed to train a neural net to the same performance on ImageNet1 classification has been decreasing by a factor of 2 every 16 months [47]."*

Parameters	FLOPs	FLOPs (in *Gopher* unit)	Tokens
400 Million	1.92e+19	1/29,968	8.0 Billion
1 Billion	1.21e+20	1/4,761	20.2 Billion
10 Billion	1.23e+22	1/46	205.1 Billion
67 Billion	5.76e+23	1	1.5 Trillion
175 Billion	3.85e+24	6.7	3.7 Trillion
280 Billion	9.90e+24	17.2	5.9 Trillion
520 Billion	3.43e+25	59.5	11.0 Trillion
1 Trillion	1.27e+26	221.3	21.2 Trillion
10 Trillion	1.30e+28	22515.9	216.2 Trillion

Figure 8: Optimum number of tokens from training megamodels, and the corresponding compute power, according to [33]

Finding the right set of hyperparameters is also an approach that could result in better networks, and new techniques are being developed to find better parameters, like uTransfer [48]. This could reduce the size of a megamodel while maintaining its levels of performance. Quantization and pruning techniques are also important to reduce the computing cost; see the HiPEAC Vision 2023 article "AI for better integrations between the physical and cyber world: Embedded AI" for further details.

The combination of learning and explicit knowledge can also be used to improve the performance of AI systems, and to verify if

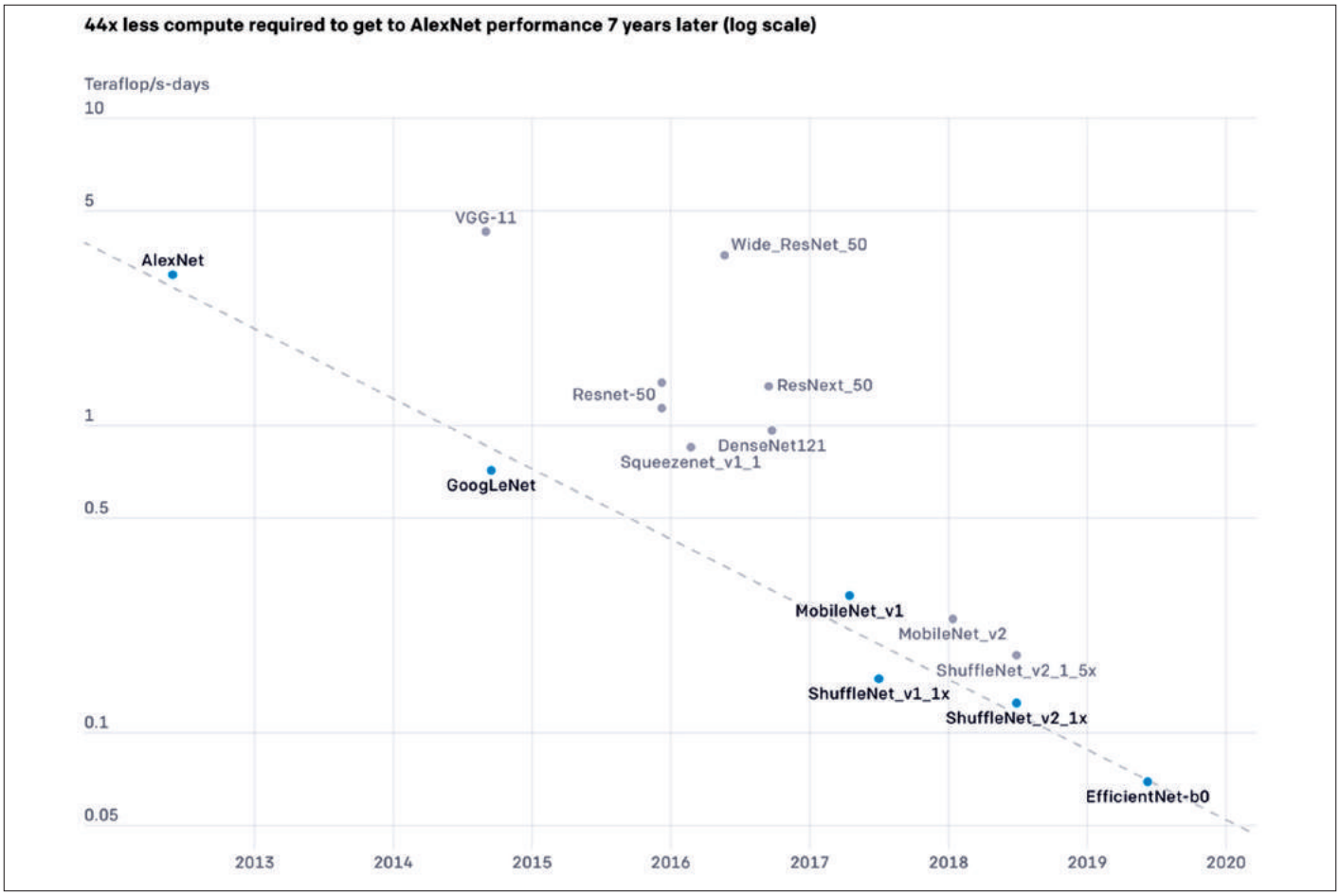


Figure 9: Decrease of the required computing power to train a Deep Neural Network for similar performances, from [47]

their results are within "acceptable" bounds (as these bounds are generally defined by rules, equivalent to "laws" for humans, well managed by systems other than deep neural networks). Multimodality [49] is also an interesting approach that could result in both more accurate and more understandable results.

New approaches could also be explored, just as transformers were an evolution of recurrent neural networks to process sequential input data. We have seen above the performance of large, monolithic networks, like in the centralized brain in mammals. But some other structures, like distributed ones, also show interesting emerging properties.

Is there proof in nature that other brain structures – different to those of humans or evolved mammals – can also show high levels of intelligence? Yes, for example in octopi. They are as intelligent as some mammals, for example they can use tools, yet they have a different brain structure: they have the equivalent to eight spinal cords, one running down each arm. The sequencing of the octopus genome shows that they possess brain-building genes called protocadherins, which were previously thought to exist only in vertebrates [50]. *"Protocadherin diversity provides a mechanism for regulating the short-range interactions needed for the assembly of local neural circuits, which is where the greatest complexity in the cephalopod nervous system appears. The importance of local neuropil interactions, rather than long-range connections, is probably due to the limits placed on axon density and connectivity by the absence of myelin, as thick axons are then required for rapid high-fidelity signal conduction over long distance" [51]. "While humans have about 60 protocadherins, the octopus genome was found to have 168, nearly three times the neural wiring capacity than humans"*. [50]

"The research team also hypothesized that cephalopod neurons don't function well over long distances, forcing octopi to evolve a "short-range brain" that works better in bursts and isn't operationally centralized [52]." [50]

Figure 10: Image generated by an AI [53] with the trigger words: friendly octopus brain , with computer on each arm. realistic

While clearly not proof, this suggests that a federation of smaller neural networks, faster and denser, can exhibit similar behaviour to a monolithic one (as far we can consider that the brain of mammals is monolithic…). This idea can also be observed in the "swarm" or "collective" intelligence in ant or bee colonies.

Therefore, research and developing distributed and decentralized federated learning techniques could be one approach to obtaining good performance levels, by harnessing the contributions of a collection of smaller networks working together. This could also have advantages in terms of privacy (data remains local, only high-level concepts are exchanged) and efficiency: if a good result is identified by the local network, less communication and less overall computing power is used. If a good result is not found, we can define systems where resources (i.e. a federation of local AIs) are triggered until a good result is found, resulting in optimal resource usage.

Conclusion

We have seen that megamodels have very interesting properties, but they are not yet well tuned and their application can be problematic. They require an enormous amount of data and processing power for their training, meaning that only a few companies and organizations are able to drive their development. Progress is in the hands of very few companies without reproducibility of their results by the scientific community. Most of them are not available for inspection and analysis (they are not available in open source, for example) and are only accessible through (paid) application programming interfaces (APIs).

Europe as a whole should take this into consideration, and be able to develop models tuned for European languages, culture and ethics. This is perhaps more an organizational problem than a scientific one: a "large instrument", or piece of infrastructure, could be created to develop those megamodels, build large, curated databases for their training and have efficient computing power to develop them in reasonable amount of time (perhaps similar to the model of the CERN for nuclear research). At a different level, perhaps appropriate regulation should be considered to ensure that only AI models with appropriate safeguards may be developed and deployed for the citizens of Europe.

If European researchers cannot experiment and test their ideas at scale, they will lose their place in the scientific race of AI, meaning that Europe might be dependent on megamodels built by non-European companies or countries. On the other hand, research should be done on innovative solutions to reduce the computing requirements of those megamodels, and finding new innovative approaches to reach similar results with different, but more efficient structures.

References

[1] "Attention, from Wikipedia, the free encyclopedia," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Attention. [Accessed 29 November 2022].

[2] "NVIDIA Omniverse," NVIDIA, [Online]. Available: https://www.nvidia.com/en-us/omniverse/. [Accessed 29 November 2022].

[3] "Azure Digital Twins," Microsoft, [Online]. Available: https://azure.microsoft.com/en-us/products/digital-twins/#overview. [Accessed 29 November 2022].

[4] E. L. Andrews, "Will the Future of the Internet Be Voice? Proposing a World Wide Voice Web," Stanford University Human-Centered Artificial Intelligence, 2 November 2021. [Online]. Available: https://hai.stanford.edu/news/will-future-internet-be-voice-proposing-world-wide-voice-web.

[5] P. E. Ross, "DeepMind's New AI Masters Games Without Even Being Taught the Rules," IEEE Spectrum, 23 December 2020. [Online]. Available: https://spectrum.ieee.org/tech-talk/robotics/artificial-intelligence/deepminds-new-ai-masters-games-without-even-been-taught-the-rules.

[6] Schrittwieser, J. et al., "Mastering Atari, Go, chess and shogi by planning with a learned model," Nature, no. 588, p. 604–609, 23 December December 2020, https://doi.org/10.1038/s41586-020-03051-4.

[7] A. Krizhevsky, I. Sutskever, G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in Advances in Neural Information Processing Systems 25 (NIPS 2012) , Reno, 2012.

[8] Vaswani, A. et al., "Attention Is All You Need," in NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA, Decmber 2017, https://doi.org/10.48550/arXiv.1706.03762.

[9] Brown, T. et al., "Language models are few-shot learners," in NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems, Vancouver, Canada, December 2020, https://doi.org/10.48550/arXiv.2005.14165.

[10] "OpenAI models - GTP-3, Codex, Dall-E," OpenAI, [Online]. Available: https://openai.com/api/. [Accessed 23 November 2022].

[11] C. Li, "OpenAI's GPT-3 Language Model: A Technical Overview," Lambda, 3 June 2020. [Online]. Available: https://lambdalabs.com/blog/demystifying-gpt-3/.

[12] "OpenAI API Models," OpenAI, [Online]. Available: https://beta.openai.com/docs/models/gpt-3. [Accessed 29 November 2022].

[13] "OpenAI Codex," OpenAI, 10 August 2021. [Online]. Available: https://openai.com/blog/openai-codex/. [Accessed 23 November 2022].

[14] "DALL·E 2," OpenAI, [Online]. Available: https://openai.com/dall-e-2/. [Accessed 23 November 2022].

[15] "Imagen Video," Google, [Online]. Available: https://imagen.research.google/video/. [Accessed 23 November 2022].

[16] "Democratizing access to large-scale language models with OPT-175B," Meta AI, 3 May 2022. [Online]. Available: https://ai.facebook.com/blog/democratizing-access-to-large-scale-language-models-with-opt-175b/. [Accessed 23 November 2022].

[17] J. Vincent, "Microsoft is giving businesses access to OpenAI's powerful AI language model GPT-3," The Verge, 2 November 2021. [Online]. Available: https://www.theverge.com/2021/11/2/22758963/microsoft-openai-gpt-3-azure-cloud-service-ai-language. [Accessed 23 November 2022].

[18] "Hugging Face transformers GitHub," Hugging Face, [Online]. Available: https://github.com/huggingface/transformers. [Accessed 23 November 2022].

[19] "Hugging Face models," Hugging Face, [Online]. Available: https://huggingface.co/models. [Accessed 23 November 2022].

[20] "Ground truth, from Wikipedia, the free encyclopedia," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Ground_truth. [Accessed 29 November 2022].

[21] "Copilot – Your AI pair programmer," GitHub, [Online]. Available: https://github.com/features/copilot/. [Accessed 23 November 2022].

[22] "NVIDIA NeMo Megatron," NVIDIA, [Online]. Available: https://developer.nvidia.com/nemo/megatron. [Accessed 23 November 2022].

[23] M. Asay, "Will the cloud eat your AI?," Tech Republic, 6 October 2021. [Online]. Available: https://www.techrepublic.com/article/will-the-cloud-eat-your-ai/. [Accessed 23 November 2022].

[24] Sevilla, J. et al., "Compute Trends Across Three Eras of Machine Learning," in 2022 International Joint Conference on Neural Networks (IJCNN), Padua, Italy, 2022, https://doi.org/10.48550/arXiv.2202.05924.

[25] V. Kosar, "Google's Pathways Language Model and Chain-of-Thought Prompting," Vaclav Kosar's Software & Machine Learning Blog , 11 June 2022. [Online]. Available: https://vaclavkosar.com/ml/googles-pathways-language-model-and-chain-of-thought. [Accessed 23 November 2022].

[26] S. Narang and A. Chowdhery, "Pathways Language Model (PaLM): Scaling to 540 Billion Parameters for Breakthrough Performance," Google, 4 April 2022. [Online]. Available: https://ai.googleblog.com/2022/04/pathways-language-model-palm-scaling-to.html. [Accessed 23 November 2022].

[27] L. Heim, "Estimating PaLM's training cost," 5 April 2022. [Online]. Available: https://blog.heim.xyz/palm-training-cost/. [Accessed 23 November 2022].

[28] "Megatron-Turing Natural Language Generation," NVIDIA, [Online]. Available: https://developer.nvidia.com/megatron-turing-natural-language-generation. [Accessed 23 November 2022].

[29] T. R. Weiss, "Google Launches TPU v4 AI Chips," HPC Wire, 20 May 2021. [Online]. Available: https://www.hpcwire.com/2021/05/20/google-launches-tpu-v4-ai-chips/. [Accessed 23 November 2022].

[30] Kaplan, J. et al., "Scaling Laws for Neural Language Models," OpenAI, 23 January 2020. [Online]. Available: https://arxiv.org/abs/2001.08361. [Accessed 23 November 2022].

[31] P. Kharya and A. Alvi, "Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, the World's Largest and Most Powerful Generative Language Model," NVIDIA, 11 October 2021. [Online]. Available: https://developer.nvidia.com/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest-and-most-powerful-generative-language-model/. [Accessed 23 November 2022].

[32] A. Romero, "A New AI Trend: Chinchilla (70B) Greatly Outperforms GPT-3 (175B) and Gopher (280B)," Towards Data Science, 11 April 2022. [Online]. Available: https://towardsdatascience.com/a-new-ai-trend-chinchilla-70b-greatly-outperforms-gpt-3-175b-and-gopher-280b-408b9b4510. [Accessed 23 November 2022].

[33] Hoffmann, J. et al. , "Training Compute-Optimal Large Language Models," DeepMind, 29 March 2022. [Online]. Available: https://arxiv.org/abs/2203.15556.

[34] B. Dickson, "Don't expect large language models like the next GPT to be democratized," The Next Web, 21 May 2022. [Online]. Available: https://thenextweb.com/news/dont-expect-large-language-models-like-the-next-gpt-to-be-democratized. [Accessed 23 November 2022].

[35] "Stable Diffusion Public Release," Stability AI, [Online]. Available: https://stability.ai/blog/stable-diffusion-public-release. [Accessed 23 November 2022].

[36] "Hugging Face," [Online]. Available: https://huggingface.co/. [Accessed 24 November 2022].

[37] S. AI, "Stable Diffusion on GitHub," GitHub, [Online]. Available: https://github.com/CompVis/stable-diffusion. [Accessed 24 November 2022].

[38] Meta AI, "Galactica GitHub," GitHub, [Online]. Available: https://github.com/paperswithcode/galai. [Accessed 24 November 2022].

[39] B. Edwards, "With Stable Diffusion, you may never believe what you see online again," Ars Techica, 6 September 2022. [Online]. Available: https://arstechnica.com/information-technology/2022/09/with-stable-diffusion-you-may-never-believe-what-you-see-online-again/. [Accessed 24 November 2022].

[40] B. Edwards, "Fake video threatens to rewrite history. Here's how to protect it," Fast Company, 10 March 2020. [Online]. Available: https://www.fastcompany.com/90549441/how-to-prevent-deepfakes. [Accessed 24 November 2022].

[41] R. Price, "Microsoft Took Its New A.I. Chatbot Offline After It Started Spewing Racist Tweets," Slate, 24 March 2016. [Online]. Available: https://slate.com/business/2016/03/microsoft-s-new-ai-chatbot-tay-removed-from-twitter-due-to-racist-tweets.html. [Accessed 24 November 202].

[42] "Galactica mission," Meta AI, [Online]. Available: https://galactica.org/mission/. [Accessed 23 November 2022].

[43] B. Wodecki, "Update: Meta's Galactica AI Criticized as 'Dangerous' for Science," AI Business, 17 November 2022. [Online]. Available: https://aibusiness.com/nlp/meta-s-galactica-ai-criticized-as-dangerous-for-science. [Accessed 23 November 2022].

[44] W. D. Heaven, "Why Meta's latest large language model survived only three days online," MIT Technology Review, 18 November 2022. [Online]. Available: https://www.technologyreview.com/2022/11/18/1063487/meta-large-language-model-ai-only-survived-three-days-gpt-3-science/. [Accessed 24 November 2022].

[45] E. M. Bender, "@emilymbender," Twitter, 16 November 2022. [Online]. Available: https://twitter.com/emilymbender/status/1592992842976489472. [Accessed 23 November 2022].

[46] J. Ryan, "Meta Trained an AI on 48M Science Papers. It Was Shut Down After 2 Days," CNET, 20 November 2022. [Online]. Available: https://www.cnet.com/science/meta-trained-an-ai-on-48-million-science-papers-it-was-shut-down-after-two-days/. [Accessed 24 November 2022].

[47] "AI and Efficiency," OpenAI, 5 May 2020. [Online]. Available: https://openai.com/blog/ai-and-efficiency/. [Accessed 24 November 2022].

[48] Hu, E. et al., "µTransfer: A technique for hyperparameter tuning of enormous neural networks," Microsoft Research, 8 March 2022. [Online]. Available: https://www.microsoft.com/en-us/research/blog/%C2%B5transfer-a-technique-for-hyperparameter-tuning-of-enormous-neural-networks/. [Accessed 24 November 2022].

[49] Goh, G. et al. , "Multimodal Neurons in Artificial Neural Networks," OpenAI, 4 March 2021. [Online]. Available: https://openai.com/blog/multimodal-neurons/. [Accessed 24 November 2022].

[50] S. t. Rocket, "Science Says: Octopus Brains are Something More Than Human…," Tor, 4 December 2015. [Online]. Available: https://www.tor.com/2015/12/04/octopi-giant-brains-alien-genome/. [Accessed 24 November 2022].

[51] C. B. A. e. al., "The octopus genome and the evolution of cephalopod neural and morphological novelties," Nature, no. 524, pp. 220–224, https://www.nature.com/articles/nature14668 , 2015.

[52] K. Jiang, "Landmark sequencing of octopus genome shows basis for intelligence, camouflage," University of Chicago, 13 August 2015. [Online]. Available: https://news.uchicago.edu/story/landmark-sequencing-octopus-genome-shows-basis-intelligence-camouflage. [Accessed 24 November 2022].

[53] "Playground AI Create", Playground AI [Online]. Available: https://playgroundai.com/create. [Accessed 24 November 2022].

Marc Duranton is a researcher in the research and technology department at CEA (the French Atomic Energy Commission) and the coordinator of the HiPEAC Vision 2023.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2023.

Cite as: M. Duranton. Gigantic transformers and megamodels: The next El Dorado for AI? Or Is "the bigger, the better" the path to AGI? In M. Duranton et al., editors, HiPEAC Vision 2023, pages 70-78, Jan 2023.

DOI: 10.5281/zenodo.7461855

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

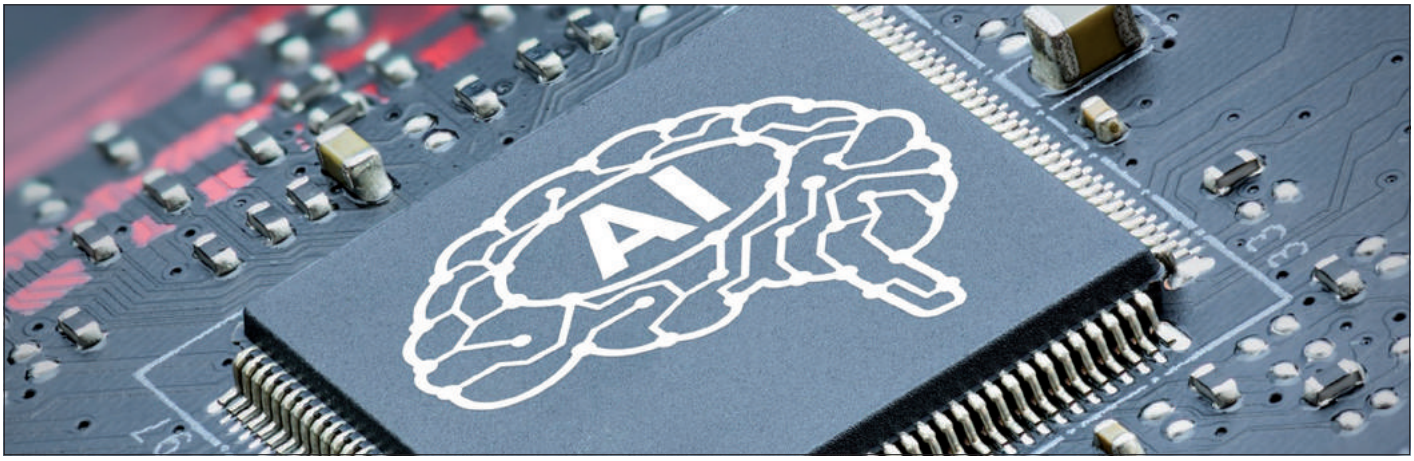
© HiPEAC 2023

Today's cutting-edge artificial intelligence (AI) techniques are making their way into the edge devices of tomorrow. Europe has the potential to be a leader in low-power, high-performance edge systems that combine the best of all AI approaches and serve the sectors that are most important to Europe.

AI for better integration between the physical and the cyber world: Embedded AI

By MARC DURANTON

Artificial intelligence allows computer systems to better interact with the physical world, either by interpreting (sequences of) images, sound, signals etc, or by generating more realistic audio, speech or video signals. This is the main ingredient in the new generation of cyber-physical systems (CPS) that, in turn, will be key for the "next web", ambient intelligence, the physical web, etc. For reasons of privacy, efficiency, etc., more and more of this processing will be run on edge devices, thanks to algorithm improvements and better hardware performance. Often, the first layers are made by deep neural network (DNN) systems, while the higher levels are often performed by more "traditional" algorithmic approaches.



Key insights

- The interfaces between the physical world and computing systems are key for the next generation of systems, such as the "next web", ambient intelligence, autonomous systems, etc.
- For reasons of privacy, latency and energy efficiency, it is often desirable for computations to be performed locally, "at the edge".
- Both DNN and more algorithmic approaches (where knowledge is explicitly introduced into the system) should be used.
- Voice interfaces are becoming more and more important for consumer systems (both recognition and synthesis) as natural means of interaction.
- Impressive new levels of performance will be achieved thanks to the fusion of different modalities and continuous signals, e.g. a sequence of images instead of a single image.
- Europe has the potential to be leader in this kind of system, but Chinese and American companies are already in the race, and are developing integrated circuits (ICs) for edge applications.

Key recommendations

- Europe should aim to retain its leading position in the field of low-power, local processing, using both DNN and algorithmic approaches for the edge use cases prevalent in Europe.
- Development environments and tools should be developed in order to translate/map complex DDN into embedded chips (using pruning, quantization, etc…)
- The diversity of European languages is an opportunity to develop more efficient integrated solutions suitable to the European market. These should be in accordance with European requirements and ethical standards.
- Europe should be at the forefront of making accelerators using state-of-the-art AI algorithms.
- Europe has good levels of knowledge and education relating to technologies required to make edge AI devices, but concretization into products is still lacking.

Artificial intelligence allows computers to interact better with the world and humans

Deep learning provided breakthroughs as a way of analysing unstructured data such as images and sound, as well as allowing an efficient interface between computers and the world, facilitating cyber-physical applications. This has really opened up possibilities for new solutions and business propositions, like self-driving cars, personal assistants and so on. Image and sequence analysis allow autonomous vehicles to check their surrounding and to take appropriate decisions. Most of the time, several modalities (images, Lidar data, ultrasonic data) can be used and merged to get a better analysis. The fusion of different modalities is an important topic for AI at the edge.

Personal assistants

In addition to image classification, progress in AI, particularly deep learning, has been made very visible by advances in voice recognition, which paved the way for the emergence of voice-activated personal assistants like Siri from Apple, Google Assistant, Alexa from Amazon, Cortana from Microsoft, Bixbi from Samsung, Duer from Baidu, Viv, etc. Indeed, for most people today the visible part of AI is embodied by these personal assistants. Voice is a natural way of communicating for humans, and, if the AI works well, it is also a way for people who are not very familiar with computers or keyboards to access digital technologies.

Improvements in the accuracy of recognition have also triggered the development of specific accelerator hardware (in the case of Google), as explained in this blog post from 2017 [2]:

"The need for TPUs really emerged about six years ago, when we started using computationally expensive deep learning models in more and more places throughout our products. The computational expense of using these models had us worried. If we considered a scenario where people use Google voice search for just three minutes a day and we ran deep neural nets for our speech recognition system on the processing units we were using, we would have had to double the number of Google data centers."

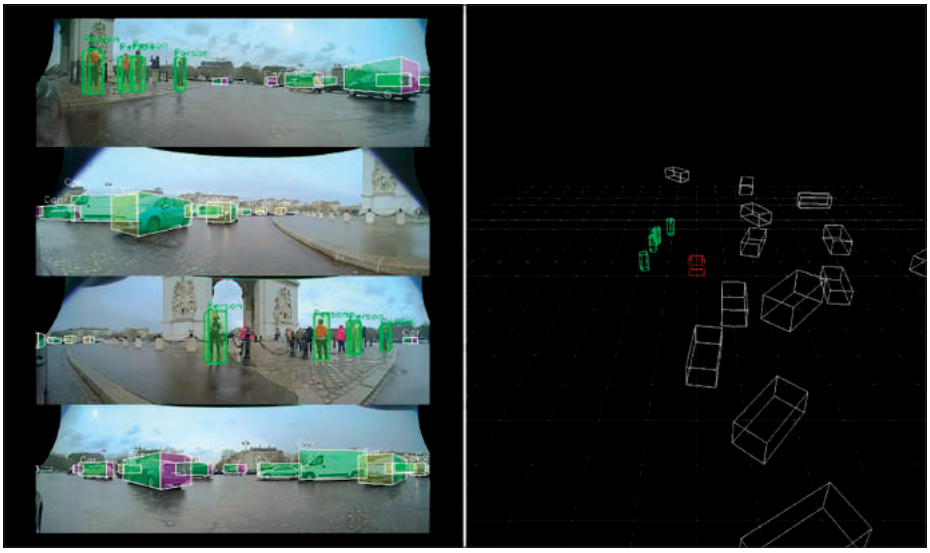


Figure 1: Deep MANTA, a neural network algorithm to perform advanced and efficient real-time analysis of video streams for autonomous vehicle perception systems (from CEA) [1]

Personal assistants first appeared in mobile phones (Google Assistant for Android, and Siri for Apple). They proved useful for some activities, like dictating and reading text messages and emails in cars, but the touchscreen interface was still more convenient for most applications. It was only when they were incorporated into a speaker for use in the home that they really found their niche.

The market for virtual assistants

First introduced by Amazon in its Echo device in November 2014, the company's Alexa assistant is a market leader in terms of smart speaker sales. (Google Assistant

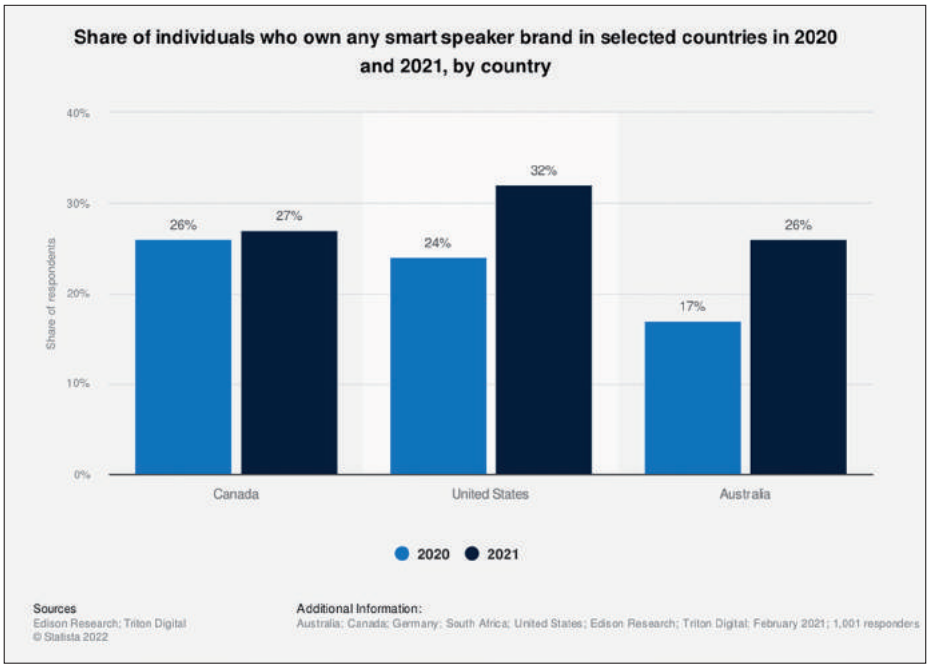


Figure 2: Proportion of people having answered to the pool who own a smart speaker (from [3])

> "Personable or not, intelligent voice *agents are poised to alter the nature of computing as we know it.*
> *Since the dawn of the PC, humans have been forced to learn an arcane, unnatural language. But in a voice-first world, mediated by artificial intelligence and machine learning ... the next 50 years will see computers learning to be more like us."* [4]

and Siri from Apple have also a very large number of instances because they are embedded in smartphones, but for the purposes of the current discussion we will focus on the specific embodiment of assistants into dedicated devices such as smart speakers).

According to Statista, Alexa can control more than 100,000 different smart home devices, and in 2019 Amazon announced that more than 100 million devices with Alexa on board had been sold. According to Voicebot.ai, there are now over 80,000 Alexa "skills" – what Amazon calls its voice apps – in the US alone and hundreds of thousands of developers in over 180 countries working on Alexa. Skills are small programs or apps developed by independent developers that run on the cloud and bring new capabilities to the Alexa personal assistant. However, there has apparently been a decline in the global growth rate of Alexa skills, which Voicebot has linked to waning developer enthusiasm for the platform [6] and a difficult economic model because most of them are free.

Google followed Amazon in 2016 with its Google Home speaker, which is rapidly growing in the market. Apple was next, introducing its HomePod embedding Siri in 2018.

Ecosystems are being created around them, with more and more devices becoming interoperable and therefore able to be controlled by voice. The introduction of Matter (*"a standard for home automation, royalty-free, ... [that] aims at achieving interoperability among smart home devices and internet of things (IoT) platforms from different providers*" [8]) will enable interoperability between the different assistants' ecosystems.

A digest of statistics concerning smart speakers can be found at [9]. In 2020, Voicebot published information about a survey that "found that 87.7 million U.S. adults were using smart speakers as of January 2020. This means the installed base of U.S. smart speaker users is up 32% over January 2019 and is 85% higher than January 2018". The article recognizes, however, that these figures "show a slowing relative growth rate" [10].

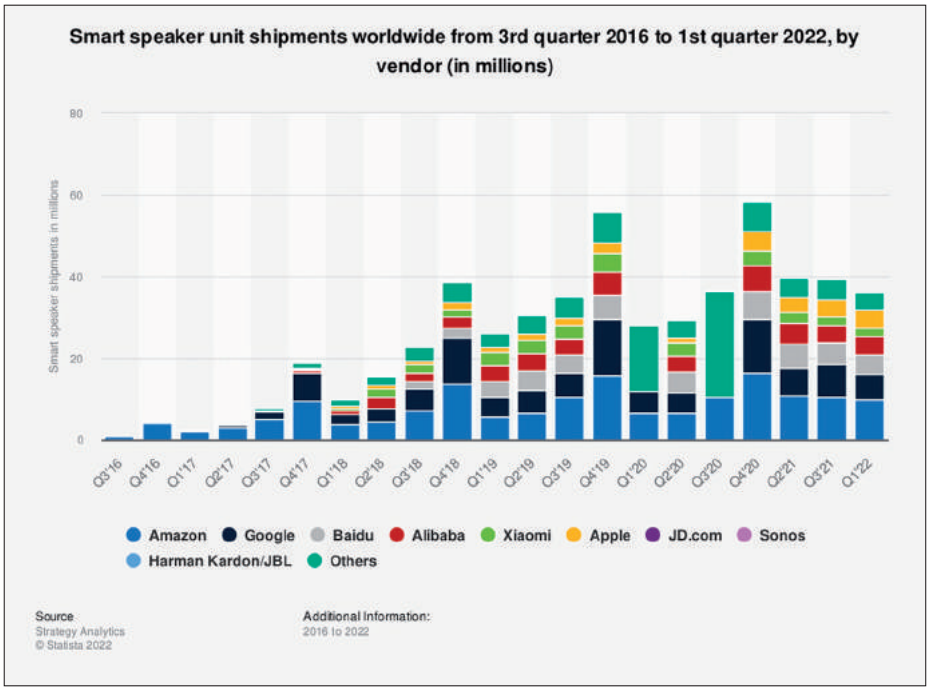


Figure 3: Smart speaker market share (from [5])

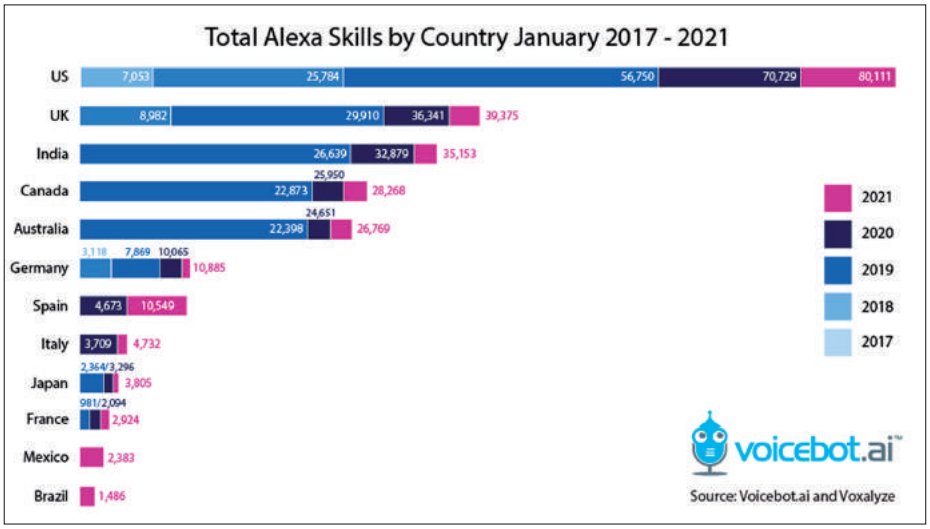


Figure 4: Total Alexa skills by country (from [6])

Figure 5: From left to right: Google Home, Apple HomePod mini and Amazon Echo speakers. From [7]

Note that, in addition to smart speakers, voice-controlled assistants are used in a range of other smart-home devices, as noted in [11].

From these figures, even with the slowdown due to the COVID-19 period, it seems that the market for voice assistants is established. However, due to the large support teams and computing resources required, virtual assistants are struggling to generate a return on investment. Most of the applications developed by their supporting communities are free, and people are not really willing to pay for this commodity. As an example, the "Worldwide Digital" group in Amazon *"lost $3 billion in just the first quarter of 2022, with "the vast majority" of the losses blamed on Alexa"* [14], *"every plan to monetize Alexa has failed"*, and buying by voice command, foreseen by Amazon as a key asset for increasing its business, was never really a success. This suggests that, despite its potential, the way to create a thriving business model for virtual assistants is still not clear.

The appeal of virtual assistants

As noted in previous editions of the HiPEAC Vision, home assistants have an attractive service offering. They show their full usefulness when connected to the IoT or smart devices such as lights and appliances, controlling them remotely. As well as providing basic functions of web interaction – such as delivering weather forecasts, traffic reports or the time, making to-do lists, setting alarms, and fetching information from Wikipedia – they provide an ideal interface for streaming music. They have also been replacing radio with their enhanced functionalities to respond to requests for a particular song, piece of music or podcast. Another use is controlling smart devices in the home (light, power plugs, etc).

Their appeal partly lies in their role as a constantly available "butler" who obeys orders, helps automate routine tasks and helps users organize their lives. With access to the internet and company data, voice assistants could also participate in meetings as required, contributing relevant information during the discussion.

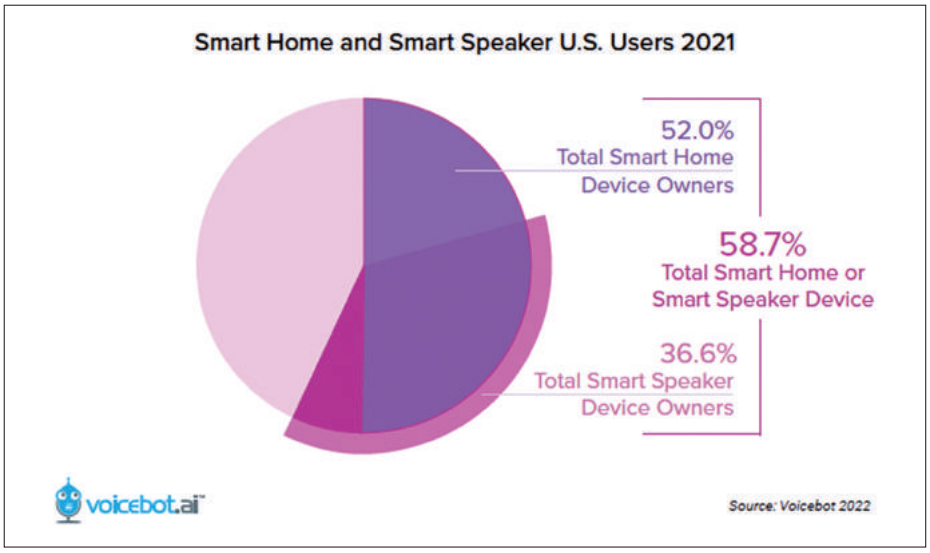


Figure 6: Smart home and smart speaker users in the US in 2021, from [9]

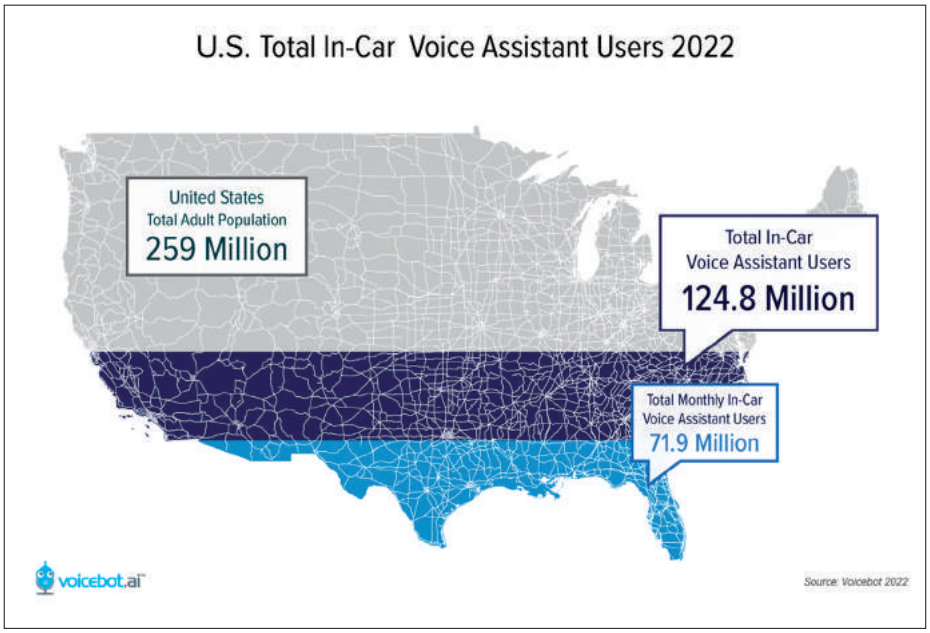


Figure 7: US total in-car voice assistant users in 2022, from [9]

As an example, in 2018 Google announced its Duplex technology which could be used for real-world tasks. The technology was demonstrated reserving a table at a restaurant and booking a haircut without human intervention; it has since been integrated into Google Assistant [15]. However, the natural-sounding conversation gave rise to concerns that the voice assistant should identify itself as AI, and not let the human with whom it is interacting believing it is also a human.

Duplex has also been promoted as a way to help mitigate the frustration of calling companies' automated response services, waiting on hold on callers' behalf, as well as screening calls [16]. This kind of application has prompted some commentators to suggest that, rather than being tailored to an individual, customer-facing AI could be used by companies to provide an "improved" service [17].

One reason for the growing acceptance of these assistants is that voice is a natural way for humans to interact. They provide a way for people unfamiliar with computers to access resources without having to know how to use a computer or a particular user interface. Current voice assistants will probably evolve towards being more

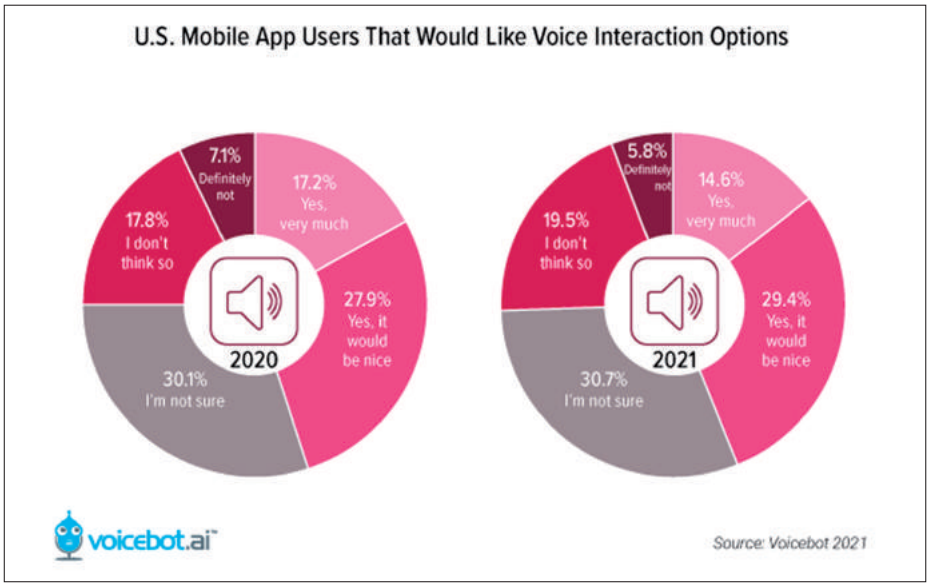


Figure 8: US mobile app users who would like voice interaction options, from [12]

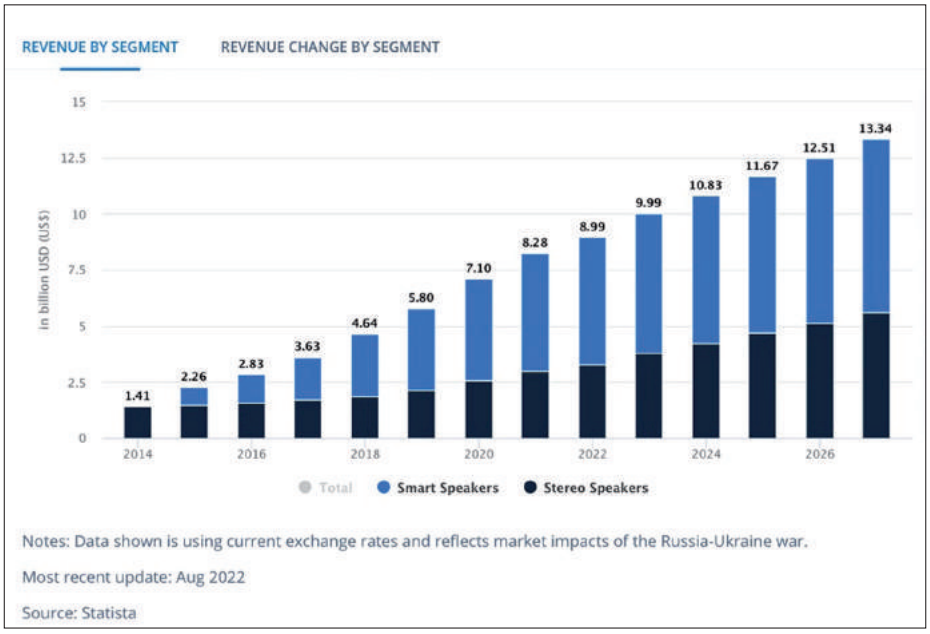


Figure 9: Smart speakers – worldwide revenue. Source [13]

personal and customizable, offering multiple modalities in addition to voice recognition, such as gesture recognition and mood analysis, and evolving to understand and operate in a larger context.

Nonetheless, commercial voice assistants have also attracted criticism for being "spies", recording everything that is said in a house. Concerns over privacy have apparently risen; according to [18], in 2020 "33% of U.S. adults cite[d] concerns about smart speakers recording what they are saying as a top reason for not purchasing the devices". Technically, this is not accurate, and the behaviour is not the same for all assistants. While devices are always "listening" for the "wake" (trigger) word, the processing of this word is normally done locally, on the device. The streaming of sound recorded to external servers only takes place when the "wake" word is detected.

Of course, false triggering can happen. Apparently, in the case of Google Assistant, samples around the supposed triggering word are sent to the server for verification. Some samples of interactions are also analysed by humans to improve the service, which has led to privacy and ethical concerns [19]. More generally, given that the devices have the capacity to listen and record, they have the potential for unprecedented surveillance in the absence of robust protections for users. Some of these issues are explored further in the article on privacy in this HiPEAC Vision.

However, the progress in AI algorithms and the development of new accelerator chips will decrease the need for assistants to use off-site servers, and more and more of the voice processing will be done locally. Using open-source technology that can be inspected, running locally on edge devices, could alleviate users' fears.

A European flavour of virtual assistant

It is unfortunate that Europe does not have a strong presence in this domain, at least on the business-to-consumer (B2C) side. Even the European start-ups which do exist in this field, such as Snips [20], are generally bought by non-European companies (Snips was acquired by Sonos, and allows Sonos to have their own assistant running locally in their smart speakers). It seems that voice is a very good human-to-machine interface, one which could be widely accepted if users could trust the systems (in terms of privacy, cybersecurity, and the absence of manipulation by corporate interests), so it should be of common interest to have a European flavour of this technology.

Collaborative work around open source could be a way to integrate the different technologies required to make a trustable vocal assistant. Professor Monica Lam's group at Stanford University has demonstrated that it is feasible to develop innovative concepts for designing a virtual assistant [21]. With the collaboration of several open-source initiatives in this field (including European ones such as LinTO [22]) and databases of various languages (such as Common Voice [23]), a fully functional, efficient, ethical, inspectable, trustworthy virtual assistant technology platform could be developed. The plurality of languages, accents and dialects in Europe could be a driver to develop European databases, speech-to-text and text-to-speech technologies tuned for European needs.

It would also be important to investigate different business models to make sure this technology is sustainable, or to ensure that it has a societal impact [24] allowing people uneasy with digital technology to easily access it, so that it should be developed with European funding to be available to all Europeans.

Dedicated chips allow local processing

Voice technology is advanced enough to have part or even most of the voice processing done locally: voice-to-text software is efficient on computers (for example, there is a native dictation application on Mac computers) and companies like Snips promote local processing to help keep privacy intact. This might require more processing power, but voice assistants already embedded in loudspeakers have the processing power of a medium-range smartphone. The development of AI accelerators and the new improvements and compression of voice recognition models allow more and more local processing in the devices themselves, minimizing communication with the cloud.

"The AZ2 chip builds on the machine learning interface that premiered with the AZ1, which allowed Amazon devices to better recognize your voice, but has extended this capability to facial recognition as well. This comes with Amazon's new focus on what it calls "Ambient Intelligence." [25]

Amazon, Google and Apple are increasingly tending towards local processing for their AI, and they have all integrated hardware accelerators for deep learning into their devices. For example, Google Translate can now run locally on a smartphone.

Chinese and Taiwanese companies also sell chips designed for local voice processing, mainly for voice commands. The Amazon AZ line of chips were developed in collaboration with Mediatek. Chips from iFlytek [26] are also used for smart assistants in China [27].

Examples of other companies in this field include Unisound with its US516P6 chip and Espressif. Espressif Systems is a public multinational, fabless semiconductor company established in 2008, with offices in China, the Czech Republic, India, Singapore and Brazil [28]. According to their Q1 2022 report, the company generated revenue of about €202 million in the first quarter of 2022, while the company employs around 520 people (as of June 2022).

Figure 10: Examples of chips able to carry out (simple) voice recognition locally

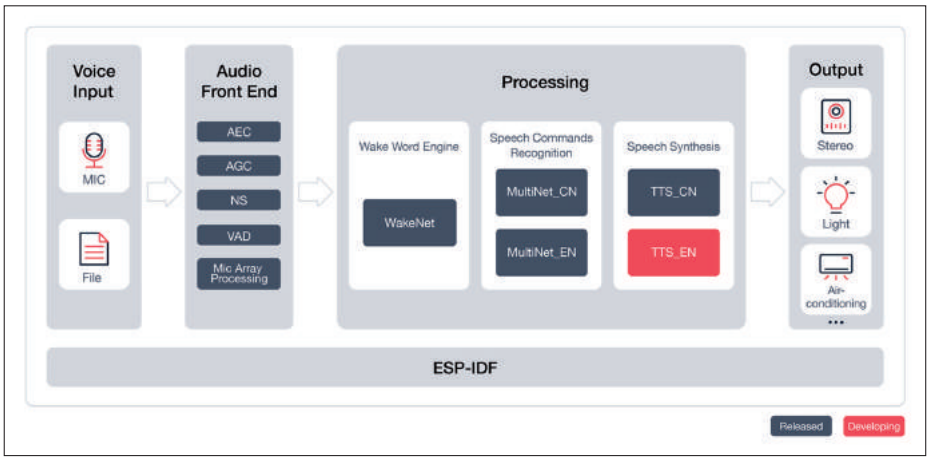


Figure 11: Espressif's voice system, according to [30]

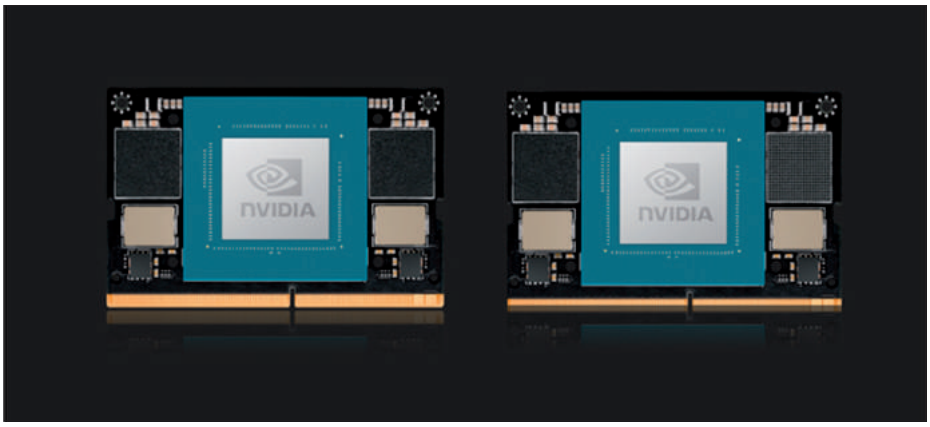


Figure 12: the new Jetson Orin nano from NVIDIA, from [34]

The Espressif model is interesting because it heavily relies on open source for software development (and the company hires the most active GitHub contributors). Espressif is also moving to the RISC-V open-source instruction set architecture, one of the building blocks of open-source hardware, and supports open-source developments (e.g. CircuitPython). One particu-

	Jetson Orin Nano 4GB	Jetson Orin Nano 8GB
AI Performance	20 Sparse TOPs | 10 Dense TOPs	40 Sparse TOPs | 20 Dense TOPs
GPU	512-core NVIDIA Ampere Architecture GPU with 16 Tensor Cores	1024-core NVIDIA Ampere Architecture GPU with 32 Tensor Cores
GPU Max Frequency	625 MHz	
CPU	6-core Arm Cortex-A78AE v8.2 64-bit CPU 1.5 MB L2 + 4 MB L3	
CPU Max Frequency	1.5 GHz	
Memory	4GB 64-bit LPDDR5 34 GB/s	8GB 128-bit LPDDR5 68 GB/s
Storage	– (Supports external NVMe)	
Video Encode	1080p30 supported by 1-2 CPU cores	
Video Decode	1x 4K60 (H.265) | 2x 4K30 (H.265) | 5x 1080p60 (H.265) | 11x 1080p30 (H.265)	
Camera	Up to 4 cameras (8 through virtual channels*) 8 lanes MIPI CSI-2 D-PHY 2.1 (up to 20 Gbps)	
PCIe	1 x4 + 3 x1 (PCIe Gen3, Root Port, & Endpoint)	
USB	3x USB 3.2 Gen2 (10 Gbps) 3x USB 2.0	
Networking	1x GbE	
Display	1x 4K30 multimode DisplayPort 1.2 (+MST)/e DisplayPort 1.4/HDMI 1.4*	
Other I/O	3x UART, 2x SPI, 2x I2S, 4x I2C, 1x CAN, DMIC and DSPK, PWM, GPIOs	
Power	5W – 10W	7W – 15W
Mechanical	69.6 mm x 45 mm 260-pin SO-DIMM connector	
Price	$199†	$299†

Table 1. Jetson Orin Nano series specification

Figure 13: The specifications of Jetson Orin Nano, from [34]

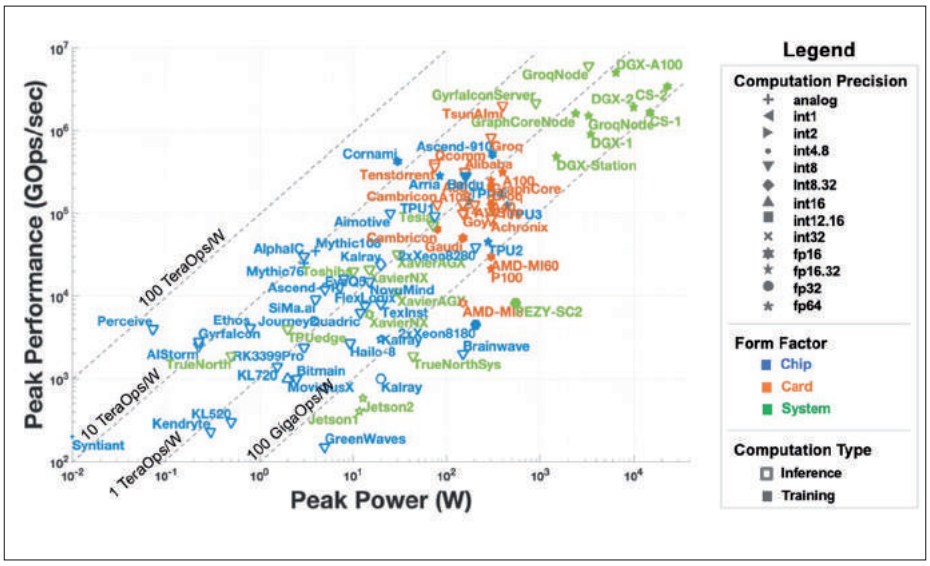


Figure 14: Peak performance vs. power scatterplot of publicly announced AI accelerators and processors, from [37]

larity of the company is that they provide a longevity commitment of their products up to 12 years. They provide a local voice command system, offering the repository on Github [29, 30].

Their multinet [31] system can understand up to 200 speech commands (Chinese and English). More details on the wake word approach can be seen in [32]. Their model shows that relying on open source can be valuable for a company. Their chips are qualified platforms for ACS, "Amazon Common Software (ACS) for Devices", which, as Espressif notes, is "Amazon's optimized software that simplifies the integration of various Amazon Device SDKs in connected products" [33].

NVIDIA is also present in the edge market: in September 2022 the company introduced its Jetson Orin Nano, which, according to [34], reaches processing power of up to 40 sparse TOPS (float8?) on less than 15 W of power and at a cost of US$ 299 for the complete system. Medium size voice recognition models should be able to run in real-time on such hardware.

European companies are also present in this field of ASICs combining a processor and AI accelerators. Companies like GreenWaves [35], ST Microelectronics, Infineon and NXP [36] all have products on the market. In fact, the market seems to be very crowded, according to a survey [37] from September 2021.

Local processing is the key to success for most "ambient intelligence" systems

Ambient intelligence is coming back into prominence according to Amazon [38]; indeed, the company chose it as the topic of a recent keynote talk about the Alexa ecosystem [39]. The term ambient intelligence is quite old, having been used e.g. by Philips 20 years ago [40]. Artificial intelligence is becoming more and more present in edge devices, not only for voice assistants and image recognition but also for tasks that were only available in the cloud until recently. For the forthcoming "metaverse", 3D face reconstruction can be made locally, see for example [41] which can achieve 150 frames per second

on a single central processing unit (CPU) thread. This kind of work will allow direct animation of avatars in real time on edge devices (currently it can operate on PCs) without requiring expansive graphics processing units (GPUs) or access to the cloud. However, it could also make it easier to create deep fakes in real time.

Even complex models, like those that can generate images from text description, can be now run on consumer laptops. For example, DiffusionBee [42] can run models of Stable Diffusion [43] locally to generate images.

According to Wikipedia, *"the model of Stable Diffusion was trained using 256 Nvidia A100 GPUs on Amazon Web Services for a total of 150,000 GPU-hours, at a cost of $600,000"* [44]. It uses a *"latent diffusion model, a variety of deep generative neural network developed by the CompVis group at LMU Munich"*. Since it was released in the public domain, many developers have worked on it, making different ports and interfaces, and creating websites [45] to display the results of the algorithms. In inference (to generate new images), it runs in a few seconds on a MacBook with the M1 processor, showing that this kind of big model can be executed at the edge (not yet in embedded systems), leading to new potential usages.

AI at the edge: combining DNN and algorithm-driven approaches

As seen with voice systems, the overall approach to AI often involves a combination of DNN for phoneme detection, and more "traditional" approaches (or statistical) to determine the probability of the words. In autonomous vehicles, the detection and analysis of the environment is often performed by deep learning techniques, while data fusion can use other techniques (Bayesian, etc) and decisions are taken by rule-based or algorithm-driven approaches (although it should be noted that deep learning is also making progress in this domain).

Together, deep learning techniques and more traditional approaches (based on explicit knowledge) are a powerful combination that can lead to interesting results.

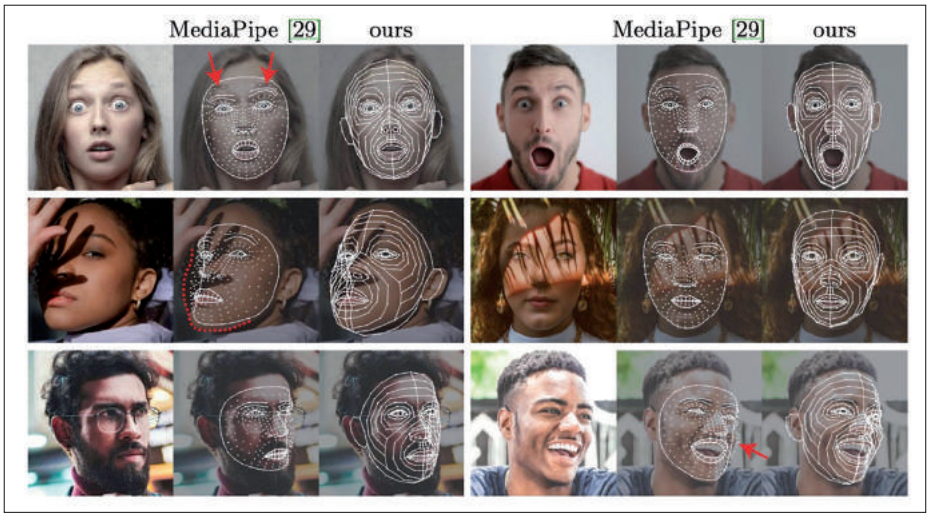


Figure 15: Results [41] allowing real-time animation of avatars at the edge

Figure 16: Image generated on a laptop with DiffusionBee (parameters: Seed: 76506 | Scale: 7.5 | Steps: 25 | Img Width: 512 | Img Height: 512 | model_version: 1.5fp16, trigger sentence: "fantasy landscape photography by Saudkova").

Figure 17: Image generated on a laptop with DiffusionBee (parameters: Seed: 58862 | Scale: 7.5 | Steps: 25 | Img Width: 512 | Img Height: 512 | model_version: 1.5fp16, trigger sentence: "cute otter, disney pixar, detailled fur, lot of details")

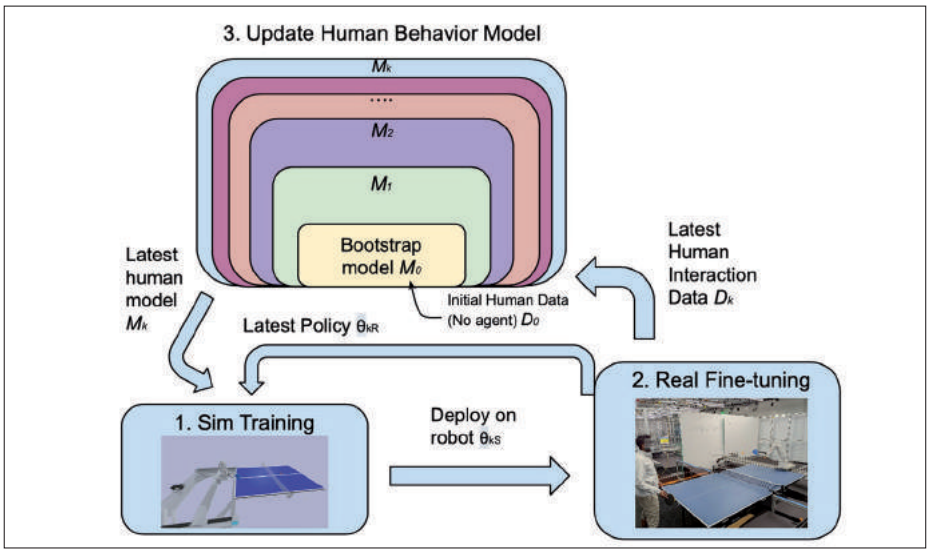


Figure 18: the reinforcement technique with simulation in the loop allow to learn and adapt with minimum numbers of real data (from [46], also explained on [47])

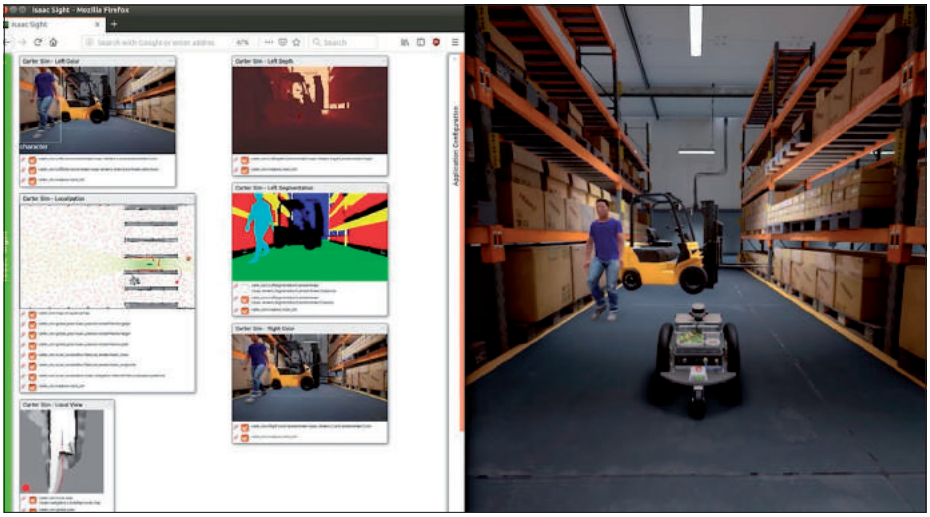


Figure 19: Nvidia Omniverse [49]

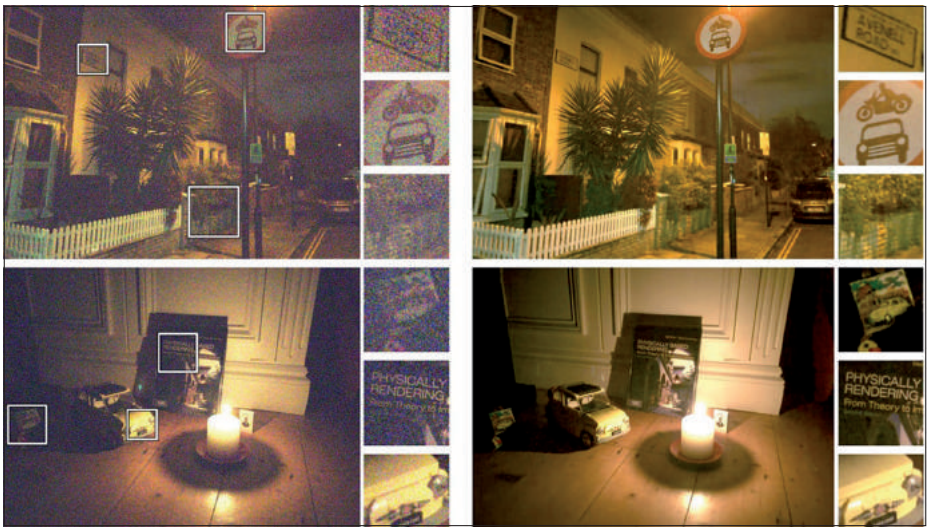


Figure 20: Low luminosity image improvement from [51]

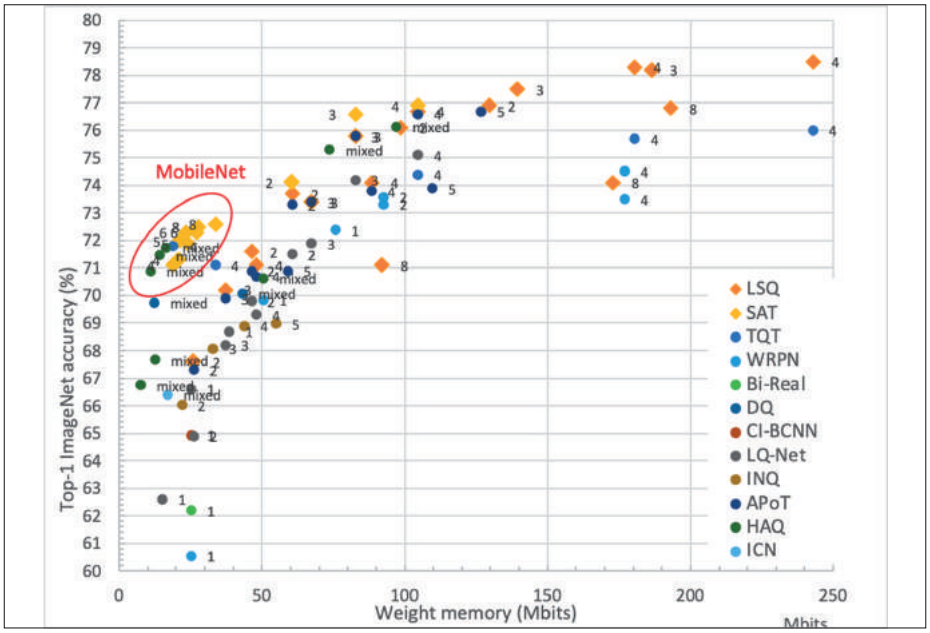


Figure 21: Various trade-off due to quantization, according to CEA [53]

For example, having simulation in the loop, as shown by the Google research paper [46], allows a robot to learn to play table tennis against a human with minimal training; it can also generalize to play with other humans.

For the development of applications using deep learning, the user still has to provide a large database of data to train the networks (if using supervised learning approach). This can be solved by creating datasets through simulation in a life-like environment. For example, NVIDIA's Omniverse environment allows the creation of a photorealistic environment to train and simulate AI models [48].

These virtual environments allow the validation of AI models that use smaller datasets or only "reward" functions, as in the case of reinforcement learning or self-supervised learning. Later, these AI models can be used at the edge (inference phase), e.g. in robots or autonomous vehicles.

Data fusion and sequences result in better performance

Deep learning has provided breakthroughs in analysing unstructured data such as images. The analysis of the environment can be improved by adding more information, e.g. a sequence of images or other modalities. For example, the paper [50] [51] shows how dark images can be improved thanks to a temporal sequence of images. Other systems, fusing different modalities (visual and Lidar, for example [52]), allow better accuracy without needing more accurate input data. This kind of fusion is used in autonomous vehicles for example.

Tools are required to squeeze high-performing approaches into embedded systems

The key to implementing DNN into embedded devices, with limited resources, is being able to "simplify" the network, without degrading the performance (too much). This can be achieved either by pruning the topology of the networks, and/or by reducing the number of bits where weight and neuron values are stored. This technique, known as "quantization" is progressing quite rapidly, mainly driven

by the TinyML community. There are now many tools and libraries for reducing the size of DNN (including CEA's N2D2 [53], TensorFlow Lite [54], microTVM [55], NNoM [56] and TinyMaix [57], among others).

Developing approaches to reduce both the memory footprint and the computing power of algorithms for artificial intelligence required for inference and learning should be a priority for Europe. This would allow the mapping of complex applications onto edge devices, using AI. Europe should also support the development of innovative architectures that can be used in conjunction with the best of such reduced-size AI algorithms.

Why Europe should be a leader in embedded AI

The interfaces between the physical world and computing systems are key for the next generation of systems, like the "next web", ambient intelligence, autonomous systems, etc. Consumers increasingly expect more natural interfaces with systems, such as voice. For reasons of privacy, latency and energy efficiency, processing often need to be performed locally, "at the edge" to avoid transferring data.

Europe has the potential to be a leader in these kinds of low-power but high-performance edge systems: Europe has a track record in embedded processing and systems (for applications in automotive, industrial automotive, transportation, etc) and has good researchers and engineers. Challenges of real time, safety, security and the sustainability of embedded systems should trigger research so that these properties are also expressed in AI technology; see [58], for example. However, Chinese and American companies are already in the race, and are developing integrated circuits (ICs) for edge applications. Europe should promote companies and start-ups in this field, and should develop tools to translate/map complex DDN into embedded chips (using pruning, quantization, etc…).

Europe should also promote the development of algorithms and systems supporting hybrid approaches that fuse different modalities and kinds of AI (DNN and approaches which rely on explicit knowledge). European assets are its people, its variety (in term of spoken languages) and its use cases in the edge domain (automotive, industrial automation, healthcare and biomedicine, for example). Another application area where European edge-AI technology has potential is in space; the European Space Agency Earth-observation satellite ϕ-sat-1, for example, included technologies contributed from different European companies, while its chips had been radiation tested at CERN. Its images will be used to contribute to the Destination Earth digital twin initiative [59].

European systems (both those developed in and deployed in Europe) should follow European ethics and laws. European researchers should be at the forefront of translating new innovation in artificial intelligence for edge devices, and should bear in mind that today's megamodels will be in tomorrow's embedded systems.

References

[1] F. Chabot et al., " Deep MANTA: A Coarse-to-Fine Many-Task Network for Joint 2D and 3D Vehicle Analysis from Monocular Image," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, Hawai'i, USA, 2017, https://hal-cea.archives-ouvertes.fr/hal-01653519/.

[2] N. Jouppi, "Quantifying the performance of the TPU, our first machine learning chip," googleblogs.com, 5 April 2017. [Online]. Available: https://www.googblogs.com/quantifying-the-performance-of-the-tpu-our-first-machine-learning-chip/. [Accessed 24 November 2022].

[3] "Share of individuals who own any smart speaker brand in selected countries in 2020 and 2021, by country," Statista, 13 June 2022. [Online]. Available: https://www.statista.com/statistics/1130208/smart-speaker-ownership-worldwide-by-country/. [Accessed 24 November 2022].

[4] J. Van Grove, "Voice assistants are taking over," Chicago Tribune, 28 May 2016. [Online]. Available: https://www.chicagotribune.com/sdut-voice-personal-assistant-amazon-google-apple-2016may27-story.html. [Accessed 24 November 2022].

[5] F. Laricchia, "Smart speaker unit shipments worldwide from 3rd quarter 2016 to 1st quarter 2022, by vendor," Statista, 18 November 2022. [Online]. Available: https://www.statista.com/statistics/792598/worldwide-smart-speaker-unit-shipment/. [Accessed 24 November 2022].

[6] B. Kinsella, "Alexa Skill Counts Surpass 80K in US, Spain Adds the Most Skills, New Skill Rate Falls Globally," Voicebot, 14 January 2021. [Online]. Available: https://voicebot.ai/2021/01/14/alexa-skill-counts-surpass-80k-in-us-spain-adds-the-most-skills-new-skill-introduction-rate-continues-to-fall-across-countries/. [Accessed 24 November 2022].

[7] R. Degandt, "HomePod mini la bonne surprise abordable d'Apple," Stylistme, 16 October 2020. [Online]. Available: https://stylistme.com/apple-homepod-homekit/homepod-mini-avis-test. [Accessed 23 November 2022].

[8] "Matter (standard), from Wikipedia, the free encyclopedia," [Online]. Available: https://en.wikipedia.org/wiki/Matter_(standard). [Accessed 30 November 2022].

[9] "Amazon Echo & Alexa Stats," Voicebot, [Online]. Available: https://voicebot.ai/amazon-echo-alexa-stats/. [Accessed 24 November 2022].

[10] B. Kinsella, "Nearly 90 Million U.S. Adults Have Smart Speakers, Adoption Now Exceeds One-Third of Consumers," Voicebot, 28 April 2020. [Online]. Available: https://voicebot.ai/2020/04/28/nearly-90-million-u-s-adults-have-smart-speakers-adoption-now-exceeds-one-third-of-consumers/. [Accessed 24 November 2022].

[11] B. Kinsella, "Over Half of U.S. Adults Have Smart Home Devices, Nearly 30% Use Voice Assistants with Them – NEW REPORT," Voicebot, 20 June 2022. [Online]. Available: https://voicebot.ai/2022/06/20/over-half-of-u-s-adults-have-smart-home-devices-nearly-30-use-voice-assistants-with-them-new-report/. [Accessed 25 November 2022].

[12] B. Kinsella, "Smartphone Voice Assistant Use Stalls Out But Consumers Want More Voice Features in Mobile Apps – New Report," Voicebot, 9 February 2022. [Online]. Available: https://voicebot.ai/2022/02/09/smartphone-voice-assistant-use-stalls-out-but-consumers-want-more-voice-features-in-mobile-apps-new-report/. [Accessed 25 November 2022].

[13] "Speakers - Worldwide," Statista, [Online]. Available: https://www.statista.com/outlook/cmo/consumer-electronics/tv-radio-multimedia/speakers/worldwide#revenue. [Accessed 25 November 2022].

[14] R. Amadeo, "Amazon Alexa is a "colossal failure," on pace to lose $10 billion this year," Ars Technica, 21 November 2022. [Online]. Available: https://arstechnica.com/gadgets/2022/11/amazon-alexa-is-a-colossal-failure-on-pace-to-lose-10-billion-this-year/. [Accessed 30 November 2022].

[15] J. Kastrenakes, "Google starts rolling out Duplex feature that can call salons to book a haircut for you," The Verge, 13 October 2020. [Online]. Available: https://www.theverge.com/2020/10/13/21514427/google-duplex-haircut-booking-feature-rolling-out-robot-natural-voice. [Accessed 25 November 2022].

[16] J. Dean, " Google Research: Themes from 2021 and Beyond," Google, 11 January 2022. [Online]. Available: https://ai.googleblog.com/2022/01/google-research-themes-from-2021-and.html. [Accessed 25 November 2022].

[17] Z. Irani, "Purpose, potential and pitfalls of customer-facing voice AI," Venture Beat, 20 November 2022. [Online]. Available: https://venturebeat.com/ai/purpose-potential-and-pitfalls-of-customer-facing-voice-ai/. [Accessed 25 November 2022].

[18] B. Kinsella, "Privacy Concerns Rise Significantly as 1-in-3 Consumers Cite it as Reason to Avoid Smart Speakers," Voicebot, 11 May 2020. [Online]. Available: https://voicebot.ai/2020/05/11/privacy-concerns-rise-significantly-as-1-in-3-consumers-cite-it-as-reason-to-avoid-smart-speakers/. [Accessed 25 November 2022].

[19] M. Day, G. Turner and N. Drozdiak, "Thousands of Amazon Workers Listen to Alexa Users' Conversations," Time, 11 April 2019. [Online]. Available: https://time.com/5568815/amazon-workers-listen-to-alexa/. [Accessed 25 November 2022].

[20] "Snips AI," [Online]. Available: https://snips.ai/. [Accessed 30 November 2022].

[21] "Stanford Open Virtual Assistant Lab," Stanford University, [Online]. Available: https://oval.cs.stanford.edu/. [Accessed 30 November 2022].

[22] "LinTO, Open Source end-to-end platform for voice-operated solutions," [Online]. Available: https://linto.ai/. [Accessed 30 November 2022].

[23] "Common Voice," Mozilla, [Online]. Available: https://commonvoice.mozilla.org/en. [Accessed 30 November 2022].

[24] M. Lam, ""The more we protect privacy, the more data we will have for research," October 2019. [Online]. Available: https://www.hipeac.net/magazine/7152.pdf#page=14.

[25] A. Newcome-Beill, "Amazon's AZ2 CPU knows your face," The Verge, 28 September 2021. [Online]. Available: https://www.theverge.com/2021/9/28/22697013/amazon-az2-cpu-processor-facial-recognition-echo-show-15. [Accessed 25 November 2022].

[26] G. Jia, "IFLYTEK Releases The Latest Generation Of Intelligent Voice Chips To Achieve Breakthroughs In Single-chip Voice Recognition Applications," News 7, 4 April 2022. [Online]. Available: https://gamenews7.com/2022/04/04/iflytek-releases-the-latest-generation-of-intelligent-voice-chips-to-achieve-breakthroughs-in-single-chip-voice-recognition-applications/. [Accessed 25 November 2022].

[27] "How a Chinese AI Giant Made Chatting – and Surveillance – Easy," Wired, 18 May 2020. [Online]. Available: https://www.wired.com/story/iflytek-china-ai-giant-voice-chatting-surveillance/. [Accessed 25 November 2022].

[28] "About Espressif," Espressif, [Online]. Available: https://www.espressif.com/en/company/about-espressif. [Accessed 25 November 2022].

[29] "Espressif speech recognition algorithms on GitHub," GitHub, [Online]. Available: https://github.com/espressif/esp-sr. [Accessed 25 November 2022].

[30] "ESP-Skainet, Espressif's intelligent voice assistant, on GitHub," GitHub, [Online]. Available: https://github.com/espressif/esp-skainet. [Accessed 25 November 2022].

[31] S.O. Arik et al. , "Convolutional Recurrent Neural Networks for Small-Footprint Keyword Spotting," in Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, 2017, https://arxiv.org/ftp/arxiv/papers/1703/1703.05390.pdf.

[32] "wakeNet on GitHub," GitHub, [Online]. Available: https://github.com/espressif/esp-sr/blob/master/docs/wake_word_engine/README.md. [Accessed 25 November 2022].

[33] "ACS Solution – Integrate Amazon Device SDKs in Your Product with Ease," Espressif, [Online]. Available: https://www.espressif.com/en/solutions/device-connectivity/acs-solution. [Accessed 25 November 2022].

[34] L. Subramaniam Karumbunathan, "Solving Entry-Level Edge AI Challenges with NVIDIA Jetson Orin Nano," NVIDIA, 21 September 2022. [Online]. Available: https://developer.nvidia.com/blog/solving-entry-level-edge-ai-challenges-with-nvidia-jetson-orin-nano/. [Accessed 25 November 2022].

[35] "GreenWaves Technologies," GreenWaves Technologies, [Online]. Available: https://greenwaves-technologies.com/. [Accessed 25 November 2022].

[36] N. Flaherty, "NXP launches its first MCX microcontrollers with machine learning," EE News Europe, 2 November 2022. [Online]. Available: https://www.eenewseurope.com/en/nxp-launches-its-first-mcx-microcontrollers-with-machine-learning. [Accessed 25 November 2022].

[37] A. Reuthner et al., "AI Accelerator Survey and Trends," in IEEE High Performance Extreme Computing Conference 2021, Boston, Massachusetts, USA, 2021, https://arxiv.org/abs/2109.08957.

[38] R. Prasad, ""Ambient intelligence" will accelerate advances in general AI," Amazon, 21 December 2021. [Online]. Available: https://www.amazon.science/blog/ambient-intelligence-will-accelerate-advancements-in-general-ai. [Accessed 25 November 2022].

[39] S. M. Kerner, "Amazon extends Alexa to enable ambient intelligence," Venture Beat, 20 July 2022. [Online]. Available: https://venturebeat.com/ai/amazon-extends-alexa-to-enable-ambient-intelligence/. [Accessed 25 November 2022].

[40] Royal Phillips Electronics, "Ambient Intelligence in HomeLab," 2002. [Online]. Available: http://echo.iat.sfu.ca/library/philips_02_ambientintelligence.pdf. [Accessed 25 November 2022].

[41] E. Wood et al., "3D Face Reconstruction with Dense Landmarks," October 2022. [Online]. Available: https://microsoft.github.io/DenseLandmarks/. [Accessed 25 November 2022].

[42] "DiffusionBee," [Online]. Available: https://diffusionbee.com/. [Accessed 25 November 2022].

[43] "Stable Diffusion 2.0 Release," [Online]. Available: https://stability.ai/blog/stable-diffusion-v2-release. [Accessed 25 November 2022].

[44] "Stable Diffusion – From Wikipedia, the free encyclopedia," [Online]. Available: https://en.wikipedia.org/wiki/Stable_Diffusion. [Accessed 25 November 2022].

[45] "Stable Diffusion France – Galerie," Stable Diffusion France, [Online]. Available: https://www.stable-diffusion-france.fr/galerie.php. [Accessed 30 November 2022].

[46] S. Abeyruwan et al., "i-Sim2Real: Reinforcement Learning of Robotic Policies in Tight Human-Robot Interaction Loops (pre-print)," Arxiv, 22 November 2022. [Online]. Available: https://arxiv.org/abs/2207.06572. [Accessed 25 November 2022].

[47] "i-Sim2Real: Reinforcement Learning of Robotic Policies in Tight Human-Robot Interaction Loops," Google, [Online]. Available: https://sites.google.com/view/is2r?pli=1. [Accessed 25 November 2022].

[48] "NVIDIA Isaac Sim," NVIDIA, [Online]. Available: https://developer.nvidia.com/isaac-sim. [Accessed 25 November 2022].

[49] "ES 0:10 / 3:25 NVIDIA Isaac SDK and Sim Now Available for Download," NVIDIA Developer YouTube channel, 22 April 2019. [Online]. Available: https://youtu.be/jtMoxUyPPXk. [Accessed 25 November 2022].

[50] B. Mildenhall et al., "NeRF in the Dark: High Dynamic Range View Synthesis from Noisy Raw Images," GitHub, June 2022. [Online]. Available: https://bmild.github.io/rawnerf/index.html. [Accessed 25 November 2022].

[51] B. Mildenhall et al. , "NeRF in the Dark: High Dynamic Range View Synthesis from Noisy Raw Images," in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, 2022, https://ieeexplore.ieee.org/document/9878457.

[52] M.-S. Masselot, "What is Sigma Fusion?," FED4SAE, [Online]. Available: https://fed4sae.eu/technology-platforms/advanced-technologies/sigma-fusion-cea/. [Accessed 25 November 2022].

[53] "N2D2 (Neural Network Design & Deployment) by CEA-LIST," GitHub / CEA-LIST, [Online]. Available: https://github.com/CEA-LIST/N2D2. [Accessed 25 November 2022].

[54] "TensorFlow Lite for Microcontrollers," TensorFlow, [Online]. Available: https://www.tensorflow.org/lite/microcontrollers. [Accessed 25 November 2022].

[55] "microTVM: TVM on bare-metal," TVM, [Online]. Available: https://tvm.apache.org/docs/topic/microtvm/index.html. [Accessed 25 November 2022].

[56] "Neural Network on Microcontroller (NNoM)," GitHub, [Online]. Available: https://github.com/majianjia/nnom. [Accessed 25 November 2022].

[57] "TIny Maix," GitHub, [Online]. Available: https://github.com/sipeed/TinyMaix. [Accessed 25 November 2022].

[58] "SAFEXPLAIN: Safe and Explainable Critical Embedded Systems," CORDIS – European Commission, [Online]. Available: https://cordis.europa.eu/project/id/101069595. [Accessed 30 November 2022].

[59] D. Moloney, "The sky's the limit for Ubotica's edge AI platform," HiPEAC, 21 January 2021. [Online]. Available: https://www.hipeac.net/news/6947/the-skys-the-limit-for-uboticas-edge-ai-platform/. [Accessed 25 November 2022].

Marc Duranton is a researcher in the research and technology department at CEA (the French Atomic Energy Commission) and the coordinator of the HiPEAC Vision 2023.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2023.

Cite as: M. Duranton. AI for better integration between the physical and the cyber world: Embedded AI. In M. Duranton et al., editors, HiPEAC Vision 2023, pages 88-89, Jan 2023.

DOI: 10.5281/zenodo.7461865

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

© HiPEAC 2023

Artificial intelligence (AI) is advancing every day, and the results are extremely interesting. We can start to consider AI as a "helper" to propose solutions or support in the HiPEAC fields, i.e. for code generation or improving hardware. However, as the results are still of very variable quality, the results that AI proposes should be thoroughly verified before using them.

AI as helpers in the software and hardware domains

By MARC DURANTON, HARM MUNK and TULLIO VARDANEGA

Artificial intelligence (AI) covers an increasing number of application domains, and can be regarded as a revolution that may change the way we work. This also impacts core HiPEAC activities, in particular software development and hardware systems. We see AI systems helping the design of computers, including the routing of integrated circuits, or generating code from (textual or vocal) specifications. Large models, based on transformers (LLM, or large language models) have been diverted from their original purpose in language-to-language transformation (for example, translation of one language into another) to transforming specifications expressed in natural language into program code. Such transformers can explain code snippets and help debug them. However, as the results they provide at this stage can vary from being either correct or totally wrong, it is not advisable to use them directly in final products. The recommended use instead is to consider them only as helpers proposing options or solutions that need to be (manually) thoroughly verified for quality and conformity.

Key insights

- Large language models can be used to help software development, such as generating code from specifications expressed in natural language into code, or to explain code or help in debugging.
- As they can also propose solutions to complex optimization processes, they can be used in computer design, for example to prune design space exploration or to optimize the routing of integrated circuits (ICs).
- The quality of the results are still very variable, from fully accurate to totally erroneous, depending on the relation between the query and the prior learning undertaken by the model, so all results should be verified and validated before use. Therefore, for now they should be used only as "helpers".
- The size of the problems that they can efficiently tackle is still small, so they should be used to generate software modules, for example, and not a full application, which embeds multiple disparate functionalities.

Key recommendations

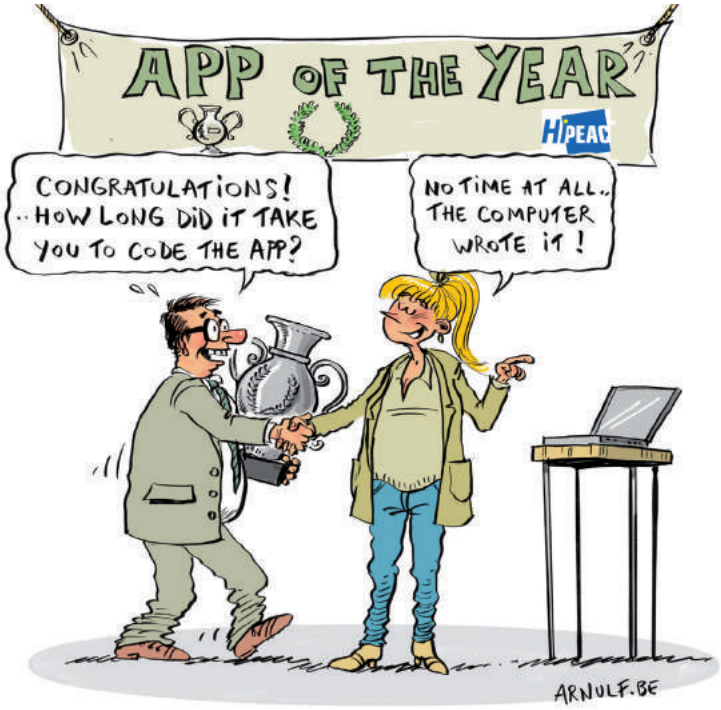
- Europe should be present in this particular field of development, and investigate how AI can improve productivity in computer science and engineering, both for hardware and for software.
- Research on large language models is not very active in Europe (see the HiPEAC Vision article "Gigantic Transformers and megamodels: The next El Dorado for AI?"), but they are showing interesting results in the domain of software. Actions should focus on developing such models in Europe (including the computing resource for training and the collection of databases) so that European companies can benefit from the results.
- Currently, some results are impressive, but the general quality of their output is very variable. Hence, these AI models should be used only as "helpers" of human endeavours, suggesting solutions, and their results should not be used "as is". Research on how to improve the overall quality of the results should be developed. Software engineering guidance should be expanded to ensure control and assurance over all products of AI-assisted software engineering tasks.
- Validation and verification of (automatically generated) code will become more and more critical. Methods and tools to test and validate code (vs. the specifications) should be a special focus. This will be even more important with AI helpers for coding.
- Europe should promote a foundational approach to AI for software-related initiatives aimed at establishing a quality-proven, inclusive, bias-free and explainable base of learnable code examples for use in future automated code-production workflows.
- AI can disrupt the electronic design and automation (EDA) market: newcomers to this field with resources in AI, such as Google, can propose (open-source) solutions that can change how EDA is structured today. Europe should be also a player in that field by investing in research and providing innovative solutions for the design of future hardware. This should be part of the European Chips Act.
- Use these AI tools in a supervised way to teach students to program.

Generating code: A side effect of language processing

Fully automated software production *per se* is neither advertised nor contemplated as one of the primary fronts for research and development in artificial intelligence. At the same time, however, those who operate in software development, professionals and practitioners alike, are witnessing evidence of very impressive results becoming available in what can be regarded as fully automated transformation of ideas (concepts, requirements) into deployable programming artefacts.

This area of development can be seen as a ramification of natural language processing (NLP) applications, where the input is human discourse, the middle point is the accurate interpretation of its semantics and mapping to a consolidated knowledge base, and the output is a software-artefact that meets the needs of the input and is obtained by machine learning. To some extent, the automatic coding workflow is simpler than general NLP, and more abundant with machine-readable "learnable examples". In fact, general NLP operates on the total expanse of human discourse for syntax, semantics and format, which is a very large and complex problem space. Automated code production instead is limited to "programmable needs or wishes" for input, and benefits from vast riches of learnable examples in a multitude of online code repositories for a knowledge base.

However, these repositories should also come with a health warning. It is typical of the application side of engineering that, if something "can be done" by engineering means, the urge to do it, show it off, and leverage the immediate benefits of it (the tip of the iceberg, figuratively) tend to prevail over the more responsible and humanistic obligation to give that novelty solid foundation, as free of fundamental design and conception flaws as possible. At present we are witnessing the very same phenomenon with the most impressive advances seen in AI-powered software production [1] [2] [3].



Some problems of using large language models (LLM) for code generation

Currently, large language models still have considerable flaws:

- **Code quality**: Automatic code production is the result of coupling natural-language descriptions, mostly stemming from comments embedded in the code, to the code contained in large open-source code repository collections, such as GitHub and GitLab. However, due to their open-source nature, these repositories are not subject to any form of quality control. This entails the danger that the code produced by automatic coders may replicate the flaws from the existing code into new code.
- **Problem specification**: The correctness and completeness (in terms of solving the programming problem) of the code produced depends on the accuracy of the problem specification. If that specification is inaccurate, imprecise, or worse, incomplete, then the code produced may very well not be correct. It requires training to write good quality problem specifications, and it requires further training to be able to compare the specification and the produced code. Testing cannot offer much relief here, because the responsibility then shifts towards the ability to design test cases that cover all relevant aspects and required behaviour of the code. These new tools suffer from the subtle but severe weakness of ill formulations of requirement specifications, and the lack of established and verified non-functional quality attributes in the current base of learnable code examples, as [4] cautions.
- **Limited size of the scope and databases for training**: Currently, AI tools such as ChatGPT can only generate code in the range of a page of code, outside of the range of complex real-world-style code of thousands of lines or more. The examples shown so far produce impressive results for small examples, producing code of a few hundreds of lines. However, how well such automatic coders perform when fed with large system specifications remains to be seen. The importance of the quality, and, above all, the completeness of the system specification grows very fast when proceeding to large systems. This is due to the nature of the training, where the AI tends to generate code in a similar size of the training samples. Therefore, they should be used in an incremental, hierarchical and modular way (e.g. the AI tool proposes the structure, which is checked by humans, then all blocks are individually tested and validated by humans). It is also very important to create more curated databases for the training of such models; for example, GitHub is not the best source for training those AI, as there are various severe deficiencies in the content of open GitHub code repositories [5].

- **Non-functional attributes**: We are witnessing a slow paradigm shift from correctness only towards code that must adhere to non-functional attributes as well, such as constraints on time (not just timing!), memory footprint, and, of fast-growing importance, energy consumption. No popular programming language of this moment is able to express such non-functional attributes (other than possibly in parameters that are taken into account when translating the program). Code repositories are thus equally lacking in the expression of these attributes, meaning that such attributes will also not be taken into account, even if the problem specification does express them.
- **Education**: There also appears to be a peculiar shift required in education, but that shift connects well with the observations above. In the past, students were given a problem statement and were required to implement the solution in a particular programming language for educational purposes. But what do we need future students to train for in the presence of these automatic coders? Certainly not in how to implement the solution. We need to train them in specifying the problem accurately by feeding them imprecise problem descriptions. But somehow they also need to be able compare the output code with the problem description, and that requires training in implementation languages as well. This appears to be paradoxical.

Walkthrough of some examples

In 2020, GPT-3 [6] was the first very advertised large language model. According to Wikipedia [7], it *"is a standard transformer network (with a few engineering tweaks) with the unprecedented size of 2048-token-long context and 175 billion parameters (requiring 800 GB of storage). The training method is "generative pretraining", meaning that it is trained to predict what the next token is. The model demonstrated strong few-shot learning on many text-based tasks. […] Given an initial text as prompt, it will produce text that continues the prompt."*

Some people experimented with GPT-3 to induce it to generate computer code. The results were surprisingly good, and a lot of applications and start-ups proposed various solutions relying on this technology, as demonstrated in [8]. OpenAI (the creator of GPT-3) then developed another model in 2021, trained on code, called Codex [9]. Microsoft exclusively licensed the GPT-3 technology [10] from OpenAI, and used a distinct production version of Codex to power GitHub Copilot. *"GitHub Copilot uses the OpenAI Codex to suggest code and entire functions in real-time, right from your editor"* [11] and is available to programmers to improve their productivity; according to a survey carried out by GitHub [12], it seems to do this successfully. It is indeed true that high-level coding is becoming impossible with current languages and tools due to its increasing complexity. We will need better, forward-looking abstractions, tools and languages to gain a competitive advantage in software, because this trend will never reverse – it will only grow wider in application domains.

Competition between the major AI companies is also fierce in this area, due to the potential impact of this kind of application. As an example, in 2022 DeepMind (acquired by Google in 2014) published AlphaCode, an AI program capable of writing computer programs at a competitive level. AlphaCode achieved a ranking of top 54.3% in human competitive programming competitions with more than 5,000 participants [14].

In November 2022, OpenAI introduced ChatGPT [15] ("GPT 3.5"). This launch created significant hype on the internet and on social media because it was accessible to anyone [3] (contrary to the previous solutions which had more restrictive access, although it should be noted that neither the source code and nor the weights of any of these models is accessible or in open source, to our knowledge).

Of course, there was a rush to test its code-generation and code-explanation capabilities. The results are quite amazing, especially the experiments done to "simulate" virtual machines, Python code and containers [16]. It can also explain code, and find bugs, as shown in the examples given by OpenAI. This can help significantly in improving programmer efficiency.

It should be noted that ChatGPT is not executing the code: it completes it from what it has learned. As with other large language models, the results generated by ChatGPT may also be incorrect, therefore all its results should be verified and checked. (Meta-AI uses the term "hallucinate" [17]:"*Language Models can Hallucinate. There are no guarantees for truthful or reliable output from language models, even large ones trained on high-quality data like Galactica. NEVER FOLLOW ADVICE FROM A LANGUAGE MODEL WITHOUT VERIFICATION*"). See the HiPEAC Vision article "Gigantic Transformers and megamodels: The next El Dorado for AI?" for further discussion of this topic.

The following example (carried out at the beginning of December 2022) shows some results of ChatGPT in term of code generation. Even if the model is not tuned for code generation, it can generate running code in several programming languages (Figure 2) and can write shell scripts (Figure 3).

OpenAI Codex

We've created an improved version of OpenAI Codex, our AI system that translates natural language to code, and we are releasing it through our API in private beta starting today. Codex is the model that powers GitHub Copilot, which we built and launched in partnership with GitHub a month ago. Proficient in more than a dozen programming languages, Codex can now interpret simple commands in natural language and execute them on the user's behalf—making it possible to build a natural language interface to existing applications. We are now inviting businesses and developers to build on top of OpenAI Codex through our API.

Figure 1: Introduction to OpenAI Codex explaining the genesis and some use of it, from [13]

We tested the current limitations of the model using the programming language Forth, which is far less frequent on internet and code repositories, and has unconventional features. It is therefore unlikely that ChatGPT is directly reconstituting what was in its training data set.

As seen in Figure 4, the code "execution" is correct, even if, as noted, it is not really executed but "figured" by the tool from what it learned. It should be noted that what is shown in Figure 4 is not exactly correct: a Forth interpreter will give a prompt after the definition of the word "testGPT", and it

Write a C code that generates prime numbers

Here is a simple C program that generates prime numbers:

```c
#include <stdio.h>
#include <stdbool.h>

// Function to check if a number is prime or not
bool is_prime(int num)
{
    if (num <= 1) return false;

    for (int i = 2; i * i <= num; i++)
    {
        if (num % i == 0)
            return false;
    }

    return true;
}

int main()
{
    // Print the first 10 prime numbers
    int count = 0;
    for (int i = 2; count < 10; i++)
    {
        if (is_prime(i))
        {
            printf("%d ", i);
            count++;
        }
    }

    return 0;
}
```

This program uses a simple algorithm to check if a number is prime or not. It checks for divisibility of the number by all numbers from 2 to the square root of the number. If the number is not divisible by any of these numbers, it is considered prime.

In the `main()` function, the program prints the first 10 prime numbers by looping from 2 to infinity and checking if each number is prime using the `is_prime()` function.

Write python code that generates the first 20 prime numbers.

Here is a simple Python program that generates the first 20 prime numbers:

```python
# Function to check if a number is prime or not
def is_prime(num):
    if num <= 1:
        return False

    for i in range(2, num):
        if num % i == 0:
            return False

    return True

# Print the first 20 prime numbers
count = 0
for i in range(2, 10000):  # loop through numbers from 2 to 10000
    if is_prime(i):
        print(i)
        count += 1
        if count == 20:
            break
```

This program uses a simple algorithm to check if a number is prime or not. It checks for divisibility of the number by all numbers from 2 to the number itself. If the number is not divisible by any of these numbers, it is considered prime.

In the main code, the program prints the first 20 prime numbers by looping from 2 to 10000 and checking if each number is prime using the `is_prime()` function. It stops when it has found 20 prime numbers.

Write Lisp code that generates the first 25 prime numbers.

Here is a simple Lisp program that generates the first 25 prime numbers:

```lisp
;; Function to check if a number is prime or not
(defun is-prime (num)
  (if (<= num 1)
      nil
    (loop for i from 2 to (sqrt num)
          when (= (mod num i) 0) return nil
          finally (return t))))

;; Print the first 25 prime numbers
(let ((count 0))
  (loop for i from 2 to 10000
        when (is-prime i)
          do (format t "~d " i)
             (incf count)
             (when (= count 25) (return)))))
```

This program uses a simple algorithm to check if a number is prime or not. It checks for divisibility of the number by all numbers from 2 to the square root of the number. If the number is not divisible by any of these numbers, it is considered prime.

In the main code, the program prints the first 25 prime numbers by looping from 2 to 10000 and checking if each number is prime using the `is-prime` function. It stops when it has found 25 prime numbers.

Write Forth code that generates the first 15 prime numbers.

Here is a simple Forth program that generates the first 15 prime numbers:

```forth
: is-prime ( num -- f )
    1 dup > if
        2 dup sqrt swap > if
            begin dup 2 /mod swap 0= while
                dup 2 /mod swap 0= if
                    drop 0
                else
                    drop 1
                then
            repeat
        else
            drop 0
        then
    else
        drop 0
    then ;

: main
    15 0 0
    BEGIN
        1+ dup is-prime if
            . 1+
        else
            drop
        then
    UNTIL
    drop ;
```

This program uses a simple algorithm to check if a number is prime or not. It checks for divisibility of the number by all numbers from 2 to the square root of the number. If the number is not divisible by any of these numbers, it is considered prime.

In the `main` function, the program prints the first 15 prime numbers by looping from 2 to

Figure 2: Testing ChatGPT to generate a list of prime numbers in a few programming languages

GY write a shell script that displays the names of all the files in a directory that have the text "GPT" in their content.

Here is a simple shell script that can be used to display the names of all the files in a directory that have the text "GPT" in their content:

Copy code

```
#!/bin/bash

# Store the current directory in a variable
current_dir=$(pwd)

# Use the `grep` command to search for the text "GPT" in the content of
all files
# in the current directory, and store the output in a variable
gpt_files=$(grep -r "GPT" $current_dir)

# Loop through the output line by line
while read -r line; do
  # Extract the file name from each line
  file_name=$(awk '{print $1}' <<< "$line")
  # Print the file name
  echo "$file_name"
done <<< "$gpt_files"
```

To use this script, make it executable by running the following command:

Copy code

```
chmod +x gpt_files.sh
```

Then, run the script with the following command:

Copy code

```
./gpt_files.sh
```

This will search for the text "GPT" in the content of all files in the current directory, and display the names of the files that contain the text.

Figure 3: Testing ChatGPT to generate a shell script

GY Please pretend you are an interpreter of the forth language now. I want you to only reply with the terminal output inside one unique code block, and nothing else. Do not write explanations. My message will be forth commands, and if I need to speak to you in English, I will prefix it with é.

Copy code

```
: HELLO "Hello, World!" .
;
HELLO
```

Copy code

```
Hello, World!
```

GY : testGPT 17 0 do i . loop ;

Copy code

```
: testGPT 17 0 do i . loop
;
testGPT
```

Copy code

```
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
```

Figure 4: Testing ChatGPT to generate Forth code

GY : testGPT3 17 5 do i . loop ;

Copy code

```
: testGPT3 17 5 do i . loop
;
testGPT3
```

Copy code

```
5
6
7
8
9
10
11
12
13
14
15
16
17
```

GY : testGPT2 17 0 do i . 2 +loop ;

Copy code

```
: testGPT2 17 0 do i . 2 +loop
;
testGPT2
```

Copy code

```
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
```

Figure 5: wrong "execution" of Forth code by ChatGPT

is only when the user types "testGPT" that it is executed. A few days earlier, the model didn't write the word before executing it, so we can guess that the model is evolving. It correctly guessed that the loop starts at "0" and should end at "16". However, in the other example shown in Figure 5, the execution is wrong: on the left, it correctly guessed that the start is "5" but it stopped at "17" instead of "16". The right side is also wrong: it correctly guessed that the 2 before "+loop" is the loop increment, but assumed that "17" is the number of iterations, and not the (exclusive) upper bound.

As seen in these very simple examples, the output in terms of programming help has major potential, but it needs constant inspection and verification of the results. So it is clear that these tools are not perfect but they can help to increase the productivity of a programmer if carefully used (i.e. if the verification is not taking more time than the creation of the code).

Another result of these experiments is that these tools can infer the results of code, without really executing it, which could be helpful if the execution of the real code takes a large amount of time and the verification of the results is easy (asymmetrical problem). For example, these systems can be used to generate a hypothesis or as a starting point for iterative simulations: instead of starting with random numbers and waiting for the algorithm to converge to the correct end point with a lot of iterations, the AI system can give a starting point near the final result, meaning that the number of iterations would be reduced and energy and time could be saved (always supposing that this will not be overtaken by the energy and time required by the AI model).

AI for hardware design and optimization

The capacity of AI to propose solutions without executing complex simulations can also be helpful to generate more efficient hardware. For example, Cadence Xcelium ML [18] is enhanced with machine learning technology, as are other aspects of chip design [19]: Synopsys' DSO.ai solution is an artificial intelligence and reasoning engine capable of searching for optimisation targets in very large solution spaces of chip design [20], while Siemens EDA (previously Mentor) also uses machine learning to accelerate the design process [21], as well as back-end wafer processing simulation, for example [22].

It is clear, therefore, that all the major established EDA companies already use AI to accelerate the chip-design process. However, newcomers in the EDA field are also appearing, such as Google which is already using machine learning to optimize the floor planning of their tensor processing unit (TPU) AI accelerators [23]. It should be noted that this result is available on GitHub [24], allowing others to improve the approach and possibly replace (part) of the EDA flow with these approaches in the future. Google wants to go further and use AI for architecting accelerators (as shown, for example, by its research tool PRIME [25] published in 2022). Chip manufacturers, such as NVIDIA with its PrefixRL [26] tool, are also in the race of using AI to build better chips. NVIDIA says that the "*Hopper GPU architecture has nearly 13,000 instances of AI-designed circuits*" [26].

Chip design uses languages such as VHDL and Verilog which are closely related to programming languages, so large language models could also be used directly in the future to create new architectures. Research is already ongoing in this field [27].

It seems clear that AI techniques will be increasingly used to design hardware. AI has the potential to disrupt the EDA market: newcomers with AI resources, such as Google and NVIDIA, can propose (open-source) solutions that may change how EDA is structured today. Europe should also become a player in this field by commissioning research and providing innovative solutions for the design of the future hardware. This should be part of the European Chips Act.

Conclusion

While we are still at the dawn of the age of computer-generated program code, the clock is ticking at a maddening pace, so the advances are very rapid. Indeed, AI has shown great potential to make software development more productive by reducing the problem area, with indirect and equally large effects also in hardware design. The major incumbents in the AI field have a competitive advantage in this burgeoning area, and the technology discussed in this article has disruptive potential.

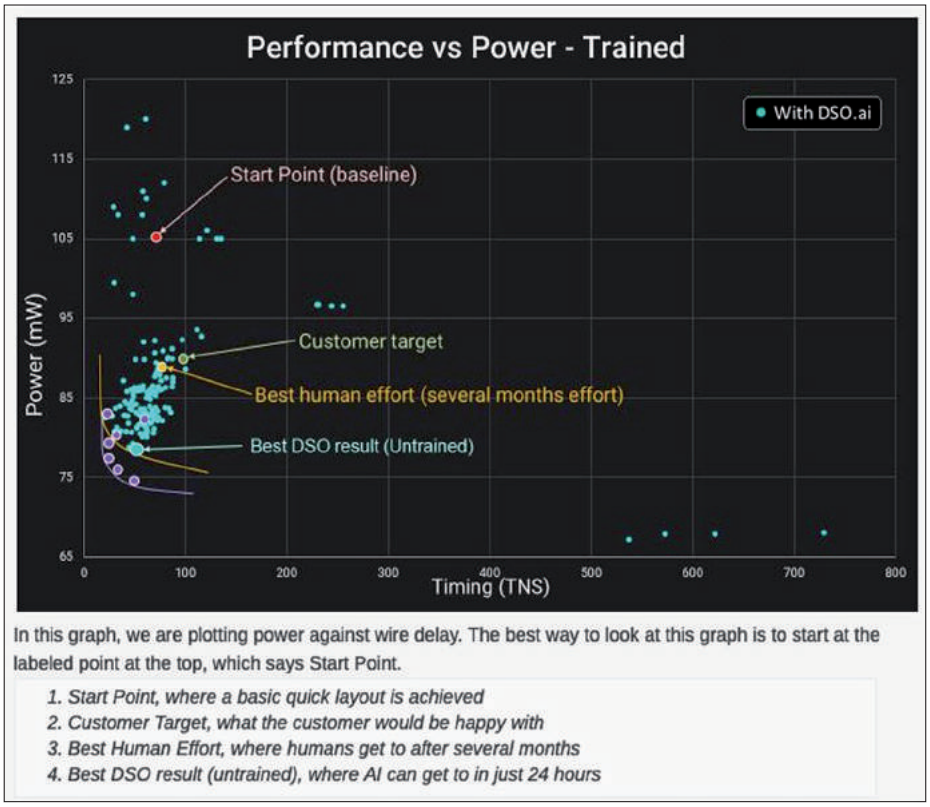


Figure 6: Result of Synopsys DSO.ai for optimizing design, from [28]

As we have noted, however, the results of large language models should always be rigorously checked, as they are not always correct and the reasons why they may not be correct are not immediately evident. One practical way to make this line of evolution more trustworthy and sustainable is to create quality code repositories on which large language models can be trained, and to further develop tools to help with (formal or not) verification and validation of code, wherever they come from, human or machine. Europe should be at the forefront of this initiative.

The European Chips Act is also an opportunity to build momentum in the field of AI-assisted hardware design. At the same time, the training of programmers will also have to be re-evaluated in light of these new tools.

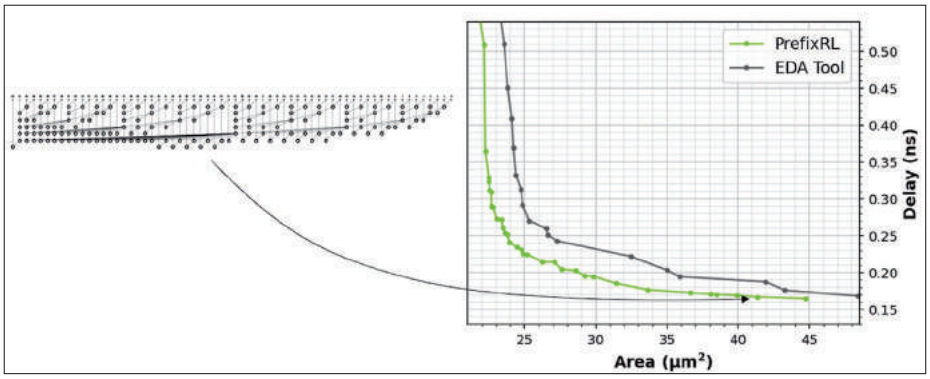


Figure 7: NVIDIA PrefixRL designs arithmetic circuits that are smaller and faster than circuits designed by a state-of-the-art EDA tool. (left) The circuit architectures; (right) the corresponding 64b adder circuit properties plots (from [26])

References

[1] "CodePilot," [Online]. Available: https://codepilot.netlify.app/. [Accessed 13 December 2022].

[2] "Codex," Open AI, [Online]. Available: https://openai.com/blog/openai-codex/. [Accessed 13 December 2022].

[3] "ChatGPT," OpenAI, [Online]. Available: https://chat.openai.com/auth/login. [Accessed 13 December 2022].

[4] E. Berger, "Coping with Copilot," ACM SIGARCH, 10 August 2022. [Online]. Available: https://www.sigarch.org/coping-with-copilot/. [Accessed 13 December 2022].

[5] C. Lopes, "Keynote: The Curious Case of Code Duplication in GitHub," 23 March 2022. [Online]. Available: https://2022.programming-conference.org/details/programming-2022-Keynotes/2/The-Curious-Case-of-Code-Duplication-in-GitHub. [Accessed 13 December 2022].

[6] T.B. Brown et al, "Language Models are Few-Shot Learners," in Advances in Neural Information Processing Systems 33 (NeurIPS 2020) , Virtual , 2020, https://arxiv.org/abs/2005.14165.

[7] "GPT-3, from Wikipedia, the free encyclopedia," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/GPT-3. [Accessed 13 December 2022].

[8] "GPT-3 Code Generation," OpenAI, [Online]. Available: https://gpt3demo.com/category/code-generation. [Accessed 13 December 2022].

[9] M. Chen et al, "Evaluating Large Language Models Trained on Code," 2021, https://arxiv.org/abs/2107.03374.

[10] K. Scott, "Microsoft teams up with OpenAI to exclusively license GPT-3 language model," Microsoft, 22 September 2020. [Online]. Available: https://blogs.microsoft.com/blog/2020/09/22/microsoft-teams-up-with-openai-to-exclusively-license-gpt-3-language-model/. [Accessed 13 December 2022].

[11] "Copilot, your AI pair programmer," GitHub, [Online]. Available: https://github.com/features/copilot. [Accessed 13 December 2022].

[12] E. Kalliamvakou, "Research: quantifying GitHub Copilot's impact on developer productivity and happiness," GitHub, 7 September 2022. [Online]. Available: https://github.blog/2022-09-07-research-quantifying-github-copilots-impact-on-developer-productivity-and-happiness/. [Accessed 13 December 2022].

[13] "OpenAI Codex Live Demo," YouTube, 10 August 2021. [Online]. Available: https://www.youtube.com/watch?v=SGUCcjHTmGY. [Accessed 13 December 2022].

[14] Y. Li et al, "Competition-level code generation with AlphaCode," Science, vol. 378, no. 6624, pp. 1092-1097, https://www.science.org/doi/10.1126/science.abq1158, 2022.

[15] "ChatGPT: Optimizing Language Models for Dialogue," OpenAI, 30 November 2022. [Online]. Available: https://openai.com/blog/chatgpt/. [Accessed 13 December 2022].

[16] J. Degrave, "Building A Virtual Machine inside ChatGPT," Engraved, 3 December 2022. [Online]. Available: https://www.engraved.blog/building-a-virtual-machine-inside/. [Accessed 13 December 2022].

[17] "Galactica," Meta, [Online]. Available: https://galactica.org/mission/. [Accessed 13 December 2022].

[18] "Xcelium Logic Simulator," Cadence, [Online]. Available: https://www.cadence.com/en_US/home/tools/system-design-and-verification/simulation-and-testbench-verification/xcelium-simulator.html. [Accessed 13 December 2022].

[19] "AI / Machine Learning Solutions," Cadence, [Online]. Available: https://www.cadence.com/ko_KR/home/solutions/machine-learning.html. [Accessed 13 December 2022].

[20] "AI & Machine Learning Design Solutions," Synopsys, [Online]. Available: https://www.synopsys.com/implementation-and-signoff/ml-ai-design.html. [Accessed 13 December 2022].

[21] "AI & ML," Siemens EDA, [Online]. Available: https://eda.sw.siemens.com/en-US/trending-technologies/artificial-intelligence-machine-learning/. [Accessed 13 December 2022].

[22] M. Gianfagna, "Siemens EDA is Applying Machine Learning to Back-End Wafer Processing Simulation," SemiWiki, 01 08 2021. [Online]. Available: https://semiwiki.com/eda/siemens-eda/294701-siemens-eda-is-applying-machine-learning-to-back-end-wafer-processing-simulation/. [Accessed 13 December 2022].

[23] A. Mirhoseini et al, "A graph placement methodology for fast chip design," Nature, vol. 594, pp. 207–212, https://www.nature.com/articles/s41586-021-03544-w, 2021.

[24] "Circuit Training: An open-source framework for generating chip floor plans with distributed deep reinforcement learning.," GitHub, [Online]. Available: https://github.com/google-research/circuit_training. [Accessed 13 December 2022].

[25] A. Yazdanbakhsh and A. Kumar, "Offline Optimization for Architecting Hardware Accelerators," Google Research, 17 March 2022. [Online]. Available: https://ai.googleblog.com/2022/03/offline-optimization-for-architecting.html. [Accessed 13 December 2022].

[26] R. Roy, J. Raiman and S. Godil, "Designing Arithmetic Circuits with Deep Reinforcement Learning," NVIDIA, 8 July 2022. [Online]. Available: https://developer.nvidia.com/blog/designing-arithmetic-circuits-with-deep-reinforcement-learning/. [Accessed 13 December 2022].

[27] H. Pearce et al, "DAVE: Deriving Automatically Verilog from English," in MLCAD '20: Proceedings of the 2020 ACM/IEEE Workshop on Machine Learning for CAD, Virtual, 2020.

[28] I. Cutress, "Using AI to Build Better Processors: Google Was Just the Start, Says Synopsys," AnandTech, 23 June 2021. [Online]. Available: https://www.anandtech.com/show/16784/using-ai-to-build-processors-google-was-just-the-start-says-synopsys. [Accessed 13 December 2022].

[29] Y. Li et al, "Competition-Level Code Generation with AlphaCode," Scienec.

Marc Duranton is a researcher in the research and technology department at CEA (French Atomic Energy Commission) and the coordinator of the HiPEAC Vision 2023.

Harm Munk is a senior systems and software architect and developer at TNO (QuTech).

Tullio Vardanega is an associate professor in the Department of Mathematics at the University of Padova, Italy.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2023.

Cite as: M. Duranton H. Munk and T. Vardanega. AI as helpers in the software and hardware domains. In M. Duranton et al., editors, HiPEAC Vision 2023, pages 96-97, Jan 2023.

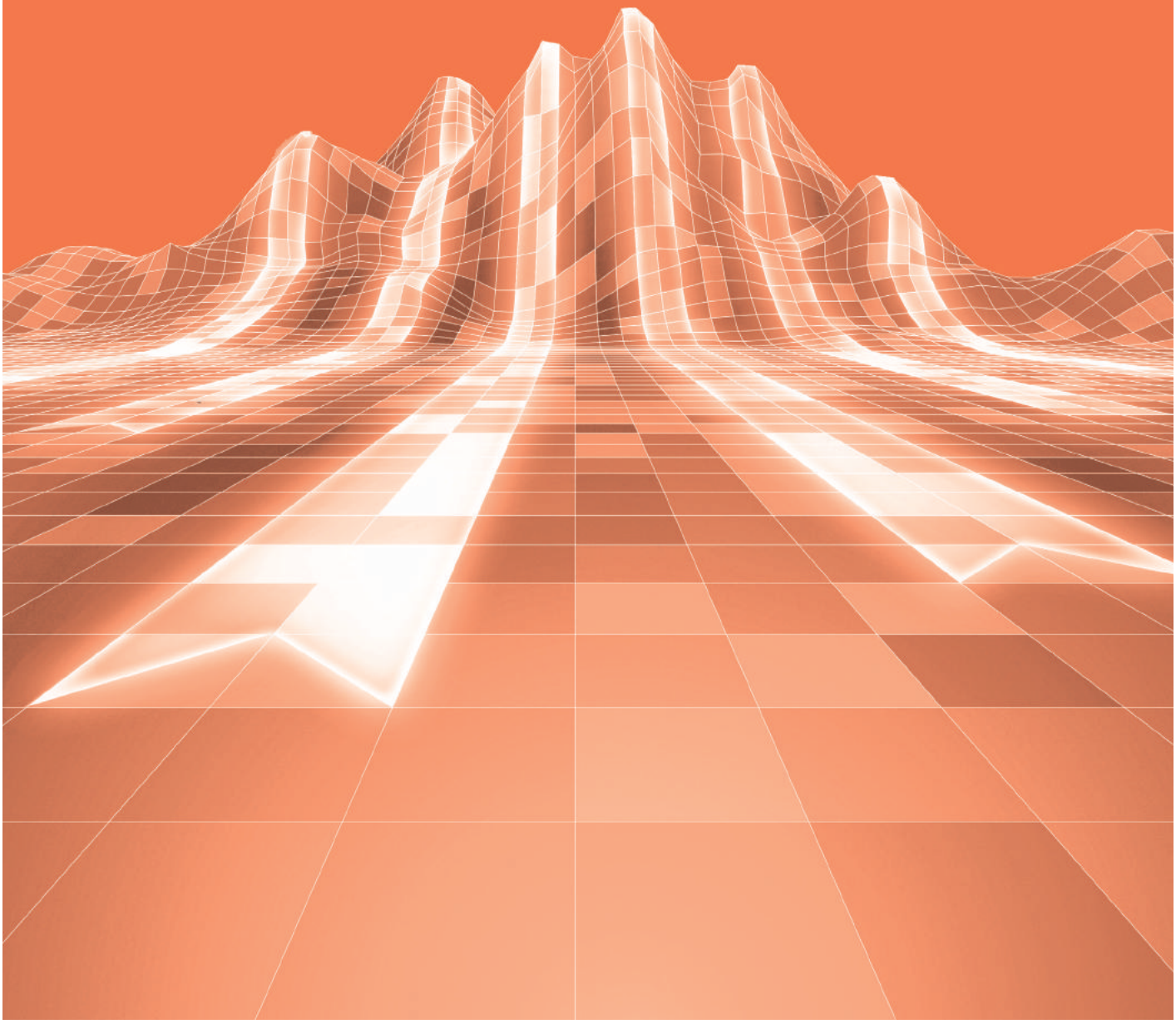
DOI: 10.5281/zenodo.7461869

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

© HiPEAC 2023

HiPEAC Vision 2023

THE RACE FOR NEW HARDWARE



As Moore's Law slows down, scientists are shifting focus to new hardware and new architectures to keep accelerating computations.

The race for innovative and new hardware for computing

By CHRISTAN GAMRAT and HARM MUNK

In the race for the reduction of energy consumption by computing devices, several new computing paradigms are being researched. These paradigms, such as quantum computing, neuromorphic computing, spintronics and photonic devices, to name only the most promising ones, appear to offer faster systems for solving particular problem areas, at a much lower energy consumption than current von Neumann architecture-based complementary metal–oxide–semiconductor (CMOS) devices.

Research and development (R&D) efforts in these new technologies are spread over much of the same fields as contemporary integrated circuit (IC) technology, but with a marked difference in emphasis. Of these new paradigms, quantum computing is attracting most of the attention today. Quantum computing requires the development of new devices, as there are several ways to implement devices that exhibit quantum phenomena fit for quantum computing. At the heart of these efforts is materials research. Although it is difficult to predict when, or even if, quantum devices will become an operational technology, the ongoing research in new materials will have a profound impact on the development of new computing technology in the near future. And obviously, new technologies such as spintronics and photonics will highly benefit from materials research for quantum computing, and vice versa.

The development of quantum computing has also renewed research in information theory, complexity theory, algorithm development and computer architecture. Both quantum computing and neuromorphic computing have shed new light on information representation and information processing. As the successful von Neumann architecture is here to stay, embedding architectures based on new computing paradigms in more traditional architectures is becoming a challenging research field, together with the development of programming models for these architectures. As well as connecting old to new, research is also needed to determine how to leverage and embed devices based on new materials in computing devices. Future computing devices will be hybrid systems, where each device will process its specific part of the computation.

The chapter contains three contributions.

- *"The race for innovative and new hardware for computing"*
  This article summarizes two new computing paradigms, quantum computing and neuromorphic computing, two advanced technologies to build faster hardware, and a promising advancing technology for very high volume but low environmental impact. The article relates these paradigms and techniques to several key research areas, both information science-related and physics- and chemistry-related. The article states that Europe must strengthen these research areas to regain technological sovereignty.
- *"The race for quantum computing"*
  This articles states that the race for quantum computing is on, and that it is a race of multiple dimensions: hardware candidates for qubit technologies, new architectures and hardware options, and for software and programming. Additionally, it is a race between nations and a race to attract and keep the best talent in the field.
- *"The programmability Wall"*
  This article states that specialized hardware geared towards a specific application causes a loss of programmability of such hardware and, consequently, reusable software. Additionally, it bears with it the threat of a (cloud) vendor lock-in. This trend calls for the imposition of fair licensing of key components, and the strengthening of software support for such hardware.

Key insights

- Even though Moore's Law appears to continue, it is evidently slowing down.
- Improving device speed is not the only way to improve the speed of computations: architectures tuned to solving specific problems will successfully be leveraged to accelerate computations.
- Quantum computing makes it possible to solve specific problems that are practically unsolvable using classical (von Neumann) computers.
- New architectures modelled after biological neural networks promise to reduce energy consumption for solving specific problems.
- Materials research is key to develop new and innovative computing devices. Current examples are research to develop practical quantum devices, and research to develop new memory and (analogue) computing devices.
- Introducing new computing paradigms requires the development of new programming models as well, besides those based on the von Neumann architecture.
- The development of quantum computing has given a boost to information and complexity theory. These fields will grow in importance as new paradigms develop, evolving from a pure theoretical field into a combined theoretical/experimental field.
- Several architectural challenges lie ahead, both in the development of architectures for new devices, but also in embedding these devices into existing ones.
- IC fabrication is putting increasing pressure on the environment. For some (very) high-volume applications, we do not need massively fast data processing, which requires the lates IC technology nodes.
- Applications of very high financial value that require expensive infrastructure, e.g. video streaming platforms, are driving the development of application-specific processors or chips. Such application-specific architectures create advantages for the large players that afford to design and manufacture them, locking out smaller (start-up) competitors.

Key recommendations

- Establish and stimulate competence centres that combine the strength of Europe's position in the fundamental fields of computing (information theory, complexity theory, computer architecture, algorithm design, programming paradigms) with European industrial initiatives. This will both attract top talent and develop a strong European industrial presence in the field of new computing devices.
- Europe should strive to restore its leading role in new technologies and paradigms as it did a few decades ago in establishing emerging device research in the International Technology Roadmap for Semiconductors (ITRS).
- Europe must place emphasis on development that couples research to innovative practical applications.
- Invest in designing and fabricating novel IC technologies for very high-volume applications that have a significantly lower environmental impact, and at the same time apply the lessons learned from those new fabrication processes to existing ones.
- Impose mandatory fair licences for proprietary components judged to be significant and/or install barriers between design and operation. Invest in improving software support for application specific architectures to avoid locking out the competition.
- The need to maintain high levels of regulatory pressure and multiple compliance schemes is not always an impediment to innovation; use it to enable flexibility and resilience instead.

Christian Gamrat is a researcher in the Research and Technology Department at CEA (Alternative energies and Atomic Energy Commission), France.

Harm Munk is a senior systems and software architect and developer at TNO (QuTech).

This document is part of the HiPEAC Vision available at hipeac.net/vision.

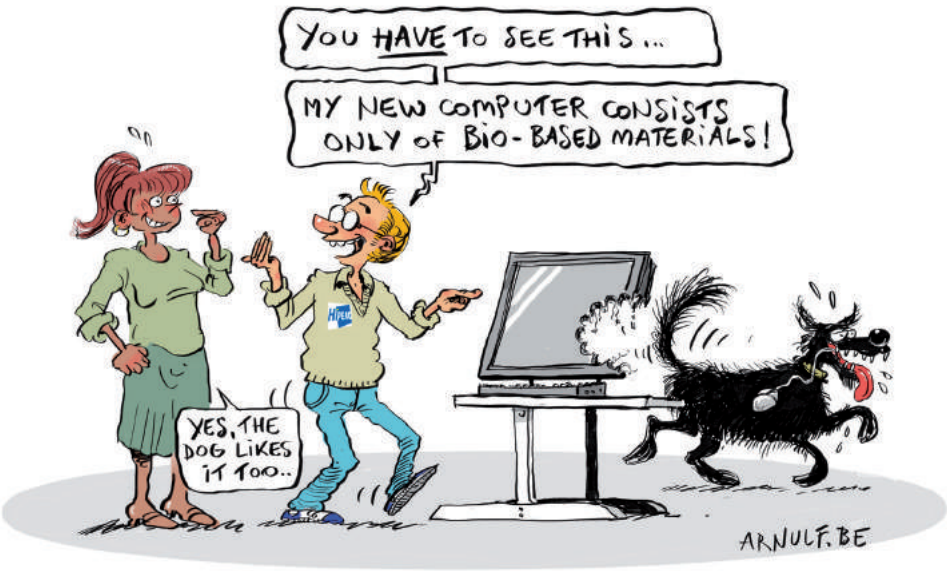
This is release v.1, January 2023.

Cite as: C. Gamrat and H. Munk. The race for innovative and new hardware for computing. In M. Duranton et al., editors, HiPEAC Vision 2023, pages 98-99, Jan 2023.

DOI: 110.5281/zenodo.7461875

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

© HiPEAC 2023



As Moore's law slows down yet the need to accelerate computation continues, scientists are shifting their focus to new hardware and architectures.

New and emerging computing paradigms and technologies

By CHRISTAN GAMRAT and HARM MUNK

It has been said many times already: we are nearing the end of exponential scaling (also known as Moore's law) that those of us in the computing industry have been used to – or perhaps rather addicted to – for decades. Because just as with an addiction, it can be very hard to sober up. Fortunately, nature does not enforce "cold turkey" on the chips industry. It started giving clear warning signals when Dennard scaling ended around 2004-2007 [18]. The cost of transistors started rising when the 28 nm technology node was introduced around 2015: a transistor at 7 nm is now 17% more expensive than its 28 nm counterpart. We had been used to transistor costs dropping, but since chip fabrication technology has become increasingly expensive, the cost has been rising noticeably [19]. Although researchers and chip manufacturers are still able to come up with clever techniques to keep Moore's law going for some years to come, we will "run out of money before we run out of physics", as Dan Reed, former vice president at Microsoft said not too long ago in 2016 [20].

It is therefore not surprising that the computing industry is looking for ways other than Moore's law to keep the speed of computations up. One way forward appears to be to tailor the hardware to the computational problems at hand, instead of building faster universal machines based on the von Neumann architecture which tackle all computational tasks with one architecture. It is important to keep in mind that these new hardware techniques are being developed to add to, rather than replace, this architecture which has been successfully in use for over 75 years.

It is not only speed that has increased sharply: in recent years: particularly in the last few years, the cost of energy has risen considerably. It was initially caused by the economy's surprisingly fast recovery from the downturn caused by the COVID-19 pandemic. Even more recently, the war in Ukraine has made the energy market unpredictable, and this has increased pressure for computer system architecture to be as energy efficient as possible. Again, one way to accomplish this is to tune the architecture to the specific task, rather than developing one architecture that handles all the tasks.

In this section of the HiPEAC Vision, "The race for new hardware" we focus on several hardware approaches that on the one hand tackle a specific class of computational problems, and on the other hand promise to make central processing units run faster. In this introductory chapter, we briefly introduce these techniques to help readers better understand the remaining chapters.

There are many new computing paradigms but, for the sake of space, we will focus on those that we think are currently the furthest developed and/or the most promising.

Key insights

- Although Moore's law appears to continue, it is evidently slowing down.
- Improving device speed is not the only way to improve the speed of computations: architectures tuned to solving specific problems will be successfully leveraged to accelerate computations.
- Quantum computing makes it possible to solve specific problems that are practically unsolvable using classical (von Neumann) computers.
- New architectures modelled on biological neural networks promise to reduce energy consumption for solving specific problems.
- Materials research is key to developing new and innovative computing devices. Current examples include research to develop practical quantum devices, and to develop new memory and (analogue) computing devices.
- Introducing new computing paradigms requires the development of new programming models in addition to those based on the von Neumann architecture.
- The development of quantum computing has given a boost to information and complexity theory. These fields will grow in importance as new paradigms develop, evolving from pure theoretical into a combined theoretical/experimental discipline.
- Several architectural challenges lie ahead, in both the development of architectures for new devices and in embedding these devices into existing ones.
- Integrated circuit (IC) fabrication processes require large amounts of energy and fresh water, and due to their toxic waste can put high pressure on the environment.

Key recommendations

- Establish and stimulate centres of competence that combine the strength of Europe's position in the fundamental fields of computing (information theory, complexity theory, computer architecture, algorithm design, programming paradigms) with European industrial initiatives. This will both attract top talent and develop a strong European industrial presence in the field of new computing devices.
- Strive to restore Europe's leading role in new technologies and paradigms, as it did some decades ago with the establishment of emerging device research in the International Technology Roadmap for Semiconductors (ITRS).
- Put emphasis on development that interlinks research to innovative, practical applications.
- Invest in designing and fabricating novel IC technologies for very high-volume applications that have a significantly lower environmental impact, and at the same time apply the lessons learned from those new fabrication processes to existing ones.

Introducing quantum computing

Quantum computing is a radically new way to consider information processing based on the unique properties of quantum physics: superposition and entanglement of states. The idea first proposed by Richard Feynman in 1982 would allow the solving of problems that are far too complex to be tackled by current computers. Since then, progress in physics and technology has gradually transformed what used to be considered a dream into a reality. Although we are still far from being able to design and fabricate an operational quantum computer, it is now becoming a very concrete possibility.

Research and development activities at both academic and industrial levels are now engaged in a race towards quantum computing. It is a race of multiple dimensions: a race among a number of candidate qubit technologies, a race for architecture and hardware options, a race for software and programming, a race between nations and a race to attract and retain the best talent.

Read more in the HiPEAC Vision article "The race towards operational quantum computers".

Neuromorphic computing

Our reflections on neuromorphic computing are based on "A Survey of Neuromorphic Computing and Neural Networks in Hardware" by Catherine Schuman et al [21], and the "2022 Roadmap on Neuromorphic Computing and Engineering" by Dennis V. Christensen et al [22].

When we compare the von Neumann architecture to the human brain, we notice differences not only in organizational structure, but also in processing capabilities and power requirements. One thing that stands out is the way data is moved in the von Neumann architecture: processing and data storage are separated, requiring data to be moved from memory to the processor for processing, and then moved back again to memory for storage. The bandwidth of the channel between memory and processor is one of the determining factors in the overall performance of the system. Another notable difference is the intrinsic capability of the human brain to learn, which is lacking in von Neumann-based computer architectures. Not unexpectedly, researchers have therefore turned to biological systems for inspiration for new approaches to data processing. One field that sprang from this is neuromorphic computing, which takes its inspiration from the structure of biological neural networks.

The term neuromorphic computing was coined by Carver Mead in 1990. Mead was referring to very large-scale integration (VLSI) designs consisting of analogue components that were more or less a copy of biological neural systems. Nowadays it means biologically *inspired* neural networks, implemented using either analogue or digital circuits. These networks are notable for being highly connected and parallel, co-locating computing and storage, and for their low energy consumption. Neuromorphic architectures can solve specific types of problems faster, using less chip real estate, and less power. More recently, two directions in neuromorphic computing have begun to emerge: spike-based processing systems and dedicated neural electronic architectures that form neuron and synaptic circuits, called artificial neural networks.

Operations are orchestrated by a clock in von Neumann architectures[1]. Neuromorphic systems on the other hand are driven not only by time but also by data and events. That holds in particular for spike-based neural networks, while artificial neural networks are predominantly time-driven. A strong cross fertilization between these two prevailing architectures exists.

The field of neuromorphic computing is highly interdisciplinary. Disciplines such as materials research play an important role in the search for efficient implementations of neuromorphic circuits, both analogue and digital. Materials research is a key shared interest of the fields of neuromorphic computing and quantum computing. There are indeed more similarities between these research fields. Embedding new accelerator devices in existing von Neumann architectures requires careful architectural design to reap the benefits of combining them. To enable efficient programming of these embedded neuromorphic compute devices, new programming models need to be developed. What is more, although maybe less prominent than in quantum computing, neuromorphic networks have renewed interest in both complexity theory and algorithm design.

Intel and IBM have developed chips based on neuromorphic architectures:

1 There are however some hardware architectures based on self-timed circuits that do not require a clock signal. Such architectures are not widely in use.

"loihi" in the case of Intel and "TrueNorth" at IBM. There is also a number of start-ups in the field: brainschip [23], GrAI Matter Labs [24], Innatera [25], iniLabs [26], INSTAR Robotics [27], Intrinsic Semiconductor Technologies [28], Koniku [29] and MemComputing [30].

Forecasts for the neuromorphic computing market vary from $12 billion in 2027 to $32 billion in 2035, rising from $200 million in 2025, provided that fundamental technical issues can be solved. Even conservative market research forecasts a market worth at least $8.6 billion. It would mean a growth in 2035 of close to two orders of magnitude in 10 years. [32,33,34]

In summary, neuromorphic computing is an interesting field with great technical potential, yet many technical challenges to be solved. It is a field that we think will play an increasingly important role in information processing in the years to come. We will revisit neuromorphic computing in future updates of the HiPEAC Vision.

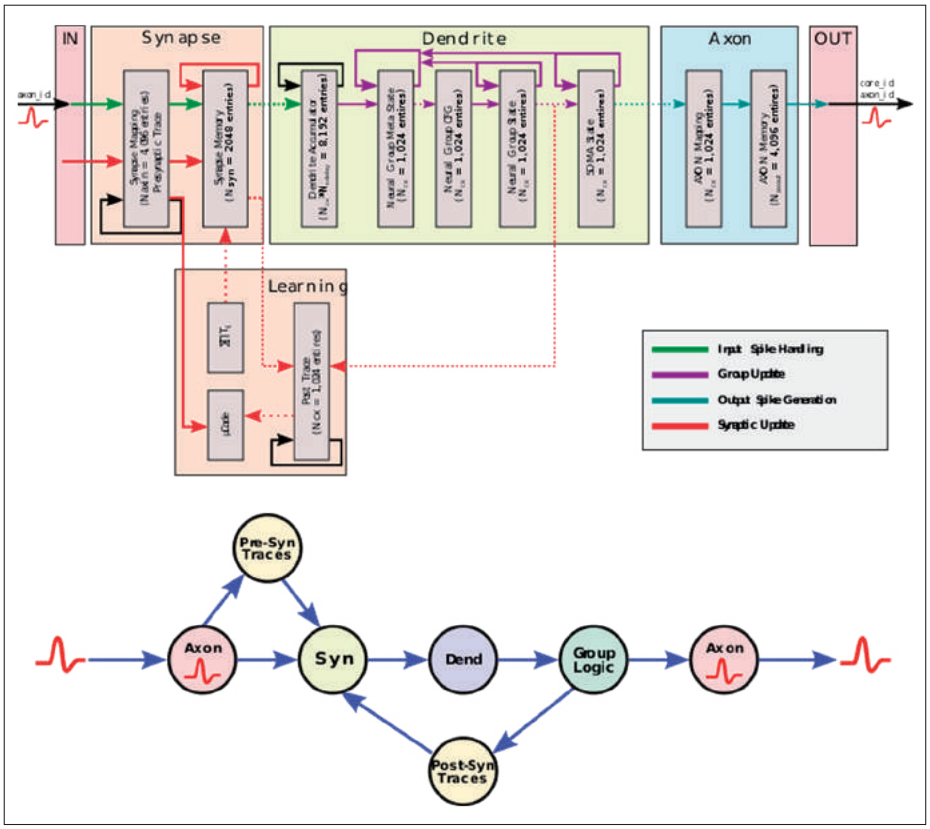


Figure 1. Block diagram of the major memory blocks of Intel's Loihi chip, where the various connectivities, configurations and dynamic states of all the neurons that are mapped to the neuromorphic core. Each core incorporates a total of 2 Mib (including ECC). From [31].

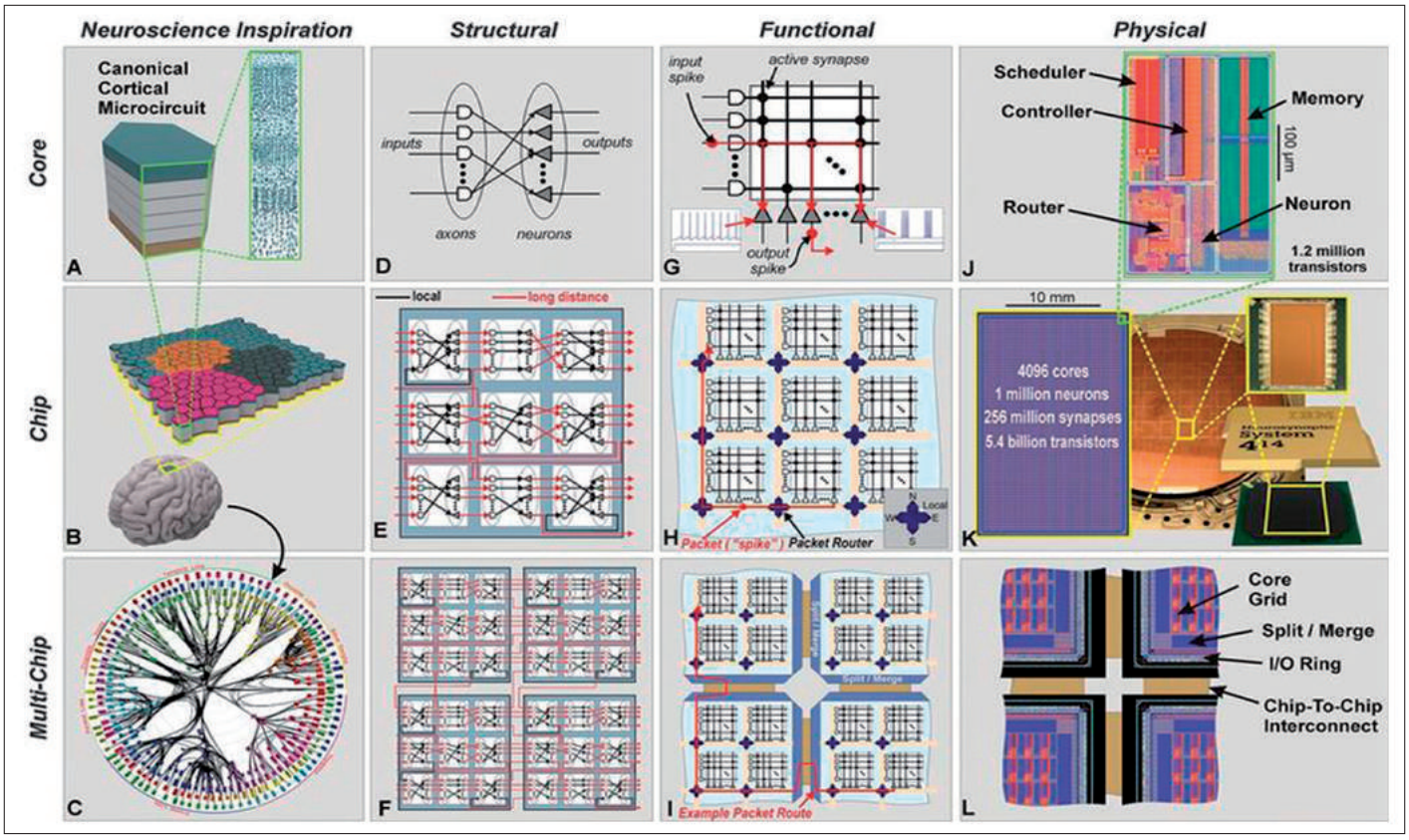


Figure 2. TrueNorth architecture. Panels are organized into rows at three different scales (core, chip, and multichip) and into columns at four different views (neuroscience inspiration, structural, functional, and physical).
(A) The neurosynaptic core is loosely inspired by the idea of a canonical cortical microcircuit. (B) A network of neurosynaptic cores is inspired by the cortex's two-dimensional sheet. (C) The multichip network is inspired by the long-range connections between cortical regions shown from the macaque brain. Merolla et al. [35]

Neuromorphic computing, neurons and synapses

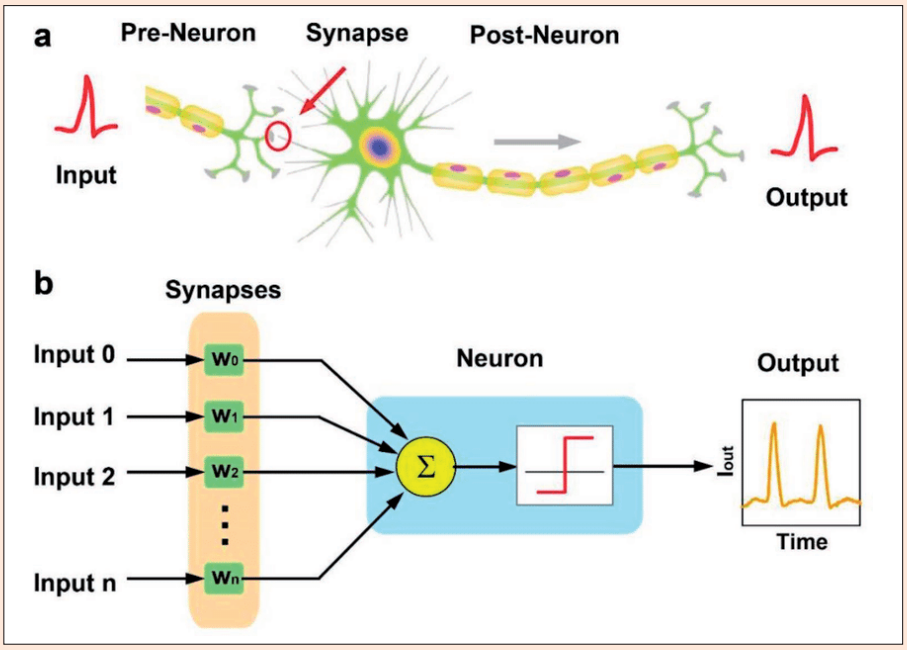
Neuromorphic devices are modelled on biological neurons, made up of neural circuits.

Nervous tissue in almost all animals consists of electrically excitable cells, called neurons, that communicate with other neurons through synapses. Many interconnected neurons together form a neural network circuit. Typical neurons consist of a cell body, the soma, and two types of filaments that protrude from the soma: dendrites and an axon. Dendrites are strongly branched, and end very close to the soma, usually only a few micrometres away. The axon, on the other hand, can be very long, of the order of one metre, or sometimes even longer. Axons are less branched, and end in synapses. The dendrites serve as the receivers of signals, the axon functions as the transmitter. The end of an axon consists of a synapse, which transmits the signal to the dendrite of another neuron, but axon-axon synapses and dendrite-dendrite synapses are also found [43].

Neurons process information through electrical and chemical processes. Neurons are electrically excited when the voltage across their membrane changes and passes a particular threshold within a certain time interval. When that happens, an all-or-nothing pulse, called an action potential is sent down the axon, activating the synaptic connections of the axon. Depending on the synapse, some increase the net voltage that reaches the next neuron, others decrease the net voltage.

Neuromorphic computing devices mimic the neurons and the connections between neurons: the synapses. The soma, axons and dendrites are abstracted away, but not the interaction between neurons through synapses, which differentiate neuromorphic devices from classical digital circuits.

Several main streams can be distinguished in neuromorphic computing: in analogue neuromorphic devices, the characteristics of specific materials are exploited to simulate the behaviour of synapses and the reaction of neurons to the synaptic stimuli. In digital neuromorphic devices, this behaviour is implemented using digital circuits. Of course, there are also mixed analogue/digital neuromorphic systems. Several packages exist that exhibit the behaviour of neuromorphic systems implemented in software. [44]



Schematic illustration of neuromorphic system. a) Biological model: the biological neuron receives inputs from other neurons by interconnected synapses; b) Equivalent electronic model: the electronic neuron for accumulating inputs generated by different pre-neurons through resistive switching memristor (RS) synapses to implement the functions of spiking neural network [45].

Introducing spintronics

Spintronics is a marriage between electronics and magnetism that exploits the spin of electrons, in addition to their charge, to reveal new phenomena. These are then exploited in innovative components and circuits with improved performance, particularly in terms of power consumption, or offering new functions.

Until now, the main application of spintronics in computer design has been in the field of storage and memory. A number of technologies have been developed for the various needs of the computer memory hierarchy (Figure 3), technologies that are always seeking the "ideal" memory device: fast to read and write, providing high-densities, economically viable and non-volatile.

A new type of magnetic memory based on magnetic tunnel junctions, magnetic random access memory (MRAMs), has entered industrial production at all the major microelectronics companies. These memories combine the advantages of non-volatility (ability to keep information without being powered), writing speed (between 0.5ns and 50ns depending on the version), endurance (10^6 to 10^{15} write cycles depending on the version), low power consumption (~100fJ per write), and integration density. The entry into industrial production of these memories is a very important step because it marks the adoption of this hybrid complementary metal–oxide–semiconductor (CMOS)/magnetic technology by the industry. The first industrial uses of MRAMs concern the replacement of embedded flash memories. Another major application is the replacement of static RAMs (SRAMs) used as cache memories in processors. At advanced technology nodes (sub-20nm), MRAMs are at least a factor of 4 smaller than SRAMs. Replacing SRAMs with MRAMs would increase the capacity of cache memories and thus increase the speed of processors while reducing their power consumption.

To reduce static power consumption, the concept of "normally-off/instantly-on computing" has been proposed. It consists of finely distributing MRAMs among the logic blocks so that the results of logic operations can be saved almost continuously.

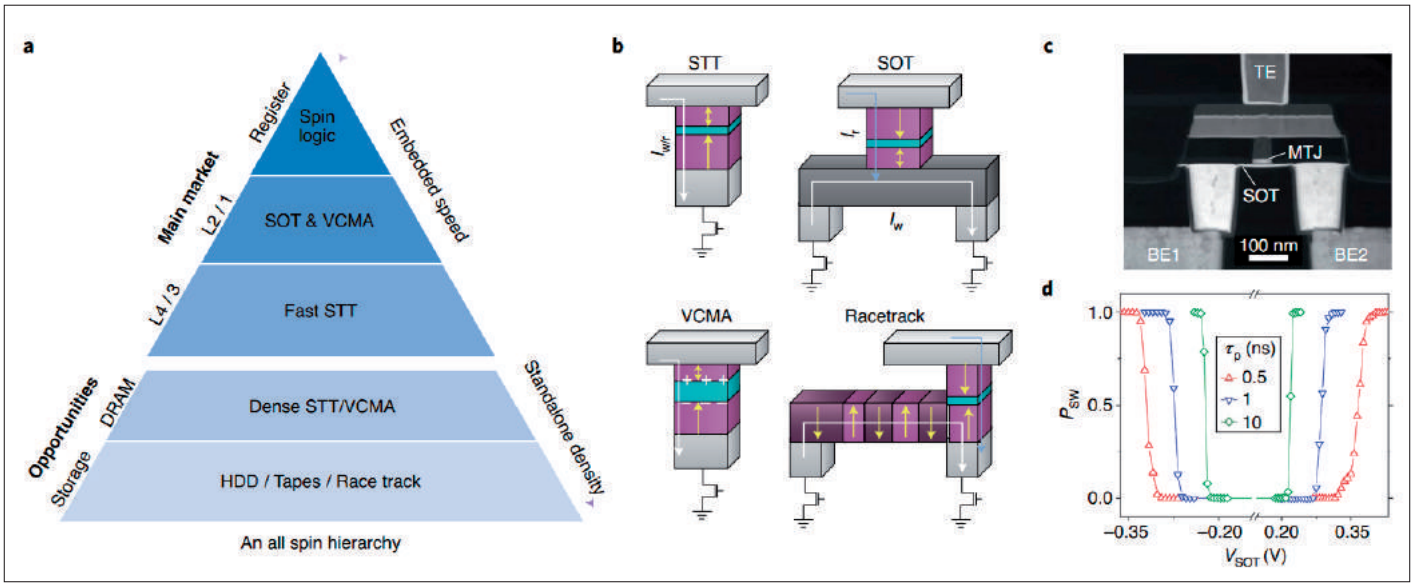


Figure 3. Non-volatile magnetic memories and their position in the computing memory hierarchy (from [16]).

This makes it possible to cut the power supply to the logic blocks and memories as soon as they are temporarily not in use and to reactivate them almost instantaneously, thus reducing the power consumption associated with leakage currents. Some architectures also allow logic operations to be performed within memory blocks (logic-in memory). All this reduces data transfers along the interconnections and thus dynamic power consumption.

Other approaches are also being studied for application in computing with much more innovative concepts:

- Magnetic materials whose properties are controlled by an electric field
- 2D spintronic materials
- Spintronic neuromorphic circuits
- Logic using electron spin directly (logic operations by spin wave manipulation. See Figure 2)
- Non-volatile logic gates combining magnetoelectric materials and spin current/charge current interconversion by spin-orbit coupling, etc.

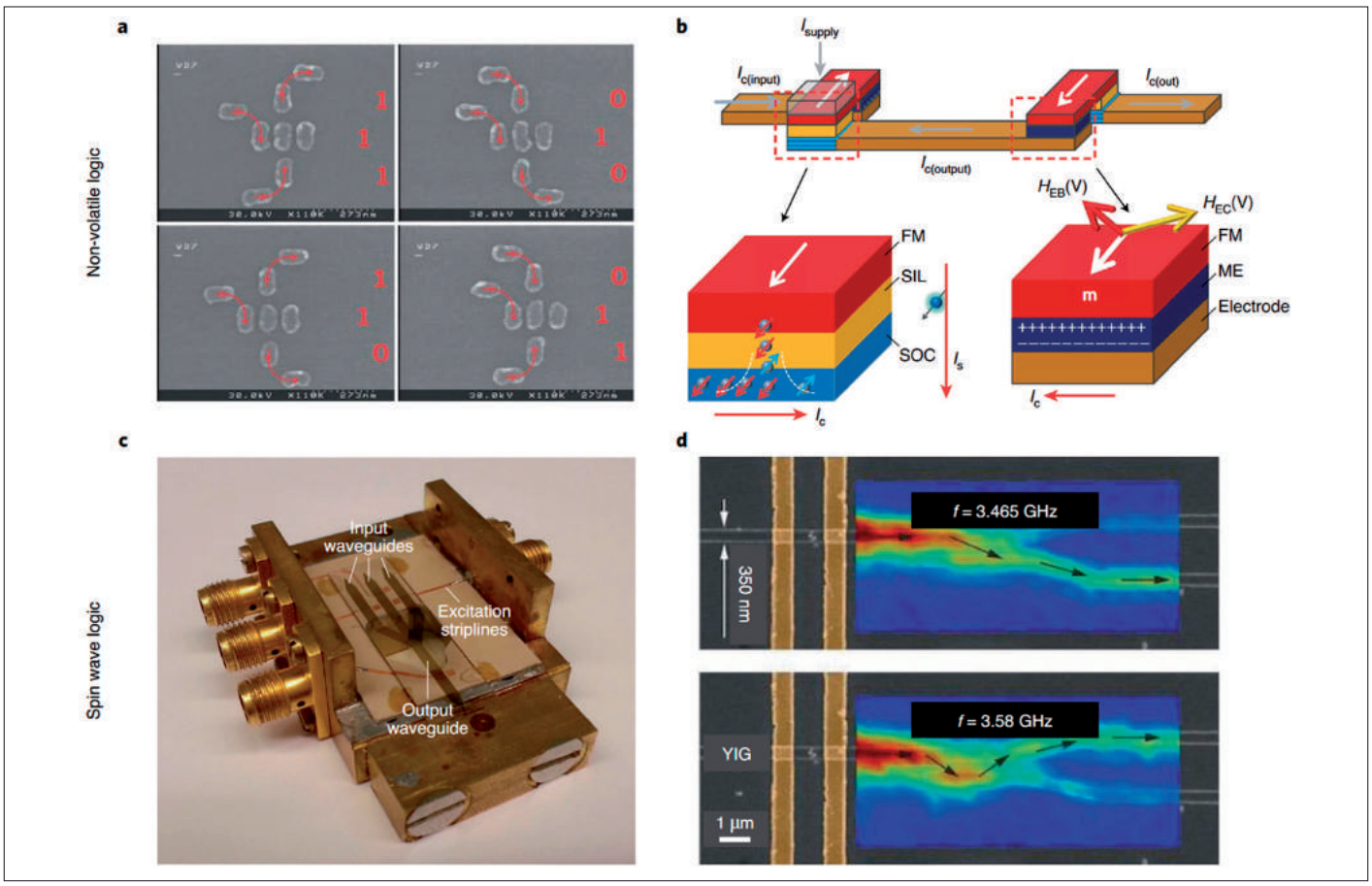


Figure 4. Spintronic approaches for logic: non-volatile logic (top), spin-wave logic (bottom) (from [16]).

All of those very innovative approaches to computing using spintronics have the potential to lead to much more energy efficient circuits and architectures [16].

Photonics computing

Using photons for computing is not a new thing: they are already at work in high bandwidth communication links and photonics technologies have been critical in the growth of the internet. Attempts have been made to use integrated photonics technologies such as semiconductor laser diodes, photodiodes and light guides to implement fast data communication inside a computer cabinet, between PC board or even on chips. However, photonics technology has faced major engineering issues including power, photonic/electronic conversion, and co-integration with CMOS. These have hindered its adoption in core computing functions.

Photonics computing can also benefit from the steady progress made in basic optical components: integrated lights sources such light emitting diodes (LEDs) and coherent light sources (lasers) of various wavelength and power, digital micromirror devices (DMD), display technologies like liquid crystal display (LCD) or organic LED (OLED) and integrated camera technologies, high-speed photodiodes arrays and time of flight sensors (TOF). By way of example of this trend, a small company, Lighton [36], designed a photonic system for speeding up matrix-vector products [37] using laser light and off-the-shelf cameras and DMD. The optical processing unit (OPU) [38] is designed to replace power-hungry graphics processing units (GPU) and is to be integrated in data centres.

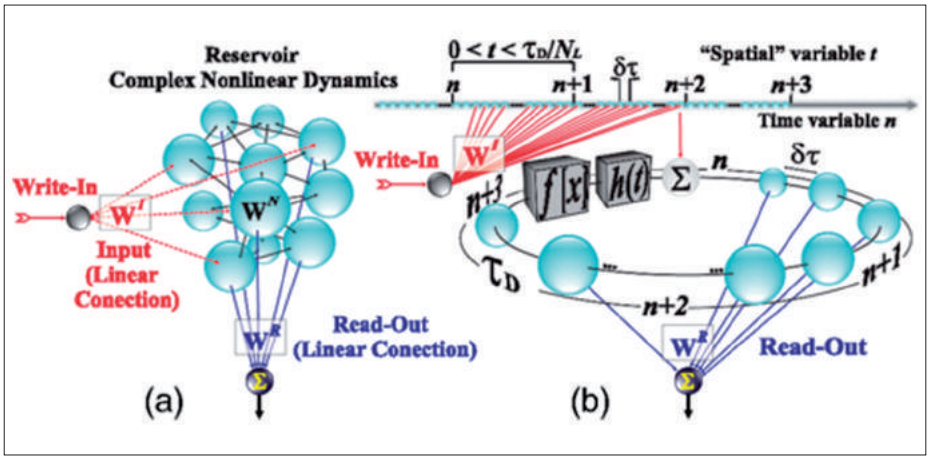


Figure 5. Concept of a fully photonics analogue reservoir computer (from [17]).

The concept of reservoir computing has also been demonstrated using photonics hardware for a dynamical system, which opens up the path to ultrafast brain-inspired computing. One implementation involves an electro-optic phase-delay dynamics designed with off-the-shelf opto-electronic telecom devices, thus providing the required wide bandwidth. The efficiency of the implementation has been demonstrated experimentally on speech-recognition tasks [39].

Quantum computing is the field in which photonics technologies are progressing the most. This is not a surprise since photons are probably the best example of quantum objects and were instrumental

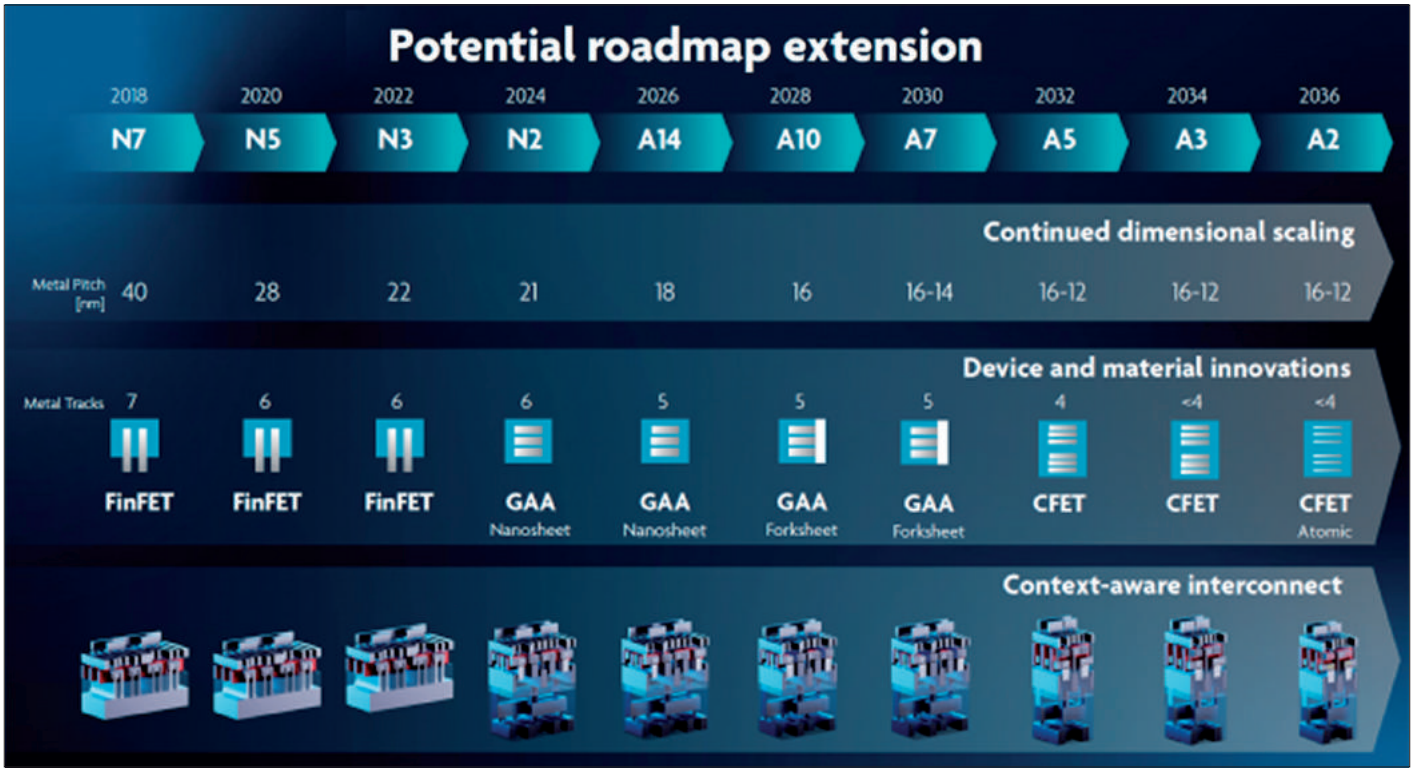


Figure 6. A view on the potential roadmap for continued dimensional scaling, new transistor architectures, new materials introduction, combined with innovative interconnect architectures. (from IMEC "20-year semiconductor roadmap" [49])

in demonstrating the properties of quantum physics [40]. However, implementing photonic quantum computing gets very difficult in practical terms because of the engineering complexity of photon sources, channels, operators and detectors. Things might change thanks to recent advances in integrated photonics. The evidence of strong coupling between the electron spin in a silicon qubit and a microwave photon in a resonator [41], shows that photonics could be used to coherently couple qubits on a silicon substrate. Photonics could thus help solve the challenge of scaling-up the number of silicon-based spin qubits and open up new architecture and design opportunities. This is a further example of heterogeneous technology integration.

Silicon photonics has made tremendous progress in recent decades. Although not yet ready for general use in computing systems (for core computing tasks), it will be a key technology for the success of novel computing paradigms currently in development, in particular in high throughput machine learning tasks for artificial intelligence and in quantum computing engineering.

Progress in semiconductor technology

The mainstream evolution of digital and memory devices is very much following the roadmap and announcements on the topic that have been seen over the last few years. Look for example at the memory technology trends exposed in [48]. Last generation FinFET at 2nm is around the corner at TSMC and, with a likely lag of about one year, at Intel. First generation gate all around (GAA, also called nanosheets or nanoribbons) are due to come in 3nm node from Samsung. Micron has started shipping their 1α technology [47], similar to 14nm logic pitches, breaking the barrier of 0.3 Gb/mm2, at 0.315 Gb/mm2.

This technology trend is likely to maintain its pace for the next five years until it reaches the physical limits of the GAA structures and the fabrication technology limits related to the extreme ultraviolet (EUV) lithography and deposition tools of the current generation.

Significant activity is underway on both fronts, looking into new ways of arranging

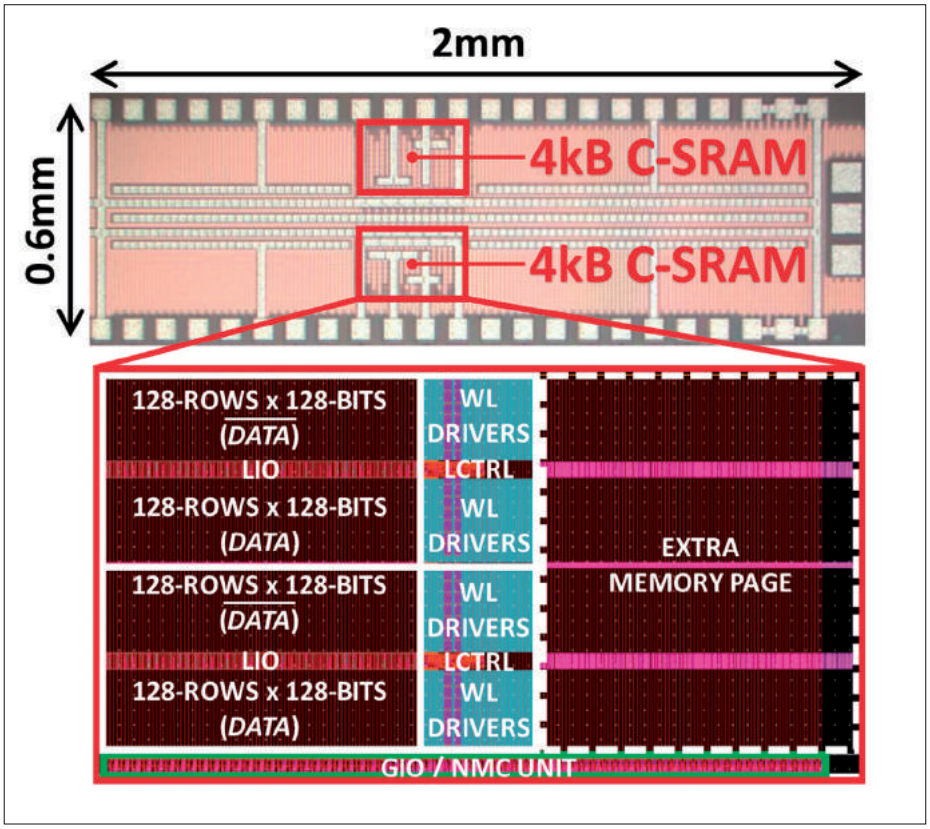


Figure 7. in-memory computing computational RAM proof of concept: Die photograph and layout view of IMC-based C-SRAM test chip (image from [46])

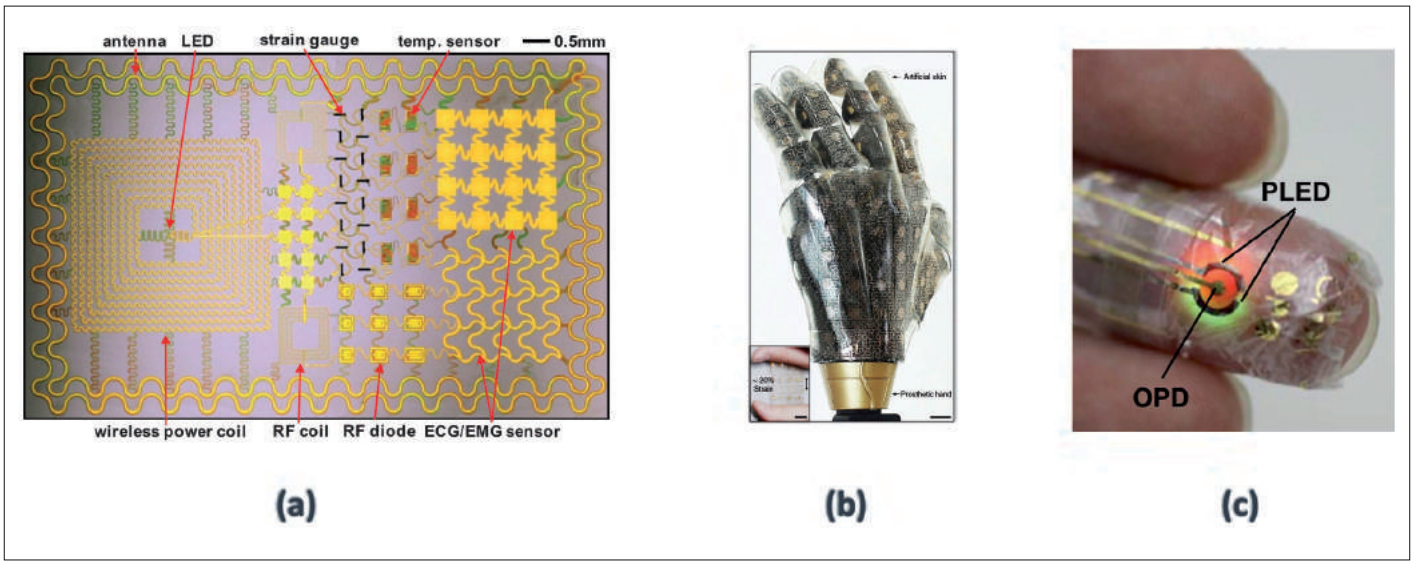


Figure 8. a) E-skin (Kim et al [6]), b) Prosthetic skin with multiple sensors (Kim et al [7]) and c) E-skin pulse oximeter using OLEDs and OPDs (organic photodiodes) (Yokota et al [8])

the devices with increasing use of stacking and in developing a higher numerical aperture EUV system and other processing tools.

It seems very likely that some of these technologies will lead to a trend of about 25% improvement in density and power for the next three or four nodes. Clearly the naming of the nodes will become less and less related to the real physical dimension of the devices.

Stacking will be the trend not only at device level: it will continue to extend to assembly of multiple chips in packaging to get specialized functions closer together in order to increase speed and reduce power consumption. Before reaching true IMC, the trend demonstrated thus far by the Intel or Apple chips will continue with increasing packaging density and a larger application of the "chiplets" concept.

Buried interconnects, monolithic 3D and "in-memory computing"

With the emergence of so-called hyper dimensional computing applications (e.g. image recognition by learning, post-quantum cryptography, etc.), that is to say, applications based on data sets that can be represented by large vectors (ranging from 128 bits to several thousand bits), current computing architectures are no longer suitable for processing them efficiently. Indeed, these applications require the processing of large amounts of data with very low computational latency to ensure an acceptable execution time. However, more than 90% of the energy of a calculation instruction can be dedicated solely to the transfer of data between the computing and the storage units.

To overcome this problem, a new concept of computing has emerged in recent years, namely in- or near-memory computing (IMC). This concept, based on the execution of relatively simple operation instructions (such as addition or multiplication of integers), uses the intrinsic properties of a certain type of memory to carry them out using much less energy than a classical computing architecture, such as the von Neumann one. This is made possible by fundamentally changing the computing paradigm by bringing the instruction and computation to the place where the data to be processed is located, rather than the other way around. For the reduction in energy and computational latency to be truly significant at the system level, the microprogramming and software compilation chain must also be optimized accordingly.

For example, an SRAM-based computing technology called computational SRAM has been proposed [42]. This technology has the advantage of being able to rely on a perfectly adapted and customizable software compilation chain developed in collaboration with the software and hardware teams. In addition, this technology is compatible with standard scalar processors (e.g. RISC-V) and other types of data memory present in the system and helps enhance computing performance and energy efficiency of computing systems.

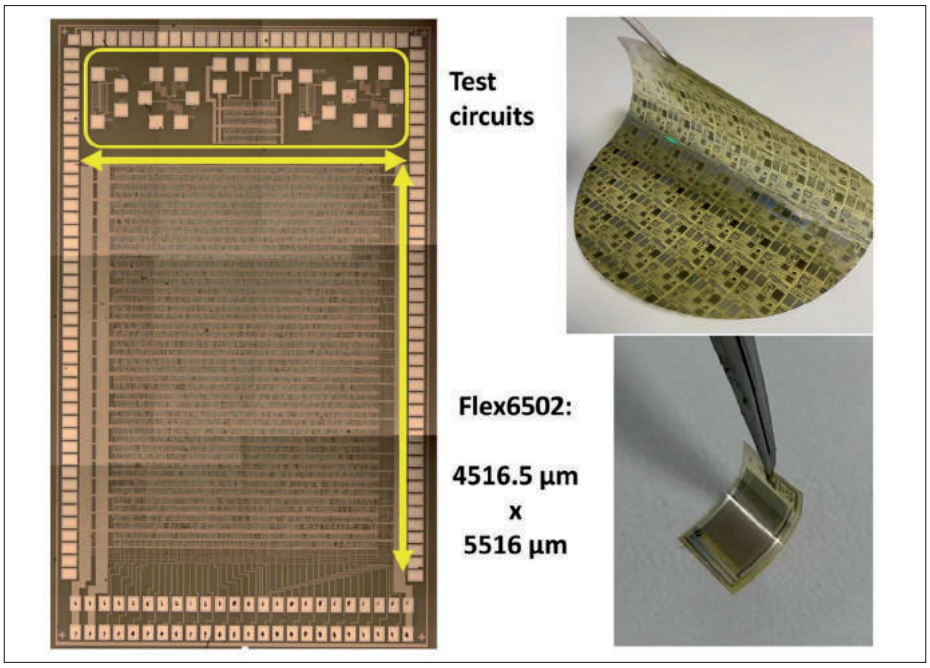


Figure 9. Flex6502: an 8-bit 6502 CPU fabricated using PragmatIC's 0.8 µm IGZO TFTs [27]

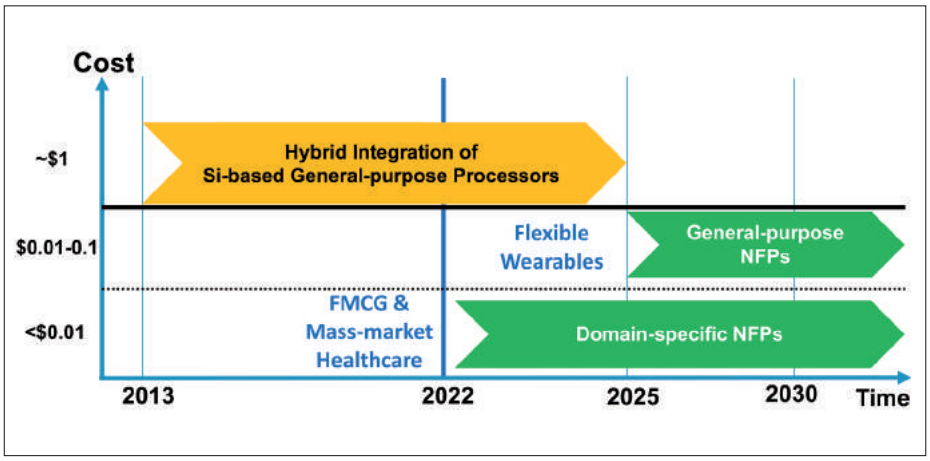


Figure 10. NFP roadmap projection

Flexible electronics: hardware technologies for low cost and sustainable production processes

Some applications require low to extreme low-cost hardware, coupled with other requirements, such as being apt for environmentally-friendly disposal. Examples of this include smart packaging, healthcare devices and wearable electronics. The technology used for such devices is called flexible electronics, and is also known as printed electronics and thin film electronics.

Flexible electronics are thin, lightweight, flexible (as the name implies) and cheap compared to silicon-based elec-

tronics, mainly because these devices are manufactured using low-cost processes, due to lower processing temperatures and a smaller number of processing steps, and thus, masks. The materials used can be insulators, conductors and semiconductor materials, such as organic, metal oxides, amorphous silicon or 2D materials. The devices are fabricated on substrates of plastic, steel foil or paper.

Flexible electronics are applied in OLEDs, flexible displays and organic photovoltaic devices. They are increasingly finding application in smart packaging: electronics embedded into products such as fast-moving consumer goods (FMCG), or in smart blister packing for pharmaceutical products. Flexible electronics devices can be used for product tracking in logistics, for environmental tracking of products (e.g. products that must be kept at a particular temperature), and for interacting with customers using illumination and sound (and even using aromas).

In healthcare, flexible electronics are used for smart systems, and are flexible, stretchable devices attached directly to the skin. The device can be used to monitor physiological activity (electrocardiogram (ECG), electroencephalogram (EEG), electromyography (EMG)), to monitor and treat wounds [2] [3] [4], and for man-machines interfaces. These devices are fully integrated, consisting of electrodes or sensors, communications, and batteries or energy harvesters.

Computing with flexible electronics: natively flexible processors

The first natively flexible processors (NFPs) were implemented with flexible electronics with thin film transistor-based (TFT) devices[18, 19]. Although already lower in cost then silicon-based processors, the cost of these devices is still too high for FMCG applications with a shelf time of only a few days. About ten years ago, the first organic TFT-based processor was developed, using 5 µm technology [6], first in unipolar p-type technology, and soon after also in complementary technology [7][8]. This 8-bit CPU consisted of 3200 TFT's and operated at 2.1 kHz. In 2019, a RISC-V processor was manufactured using carbon nanotube-based field-effect transistors (FETs), running at 10 kHz [9]. In 2021, a 32-bit ARM Cortex-M with memory and peripherals was produced using indium-gallium-zinc-oxide (IGZO) TFTs. This processor runs at 29 kHz and consists of 73,300 transistors [12], which is the most complex NFP built so far. And even a good old 6502, a microprocessor design dating back from the 1970's, was implemented using IGZO TFTs in 2022, running at 71 kHz and consuming 132 mW [13].

Machine learning hardware and a neural network, consisting of 4100 and 5700 TFTs and running at 104 and 22 kHz respectively, have also been recently developed [10][11].

At the moment, the technology of flexible electronics is transitioning from a phase where conventional Si-based processors are integrated with flexible electronics based peripheral devices to a phase where NFPs have sufficiently evolved to replace Si-based processors (Figure 10). Such NFPs will find application in wearables, but less so in FMCG with a very short shelf life. Wearables have longer lifetimes, and might thus require occasional software updates. Application-specific devices for consumer products can already be manufactured using flexible electronics due to their lower complexity. By the end of the decade, it is expected that the cost of NFPs will have dropped sufficiently to make them feasible for use in FMCG.

Sustainable fabrication processes

State-of-the-art chip manufacturing processes put a lot of strain on the environment, in terms of energy consumption (maintaining clean-room conditions, lithographic equipment, ovens), toxic chemicals, and use of vast quantities of clean water. One way to reduce toxic waste is to recycle some of the materials or find less toxic replacements. Semi, the semiconductor association, has launched a consortium to focus on (the challenges of) sustainable chip making [15].

Experimenting with chip processes can be very expensive, as upsetting a process too much can wreak havoc on the yield. Moreover, because of the many processing steps required, it can take weeks before the effects can be measured.

Flexible electronics are made with processes that have far fewer steps. Results of experiments on a chip manufacturing process may take only a few days to surface. This makes the manufacturing of flexible electronics a more viable testbed for developing and testing more sustainable chip manufacturing processes [14].

In summary, although flexible electronics are not a field that will take computing to the next speed level, it is a viable technology for producing extremely cheap small processing systems with a significantly smaller environmental footprint.

Acknowledgements

- Sebastian Feld, assistant professor at Delft University of Technology
- Laurent Larger, Director, FEMTO-ST Institute, France
- Menno Lindwer, VP IP & Silicon at GrAI Matter Labs, Eindhoven
- Ioan Lucian Prejbeanu, executive director of SPINTEC research laboratory, France
- Carlo Reita, Director, Strategic Partnerships and Planning, CEA, LETI, France
- Sanaz Mahmoodi Takaghaj, assistant teaching professor, Pennsylvania State University School of Engineering Design and Innovation
- Emre Ozer, Senior Director of Processor Development, PragmatIC Semiconductor
- Yulia Sandamirskaya, research scientist at Intel Labs in Munich

References

[1] Organic and Printed Electronics Association, www.oe-a.com, 2022.

[2] D.-H. Kim, et al, "Epidermal Electronics", Science, 333, 838–843, 2011.

[3] J. Kim, et al, "Stretchable silicon nanoribbon electronics for skin prosthesis", Nat Commun, 5, 5747, 2014.

[4] T. Yokota, P. Zalar, M. Kaltenbrunner, H. Jinno, N. Matsuhisa, H. Kitanosako, Y. Tachibana, W. Yukita, M. Koizumi, and T. Someya, "Ultraflexible organic photonic skin", Sci. Adv., 2, Apr. 2016.

[5] G. Fortunato, A. Pecora and L. Maiolo, "Polysilicon Thin-film Transistors on Polymer Substrates", Material Science in Semiconductor Processing, 627-641, 2012.

[6] H. Klauk, "Organic Thin-film Transistors", Chemical Society Reviews, 39, 2643-2666, 2010.

[7] K. Myny, S. Smout, M. Rockelé, A. Bhoolokam, T. H. Ke, S. Steudel, K. Obata, M. Marinkovic, D-Vu Pham, A. Hoppe, A. Gulati, F. Gonzalez Rodriguez, B. Cobb, G. H. Gelinck, J. Genoe, W. Dehaene and P. Heremans, "8b Thin-film Microprocessor Using a Hybrid Oxide-Organic Complementary Technology with Inkjet-Printed P2ROM Memory", IEEE International Solid-State Circuits Conference (ISSCC), 2014.

[8] K. Myny, S. Smout, M. Rockelé, A. Bhoolokam, T. H. Ke, S. Steudel, B. Cobb, A. Gulati, F. G. Rodriguez, K. Obata, M. Marinkovic, D-Vu Pham, A. Hoppe, G. H. Gelinck, J. Genoe, W. Dehaene and P. Heremans, "A Thin-film Microprocessor with Inkjet Print-programmable Memory", Scientific Reports, 4 Article Number:7398, Dec. 2014.

[9] G. Hills, et al, "Modern microprocessor built from complementary carbon nanotube transistors", Nature, 572, 595–602, 2019.

[10] E. Ozer, J. Kufel, J. Myers, J. Biggs, G. Brown, A. Rana, A. Sou and C. Ramsdale "A hardwired machine learning processing engine fabricated with submicron metal-oxide thin-film transistors on a flexible substrate", Nature Electronics, 3, 419-425, July 2020.

[11] E. Ozer, J. Kufel, J. Biggs, J. Myers, C. Reynolds, G. Brown, A. Rana, A. Sou and C. Ramsdale, "Binary Neural Network as a Flexible Integrated Circuit for Odour Classification", IEEE International Conference on Flexible and Printable Sensors and Systems (FLEPS'20), Aug. 2020.

[12] J. Biggs, J. Myers, J. Kufel, E. Ozer, S. Craske, A. Sou, C. Ramsdale, K. Williamson, R. Price and S. White, "A natively flexible 32-bit Arm microprocessor", Nature, 595, 532–536, July 2021.

[13] H. Çeliker, A. Sou, B. Cobb, W. Dehaene and K. Myny, "Flex6502: A Flexible 8b Microprocessor in 0.8μm Metal-Oxide Thin-Film Transistor Technology Implemented with a Complete Digital Design Flow Running Complex Assembly Code," 2022 IEEE International Solid- State Circuits Conference (ISSCC), 272-274, 2022.

[14] https://www.eenewseurope.com/en/making-semiconductor-fabs-more-sustainable/

[15] https://www.eenewseurope.com/en/making-semiconductor-fabs-more-sustainable/

[16] B. Dieny et al., « Opportunities and challenges for spintronics in the microelectronics industry », Nature Electronics, vol. 3, no 8, p. 446 459, août 2020, doi: 10.1038/s41928-020-0461-5.

[17] L. Larger, A. Baylón-Fuentes, R. Martinenghi, V. S. Udaltsov, Y. K. Chembo, et M. Jacquot, « High-Speed Photonic Reservoir Computing Using a Time-Delay-Based Architecture: Million Words per Second Classification », Phys. Rev. X, vol. 7, no 1, p. 011015, févr. 2017, doi: 10.1103/PhysRevX.7.011015.

[18] https://en.wikipedia.org/wiki/Dennard_scaling

[19] https://www.fabricatedknowledge.com/p/the-rising-tide-of-semiconductor

[20] https://thenewstack.io/farewell-moores-law/

[21] https://arxiv.org/abs/1705.06963

[22] https://iopscience.iop.org/article/10.1088/2634-4386/ac4a83

[23] https://brainchip.com/

[24] https://www.graimatterlabs.ai/

[25] https://www.innatera.com/

[26] https://inilabs.com/

[27] https://instar-robotics.com/

[28] https://www.intrinsicsemi.com/

[29] https://koniku.com/

[30] https://www.memcpu.com/

[31] https://en.wikichip.org/wiki/intel/loihi

[32] "Neuromorphic Computing Market", https://www.alliedmarketresearch.com/neuromorphic-computing-market-A13743],

[33] https://www.globenewswire.com/news-release/2022/07/26/2485578/0/en/Neuromorphic-Computing-Market-Size-Share-Competitive-Analysis-Growth-Status-Top-Manufactures-and-Is-Projected-To-Reach-USD-34-61-Billion-by-2029.html

[34] https://www.verifiedmarketresearch.com/product/neuromorphic-computing-market/

[35] https://www.greencarcongress.com/2014/08/20140808-truenorth.html

[36] http://www.lighton.io/

[37] http://arxiv.org/abs/1510.06664

[38] http://www.mwee.com/news/startup-integrates-optical-processing-data-center-ai-operations

[39] https://journals.aps.org/prx/abstract/10.1103/PhysRevX.7.011015

[40] A. Aspect, J. Dalibard, et G. Roger, « Experimental Test of Bell's Inequalities Using Time-Varying Analyzers », Phys. Rev. Lett., vol. 49, no 25, p. 1804 1807, déc. 1982, doi: 10.1103/PhysRevLett.49.1804.

[41] N. Samkharadze et al., « Strong spin-photon coupling in silicon », Science, p. eaar4054, janv. 2018, doi: 10.1126/science.aar4054.

[42] J.-P. Noel et al., « Computational SRAM Design Automation using Pushed-Rule Bitcells for Energy-Efficient Vector Processing », in 2020 Design, Automation & Test in Europe Conference & Exhibition (DATE), mars 2020, p. 1187 1192. doi: 10.23919/DATE48585.2020.9116506.

[43] https://en.wikipedia.org/wiki/Neuron

[44] https://en.wikipedia.org/wiki/Neuromorphic_engineering

[45] https://www.researchgate.net/figure/Schematic-illustration-of-neuromorphic-system-a-Biological-model-the-biological-neuron_fig1_335021794

[46] M. Kooli et al., « Towards a Truly Integrated Vector Processing Unit for Memory-bound Applications Based on a Cost-competitive Computational SRAM Design Solution », J. Emerg. Technol. Comput. Syst., vol. 18, no 2, p. 40:1-40:26, avr. 2022, doi: 10.1145/3485823.

[47] https://www.micron.com/about/blog/2021/january/inside-1a-the-worlds-most-advanced-dram-process-technology

[48] J. Choe, « Memory Technology 2021: Trends & Challenges », in 2021 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD), sept. 2021, p. 111 115. doi: 10.1109/SISPAD54002.2021.9592547.

[49] https://www.imec-int.com/en/articles/20-year-roadmap-tearing-down-walls

Christian Gamrat is a researcher in the Research and Technology Department at CEA (Alternative energies and Atomic Energy Commission), France.

Harm Munk is a senior systems and software architect and developer at TNO (QuTech).

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2023.

Cite as: C. Gamrat and H. Munk. New and emerging technologies and paradigms. In M. Duranton et al., editors, HiPEAC Vision 2023, pages 100-109, Jan 2023.

DOI: 10.5281/zenodo.7461887

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

© HiPEAC 2023

The race towards an operational quantum computer is on. It is a race of multiple dimensions: a race among a number of candidate qubits technologies, a race for architecture and hardware options, a race for software and programming, a race among nations and a race to attract and keep the best talent.

The race towards operational quantum computers

By CHRISTIAN GAMRAT and KOEN BERTELS

Although no one can yet forecast if or when quantum computing (QC) will become an operational technology, research and development efforts to that end will nevertheless have a profound impact on the future of computing: they are accelerating progress in important areas such as hybrid architectures, and algorithmic and computational complexity, as well as forcing us to re-think the very nature of computing. In principle, the realization of machines applying quantum principles could make it possible to tackle problems that are difficult or even impossible to solve with conventional computers. The application fields that would benefit most from quantum computing are those with problems incurring an exponentially large number of variables in relation to the size of the problem such as can be found in chemistry, pharmacology, physics, cryptography, optimization and machine learning.

Key insights

- The **race for qubit** technology rages on. Europe has pioneered several technological paths: superconducting qubits, trapped ion qubits and neutral atoms, and is very active in the development of those technologies.
- The **race for quantum hardware** is currently led by players in the United States but Europe is narrowing the gap thanks to a number of innovative solutions developed by start-up companies.
- In the **race for quantum software**, most available quantum programming frameworks originate from major US companies. European players are entering the race and are already proposing approaches for experimenting with quantum programming. In the **race between nations**, Europe as a whole (European Commission and member states) is very much in the race when levels of public funding are taken into account, but probably needs greater coordination.
- Education and training are critical for the European quantum computing ecosystem to be able to get up to speed. Europe is also entering a **race to train, attract and retain quantum talent**.
- The quantum computer will be a hybrid quantum/classical computer. A good **synergy between quan-tum and computer science communities** is required at both the hardware and software levels.

Key recommendations

- Support and advance research in the **system architecture and software stack** for quantum computing. With European assets in system, software and HPC technologies, this could be a significant opportunity for Europe.
- Invest in the development of **design tools and libraries** for quantum chip design and in their integration with classical technologies so as to provide an easier path towards prototyping and fabrication for both research and industry.
- The **chip technologies** and industries will play a key role in developing and maturing new design methods, fabrication techniques and novel packaging for the quantum computer. Recognize that they need not be supported and, in this context, the role of quantum pilot-lines will be key.
- Design support and investment programmes in the knowledge that the race for quantum computing is not only about mastering core quantum technologies (e.g. qubits). It is also about the **infrastructures and industry** environment that will help make it real.
- Since a quantum computer will be a hybrid classical/quantum system, high-performance computing **(HPC) infrastructures** will play a key role in integrating the required **hybrid quantum system stack**. Develop the integration of quantum accelerators into future exascale infrastructures.
- Position Europe as a key player in the global race:
  - Europe is the leader in the foundations of quantum physics: it should keep promoting basic research and novel ideas.
  - Promote the emergence of a European quantum cloud connected to quantum-hybrid HPC platforms
  - Startups in quantum computing are emerging. Europe is well on track but needs to increase its support for new and small enterprises in the field.
  - Promote a European quantum computing ecosystem that will encourage talented people to stay in or relocate to Europe.

Qubits technologies: the race is on

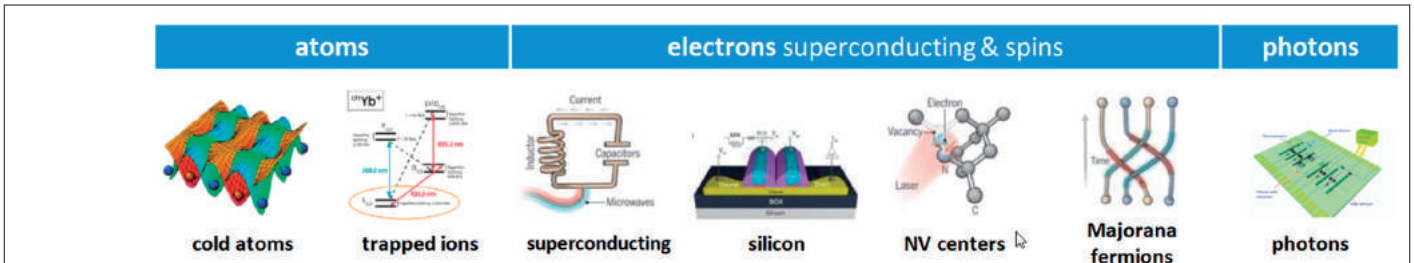
The design and realization of a quantum computer capable of operating efficiently on concrete problems remains a long-term prospect. At the hardware level, the fabrication of the indispensable qubits with sufficient fidelity and scalability potential is a major research field with several technologies being actively investigated. This is taking the form of a real race between a number of competing technologies (see figure 1 [9]).

Superconducting qubits have pioneered the development of the field and were first demonstrated in 2002 [2]. The technology more recently made the headlines with the famous paper on "quantum supremacy" by a Google/NASA team and its 54 superconducting qubits chip (the Sycamore chip) [3]. Quantum supremacy defined as "the potential ability of quantum devices to solve problems that classical computers practically cannot" was initially introduced by John Preskill [4]. The Google/NASA team actually achieved this but on a rather limited problem: they used the 54 superconducting qubits chip with nearest neighbours' interactions to run a circuit that generated near perfect random sequences of bits. Interesting, but not a game changer yet.

Although superconducting qubit technology has taken a small lead and is used by many players (Intel, IBM, Rigetti, IQM, Google, etc.), it is actively challenged by other technologies with their own strengths:

- Trapped ions [5] (developed by the Innsbruck University and later company IonQ) has record fidelities and coherence times but little potential for scaling up.
- Silicon-based spin qubits (TU-Delft, CEA-Leti, Riken, Intel, SQC) hold the promise of easier scaling up by using industry-proven microelectronics technologies [6].
- Cold atom technology [7] (developed at University of Wisconsin, CNRS, PASQAL, ColdQuanta) use laser cooling technology to manipulate atoms with the promise of 100-200 qubits processors by 2023.
- Qubits based on coloured centres in diamonds [8] (TU-Delft, Quantum Brilliance) or based on photons (CNRS, Xanadu, Quandela, PsiQuantum), which hold out the promise of systems operating at room temperature.

All these qubit technologies have their own advantages and disadvantages. Developed by numerous research laboratories and private sector companies, they are the subject of a race towards increasing the merit factors (fidelity, speed, coherence and scalability) that will make them usable in a real system. Today, the superconducting qubit technology is holding the line but has not yet demonstrated an operational quantum computer.



	atoms		electrons superconducting & spins				photons
	cold atoms	trapped ions	superconducting	silicon	NV centers	Majorana fermions	photons
qubit size	about 1 µm space between atoms	about 1 µm space between atoms	$(100\mu)^2$	$(100nm)^2$	$<(100nm)^2$	$(100\mu)^2$	$(100\mu)^2$
two gates fidelities	98%	99,9%	99,4%	>98%	99,2%	N/A	98%
readout fidelity	99%	99,9%	99,4%	98%	98%	N/A	50%
gate time	1 µs	100 µs	20 ns - 300 ns	≈5 µs	10-700 ns	N/A	1 ms
coherence		0,2s-10mn	100-400µs	20-120µs	2.4 ms	N/A	
qubits temperature	< 5mK	<1mK to 10K	15mK	100mK-1K	4K-ambiant	15mK	ambiant & 4K/10K photons generators & detectors
operational qubits	100-196 (simulator)	32 (IonQ)	65 (IBM) 56-66 (China)	4 (Delft)	5 (Quantum Brilliance)-10	N/A	70 (China)
scalability	1000	<50	100s	millions	100s	?	100s-1M

Figure 1: Qubits Technologies in the race ([1] p 240)

Besides progress in qubit technologies, one of the main questions that need to be tackled on the road to operational QC is that of scalability. Depending on the qubits' implementation technology, the challenges may vary but one that sticks is the problem of noise. The term "noise" refers to the impact of various factors leading to a loss of precision of the qubits' states (see figure 1 [9]). Those factors can come out of fabrication imperfections, sensitivity to parasitic signals, crosstalk, material degradation, etc. Digital computing technologies such as complementary metal-oxide-semiconductor (CMOS) are facing many of the same problems: after all, a CMOS transistor also suffers from noise, parasitic interference and crosstalk - but the difference is that the classical binary coding scheme provides an efficient shield against those underlying sources of noise.

From qubits to system, the hardware race

While the development of qubits of the highest possible quality, i.e. quantum superposition, is a crucial step, a large number of interacting qubits (through a phenomenon called entanglement) is necessary in order to be able to perform a useful computation: the technology needs to be scalable.

The problem with the quantum nature of the qubits' states is that they are extremely fragile. They are subject to the phenomenon of decoherence which destroys the quantum superposition and their interactions. In order to control this adverse effect, qubits should be placed in isolation from all sources of decoherence: environment, temperature, parasitic effect and even control and measurement apparatus. This is a real engineering problem given the complexity of the whole system stack as will be explained in the next section. To make things event trickier, the more qubits, the more sources of decoherence. As a result, the art of QC engineering and architecture is to design an ensemble of entangled qubits out of components that would be better isolated.

There exist several ways of using qubits for quantum computing:

- We can use them "as is" and try to exploit their quantum states directly by implementing an ensemble of qubits in which we are interested in their collective quantum behaviour. With recent technological advances, reliable large-size systems can be realized, for example systems of ultracold atoms in optical lattices with ten thousand or more constituents interacting with each other, arrays of Rydberg-excited. This type of architecture, known as **quantum simulators**, is directly inspired by Feynman [10] and allows one to predict static and dynamical properties of complex quantum materials or to solve optimization problems (e.g. quantum annealers).
- We can use them in a way very similar to classical digital computers. Also known as **universal quantum computing** in which each physical qubit shall be precisely controlled through a sequence of elementary quantum operations (gates). Universal QC is the "holy grail" of quantum computing and can be considered as the quantum equivalent of a general digital instruction set architecture (ISA)-based computer. Such a concept requires a long coherence time (at least long enough to perform useful computation) and **needs quantum error correction** to mitigate the inevitable decoherence of physical qubits. The problem with this approach is that a very large number of physical qubits is required to correct one logical qubit (an exploitable one). Although the principle of error correction has been demonstrated, it is far from being readily operational.
- Another way to use qubits, as proposed by John Preskill [11], is to set up a gate-based architecture (in a manner similar to universal QC) but without error correction. This is known as the **noisy intermediate-scale quantum** (NISQ) approach. This has been the approach used for all existing quantum computers experimented with to date. Superconducting qubits chips from IBM or Google fall into this category.

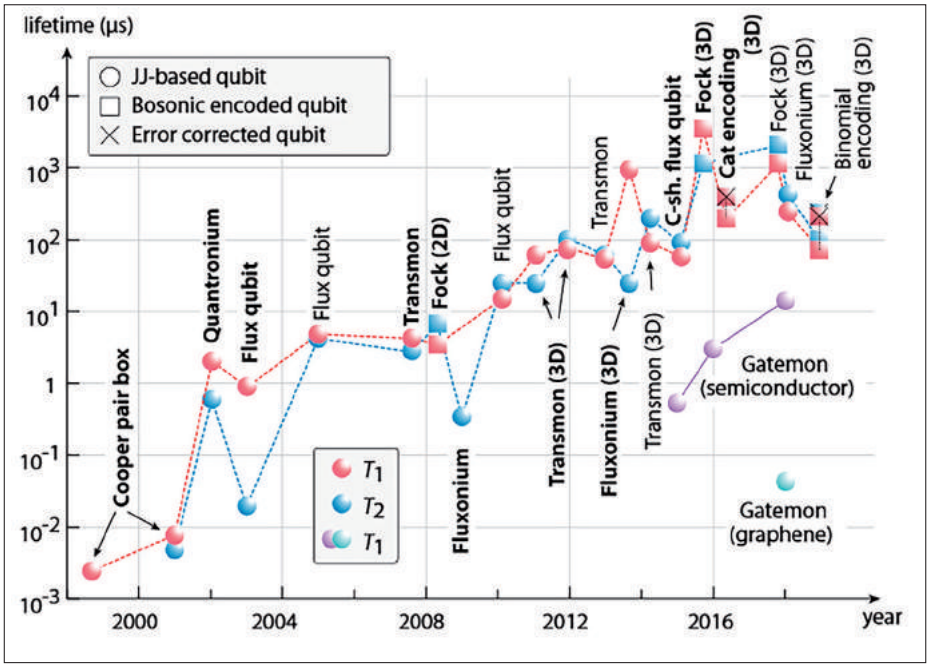


Figure 2: Evolution of lifetime and coherence time of superconducting qubits. From [9]

Regardless of the QC hardware architecture and the method of control of the many qubits involved, their initialization and readback will require a large amount of supporting electronics. The resulting QC machine will combine a quantum engine (i.e. an ensemble of qubits) surrounded by a complex control machinery and a classical digital computer. This can be envisaged through the work of Qutech [12] in which the architecture of such a hybrid QC architecture is proposed. In the diagram of the physical architecture (Figure 3), the quantum part is coloured pink, the interface circuits are in green and the conventional computing (by far the largest) part is blue.

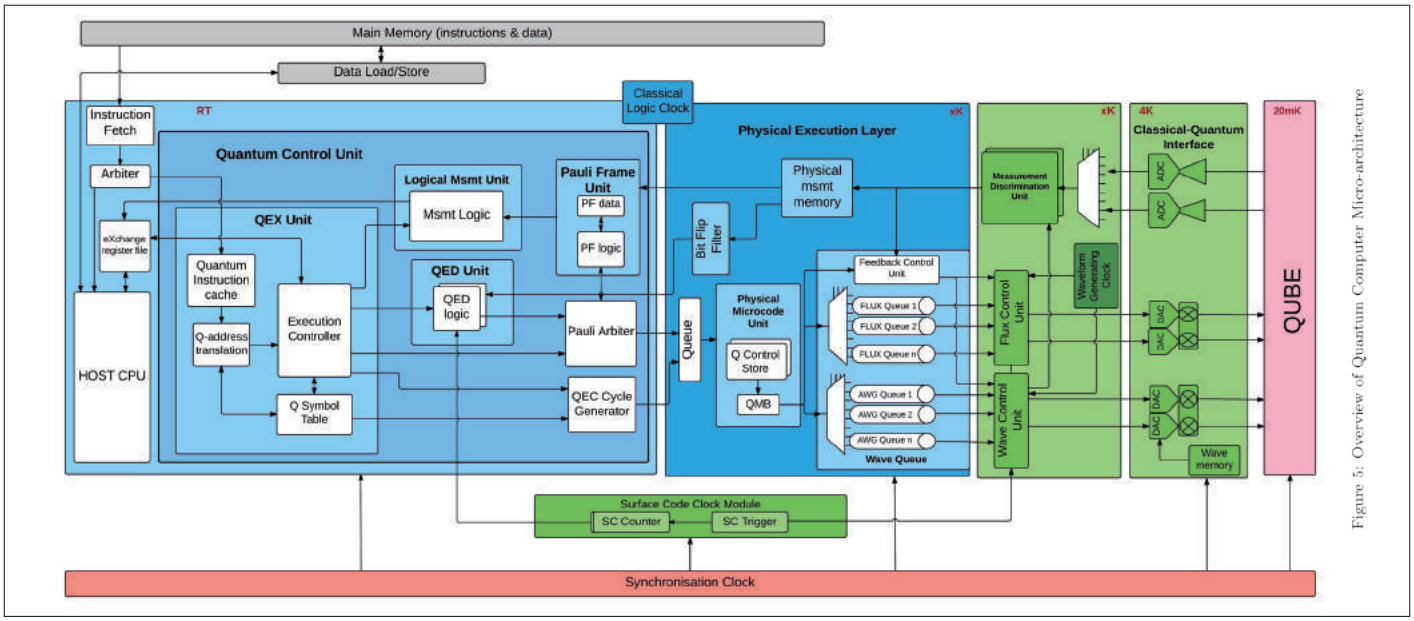


Figure 5: Overview of Quantum Computer Micro-architecture

Figure 3: Overview of a quantum computer architecture, from [12]

Quantum computing hybrid HPC integration

Let's be clear: a quantum computer cannot work without a classical von Neumann-type computer. Indeed, if quantum information processing can bring benefits in terms of parallelization of massive calculations (superposition of states), it will not bring anything significant in logic and control operations. Moreover, solutions to the problems we are interested in solving with a quantum computer must be exploited in our very classical world. Therefore, a quantum computer is and will remain a hybrid-computing machine with a "boring" classical computer doing the input/output interface and probably a big bulk of the processing logic. We then have the concept of a quantum accelerator core just like a graphics processing unit (GPU) core can speedup graphical computations or a field-programmable gate array (FPGA) can speedup specific tasks hand-in-hand with their host processor.

Advanced chips and supporting technologies are key

Quantum computing hardware is not just about qubits. In fact, the actual realization of a quantum computer, let alone its industrialization, requires a number of developments and much progress in very classical technologies. For example, the sheer number of connections needed for the control of more than one hundred qubits in a cryogenic environment will require R&D efforts in advanced chip technologies, wiring and packaging at the very edge of today's capabilities. Indeed, mastering such technologies is a critical step on the road towards the practical realization of quantum computing systems. What's more, a number of those technologies are very likely to be utilized regardless of the core qubit technology at work: superconducting, photonics, spins in silicon, etc…

The recently published *Strategic Research and Industry Agenda* of the European Union's Quantum Flagship [13] stresses that access to advanced chip technologies will be a requirement if Europe wants to be a serious contender in the race for quantum hardware.

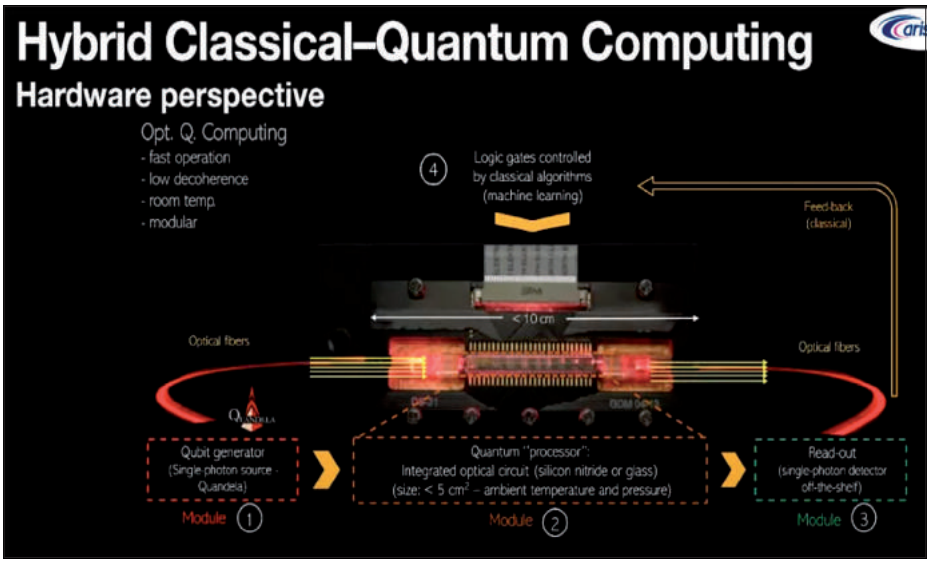


Figure 4: Perspective of a hybrid classical-quantum computing architecture based on single photon source qubits as proposed by the Quandela startup company [34]

Heterogeneous integration

Integrating various signals (digital, analogue, photonic, RF, etc.) and implementation technologies (various substrates), also known as "more than Moore", will be key for the integration of quantum technologies (QT) within systems and application

Advanced packaging

The development of quantum technologies requires specific chip packaging to accommodate for dense signal pathways, various signals (RF, digital, photonics), cryogenics or/and ultra-high vacuum interfaces. Advanced packaging techniques

such as 2D and 3D stacking developed in the context of the chips-act will be a major asset for QT.

Cryo-chips

In order to allow for the efficient control and scaling of quantum circuits, the control electronics shall be co-located with the actual quantum chips. We shall see an increase in the number of qubits that can be simultaneously controlled in line with the development of quantum processors (qubits) over the next three, six, and nine years. The fabrication of classical control chips (CMOS) able to operate at cryogenic temperatures as well as the fabrication of fast classical superconducting control chips, and highly efficient single photon detectors, are critical for the development of quantum computing hardware, in order to optimize signal routing, and increase qubit readout speed and efficiency.

In order to operate such a complex hybrid machine from qubits up to higher level programming, a complete system stack will have to be built. It could resemble the layered scheme presented in Figure 5. The overall system architecture will implement all layers ranging from qubit initialization and control, quantum interfaces, analog circuits, microarchitecture, quantum and classical instruction sets, system orchestration and programming interface (languages) right through to implementation of the quantum algorithms.

Programming the quantum computer: the software race

Due to the hybrid nature of the quantum computer, the program that would run on such a machine would need to combine at least two computing models: a classical part, to prepare data and process results and a quantum part to actually execute quantum operations. This will require a tight connection between the two programming models. Some preliminary ideas have been put forward to tackle this problem [14], but there is still a lot of research and development to be done. However, the classical computers with e.g. FPGAs also have a different way of programming. It is important to know what part is executed by the FPGA and what part is running on the classical processor. Something similar with

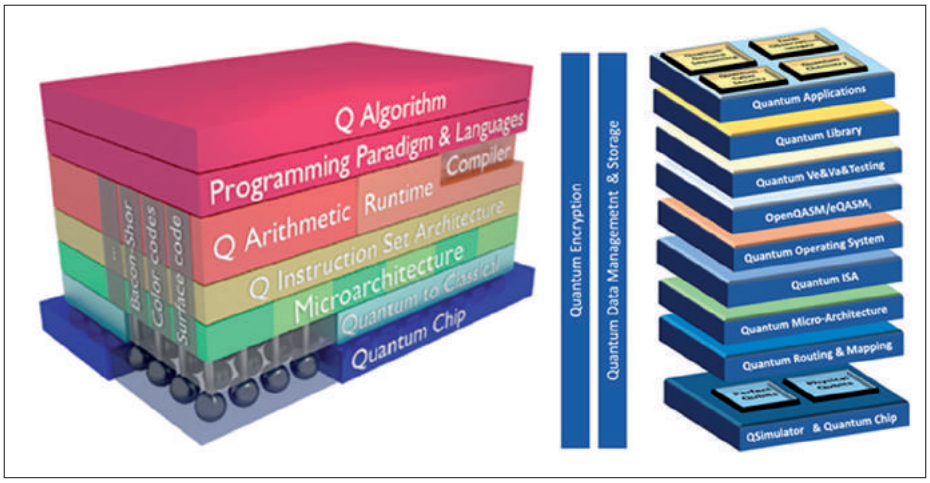


Figure 5: Possible structure of a hybridclassical/quantum computer (from [12] and [24])

a quantum accelerator will have to be developed. There is one big difference: quantum logic is substantially different to classical logic, regardless of which co-processor and non-quantum technology is used.

For the time being, a programming model for a quantum computer could be summarized in these four steps:

1. **Prepare** a set of qubits in an initial (quantum) state for the problem at hand;
2. Apply a sequence of **quantum operations** on the set of qubits;
3. **Measure** a final (classical) state with a given probability;
4. **Iterate** until readout probability builds up.

If everything goes well (and a lot of things can go wrong!) the final measurement will yield a solution to the problem with an even higher probability if the fidelity of the quantum calculation is maintained throughout the sequence of steps. It is however obvious that to increase the reliability of the final result, this succession of steps will have to be repeated a sufficient number of times.

We observe too that this succession of steps implies a part devoted to classical computation for the preparation and measurement phases: any quantum computer program is necessarily hybrid. Its quantum part can be seen as an accelerator and its classical part as an operations supervisor as shown in Figure 6. Consequently, the overall performance of the quantum computer will be limited by the performance of its classical part. This is a quantum version of the famous Amdahl's law.

The way quantum computing works is radically different from Von Neumann's classical computing. By way of illustration, on a classical computer the result is deterministic, whereas it is probabilistic in the case of a quantum computer. Another notable difference is that the value of a qubit cannot be copied. Thus quantum programming is a total break from a classical programming approach and it is therefore necessary to completely rethink the way to solve the initial problem.

Paradoxically, by rethinking the way problems are solved in order to adapt them to quantum computation, we are also advancing classical algorithms and pushing their limits a little further [15].

Research in quantum algorithms can also improve classical algorithms: as an example, the young Ewin Tang has proven that classical computers can solve the "recommendation problem" nearly as fast as quantum computers. The "recommendation problem" relates to how services like Amazon and Netflix determine which products you might like to try. Computer scientists had considered it to be one of the best examples of a problem that's exponentially faster to solve on quantum computers – like Kerenidis and Prakash's algorithm [16], Tang's algorithm ran in polylogarithmic time – meaning the computational time scaled with the logarithm of characteristics like the number of users and products in the data set, and was exponentially faster than any previously known classical algorithm [17].

Although true QC has not yet come to fruition, there is no shortage of programming languages, software simulators and development tools for it. For example, the quantiki website [29] lists over 150 quantum simulators written in more than 21 different languages. Similarly, most companies actively carrying out QC research (IBM, Google, Microsoft, etc.) often provide software simulators in order for users to get their hands into quantum programming.

On another level, QuTech in the Netherlands (TNO and TU-Delft) has introduced a quantum computing platform [30] to help run and evaluate quantum algorithms on a variety of simulators or hardware [18] [5]. With this platform the focus is put on the ease of access to a quantum simulation environment for experimenting with quantum processing on both simulated and real hardware platforms.

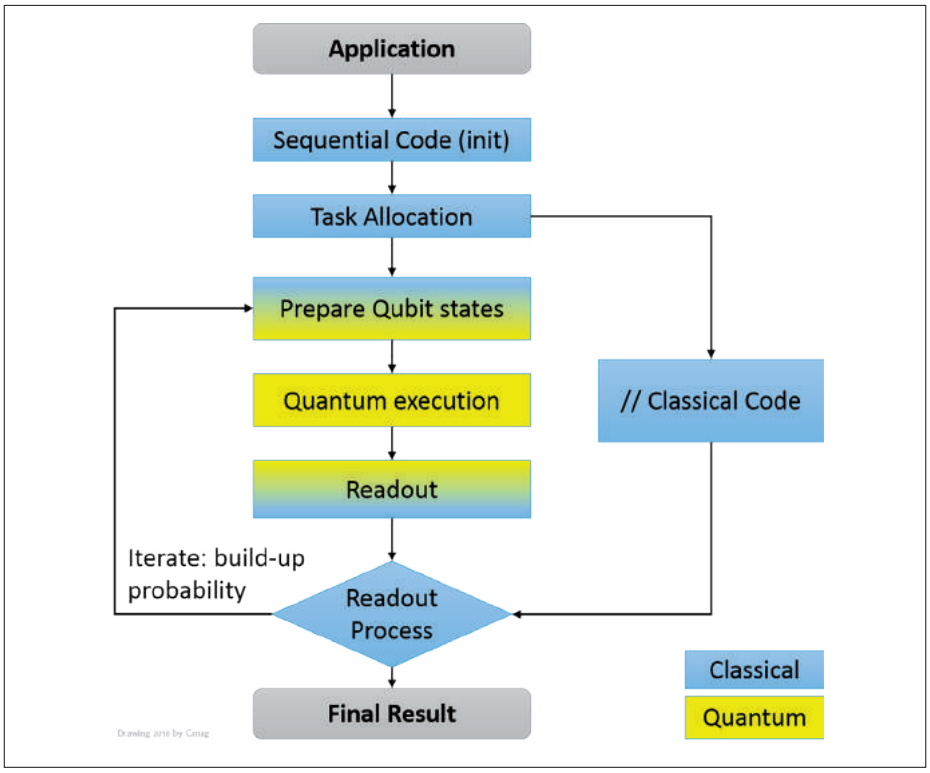


Figure 6: The hybrid programming flow of a quantum computer. In yellow are the quantum parts and in blue the classical code, either sequential or parallel.

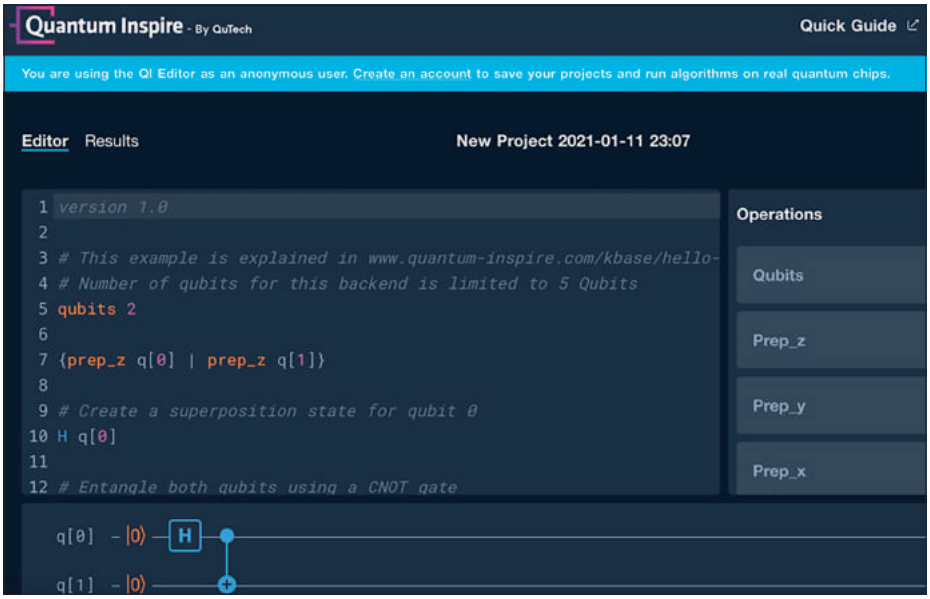


Figure 7: The QI Editor from Quantum Inspire [30] allows users to write, run and test quantum algorithms online

Those Europe-based QC platforms represent a bold move by offering state of the art quantum hardware and quantum programming tools to both industry and academic users in addition to the online platforms already proposed by the likes of Google (Cirq) [19], IBM (QisKit) , DWave, IONQ or Microsoft (QDK).

The advent of European cloud-based quantum computing platforms might also mitigate some concerns surrounding confidentiality and sovereignty. The recent announcement by Atos of the quantum metrics reference "Q-Score" [31] that aims to benchmark and measure the capacity of quantum processors to solve real problems will also help test out the effectiveness of quantum computing.

As the trend towards integrating quantum accelerated processors with classical (CMOS-based) computers accelerates, the need arises to develop specific system architectures and software layers that will orchestrate the execution of both the classical and quantum circuits required to run real applications.

Quantum computing and artificial intelligence (AI) are often cited as a winning combination. A number of recent scientific works emphasized the interactions between QC and AI [20]. It is hard to tell if this will become a reality or if it is just the result of a buzzword-heavy strategy.

Machine learning (ML), which is at the heart of the current incarnation of AI, is a multi-parametric optimization problem and it can in principle benefit from advances in QC or quantum annealing. At the same time progress in ML and AI can also benefit QC. This is particularly true in the field of quantum state measurements. Machine learning based measurement protocols have shown interesting results with both supervised [21] and unsupervised [22] learning methods.

Following this trend of AI/quantum "entanglement", Google released an open-source library aimed at bridging the gap between AI and quantum computing: the "TensorfLow Quantum" [32].

Variational heuristics (variational quantum eigensolver (VQE), quantum approximate optimization algorithm (QAOA)) are hybrid methods, often requiring AI techniques, used in linear algebra or combinatorial optimization that put a quantum "core" into a mostly classical loop. The architecture of such a hybrid quantum/classical method as proposed by [23] is presented in Figure 8.

In some ways, AI techniques with their abilities to capture hidden mechanisms might be part of the path to a solution. However, even AI is not magic, and it has been shown many times that, in order to be efficient, AI methods needs a little help from experts. That is to say: some physical knowledge (in our case quantum mechanics) has to be added into the overall equations.

Like we said earlier, it seems obvious that a quantum computer will be a hybrid classical-quantum machine, but we are also coming to realize that QC might even be more hybrid than that! It is quite possible that quantum computing could only be achieved as the result of a close coupling between quantum and machine learning techniques and algorithms: the quantum-AI hybridization.

One problem with developing a system and software stack for the quantum computer is that the race for the ideal qubit is still raging. No one can yet affirm when and how operational quantum hardware (i.e. with thousands of logical qubits of sufficient fidelity) will be available. One idea proposed in [24] suggests the introduction of an abstraction layer: the perfect intermediate-scale quantum computing (PISQ) as a complementary approach to NISQ as illustrated on figure 9. **The PISQ approach** provides an abstraction layer, so that no direct reliance on the roadblocks and/or progress of the quantum physical chip development efforts would be necessary. Quantum applications formulated using perfect qubits can be executed and tested on quantum simulators running on powerful classical computers. Such computers are dedicated to the emulation of a limited number of perfect qubits (up to ~50 qubits) using a very large memory space computer like the Atos Quantum Learning Machine [25].

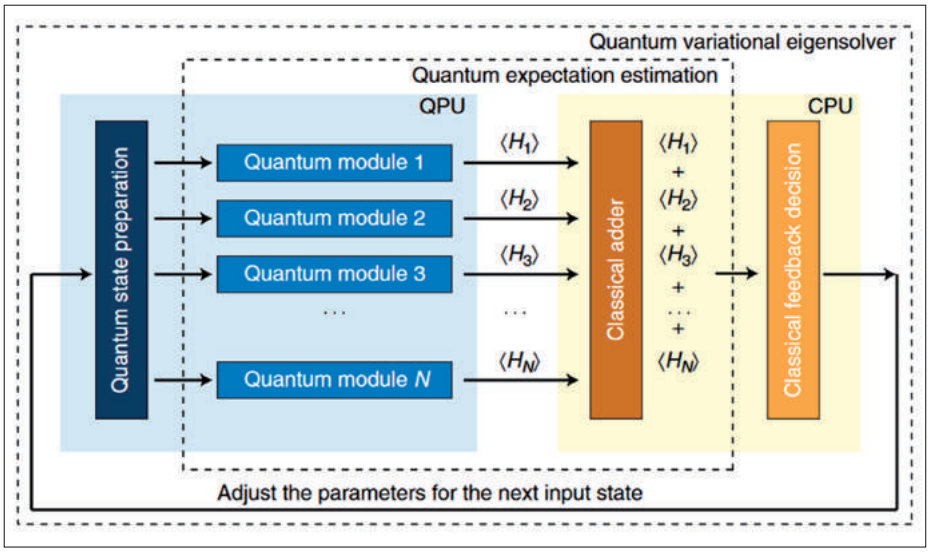


Figure 8: The typical architecture of a hybrid quantum/classical algorithm: the quantum part (blue) and the classical part (yellow) are linked in the execution loop (quantum-variational eigensolver from [23])

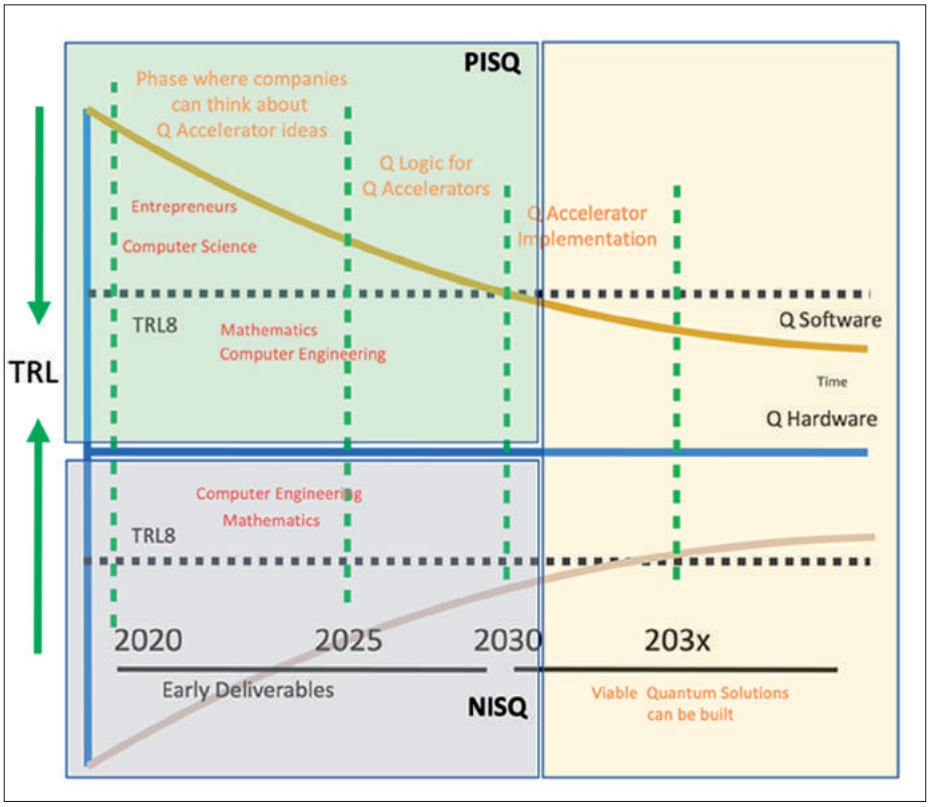


Figure 9: The complementary PISQ and NISQ approaches to quantum computing (from [24])

Although limited in terms of speedup for the time being, the PISQ approach will ultimately integrate NISQ and fault-tolerant quantum computing approaches for real quantum advantage. In the process it would have helped the software communities to enter the race to develop quantum software.

Quantum computing: a global race

European scientists built the foundation of quantum physics at the beginning of the 20th century. The number of eminent European physicists that were awarded the Nobel Prize for their contribution to quantum physics is simply amazing: 13 out of 16 Nobel Prizes [35]. Yet again in 2022 two European physicists, Professors **Alain Aspect** and **Anton Zellinger** together

with American colleague Professor John Clauser, were awarded the Nobel Prize in Physics for the experimental demonstration of quantum state entanglement, a result published several decades ago[26] [27].

In the **race for qubits**, Europe has pioneered several technological paths including the first experimental demonstration of quantum gate with superconducting qubits [2] and state-of-the-art demonstration of the capabilities of trapped ion qubits [5].

In the **race for quantum hardware**, Europe must get up to speed. If concept proofs and innovative solutions are being developed by pioneering European startup companies such as PASQAL, PsiQUantum, IQM, and Alice & Bob, and research labs like TU-Delft, ETH, and CEA Leti, the truth remains that quantum chips have so far been realized on a meaningful scale by North American companies such as Intel, Google, IBM, Dwave, and Riggeti. Yet Europe is now changing up a few gears with the roadmaps from the Quantum Flagship Initiative (QFLAG) and the association of European Industrial for Quantum Technologies (QUIC). The European Chips Act will also be an important actor in the European QC landscape and help the European microelectronics industry to join the race for quantum hardware.

Figure 10: winners of the Nobel Prize in Physics 2022, John F. Clauser, Alain Aspect and Anton Zellinger, from [35]

In the **race for quantum software**, the situation is very similar. Most available quantum programming frameworks originate from major US companies: Google, Microsoft, IBM, Rigetti, IONQ and DWave. The fact is that, for most software developers, the route to quantum software often starts with downloading a quantum programming library (Python, C/C++) and the choice is made by looking at the number of contributors, the support pages, number of forks and number of supported hardware platforms.

The world of software developers is sometimes disconnected from the world of quantum physicists.

In Europe, players like Atos, VeriQloud and Quantum Inspire, are entering the race and are already proposing approaches to experiment with quantum programming. One must also mention the efforts of hardware start-ups that allow experimentation with their platforms by providing cloud access to their progamming tools like Quandela cloud [33].

In terms of the **race between nations**, from the point of view of total public funding, the European Union as a whole (European Commission plus member states) is still in the race, being second behind China and ahead of the US. But this hides the fact that each member state has its own agenda and funding strategy, and these are

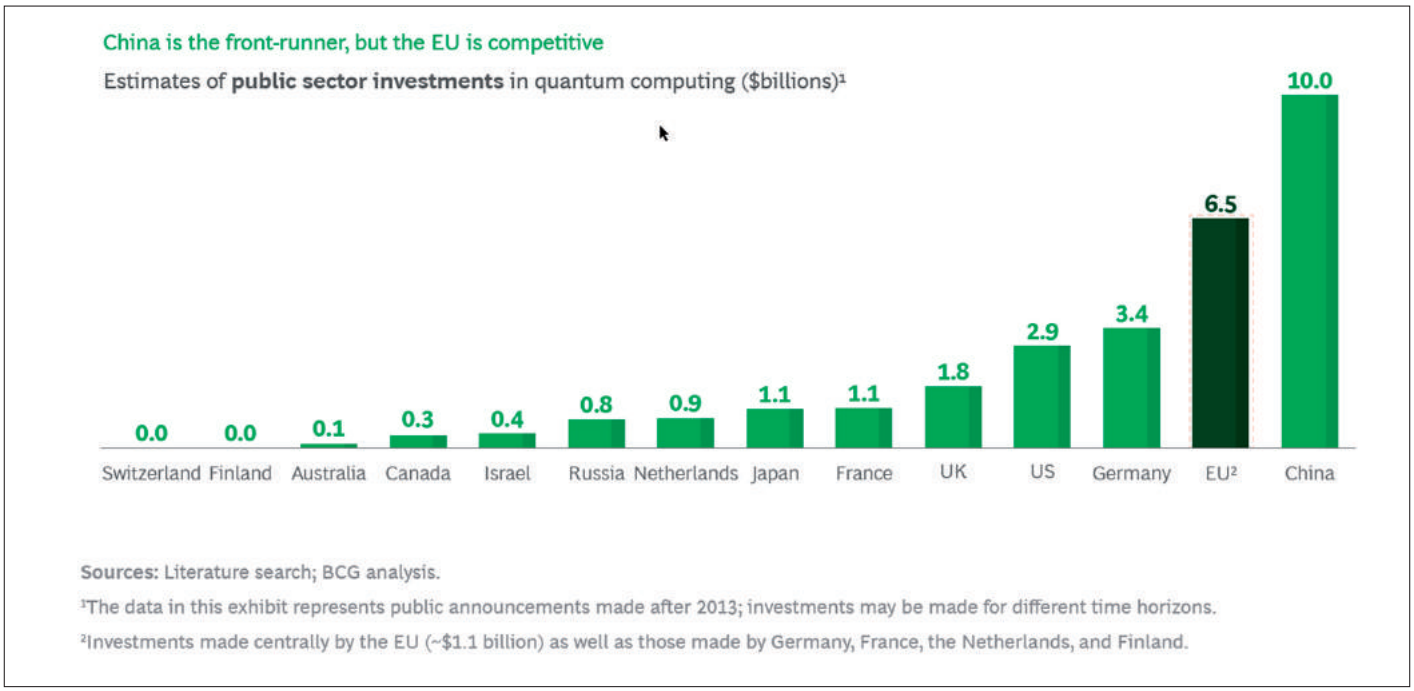


Figure 11: Estimates of public sector investments in quantum computing [36]

not always well coordinated. It is expected that community initiatives such as the EU QFLAG could bring greater coordination across the European QC landscape.

Innovative ideas originally developed in academic labs or by individuals often surface within start-up companies. Quantum computing is no exception and recent years have seen the emergence of a number of start-ups [28]. Even if some are just consultancy companies, a number of them are developing and marketing hardware, software or communication products. Their areas of interest range from single photon sources that could fuel quantum networks or photonic based quantum computing (Qandela) to designing complete quantum processors with atomic arrays (PASQal), and from developing software IPs for enterprise and industrial customers (AppliedQubit), to developing quantum software platforms for drug discovery (ApexQubit). There are companies in the field of quantum-secured communications (ID Quantique) and quantum network technologies and distributed quantum computing (VeriQloud). Europe is in a good position in this sector, being second to the US, with a good number of companies (see Figure 12). However, due to the lack of available risk funding opportunities in Europe, those companies might not secure the resources to support their growth. Some might find better opportunities across the Atlantic, while some might simply disappear.

In order to allow the European quantum computing ecosystem to get up to speed, Europe is also entering a race for quantum talent. To achieve this, twenty universities across ten European countries are setting up an education and training program for quantum technologies in Europe with 16 new specialized master's degrees. DigiQ (Digitally Enhanced Quantum Technology Master), the primary workforce development project of the Quantum Flagship, is funded by a Euro €17.6M grant over four years through the European Commission's Digital Europe Programme. The Quantum Technology Open Master, coordinated at Aarhus University and Czech Technical University of Prague, has set up an open ecosystem of specialist quantum technology courses among 26 partners. Empowering the Future Experts in Quantum Technology (Strasbourg University and Heidelberg University) has developed and trialed the network structure which will form the backbone of the student experience in the DigiQ master's degrees. The project will also be run in close synergy with the wider coordination efforts of the ongoing European Commission-funded QUCATS coordination and support action.

This initiative will lay the ground for future QC research and development, and should put Europe in a far better position in the numerous and lengthy races towards quantum computing.

References

[1] O. Ezratty, Understanding Quantum Technologies 2021: The most comprehensible 360° view on quantum technologies, from science and technology to use case, vendors and market. Independently published, 2021.

[2] D. Vion et al., "Manipulating the Quantum State of an Electrical Circuit," Science, vol. 296, no. 5569, pp. 886–889, May 2002.

[3] F. Arute et al., "Quantum supremacy using a programmable superconducting processor," Nature, vol. 574, no. 7779, pp. 505–510, Oct. 2019.

[4] J. Preskill, "Quantum computing and the entanglement frontier," 26-Mar-2012.

[5] J. I. Cirac and P. Zoller, "Quantum Computations with Cold Trapped Ions," Phys. Rev. Lett., vol. 74, no. 20, pp. 4091–4094, May 1995.

[6] R. Maurand et al., "A CMOS silicon spin qubit," Nature Communications, vol. 7, p. 13575, Nov. 2016.

[7] M. Saffman, "Quantum computing with atomic qubits and Rydberg interactions: Progress and challenges," J. Phys. B: At. Mol. Opt. Phys., vol. 49, no. 20, p. 202001, Oct. 2016.

[8] L. Childress and R. Hanson, "Diamond NV centers for quantum computing and quantum networks," MRS Bulletin, vol. 38, no. 2, pp. 134–138, Feb. 2013.

[9] M. Kjaergaard et al., "Superconducting Qubits: Current State of Play," Annu. Rev. Condens. Matter Phys., vol. 11, no. 1, pp. 369–395, Mar. 2020.

[10] R. P. Feynman, "Simulating physics with computers," Int J Theor Phys, vol. 21, no. 6–7, pp. 467–488, 1982.

[11] J. Preskill, "Quantum Computing in the NISQ era and beyond," Quantum, vol. 2, p. 79, Aug. 2018.

[12] X. Fu et al., "A Heterogeneous Quantum Computer Architecture," in Proceedings of the ACM International Conference on Computing Frontiers, 2016, pp. 323–330.

[13] T. Q. Flagship, "Preliminary Strategic Research and Industry Agenda," 2022.

[14] B. Valiron, N. J. Ross, P. Selinger, D. S. Alexander, and J. M. Smith, "Programming the Quantum Future," Commun. ACM, vol. 58, no. 8, pp. 52–61, Jul. 2015.

[15] C. S. Calude, "De-quantizing the solution of deutsch's problem," Int. J. Quantum Inform., vol. 5, no. 3, pp. 409–415, Jun. 2007.

[16] I. Kerenidis and A. Prakash, "Quantum gradient descent for linear systems and least squares," 17-Apr-2017.

[17] E. Tang, "A quantum-inspired classical algorithm for recommendation systems," 10-Jul-2018.

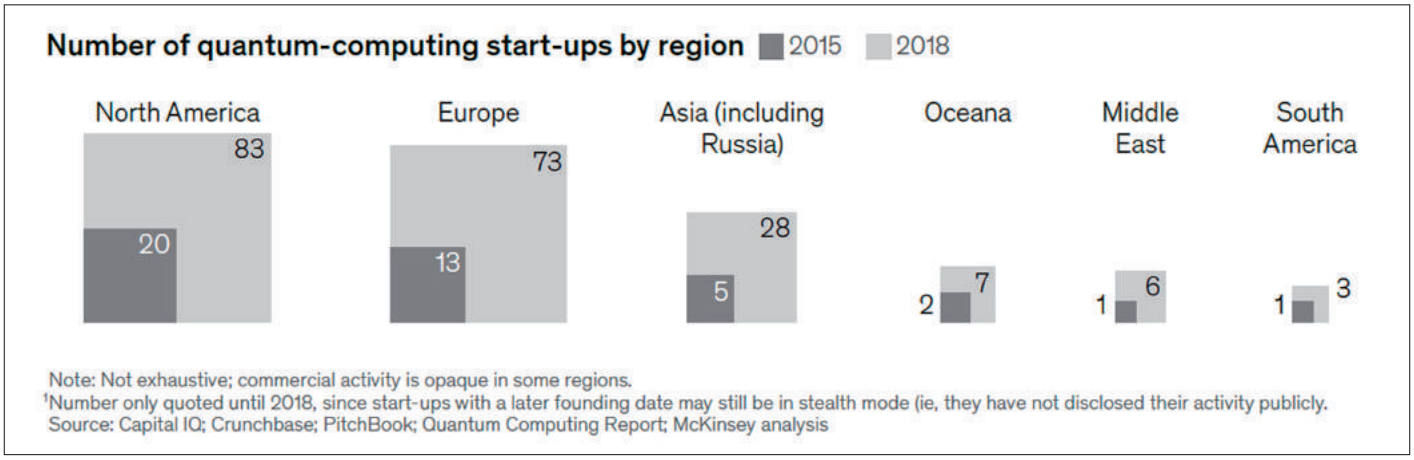
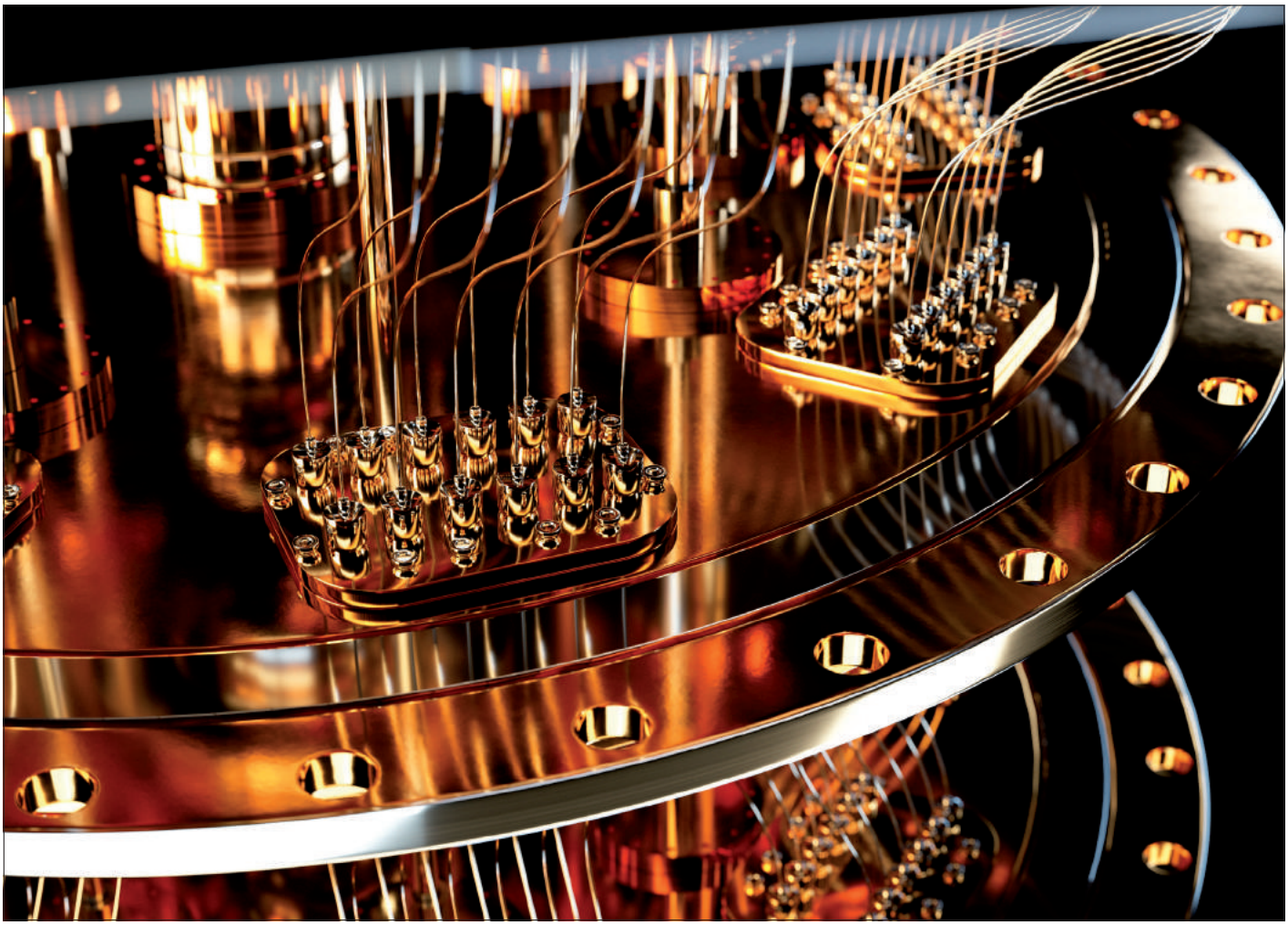


Figure 12: Number of quantum start-ups by region [37]



[18] TNO, "Minister Ingrid van Engelshoven and European Commissioner Mariya Gabriel launch Europe's first quantum computer in the cloud: Quantum Inspire," Apr-2020. [Online]. Available: /en/about-tno/news/2020/4/europe-s-first-quantum-computer-in-the-cloud-quantum-inspire/. [Accessed: 12-Jan-2021].

[19] M. Broughton et al., "TensorFlow Quantum: A Software Framework for Quantum Machine Learning," 05-Mar-2020.

[20] R. G. Melko, G. Carleo, J. Carrasquilla, and J. I. Cirac, "Restricted Boltzmann machines in quantum physics," Nature Physics, vol. 15, no. 9, pp. 887–892, Sep. 2019.

[21] V. Havlíček et al., "Supervised learning with quantum-enhanced feature spaces," Nature, vol. 567, no. 7747, pp. 209–212, Mar. 2019.

[22] K. A. McKiernan, E. Davis, M. S. Alam, and C. Rigetti, "Automated quantum programming via reinforcement learning for combinatorial optimization," 21-Aug-2019.

[23] A. Peruzzo et al., "A variational eigenvalue solver on a photonic quantum processor," Nature Communications, vol. 5, p. 4213, Jul. 2014.

[24] K. Bertels, T. Sarac, A. Sarkar, and I. Ashraf, "Quantum Computing – from NISQ to PISQ." arXiv, 08-Jul-2022.

[25] Atos, "Atos Quantum Learning Machine." 2017.

[26] D. M. Greenberger, M. A. Horne, A. Shimony, and A. Zeilinger, "Bell's theorem without inequalities," American Journal of Physics, vol. 58, no. 12, pp. 1131–1143, Dec. 1990.

[27] A. Aspect, J. Dalibard, and G. Roger, "Experimental Test of Bell's Inequalities Using Time-Varying Analyzers," Phys. Rev. Lett., vol. 49, no. 25, pp. 1804–1807, Dec. 1982.

[28] "Quantum gold rush: The private funding pouring into quantum start-ups." [Online]. Available: https://www.nature.com/articles/d41586-019-02935-4. [Accessed: 12-Jan-2021].

[29] https://www.quantiki.org/wiki/list-qc-simulators

[30] https://www.quantum-inspire.com

[31] https://Atos.net/en/solutions/q-score

[32] https://www.tensorflow.org/quantum

[33] https://www.quandela.com/cloud/

[34] https://www.nature.com/articles/s41467-020-19341-4

[35] https://www.nobelprize.org/prizes/lists/all-nobel-prizes-in-physics/

[36] https://www.bcg.com/publications/2022/can-europe-catch-up-in-quantum-computer-race

[37] https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/quantum-computing-use-cases-are-getting-real-what-you-need-to-know

Christian Gamrat is a researcher it the Research and Technology Department at CEA (Alternative energies and Atomic Energy Commission), France.

Koen Bertels is the founder of QBee and visiting professor at Ghent University, Belgium.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2023.

Cite as: C. Gamrat and K. Bertels. The race towards operational quantum computers. In M. Duranton et al., editors, HiPEAC Vision 2023, pages 110-119, Jan 2023.

DOI: 10.5281/zenodo.7461895

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

© HiPEAC 2023

Custom cloud hardware offers break-out performance, but it comes at a price: the loss of programmability and reusable software, and with a lock-in to cloud vendors.

The Programmability Wall

By THOMAS HOBERG

Cloud and internet have bypassed the Gigahertz Wall by turning nearly every use case into a scale-out workload. This has enabled bespoke technology right down to exotic data representation but it has come at a price, erecting a new wall of drastically rising programmability efforts and meagre software reuse while also creating cloud lock-in. With a clear trend ahead but many avenues to explore and obstacles in the way, Europe needs to find ways to create globally attractive technology in both soft- and hardware.

Key insights

- Global interconnectivity makes competition for valuable insights or content aggregate to global scales and relentlessly exploit ever more specialized hardware; "no bit may remain unflipped" in the quest for lower cost.
- Little was as important as abstractions, general-purpose architectures and high-level programming languages and the degree of reuse they enabled for past decades to make the extreme cost of new code affordable, but while bespoke architectures are optimal for runtime efficiency, design costs explode for hardware and software.
- Cloud giants using bespoke and proprietary hardware to replace freely accessible commodity components leads to a competitive advantage, and, without geographical or regulatory barriers, eventually to a monoculture.
- You don't have to produce chips to create significant and indispensable value for hardware and software.

Key recommendations

- Cloud giants using bespoke and proprietary hardware locks out competition and leads to monoculture or even monopolies: swiftly impose mandatory fair licences for proprietary components judged to be significant and/or instal barriers between design and operation.
- Given that the benefits of general purpose will not return, work on better software support for bespoke architectures to ensure Europe's programmers enjoy optimum productivity.
- The ability to integrate custom intellectual property (IP) blocks e.g. via Universal Chiplet Interconnect Express (UCIe) on chips or as accelerators via Compute Express Link (CXL) opens up broad opportunities for small and smart designs that can be licensed broadly: invest into design know-how, patterns, tools and avoid focusing too much on working only on entire chips.
- The need to maintain high levels of regulatory pressure and multiple compliance schemes is not always an impediment for innovation; use it to enable flexibility and resilience instead.

General purpose and high level

The first electronic computers were outrageously expensive not just because they required massive amounts of costly components like radio tubes, but also because the solution logic had to be rebuilt from hundreds of wires and dials for every new computational challenge. It's for that reason they only went to work on the most valuable problems, like winning WWII. Days or weeks of implementation work, however, were rewarded when electronic computers spewed out artillery firing solutions and tide tables in mere hours when human computers would have required months to end up with the same result.

The invention of software first meant that any configuration that had proven useful could be reloaded and used again, at practically no cost and effort. And since multiplications or sinus functions had rather generic use, the more efficient routines were quickly copied and no longer reinvented. The science and art of information technology shifted towards using the minimal number of hardware and software design patterns to cover the greatest possible problem space; the most generic and reusable function blocks were sought after because they reduced engineering manhours while maximizing the range of problems that could feasibly be solved economically with computers.

Every problem, no matter how big, was broken into tiny operations performed somewhat drudgingly, but at speeds far beyond what mechanics could match, from ENIAC's "screeching" 5KHz to today's incredible 5GHz central processing units (CPUs). Early computers might have had to operate digit-by-digit, breaking down multiplications or square roots into lengthy routines. High-level programming languages abstracted and helped hide the details of complex operations via intrinsic operators or function libraries. Today,

complex mathematics on whole vectors can be completed every clock cycle thanks to sophisticated arithmetic logic units (ALUs), vast register files, and deep pipelines. Combined with the steady increase of clock speed this approach made iteration king for decades: programmers might not have to change a single line of their Fortran code to achieve a near billion:one speed gain.

That has now changed and for more reasons than the Gigahertz Wall[1]: general purpose went missing, too.

The loss of general purpose and the rise of the special

We tend to put technical feasibility in the foreground, but it's always been rather more about the value proposition: is there sufficient evidence that performing computation on a piece of data will generate sellable insights or content[2] with a value greater than the total investment made? Economics is the true driver behind many "laws of IT".

Before the internet connected global computing and data, computers stood alone and were almost by necessity general purpose: very few problems were both valuable and big enough to warrant a bespoke solution. But one of the many side effects of the global internet is that an astonishingly high number of challenges globally now reach a scale where increasingly bespoke solutions running that single computation at scale-out is more cost effective than mixing it with others on a general-purpose platform.

Special purpose processors that sacrificed the advantages of genericity, easy programming and even reusability of code and design made an early appearance, where both the intensity of the problem and the scale of application justified the extraordinary cost of starting afresh.

The most significant and still most numerous class remains digital signal processors (DSPs), which appeared soon after the first microprocessors in the 1970s, e.g. the Texas Instruments TMS5100 for speech generation. DSPs played a crucial part in the birth of the internet, as they were employed in the dial-up modems that converted a telephone line designed for human voice communication into a pure digital high-speed communication channel. They did this by translating bits into a harmonic sing-song and back with line adaptation and fault resilience via sophisticated modulations, discrete fast Fourier transforms and Huffman compression that exploited them to very near the Shannon limit. Their successors keep doing that at much higher speeds on coax and twisted pair copper cables, because binary signals alone would not be nearly as fast or as reliable. Close and remote cousins did all the heavy lifting at minimal cost and use of power inside your CD and DVD players, They work on anything from audio, video, sonar or lidar signal processing and even inside your smartphone's photo sensor for various stages of exposure management, auto-focus, pixel binning, offsetting deficits in lighting or stability while other more distant relatives may graduate to inventing a realistically looking piece of art, inferred with significant help from artificial intelligence (AI) inspired by what was actually measured by the photo sensor.

These DSPs and their brethren are designed to perform a very limited set of pre-defined operations on data that is continuously flowing through for a complete transformation. The code size is relatively tiny and may remain near constant for the lifetime of the chip. While many of their operations or instructions may in fact use ASIC blocks, they retain some programmability because there is sufficient variability and complexity in the transformations to make that a benefit, or because new variants or optimizations may be required during their lifetime, resulting in new logic. For the sake of low power and cost, they will sacrifice generic capabilities and instructions, numerical precision and address space size, and add novel memory spaces, odd access patterns and powerful single instruction, multiple data (SIMD) or multiple instruction, multiple data (MIMD) instructions that only recently also found their way into general-purpose CPUs or graphics processing units (GPUs) before that. DSPs are everywhere, including really powerful ones in your earbuds and there are probably a dozen for every CPU in a smartphone or a modern car.

In DVD players DSPs could quietly do on a few watts and dozens of megahertz what used to require running a general purpose x86 CPU already augmented via "DSP inspired" MMX instructions at full hilt on active cooling at hundreds of megahertz. A pure general-purpose 80386 could not possibly decode DVD video even when running meltingly hot at 5 GHz, while modern smartphone SoCs manage far more complex codecs and higher resolutions at mere milliwatts.

But you can't take a pure C++ or Rust implementation of a software video decoder and use that to operate the smartphone DSP or indeed the whole pipeline. Instead, quite a few very distinct bits of software running on the smartphone's main CPUs, its GPUs, its DSPs and perhaps even NPU[3] and IPU[4] blocks for post-processing require not only careful interfaces, orchestration and collaboration but also different compilers, assemblers and even binary coding by true specialists to deliver that result. **The Gigahertz Wall to general purpose scaling has been broken via special purpose processing units or xPUs, but that has erected a new Programmability Wall to software engineering productivity that eats heavily into the ever-tinier value margins we obtain from computing**.

Large media production companies like Netflix have long used (DSP augmented) CPU "software" encoders[5] to generate their content, because even if software encoders run much more slowly than pure application-specific integrated circuits (ASICs), they can analyse the source material much more deeply and obtain higher compression. For them any megabyte saved on a

1 Also known as the end of Dennard scaling.
2 Personalized content generation, which includes gaming, is by far the biggest driver of IT evolution today.
3 Neural processing unit, optimized for low precision inference of machine learning models.
4 Image processing unit, optimized for large tile image manipulation tasks.
5 In quotes, because even these software encoders actually use special instructions and hardware accelerator blocks wherever possible.

gigabyte of video is worth the extra effort when the result is sent and thus copied to billions of users. Google is replacing that with a new generation of visual processing units (VPUs)[6] to enable a new quality in its video content distribution services. Its custom Argos VPUs [1] enable personalized real-time stitching of source materials and client specific bandwidth and codec adaptations, including the currently most bandwidth efficient but also very complex AV1 codec. Software encoders cannot economically handle batch encoding of AV1, let alone real-time conversions. Like most Google chips, Argos is a proprietary design, meant to keep competitors at a distance. Fortunately NETINT [2] has a similarly capable solution and is ready to sell to a broader public.

A GPU intermezzo

Perhaps best-known today, is the use of graphical accelerators or GPUs originally designed for 3D gaming, becoming enriched with enough general-purpose capabilities to run loopy workloads at perhaps one tenth of the speed of a scalar CPU, but with thousands of iterations in parallel for an effective 100:1 improvement in speed and power consumption. This step change was the main driver behind the AI revival a decade ago.

Neural network simulation had languished for decades before the invention of GPUs, because classical scalar architectures made them computationally too expensive for practical applications. An opportunity for change arose when GPUs for the giant personal computer gaming market transitioned from graphics specific data types like 'paletted color'[7], or game-geometry optimized 24-bit floating point numbers to IEEE single and double precision format and standard integer formats.

Because GPUs calculate individual triangles and pixels highly independent of any other, they achieved their fluid motion performance with vast numbers of computational units, exactly what GPGPU-based HPC could then exploit where data independence, locality and achievable parallelism were high enough.

As NN simulation at initial scale was still a classic HPC problem, general-purpose GPU (GPGPU) propelled the current AI spring through explosive growth for a decade, especially as long as technological progress and process shrinks could be sold equally to the various growth markets of personal gaming, HPC and now cloud giant AI: in that it followed the path of x86 CPU technology, which gained its competitive advantage versus other CPU architectures from fuelling both camps, PCs and servers, until recently.

That gain came at a price. While very smart compilers with the help of some hints were still able to extract loop level parallelism for SIMD and multi-processor architectures, GPUs required the complete replacement of the inner loops. What was often as much the description of a solution approach as well as the code to solve it, had to be replaced with something of a "software-ASIC": HDL like orchestration code that carefully marshals the GPUs with deep pipelines and huge register files in wavefronts to ensure they'd never drop a beat or stall an entire wavefront of GPU cores for dozens of cycles to fetch a waylaid bit of data from far away VRAM.

Because of that tight orchestration such code is very hardware dependent, unportable and needs to be adjusted to the specific resources available on the underlying hardware. Moreover, with every generational update of such hardware, it needs to be redesigned to the new functional capabilities and specific resources available. Frameworks and libraries like NVIDIA's CUDA, AMD's HIP or Intel's OneAPI offer applications the ability to inquire into the availabilities and reconfigure themselves accordingly, but it still comes down to having to support distinct code variants to cover the hardware variability within the application and having to test that logic on hardware variants to ensure the orchestration comes close to take proper advantage of the hardware resources available. The advantage of a single high-level language code base which can be maintained and reused across vendors and hardware generations is lost, with a huge impact on software development cost.

For GPGPU or the new GPGPU computing in AI training, HPC applications and game engines, the constant rewrites and adaptations to new hardware could still be limited to the inner loops of abstracted library routines, while much of the main logic and outer loops could continue to run on CPUs yet profit from the significant speed update.

But just as NVIDIA, AMD and Intel are betting their future on GPGPUs, dark nimble-scalability clouds are gathering from cloud giant use cases and the strong competition to sell content and attention at minimal cost is erupting into novel hardware architectures that leave virtually none of the legacy assets in operating systems and seven decades of code libraries standing.

The revenge of the xPU

AI remains a main driver with its need for quality, speed and wide applicability pushing technical evolution forward and the costs of energy, and hard- and software design stalling progress. But AI is two very different things most of the time: training and inferencing, which not only have very distinct computing requirements, but also tend to execute at very different locations, scale, precision, efficiency and latency requirements.

Like software and hardware engineering, AI models recoup their giant development and training cost by being deployed at scale and low cost for a sufficient value of run-time inference. And much of the time their sole responsibility is to decide if a piece of insight gained from data via inferencing is worth forwarding or discarding. One key to operating AI at a profit is to discard everything without potential value as early as possible, because transporting or even storing useless data wastes energy. And the other part is to turn any insights gained into action before a consumer

6 Video processing unit, sometimes also called VCU or video compression unit because VPU has been previously used for computer vision accelerators, which are more like computer vision optimized NPUs.

7 With strings of bits encoding red, green and blue levels

notices a response lag or that this insight has just been auctioned off to the highest bidder, who now determines what they will see next.

Whether it's social media content from TikTok/Youtube/Meta or a game with embedded purchases, AI models constantly observe and assess what might trigger the consumer to make a purchase. They then stitch an ad clip or even replace the accessories being worn by the main character of a live video streams (via VPUs) with something they want the consumer to buy. Alternatively, they personalize the game in a similar manner to maximize consumption. To make this happen both in real time and using an amount of energy that costs less than the value of the induced purchase, the business decision-making traditionally coded in high-level code and run on a server processor as a user-space application needs to move into the midst of the network data plane: nobody can wait for an enterprise resource planning (ERP) tool to catch on when the content stitching and transformation needs to start perhaps 100-200 ms after a meaningful user response was registered.

Somewhat like DSPs, data processing units (DPUs) treat, change and transform live data streams, but more importantly they decide where data will go in a network; if it's to be multiplied to several interested parties or should indeed be stopped before it wastes more energy. Those decisions are made while data flows through them at 400Gbit/s per port and they are often coded in P4 [3], a dataflow programming language designed to avoid any latency overhead. Where that paradigm isn't powerful enough or a deep processing escalation is required, DPUs currently offer several dozen CPU cores capable of running Linux applications from within the 'network chip', to which these more complex messages can be delegated. **Sun Microsystem's old adage of "the network is the computer" has never been more true, when the highest level business decisions which would be the core of an enterprise resource planning system move into the very fabric which originally only transported data, and whole clusters of Unix machines far more powerful than any Sun server are sprinkled into a DPU chip.**

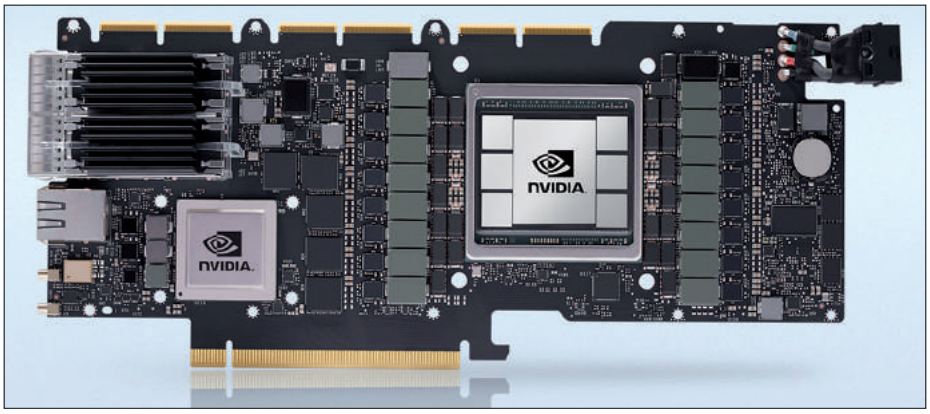


Figure 1: NVIDIA's converged accelerator [5]

Modern video surveillance hardware combines DPU and GPU blocks into an "AI network interface card (NIC)" that also contains whole servers in a manner that seems more like an afterthought than the CPU of old [4]. The entire processing pipeline from the camera inputs through to the resulting security alerts for suspects never leaves the system on chip, which runs its services on a variety of very distinct processing units, coded in domain specific languages using proprietary libraries, runtimes and operating systems with little if any portability between vendors or even hardware generations.

VDCD (Very Deep Custom Design)

Crypto currencies, BitCoin in particular, achieved such computational scale and ephemeral value that their proof-of-work algorithm soon went from general purpose CPUs in 2009 via GPUs in 2010 to more specialized hardware like FPGAs in 2011, and then exclusively to ASICs since 2013. In a sense, it's an entirely artificial computational challenge; the generation of SHA-256 based one-way cryptographic hashes in particular was designed for minimal code complexity but strong, in fact ever stronger, computational effort.

Tiny code specifically designed for temporal dependency to enforce loops results in chips that are very simple and thus small, require extremely little data flow due to near ideal data locality and are typically deployed in huge numbers with minimal networking or operational overhead. The energetic efficiency gain of a bitcoin ASIC vs. a 2009 general purpose CPU has been estimated at 100 billion to one.

In practically all regards, this made the BitCoin computational challenge an ideal candidate for a custom ASIC; it's almost never that easy in less artificial real-world contexts. Still, to progress, future mining hardware generations would need to continue to generate more value than they consume and that seems to change drastically: doubt in their value seems the only upward trend currently being observed.

In comparison, the perceived value of machine learning has grown more slowly, but still continues to accelerate. The problem domain, however artificially simulated, is a model of a typically chaotic real-world challenge, where programmers have given up on finding purely algorithmic solutions: AI approaches continue to deliver sufficient quality at lower cost than human effort or coding.

It has the computational challenge not just at the other end of the complexity spectrum exemplified by crypto mining; AI simulations also tend to have far worse locality than even "tricky" HPC workloads, yet offer unprecedented efficiency gains through reduced numerical precision or if sparsity can be properly exploited.

Furthermore, after everything else has been optimized, the biggest remaining dependency for the quality of AI is the size of the model that can be afforded, adaptive precision and optimal sparsity support are the strongest focus of evolutionary pres-

sure for hardware and software. New large-range/low-precision bfloat16 and even bfloat8 data types were invented and quite often even 16/8/4/2/1 bit integer numbers can be used, in what seems an atavism of the early GPU specific data types. Supporting all of these formats efficiently with a single set of ALUs and in the same memory space seems "conflicted" and finding a gaming benefit for them to pay for the design and retain the scale, is getting harder. It results in partitioning of resources and hard choices as models grow ever bigger and the overhead of RAM access wrecks the carefully marshalled wave fronts.

The larger the models, the more sparsity becomes an issue, as most coefficients in a neural net will actually be zero, yet require between 1 to 64 bits for storage. As a consequence, one of the biggest challenges for the growth of bigger models is reducing the heavy penalty of going off chip for data that no longer fits locally.

Individual GPU chips have reached the reticle limit and so condiderable effort is going into scale-out architectures like NVIDIA NVlink or Tesla's DOJO [6] and fast and smart interconnects like Lightmatter [7]. Yet that still puts the main effort of managing the non-linear access time among the billions of coefficients onto the software that manages, optimizes, sparsifies and distributes the huge neural nets, and onto the designers that create them, a similar hardware, topology and generational dependency we've seen for xPUs, not the high-level abstraction that makes things economical.

As a consequence, the happy GPGPU marriage of synergies between GPUs designed and made economical through gaming and then used and made profitable via HPC and AI is suffering, just like x86 servers and PCs no longer enjoy similar bliss. We are seeing the emergence of true unicorn designs that will also require extensive software adaptations.

Cerebras [8] looked at the approaching end of the GPGPU-AI synergy phase nearly a decade ago and started a completely novel approach to hardware scaling via wafer scale integration and the latency issues of model weights, designing something of an "AI Harvard architecture" for which many details remain unknown, while the ability to scale models with very near linear efficiency across clusters with hundreds of wafer scale chips has been demonstrated.

Wafer scale integration's brilliant idea is not to break into pieces what you will eventually need to be whole again, to skip cutting wafers into chips, when right after you'll need to join them up into a fabric using expensive interconnects. While that is already much more complex than it sounds, it is the weight and gradient streaming to single cycle local static RAM which allows avoiding the need to partition data sets and averaging the gradients that cause badly diminishing returns for scale-out GPU clusters [10]. This allows Cerebras to scale to near 100 billion parameters with a cluster of 100 systems. A data flow engine design for the CPU suppresses forwarding and thus executing on sparse/zero data [11], naturally and dramatically reducing energy consumption.

While each of the 850,000 cores in a wafer scale Cerebras 'chip' may be even smaller and 'simpler' than Bitcoin ASICs, the bandwidth needs and capabilities are at opposite ends of the scale. But one can't help but admire how the company has managed to put the seemingly chaotic nature of AI into such regular structures, while at the software layer it maintains all the benefits of the single GPU abstractions most machine learning frameworks started with.

Whither Europe

The ever-greater diversification into functionally specialized ASICs and processors follows the giant cumulative pressure on hyper-scalers to improve their economy and efficiency. The main hurdle is the need to develop and operate these specialties. While cloud giants are constantly conflicted between advertising their secret sauce designs to gain customers and giving away too much to potential competitors, some silicon start-ups will try to fill significant niches with alternatives,

Figure 2: Cerebras WSE-2 chip [9]

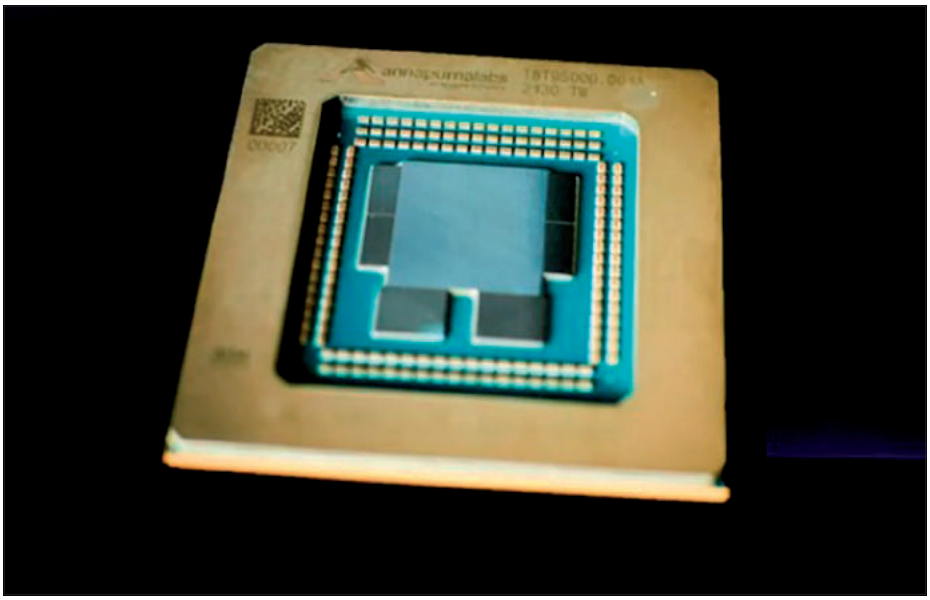


Figure 3: Amazon's Graviton3 Arm server processor, from [14]

as seen by NETINT's efforts facing up to Google's Argos VPU. Yet such an approach doesn't quite commoditize such technology even if the hardware and software is freely sold. The scale threshold required to make these highly specialized products integrate and work together within smaller companies e.g. inside European countries keeps getting bigger, creating an environment where they are no longer able to compete, especially when you need most of them for a complete stack.

Worse still is the trend that already supply of last generation 'legacy' components suffers from the preferential access only US cloud giants enjoy to US designed chips and often a year before their official release. Meanwhile, the quickly melting customer base in more traditional commodity servers stops their evolution, increasing the price pressure for IT service customers to move into the cloud, even if they desire nothing but commodity services. Software and hardware offerings for on-premise corporate data centres are disappearing and the trend is quickly accelerating as cloud vendors are increasingly turning towards custom designs even for CPU chips. It allows them to insert smaller xPU IP blocks on-chip and thus creates a significant competitive advantage versus commodity [12, 13].

Europe's ability to access the latest high-performance chips already suffers from lack of scale, but the US near complete chip manufacturing boycott against China also makes Europe's IT customers and vendors vulnerable to the whims of a US government department, because any third party levelling technology will find itself politically constrained to its country of origin. The decades old ecosystem of x86 commodity servers could collapse rather quickly over the coming years, with Europe finding itself without secure or timely access to what will replace them.

Admittedly, the ability of vendors to create their own hardware creates a lot of synergies and efficiency, but at the risk of market dominance. Dual market platforms work best without competition so the big chip game is very similar to what we see in the gaming industry or with Amazon: game engine (EPIC) and game distribution (Valve) attempt to force each other out, combining the leverage of originally completely distinct home markets, while Amazon competes with everyone who is only a shop, a producer, a transport or storage company, a media publisher, a big data company, a space travel or a cloud service provider or doesn't yet operate at global scale and with a universal spectrum.

To retain the ability to compete fairly within Europe, this type of cross domain leverage needs to be cut, just as it had to be cut between the oil production and oil refinery and gas distribution parts of SOHIO [15]. Google may design tensor processing units, Amazon may design Gravitons, everyone should work on xPUs, but they then need to be sold or fairly licensed at the same price to everyone, including the open-source tools required to use them. When the predictable result of innovation nears a monopoly, it's time to "move fast and break things", before it happens.

Europe and indeed every other global region needs to make sure it creates and guards as many other critical pieces of the pie in order to remain or to become so relevant [17] they cannot be ignored or bullied by silicon superpowers.

Meta's stumbles and Twitter's potential meltdown hint that, in the cloud, CEO-driven corporate ethea may be more likely to crash into than overcome political barriers, when their services become critical enough via digitalization to challenge political rulers on the ground. And that could very quickly bring their innovation machine to a stop, creating opportunity for smaller players, including those in Europe. Europe would do very well and rightly so in making sure that regulation on privacy and anti-competitive behaviour is actually enforced, now that Meta and friends have less money to spend on lobbyists.

Bitcoin ASICs are not built in a garage and Cerebras' wafer-chips require facilities and resources far beyond what anyone in Europe can consolidate. However, in both cases their design and the necessary feasibility studies were possible at much more modest scales and budgets: thinking big and forward enough would have been sufficient to develop technology that could be licensed even to the biggest partners. Point in case: none of Google's IP blocks used in their groundbreaking Argos VPUs was developed by Google. Only the final silicon is proprietary.

Formerly a European hardware superpower, Arm currently seems at risk of self-destruction due to the financial stress of its Japanese owner. But most of the work done to create CHERI [18], a security extension to its base architecture that manages to make it much more resilient to exploitation by malware while achieving very low execution and energy overhead, shows how long-running European research projects can obtain results that may become mainstream on all devices with direct internet exposure. Much like the Morpheus [19] augmentation and the similar RISC-V both prove that you don't have to be a CPU vendor to create significant value at the very core of the CPU space.

While the EU struggles to be relevant in chip production, there is that huge and growing issue of creating value from hardware that increasingly seems as solid as quicksand and compromises on format, function and reliability of data types as the most fundamental base of all programming: current high-level programming languages have long lost the benefits of abstraction or providing type safety. Quite often they are just used for boot-strapping compute kernels into DSPs or xPUs or in fact the new operating systems [20] and hypervisors are actually running on DPUs rather than CPU cores.

Newer languages and compilers which allow developers to manage and safely program with variable precision data types or the DSP, SIMD, MIMD inspired "super CISC" instructions; others that support a better integration of data-flow and data-plane processing with functional or even imperative stiles, may help to empower new generations of programmers to work efficiently with newer generations of hardware even before they happen. Expanding the scope of economically computable functions with the advances obtainable from chip technology requires us to look

at and doubt every bit we haven't flipped over yet, so we clearly know that general purpose CPUs won't ever regain a lead role. RISC-V was built around that realization more than a decade ago and opened CPU cores and instruction sets as widely as possible for a general-purpose base. We need the same foresight for programming languages, software development tools, frameworks and libraries, because readable and reusable code is still a key ingredient to create value in IT services.

Likewise, CHERI and Cerebras demonstrate that trends can be anticipated and prepared for many years in advance. There are few complete surprise changes of direction in the basic physics or the fundamental technological challenges. Even if IT is more politics than ever, upheavals are another trend that isn't likely ever to decline, while a solar super flare could stop digitalization rather abruptly. And since both are sure to happen and the only variables are when and to what extent, perhaps that's something else worth preparing for. The need to cover various distinct levels of resilience, security, and energy efficiency should not require having to code for each permutation.

Europe's seeming inability to have nearly 30 countries rapidly agree on anything and its strong focus on regulation is often painted as an obstacle to innovation. But the giant innovators East and West often create monocultures, which can prove risky. Europe's need to design for heterogeneity needs to be exploited to support flexibility, resilience, multi-compliance and the ability to adapt to large changes in obtainable value versus. fluctuating energy cost at the lowest possible software engineering cost. That won't always win in in terms of economics, but it offers opportunities as economic and political cycles occur.

References

[1] https://www.semianalysis.com/p/google-new-custom-silicon-replaces

[2] https://www.semianalysis.com/p/meet-netint-the-startup-selling-to

[3] https://p4.org/

[4] https://www.barkhauseninstitut.org/fileadmin/user_upload/Publikationen/2022/202204_Vilanova_EuroSys_FractOS.pdf

[5] https://www.nvidia.com/content/dam/en-zz/Solutions/gtcs22/converged-accelerator/nvidia-converged-accelerators-hero-bb580_440-d@2x.jpg

[6] https://www.servethehome.com/tesla-dojo-custom-ai-supercomputer-at-hc34/

[7] https://www.semianalysis.com/p/beyond-advanced-packaging-lightmatter

[8] https://www.scribd.com/document/477818755/HC31-1-13-Cerebras-SeanLie-v02-pdf

[9] https://www.cerebras.net/product-chip/

[10] https://www.nextplatform.com/2021/08/24/cerebras-shifts-architecture-to-meet-massive-ai-ml-models/

[11] https://www.servethehome.com/cerebras-wafer-scale-engine-wse-2-and-cs-2-at-hot-chips-34/

[12] https://www.nextplatform.com/2022/01/04/inside-amazons-graviton3-arm-server-processor/

[13] https://www.semianalysis.com/p/is-ampere-computings-cloud-native

[14] https://www.nextplatform.com/2022/01/04/inside-amazons-graviton3-arm-server-processor/

[15] https://en.wikipedia.org/wiki/Standard_Oil

[16] https://en.wikipedia.org/wiki/Meta_Platforms#History

[17] https://www.semianalysis.com/p/austrias-silent-monopolies-on-advanced

[18] https://www.cl.cam.ac.uk/research/security/ctsrd/cheri/

[19] https://web.eecs.umich.edu/~barisk/public/morpheus.pdf

[20] https://dl.acm.org/doi/pdf/10.1145/3492321.3519569

[21] https://www.servethehome.com/a-cerebras-cs-2-engine-block-bare-on-the-sc22-show-floor/

Thomas Hoberg is the technical director at Worldline Labs in Frankfurt, Germany.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

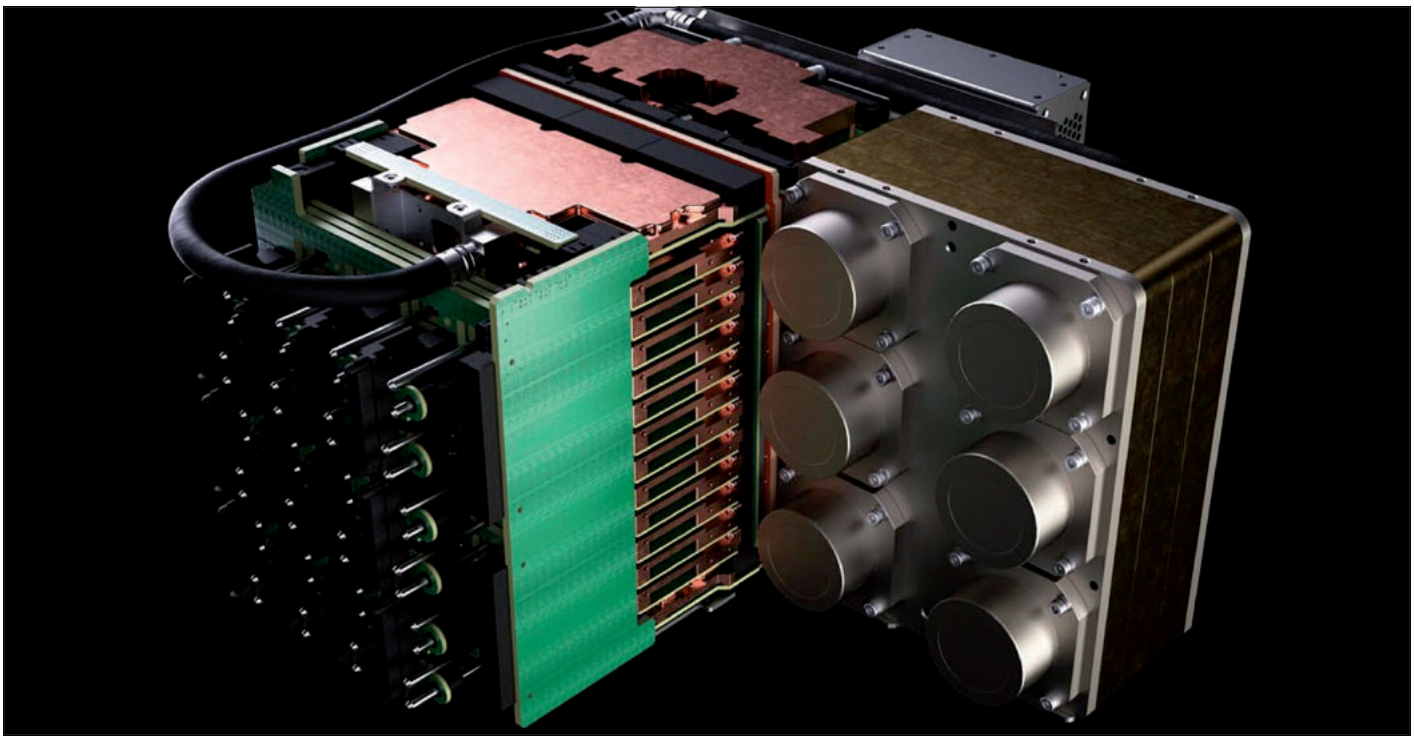
This is release v.1, January 2023.

Cite as: T. Hoberg. The programmability wall. In M. Duranton et al., editors, HiPEAC Vision 2023, pages 120-126, Jan 2023.

DOI: 10.5281/zenodo.7461899

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

© HiPEAC 2023



Cerebras CS 2 WSE 2 Engine Block, from [21]

HiPEAC Vision 2023

THE RACE FOR CYBERSECURITY

As privacy breaches, cybercrime and cyberwarfare have been soaring, significantly improved cybersecurity has both become crucial, and an opportunity for the EU to be a world leader.

The race for cybersecurity

By OLIVIER ZENDRA and BART COPPENS

After decades of digitalization spreading into every area of our lives, with very little attention given to the aspects linked to cybersecurity, information technology (IT) had essentially become an "open bar" for cybercriminals. For a few years, with a marked degradation during the peak of the COVID-19 pandemic, the news has been rife with reports of privacy breaches and cyberattacks (mainly ransomware) on companies and institutions, especially local governments and hospitals. In addition, cyberwarfare has been making the news too, especially in relation to the conflict in Ukraine.

Thus, the era of blissful ignorance and naiveté has ended. Although the wake-up call was abrupt, knowledge of these issues has expanded, and governments and to some extent businesses have taken first moves to enhance the cybersecurity frontline. However, cybersecurity is a highly competitive race between nations, between defenders and attackers, with enormous stakes. The pervasiveness of IT provides a broad attack surface, and attacks can be economically devastating, but they can also have tangible or even lethal repercussions on the physical world.

Despite several highly acclaimed advancements (e.g. the General Data Protection Regulation-GDPR), the EU still has a great deal of work to do in this regard, particularly to maintain its sovereignty and become a leader in the global competition. Cybersecurity is indeed a matter of both economic leadership and national sovereignty.

This chapter contains two contributions.

- *"From cybercrime to cyberwarfare, nobody can overlook cybersecurity any more"*
  This article describes the current state of IT system cybersecurity, showing how vulnerable systems are to the numerous dangers and challenges posed by cybercrime and cyberwarfare. It goes on to present a few concrete ways to remedy the issue, whether by technical, legal, sociological, or political means. Indeed, although the EU has weaknesses, linked to its extremely high reliance on IT systems, it also has the potential to become a world leader in cybersecurity, owing to both its strong technical culture and its regulatory capabilities.
- *"Is privacy possible in a digital world?"*
  Over the last few years, privacy has become a hot topic. However, this is in large part due to the fact that ever more data is being collected, not only by governments, but also by companies. It is often unclear for which purposes this data ends up being used; worse, it can even be leaked to third parties by attackers. Furthermore, even if this collected data would appear not to be sensitive in and of itself, sometimes sensitive information can be deduced from it. In this article, we present a summary of some of the ways in which data is gathered; how additional information can be inferred from it and how this is problematic; and how we can try to protect our privacy.

Key insights

- Ever more data is being sent to and collected from users, which has serious implications for privacy, especially when this data is analysed with big-data methods and artificial intelligence (AI).
- The scope and volume of this data is often not clear to consumers, who are sometimes even completely unaware that it is happening.
- Cyberattacks are constantly increasing, and the cost of damage caused by lack of cybersecurity is soaring. Cybercrime has become an industry, with tools and automation.
- Cyberwar has become prevalent, as shown in recent conflicts.
- The boundaries between cybercrime and cyberwarfare are blurred in some countries.
- Cyberattacks can be made at relatively low cost, and can cause very widespread and profound, even life-threatening, damage.
- There is increasing public concern for privacy and cybersecurity.
- Security and privacy must be first-class concerns of new programs from day one, in both specifications, source code and deployment. Methods and tools have to support this.
- However, legacy hasn't gone away: tools that are able to analyse the legacy code base, and find vulnerabilities and unwanted behaviour are needed, as are those that automatically circumvent or mitigate such vulnerabilities.
- Liability and certification in IT systems are crucial to the advance of security and privacy aspects. Security and privacy certification must become mandatory. Certification and liability reinforce one another. Regulation must play its role.
- Privacy and cybersecurity, hence IT sovereignty, which drastically impacts political sovereignty, depends on having control over hardware and software.
- Cybersecurity, privacy and blockchain specialists are direly needed, including for cyber warfare.

Key recommendations

- Invest in research on methods and tools to have security, including privacy, as a first-class citizen during the development of new IT systems (including e.g. quantitative security metrics/properties), find vulnerabilities in existing IT systems (e.g. with static analyses on source code and behavioral analyses at runtime), and automatically prevent or mitigate them (e.g. with automated refactoring tools and blocking system).
- Broaden mandatory security and privacy, EU-based, certification of IT systems, with several levels, and regulate to make IT systems providers and resellers liable in case of sub-par cybersecurity.
- Invest in research on automated methods and tools, possibly based on AI, for intrusion, attack and privacy breaches detection, and for active cyber defence.
- To reclaim IT sovereignty
  - base the critical parts of IT systems either on understood open-source software and hardware, or on EU-made, trustable because audited, hardware and software;
  - train more cybersecurity specialists and develop through education a culture of cybersecurity awareness in the greater workforce and public.
- Invest in post-quantum cryptography (PQC) in order to have EU-designed and EU-validated quantum-resistant encryption schemes.
- Invest in cyber warfare capacities, both for defence (resilience) and for counterstrike (deterrence).
- Harden both the hardware part and the software part of critical infrastructures and supply chains (utilities, transports, health, etc.) against cyber attacks, with reinforced cyber defences, monitoring and resilience capabilities (including redundancy).
- In addition to the above measures, harden edge and IoT devices against side-channel attacks and reverse engineering.

Olivier Zendra is a tenured computer science researcher at Inria, Rennes, France.

Bart Coppens is a part-time assistant professor and a post-doctoral researcher in the electronics department of Ghent University, Ghent, Belgium.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2023.

Cite as: O. Zendra and B. Coppens. The race for cybersecurity. In M. Duranton et al., editors, HiPEAC Vision 2023, pages 128-129, Jan 2023.

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

© HiPEAC 2023

Many information technology (IT) systems still lack (cyber)security. The EU has much to lose from having poorly secured IT systems, and much to gain from secure ones. The good news is that we can do it, in Europe. Here is how.

From cybercrime to cyberwarfare, nobody can overlook cybersecurity any more

By OLIVIER ZENDRA and BART COPPENS

After decades of apparently low-intensity cyberattacks, during which security was not really thought of in most information technology (IT) systems, recent years, including those marked by the COVID-19 pandemic, have brought a flurry of well-organized, larger-scale attacks that have caused billions of euros of damage.

Such attacks have been made possible by the plethora of IT systems that have been produced with no or low security, a trend that has further increased with the rise of ubiquitous computing, with smartphones, the internet of things (IoT) and smart-* spreading everywhere with extremely low control.

However, although the current situation in IT systems can still be considered as critical and very much working in favour of cyber attackers, there are paths to massive but achievable technical improvements that can lead us to a much more secure and sovereign IT ecosystem, along with strong business opportunities in Europe. Furthermore, when it comes to cyberwarfare, improved cybersecurity is now the first line of defence, as the conflict in Ukraine has shown.

Key insights

- Cyberattacks are ever increasing, and the cost of damage caused by lack of cybersecurity is soaring.
- Cybercrime has become an industry, with tools and automation.
- Cyberwar has become prevalent, as shown in recent conflicts.
- The boundaries between cybercrime and cyberwarfare are blurred in some countries.
- Cyberattacks can be made at relatively low cost, and can cause very widespread and profound, even life-threatening, damage.
- Security must be a first-class concern of new programs from day one, in both specifications, source code and deployment. Methods and tools have to support this.
- However, legacy hasn't gone away: tools that are able to analyse the legacy code base, and find vulnerabilities and unwanted behaviour are needed, as are those that automatically circumvent or mitigate such vulnerabilities.
- Liability and certification in IT systems are crucial to the advance of security aspects. Security certification must become mandatory. Certification and liability reinforce one another. Regulation must play its role.
- Cybersecurity, hence IT sovereignty, which drastically impacts political sovereignty, depends on having control over hardware and software.
- Cybersecurity specialists are direly needed.

Key recommendations

- Invest in research on methods and tools to make security a first-class citizen during the development of new IT systems (including e.g. quantitative security metrics/properties), find vulnerabilities in existing IT systems (e.g. with static analyses on source code and behavioural analyses at runtime), and automatically prevent or mitigate them (e.g. with automated refactoring tools and blocking system).
- Broaden mandatory, EU-based security certification of IT systems, with several levels, and regulate to make IT systems providers and resellers liable in case of sub-par cybersecurity.
- Invest in research on automated methods and tools, possibly based on artificial intelligence (AI), for attack and intrusion detection, and for active cyber defence.
- To reclaim IT sovereignty:
  - base the critical parts of IT systems either on understood open-source software and hardware, or on EU-made, trustable because audited, proprietary hardware and software;
  - train more cybersecurity specialists and, through education, develop a culture of cybersecurity awareness in the greater workforce and public.
- Invest in cyberwarfare capacities, both for defence (resilience) and for counterstrike (deterrence).
- Harden both the hardware part and the software part of critical infrastructures and supply chains (utilities, transports, health, etc.) against cyberattacks, with reinforced cyber defences, monitoring and resilience capabilities (including redundancy).
- In addition to the above measures, harden edge and IoT devices against side-channel attacks and reverse engineering.

Cyber threats are real and costly

Barely a few years ago, cybersecurity was, if not unheard of, at least not on the minds of many people, including policy makers. IT systems seemed to be working, attacks only seemed to target "others"; in short, cybersecurity was a non-pressing matter, hence often non-existent… Since then, cyberattacks have made headlines: the influence on the 2016 US Presidential elections; the "NotPetya" ransomware attacks in 2016-2017 with losses estimated to US$10 billion [54][8], and the "Wannacry" ransomware attacks in 2017 leading to losses estimated to up to US$4 billion [53][9].

However, if awareness has indeed risen, it is still true that most of the world, including Europe, has not yet fully awakened to the cybersecurity aspect of IT systems.

Threats are numerous: malware in all its forms (spyware, ransomware, Trojan…), sniffing, spoofing, "man-in-the-middle" attacks, backdoors in hardware or software, etc. They are also spread across most if not all IT domains, ranging from the simplest, cheapest IoT devices to the more expensive smartphones, cars, planes, banking systems, air traffic control systems, etc. They know no frontiers, as IT information can easily cross (most) frontiers in the world, and know no delays thanks to the quasi-instantaneousness of information transmission.

Cyber threats used to be considered as being limited to hacking, which itself was seen as an uncommon activity of geeky-nerdy underground teenagers (see for example the 1983 movie Wargames [10]). Now, there is plenty of evidence that this has morphed into "industrial-scale" activities, with states using cyber warfare units to perform state-sponsored ransomware attacks even in relatively peaceful times [33], and organized cybercrime groups also having their professional hacker teams. Meanwhile, cyberattacks are crucial to modern warfare tactics, with the Russia-Ukraine conflict offering numerous examples [41]. Even a period of global pandemic such as that of COVID-19 did not create any truce on the cybersecurity front; on the contrary, "[c]ybercriminals [were] developing and boosting their attacks at an alarming pace, exploiting the fear and uncertainty caused by the unstable social and economic situation created by COVID-19." [11]

Security company McAfee estimated [55] that the global cost of cybercrime was over US$300 billion in 2013, US$600 billion in 2018, and US$1 trillion in 2020. These figures include direct costs such as cyber-

security expenses and paying for damage, but also "hidden costs [of cybercrime] that organizations may not be aware of, such as lost opportunities, wasted resources, and damaged staff morale." [55]. Most of this sharp increase does not come from better reporting but clearly from cybercrime expansion.

A very striking and telling example of the cybersecurity issues society faces is the sharp rise and spread of ransomware attacks, which has even led to the creation of a new term: "RaaS". As explained in [26], "Ransomware as a Service (RaaS) is a business model used by ransomware developers, in which they lease ransomware variants in the same way that legitimate software developers lease SaaS products. RaaS gives everyone, even people without much technical knowledge, the ability to launch ransomware attacks just by signing up for a service".

Statistics about ransomware can get one's head spinning. According to [25,27], 90% of financial institutions have been targeted by ransomware attacks. More than 68,000 new ransomware trojans for mobile were found in 2019 [25,28]. The consequences of ransomware attacks are major. The cost of ransomware attacks surpassed US$7.5 billion in 2019 [25,31]. The average downtime a company experiences after a ransomware attack was 21 days [25,30] in Q4 2020, an 11% increase from Q3. As found in [25,29], 80% of victim companies who paid a ransom experienced another attack soon after, while 46% got access to their data but most of it was corrupted; 60% experienced revenue loss and 53% stated their brands were damaged as a result; and, even worse, 29% respondents stated their companies were forced to remove jobs following a ransomware attack. Lists of affected companies and institutions flourish on the internet [32], but these may be only the tip of the iceberg (or, shall we say, ransomberg).

The healthcare sector has become one of the favourite targets of ransomware attacks, even at the most critical moments of the COVID-19 pandemic. The Health Sector Cybersecurity Coordination Center (HC3) of the US Department of Health & Human Services found [33] that 34% of healthcare organizations had been hit by ransomware in 2020; 65% that were hit by ransomware in 2020 saw cybercriminals succeed in encrypting their data in the most significant attack, and 93% of affected organizations got their data back but only 69% of the encrypted data was restored after the ransom was paid. [25,31] found that in 2020, 560 healthcare facilities were affected by ransomware attacks in 80 separate incidents, while [25,34] stress that in September 2020 alone, cybercriminals infiltrated and stole 9.7 million medical records in the US. Overall, ransomware attacks on US healthcare organizations cost US$20.8 billion in 2020 [35]. Of course, attacks were not limited to the US, with for example a full-scale attack impacting most of the Irish health system [33,36], and many more being reported on hospitals all across Europe.

Another telling, recent example is the Log4j vulnerability [37] unveiled in December 2021 that could give attackers full control of a system and had cybersecurity stakeholders scrambling [38]. A logging utility for Java applications, Log4j is widely used, being a (small) part of numerous products as a common off-the shelf tool that many use without paying too much attention to it.

The potential magnitude of the vulnerabilities was thus paramount, but not fully understood at the beginning. Fortunately, the worst did not happen because mitigating factors existed [37], such as the fact that not all versions were affected, or that only specific configurations enabled tampering with the affected versions, or the fact that, despite integrating the affected application programming interface (API) of Log4j, some systems were using another non-vulnerable implementation of the API, etc. Yet this could be considered as a free warning of what could have happened when a seemingly unimportant, known but very widespread component is vulnerable.

This should further draw attention to the need for security inside the IT industry supply chain, because (unintended) vulnerabilities or (malevolent) backdoors put in systems though the supply chain can allow massive-scale attacks.

The rise and spread of supply-chain attacks

Indeed, the first category of such supply-chain attacks is that targeting digital supply chains, i.e. supply chains used to provide software and software dependencies. Many modern software projects do not create every line of code from scratch: they depend on a multitude of other projects and frameworks. This makes total sense: it means that developers can create software in their own field of speciality, while at the same time drawing upon others' specialities. This makes software development *faster and cheaper*, as one can rely on pre-existing code bases for many features; but also *more secure*, as an app developer need not hold several PhDs in cryptographic engineering and related fields in order to create a secure network connection for their app: they can rely on the expertise of countless of specialists by simply importing external code.

However, this also creates a new attack surface, as attackers can now try to insert malicious changes into commonly used frameworks. These malicious changes then automatically get pulled in by the victims who update their dependencies to the most recent version (which is in and of itself necessary to stay secure, because the most recent version will contain the most recent fixes for security vulnerabilities). For example, contributors to the Python Package Index, a large software repository containing code which is used by developers all around the world, have been targeted in phishing attacks. Attackers could then use the phished credentials to upload malicious versions of the packages maintained by the phished users [56].

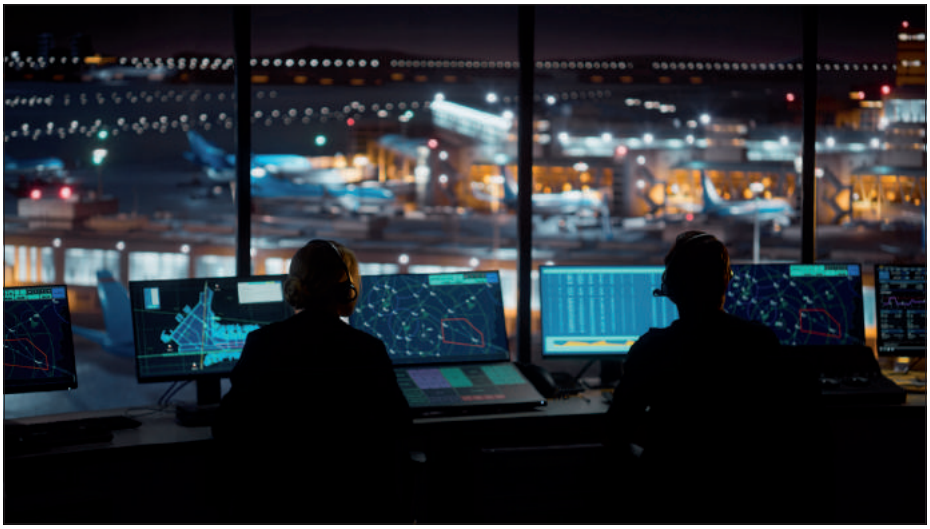
While external attackers are the most common source of this kind of supply-chain attack, there is also the possibility of a legitimate developer going rogue and inserting malicious changes themselves. An interesting recent example is that of a developer of a JavaScript library. When the war between Russia and Ukraine started, this developer inserted changes in their own code that would detect if the code was deployed on a computer with an IP address located in either Russia or Belarus; if so, it would wipe files from that machine [57].

Another famous example is the 2020 United States federal government data breach [39], which included the famous SolarWinds attack [58]. This series campaign included a supply-chain attack on Microsoft cloud services, which allowed an avenue for attackers to breach victims who had purchased these services through a reseller, and a supply-chain attack on SolarWinds' widely used Orion software. Multiple weaknesses in other products provided attackers with further access.

Due to the sensitivity and relevance of the targets as well as the length of time (8-9 months), this cyber-espionage incident is regarded the worst in US history. At least 200 businesses may have experienced data leaks as a result of the hack. Among the affected international organizations were NATO, the United Kingdom and United States governments, the European Parliament, Microsoft, and more institutions [39].

Cyberattacks can also threaten cyber-physical systems, and thus even threaten the concrete physical world, through actual, physical supply chains. One famous early example is the so-called Aurora Generator Test [48], in which in 2007 the Idaho National Laboratory demonstrated how a cyberattack could destroy physical components of the electric grid, namely a 2.5MW diesel generator.

Outside of a lab, the first publicly acknowledged successful cyberattack on a power grid was the cyberattack of 23 December 2015 on the power grid of Ukraine [49] that left 230,000 people without electricity for several hours. This is a striking example of a critical supply-chain cyberattack. This highly sophisticated attack had been carefully planned and prepared for months, patiently gaining access, putting triggers in appropriate places in the Ukrainian power grid. It involved various techniques, trojans and viruses, and even malware based on good old-fashioned, 1990s-era Microsoft Word macros. The latter shows that even ancient attack vectors, which some would have considered to be no longer a threat, can still be used to wreak havoc. It was a complete attack, since not only the systems delivering electricity to the customers were targeted, but also the backup systems in distribution centres, so as to blind the operators trying to restore power. It was even complemented with a distributed denial of service attack on the operators call centres, to prevent customers from reporting the exact extent of the problem and being informed about what was going on. Although formal attribution of the attack remains difficult, Ukraine's intelligence community has said with utter certainty that Russia is behind the attack [49].



Cyberattacks now target critical infrastructure in areas such as energy distribution and air traffic control

The aforementioned "NotPetya" malware, which some consider the "most devastating cyberattack in history" [54], was specifically designed to target "complete energy companies, the power grid, bus stations, gas stations, the airport, and banks" [8], i.e critical infrastructure and mostly supply chains. Among those, world leaders like pharmaceutical company Merck, delivery company TNT Express (the European subsidiary of FedEx), Danish shipping company Maersk, German logistics company DHL, India's largest container port JNPT, as well as food companies and scores of others. Restoring minimal operations took days, and full operations months, to the affected companies.

The EU *Action Plan on Military Mobility 2.0* [42] of November 2022 "proposes measures to enhance the protection of the transport sector against cyber-attacks and other hybrid threats", calling "for a joint civil-military endeavour to ensure a resil-

ient and robust data sharing network with a high level of cybersecurity.", and inviting EU Member States "to ensure the cyber-resilience of the future digital processes and procedures, building on the European Defence Agency's (EDA) work [43] and exploring the possibility of developing functional requirements related to security". There, the matters addressed do not only pertain to supply-chain cybersecurity, but also edge into the cyberwarfare domain.

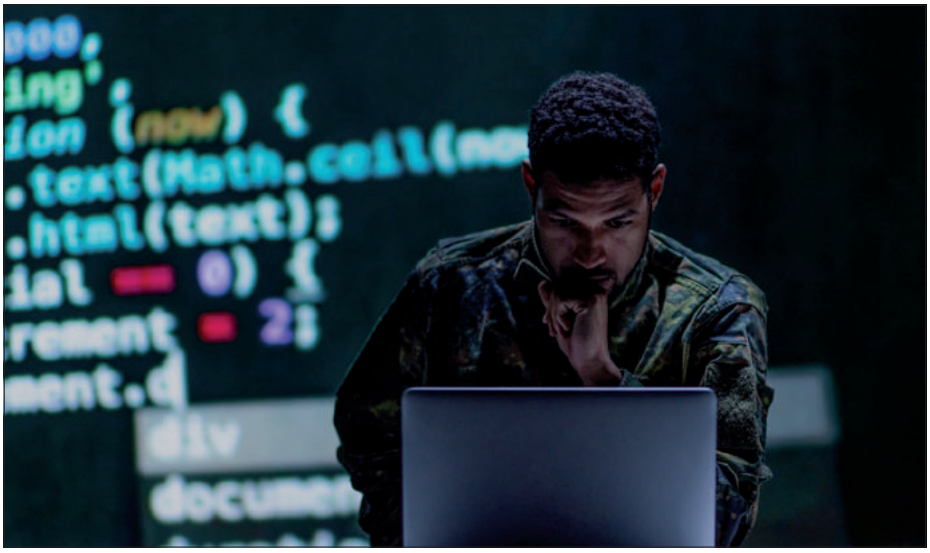
Cyberwarfare is upon us

Indeed, the **cyberattacks on critical supply chains** have now become more commonplace and are a clear part of cyberwarfare. Infiltration of computer networks used to be mostly about espionage (stealing secrets) but is now tied to destructive operations like sabotage and war preparation, as some very concrete examples show.

After the attack on their electricity grid in 2015 [49] mentioned above, Ukraine had reinforced its defences and prepared to defend against new attacks. This preparedness has shown itself to be useful, since the Ukraine power grid was again cyberattacked on 22 April 2022 by a Russian military hacking team [50], but Ukraine was this time able to prevent any damage.

Similarly, as reported [51] on 25 November 2022 by cyber security company Dragos, hacker groups Xenotime and Kamacite, which the FBI says are linked to Russian FSB and GRU secret services, have been found poking at the digital systems of Dutch LNG terminals in Rotterdam, which is a first step in enabling a possible future cyberattack.

It is thus very clear that attacks against critical energy supply chains can be, and have been, prepared, and in some case performed, in a real-world, physical war, to complement missiles and bombs. Given the criticality of such energy infrastructures, even a minor disruption of which can have very strong economical and practical impacts, not to mention the military impact, it is crucial that the EU reinforces its infrastructures and its cyberdefence teams, learning a great deal from allied countries that have already been impacted.



This **prevention and learning process through cooperation** is very effective, and is for example very much followed by the US cyber forces. Indeed, the US military has been conducting "Hunt Forward" missions in the computer networks of allied nations [47] for years. More specifically, since 2018, the United States has sent military personnel to twenty allies in Europe, the Middle East, and the Indo-Pacific. Lithuania announced in May 2022 that a three-month US deployment working on its defence and foreign affairs networks had recently been completed. Most recently [47], Croatia hosted a "thorough and successful hunt", in which malicious attacks on Croatian state infrastructure were discovered and prevented, according to its security and intelligence service. Other examples abound.

Although the majority of US cyber fighters' duties entail combating Chinese and North Korean state hackers, their most persistent opponent is Russia. The recent war in Ukraine provides us with a remarkably telling example. Indeed, in Ukraine, Russia coupled cyberattacks with a full-scale, physical conflict for the first time. In December 2021, a forty-member cyber military team from the United States had arrived in Ukraine to begin detecting Russian cyber fighters online [47]. Western intelligence officials monitored Russian military preparations, calculating that a wave of cyberattacks might precede and facilitate an invasion by damaging communications, power, banking, and government systems and destabilizing the nation prior to the invasion. Russian cyberattacks did hit Ukraine in January 2022. Hackers wrote on the website of the Ukrainian Foreign Ministry, ""Be afraid and expect the worst". Wiper malware hit multiple government sites.

On 24 February 2022, hours prior to the Russian invasion of Ukraine, a cyberattack disrupted a US satellite communications provider assisting the Ukrainian military. Critical infrastructure such as railways and the electricity system were anticipated to be disrupted by computer attacks, but this did not occur because the Ukrainians were better prepared, having learned from previous cyberattacks such as the one on their power grid on 23 December 2015 [49]. Help from their allies and the involvement of the Ukrainian commercial sector also played a role in boosting Ukrainian cyber defences and fending off the Russian cyberattacks. Ukraine's friends, led by the United States, are thus also learning from the conflict. It is providing teams with experience which can be put to use domestically. The US cyber fighters, for example, detected malware they had previously found when deployed in Europe in a government institution in the United States [47]. Since the invasion of Ukraine, offensive cyber actions have also been taken against Russia, although officials have not disclosed details.

Countries such as the United Kingdom, Germany, and France, which have their own cyber security expertise and personnel and are therefore more independent, may not experience such US deployments.

Indeed, even US allies are hesitant to allow the US to search sensitive government networks. Edward Snowden's revelations [52] a decade ago highlighted the extent to which the United States spied on allies as well as adversaries in the cyber world. This left profound traces of mistrust and continues to have repercussions. The fact that Gen. Paul Nakasone, who commands the US military's Cyber Command, also commands the National Security Agency, America's largest spy agency, can be a source of concern for allies, despite the fact that this dual command also provides a greater understanding and anticipation of potential cyber adversaries' actions. Trust has to be built that the US military are there to assist, not snoop, and relay their discoveries so that local partners can expel hackers.

Such trust should be easier to build on a EU level, although EU countries retain their sovereignty in cyberwarfare-related matters, as they do for conventional warfare. The process may thus be long and bumpy. Nonetheless, various levels of cooperation, with more or less tight integration, can be envisioned, and should thus be encouraged by EU.

The cyberwar in Ukraine also directly impacts the EU. On November 23, 2022, the **European Parliament's website** may have been subjected to its most sophisticated cyberattack ever [46]. The attack was a distributed denial-of-service (DDoS) attack, in which massive amounts of traffic are sent to servers in an attempt to bring them to a halt and make it impossible or excruciatingly slow to respond to users. This attack occurred just after the European Parliament voted to designate Russia as a state supporter of terrorism for its strikes on civilian targets in Ukraine. Killnet, a Russian-affiliated hacker organization, has claimed responsibility in a message posted to its Telegram channel. DDoS attacks are routinely used by hackers to cause chaos, and they are rumoured to be favoured by Russian hacking groups such as Killnet in the struggle against supporters for Ukraine in the war. This is a striking example of how cyberspace has become an extension of the physical fighting grounds opposing countries or blocks of countries, and how the EU is exposed to it.

As cyberattacks can impact the physical world, as we have seen for attacks on power grids, they can also be used to **directly target military assets**. The US Air Force (USAF), for example, is taking this issue very seriously. Since 2016, its Defense Digital Service has set up large-scale white-hacking competitions to find vulnerabilities in its fighter jets. As [45] reported in 2019, hackers simulated cyberattacks that could be made by opponents infiltrating the numerous suppliers of the Air Force. In doing so, they "found a mother lode of vulnerabilities that – if exploited in real life – could have completely shut down the Trusted Aircraft Information Program Download Station, which collects reams of data from video cameras and sensors while the [F-15] jet is in flight."

Even more shocking was the fact that several of these vulnerabilities had been identified and documented in previous evaluations but had not been fixed. However, the USAF, despite its huge budget, is similar to many institutions and businesses around the world in that costs and time must be limited, sometimes at the expense of thoroughness and quality, and most frequently at the risk of cybersecurity. Such sobering experimental evaluations are crucial in providing feedback on vulnerabilities that make systems more susceptible to cyberattacks, so helping in the correction of these defects.

However, attempting to rectify errors after a system has been constructed can be quite challenging. Legacy systems may be significantly out of date, and their providers may no longer have the necessary competence. Some systems are designed to be extremely difficult to modify. Some suppliers are protective of their intellectual property rights and do not want their systems to be accessed and investigated, let alone modified, which could result in liability-related legal complications. Therefore, it appears simpler for the USAF [45] to leverage the knowledge gathered from these white-hacking competitions to enforce a greater level of cybersecurity in the IT systems it acquires from suppliers.

The EU, or at least individual EU countries, should also take inspiration from this experience, and hold such white hacking events to further harden their critical, especially military, systems. Such events can very usefully complement the inside work already done by the various bodies (like DGA, the military procurement agency in France,) in charge of evaluating the quality of military materiel.

Another, extended form of cyberwarfare is **cognitive warfare** [44]. Although by far not limited to the cyber world, cognitive warfare thrives in it. The battlefield in cognitive warfare is the mind. Cognitive warfare is in a way the apex of influence, or propaganda (to use an old-fashioned word that somehow faded a bit after the collapse of the USSR). The objective of cognitive warfare is to influence people's thoughts and behaviour. It seeks to sow FUD (Fear Uncertainty and Doubt), to introduce conflicting narratives, to polarize opinions, to radicalize groups, and to motivate them to acts that can disrupt or fragment an otherwise cohesive society. It aims at influencing the thoughts and actions of individuals and groups in favour of the tactical or strategic objectives of an aggressor. In its most extreme form, it can divide a community to the point where it lacks the determination to oppose an adversary's goals, taming the community without any use of force.

Although cognitive warfare is not strictly to the cyber world, it is greatly facilitated by social networking, social messaging, and mobile gadgets. The widespread use of social media and smart device technologies render EU societies particularly vulnerable to this kind of unobtrusive yet very powerful attack.

Cybersecurity: let's build defences

The above examples clearly show that as a society all our systems, both industrial-scale and personal ones, must be as secure as possible. The stability and continuity of our daily activities, be they private or professional, and even our lives, depend heavily on the secure and continuous operation of IT systems. Until very recently, security was not a top priority for those who design and implement IT systems.

This has led to the current situation, with scores of vulnerable IT systems being used, and many still being developed with poor security.

Cybersecurity can be seen as a defence system, and has often been compared to medieval castles. Although it is true that some cyber defences can be nested like medieval castle defences were, the metaphor is inappropriate for whole IT systems. Unlike medieval castles, in which attackers must pass all the defences successively (climb the slope, and cross the outer moat, and climb or breach the outer wall, and the same for the inner wall, and for the dungeon), IT systems are in fact much less nested. They are more layered, or stacked, meaning that any breach present in either the hardware, or the operating system, or the execution environment, or the application, could be exploited by an attacker to gain information or control from the system. In that sense, one hole in IT systems may be enough. This is similar to the proverbial "forgotten, concealed postern" in a castle, or to a secret tunnel. Of course, there still may be some compartmentalization, in which case not the whole system falls but only part of it (a bit like one castle among several). But overall, IT systems and their defence appear rather more vulnerable in principle than the iconic medieval castle.

However, solutions exist to this grim situation, some of them specific to one layer of the IT system, others applicable to several layers. Many defence techniques exist, and can add their "stone" to the security walls of IT systems.

One way to create a "stronger stronghold" for IT systems is by having a smaller attack surface, by basing everything on a smaller yet highly secure base, like the trusted computing base (**TCB**) [1]. Research work has clearly shown that "From the security point of view, the monolithic OS design is flawed and a root cause of the majority of compromises." [22], and that **microkernels** make it easier to have more secure operating systems (OSes), even more so when they are verified. Indeed, OSes must have some parts that execute with the highest level of privileges, in kernel mode. But being software, kernels are also flawed. The bigger the kernel, the more flaws can be exploited in kernel mode to gain access to all information on the system. By reducing this attack surface to a bare minimum, the risk is mitigated. By reducing this kernel code to a bare minimum, it becomes easier to check it, secure it, or even verify it formally.

At the application and OS levels, a simple solution would be to use end-to-end cryptography, in which Europe is strong [40], in order to provide better protection for user data with regard to attacks, at storage level (encrypted filesystems), memory level and communication level. Recent OSes make it easy to encrypt files or the whole filesystem, and dedicated hardware support has made the cost of this largely unnoticeable in practice. At application level, several messaging or conferencing applications have introduced end-to-end crypto, some of them even putting a strong emphasis on this aspect (e.g., Olvid [23], Signal [24]). This need has only grown with COVID-19-induced teleworking, and the flurry of cyberattacks it has allowed [11].

Many more techniques exist. In the remainder of this article, we focus on a few techniques that must be supported and promoted since they provide the means for the EU to secure its IT systems and IT ecosystem. We show very promising *technical solutions* that make it possible to **find and fix vulnerabilities** in existing IT systems, to **strengthen** the security and **resilience** of IT systems, and to **express security properties** in order to be able to *verify* the security of existing systems and to *produce* new ones that are much more *secure* if not devoid of any vulnerability. However, if we do want to increase the security of our systems, we also need to *motivate* all actors to act in their best shared interest. This may imply **regulations**, based on **liability** in the face of the law and on **certification** processes. Finally, cybersecurity and IT sovereignty, hence simply **sovereignty** for the EU, depend on **trustable and auditable hardware and software**.

Finding and fixing vulnerabilities in existing systems

Since it is now infeasible to recreate all systems from scratch and redesign and rewrite them with perfect security, we will have to live with the legacy of our existing systems for a long time to come. As a result, these systems will need to be made secure to ensure that users and companies can access and store their private data for years to come.

In some cases, pragmatism can help us move forward on this path. As an example, consider the now commonplace technique of address space layout randomization (ASLR), which defends against memory-related vulnerabilities in software. Its core idea is that software will contain many bugs that can lead to exploitable vulnerabilities, but that attacks against these vulnerabilities were made exceedingly easy by the fact that program code and shared libraries were located on fixed addresses in memory. ASLR randomizes the locations of program code and libraries in memory, which makes such easy attacks fail. ASLR is a cross-layer technique, since it can be applied to the memory of application code/data, the execution environment (e.g. virtual machine, or VM) code/data, and OS code/data. Thus, applying ASLR to all systems increases the cost and difficulty of mounting attacks. While ASLR cannot prevent all possible attacks, users are nonetheless safer thanks to it.

This goes to show that in the cases where we cannot rigorously ensure the required security properties, we can still increase the overall security of systems through other means. In particular, **we need to have techniques to find, and fix or mitigate, security vulnerabilities in legacy code bases**; or techniques to at least isolate, thanks to containerization, these potentially insecure legacy code bases in hypervisor-like infrastructure, while at the same time controlling their exchanges with the rest of the system. While none of these techniques will be able to guarantee that programs are *completely* secure, each of them can lower the impact of inadvertent mistakes, and can raise the bar against specific attacks. The more of these techniques that are combined, the greater the security of the resulting system will be.

Even in large-scale systems, which are hard to deal with globally, vulnerability-finding techniques can perform a dual function. While an automated tool being unable to find security vulnerabilities in a certain amount of time does not constitute proof of the security of a system, if such an automated tool *does* find a security vulnerability, that vulnerability serves as constructive proof of the system's insecurity. Such security vulnerabilities (found automatically) can be passed as actionable items to the appropriate engineers to solve. This is a much more tractable task than trying to prove the security and correctness of an entire, large-scale system. For example, operating system kernels are complex, security-critical systems that are hard to analyse. Special-purpose techniques can be developed that target different kinds of bugs that can have a security impact in operating systems [12,13]. These analyses can still be improved in terms of how many such bugs they find in certain hard-to-analyse contexts. Furthermore, as such tools only target specific classes of bugs, more such tools need to be developed in order to find so-far under-reported classes of bugs.

For cases in which the systems still contain vulnerabilities which neither automated tools nor manual code analyses find, we need to have ways to mitigate the impact of these remaining vulnerabilities. This is typically done by targeting the ways in which an attacker typically exploits such vulnerabilities. An example of this was already touched upon in the context of ASLR: if typical attacks use hard-coded memory locations, making memory locations vary between executions will make attacks harder. Similarly, attackers can try to redirect the execution of the program in a way which the original developer did not write in the source code of the application. Then a type of defence, called control-flow integrity, inserts checks that verify that functions are executed only from the calling context of functions of which the developer intended this [14]. More such defences need to be researched and developed.

Diversity is a means of resilience against attacks

One way of mitigating and protecting against security vulnerabilities is introducing diversity. This is akin to avoiding monocultures in crops: having a more diverse gene pool in crops can make fields more resilient to diseases and infections. This was already alluded to in the context of ASLR: if every execution of a program has a totally unpredictable memory layout, it is harder for an attacker to create a single exploit that works against all these different memory layouts. An attacker would need to carefully craft an exploit that works around these limitations, or there would need to be a way for the attacker to gain knowledge of the exact memory layout.

Diversity can be introduced at many different, if not all, levels of an IT system: hardware, operating system, execution environment, application level. This can be done by having different implementations, by generating different versions at compile time [15], by randomizing the programs at install time[16], by randomizing them at load time such as with ASLR, or even by randomizing them during the execution of the program [15,16].

Diversity has even more security-related applications. For example, multi-variant techniques take their inspiration from reliability in critical systems such as aeroplanes in which multiple different implementations are compared against one another [18]. By executing (specially chosen) diversified instances of the same application, feeding these the same inputs, and then comparing their outputs, some classes of attack can be prevented from the fact that these attacks will impact these instances in a different way, which can be detected[19].

An application of diversity in a wider sense can be found in the context of security updates. When a developer discovers and fixes a security vulnerability and releases an update, users don't typically apply this update immediately. However, attackers can still compare the original application which contains the vulnerability with the just-released updated version of the application in which the vulnerability has been patched. Based on this differ-

ence, attackers can easily create attacks against the users who do not yet have the update installed [20]. This is sometimes called patch *Tuesday / exploit Wednesday*, as Microsoft typically releases their (security) updates on a Tuesday, after which attackers can try to exploit the unpatched users the day afterwards. A mitigation to this could consist of making the original version and the updated version differ more, that is, making the versions more diverse from each other. This will slow down the analysis and attack-generation by the attacker, allowing more users to apply the security update [21].

Some of these diversity techniques are already widely adopted, but not all. One reason that some proposed techniques are not yet used in practice is that they have too high an overhead to be applicable in most practical situations, or still have other limitations. Furthermore, as no technique can prevent every possible vulnerability, it is important that more such techniques are investigated and can be applied in practice. **This means researching new diversity techniques that are both efficient and effective, as well as making existing technologies more widely applicable,** by raising their maturity levels such that they can be adopted at large, rather than existing as academic prototypes.

Non-functional security properties must be included as first-class citizens in new IT systems

In addition to taking care of vulnerabilities in existing legacy systems, it is of the utmost importance to include security as a first-class citizen when building new IT systems. Far too often, security is not considered upfront in programs at the design and implementation stages, but only as an afterthought. This situation absolutely must change. The non-functional property that security is, or more precisely the set of non-functional properties that pertain to security in its various aspects, must be present in programs, in the minds of designers and implementers, from the very beginning of the creation of an IT system.

An example of a security property would be the level of threat by interception and decryption of communications in a given environment (e.g., known by its location in a military conflict). Another example would be, say, the strength of a cryptographic algorithm. An IT system could adapt its operation depending on the threat level, sending unencrypted or lightly (hence cheaply in term of time and energy) encrypted information in safe cases, choosing a stronger encryption algorithm in higher threat situations, or even avoiding transmission altogether if the level of threat is above the strength of the available algorithms.

To get to this point, security properties must be treated as first-class citizens, which means that they have to be expressed as clearly and as explicitly as the functional properties (i.e., what the program does, its algorithms), and that designers and developers should be able to reason about them, query them, manipulate them, just as for the functional aspects of the program. It is only by having security interleaved in all the fibres of the IT system that it can be solid.

To this end, first, security properties must be present in the **specifications** of the IT system from the beginning. Designers should be able to express the level of security they want, for the various facets of security (interception of communications, side-channel attacks, fault injection, malware, etc.), and reason about them. **Security contracts** must be present in the system that express and guarantee security at module or component boundaries. Developers should be able to pick up off-the-shelf **modules** or components knowing the security levels they provide for specific facets or security, and plug them in as part of the security continuum of their system.

Programming languages are also not all equal in terms of security. Many programming languages tolerate sloppy programming, where code that looks reasonable at first sight may in fact contain major vulnerabilities. Some, e.g., C/C++, tend to be harder to master and more error-prone, making it difficult to find bugs like security issues. For a compendium of programming language vulnerabilities, see the work by ISO/IEC TR 24772 Programming languages – *Guidance to avoiding vulnerabilities in programming languages* [2]. Other languages should be promoted: those that have stricter rules, more safeguards, and make it easier to develop more secure code. The Rust programming language, originally developed by Mozilla, is one example that makes many aspects of memory management and memory safety explicit in its language constructs, making it harder to leave room for security glitches that could be exploited malevolently.

Automated or semi-automated **tools** must be able to rely on these specification and security contracts **to verify security**, to prove security properties, to provide strong guarantees about the quality of IT systems with regard to security, at both specification and code level. Notable examples that can help in this endeavour in the line of program verification include Frama-C [3], Coq [5] and Ada SPARK's Discovery toolset [4]: those tools operate on the premise that the source code must conform to some formal specification.

Although **security by design** of whole IT systems seems a perfect answer to security issues, the current approach is often more limited and pragmatic, with only a limited part of the IT system being trusted. This trusted computing base (TCB), as mentioned above, comprises the system hardware, firmware and software components whose combination is intended to provide the system with mechanisms for a secure environment. The idea here is that verifying this, either automatically with verification tools or manually by human examination, on a small set of hardware and software is a more tractable and less costly task than doing it on the whole system. However, verification of large pieces of real-life software is already doable, as proven by recent research-level successes like the CompCert compiler [6] and the verified parts of the seL4 microkernel [7].

Blockchain: a technology that is still maturing

Blockchain has frequently been hailed as a disruptive, enabling, infrastructure technology for cybersecurity.

Blockchain is a potentially revolutionary technology, frequently referred to as the "trust machine," and it represents the first implementation of decentralized trust at present. According to Gartner [59], "Blockchain is clearly heading out of the Gartner Hype Cycle's Trough of Disillusionment, and now is the time to act." [60]. Nonetheless, it is worth examining the blockchains and their current status with attention.

The blockchain was born in 2009 with the emergence of Bitcoin, but its true potential was shown a few years later with the emergence of Ethereum and the concept of the "smart contract": not just money, but any digitized object may be exchanged, and this can be done according to sophisticated rules. In the past decade, blockchain technology has moved from digital currency (Blockchain 1.0) to smart contracts or programmable blockchains (Blockchain 2.0) to a vision of economically and legally sophisticated forms of decentralized collaboration (Blockchain 3.0).

Blockchains are a type of (digital) distributed ledger. Blockchain enables two or more individuals, corporations, or machines that may or may not know each other to exchange value in digital settings – in a monetary transaction, information exchange, or other asset exchange – without the need for a service provider. The recorded event could be a monetary transaction, as is generally recognized due to the initial deployment of blockchain for Bitcoin, but it could also be any information exchange.

Schematically, the functionality that a blockchain provides is confirming the pseudonymous identity of the participants, validating that the participants own the information/assets they wish to exchange, authenticating and approving the transaction, and recording the transaction information to the ledger, a copy of which is independently updated and held by each node on the network.

Blockchain integrates technologies and techniques such as distributed digital ledgers, encryption, immutable records management, asset tokenization, and decentralized governance to capture and record the data that network participants require to engage and transact. No intermediaries, like banks, validate and safeguard the transactions. Thus, the requirement for a central administration is eliminated.

Thus, the major purpose of blockchains is to facilitate user-to-user transactions in an environment where participants are "equal peers" and do not *a priori* trust one another. Instead of going via a central control organization that acts as a guarantee or "third party," the security of exchanges is achieved by maintaining a shared register that preserves the history of transactions. Using a computer protocol, each user keeps their own copy of the ledger, and each new transaction is validated locally by the user before being recorded. Due to the fact that each user has a local copy, it is evident that the ledger will be very accessible, as it is redundant on a large number of nodes.

The use of simple cryptographic techniques makes it extremely challenging, even locally, to alter the ledger's content without rendering it incorrect in the eyes of the other participants. To do so, one would have to reconstitute the contents of the registry from the modified record and corrupt the other participants into adopting the regenerated contents. As this is deemed almost impossible or ridiculous, then inadvertent or intentional modification remains nearly impossible, ensuring the ledger's integrity.

The blockchain revolution is technically predicated on the concept of "public verifiability", i.e., a technical system that allows anyone to independently verify the system's state's accuracy. In a distributed ledger, every observer can verify that each activity affecting the system's state is valid, that is to say, in line with the rules that govern the system, which are recognized by all. This verifiability is only achievable if the information required for validation is accessible, verifiable in a reasonable amount of time, and if the actions taken are observable.

Formally, a complete blockchain consists of five components.

1) **Tokenization**: Value is traded in the form of tokens that can represent a range of asset kinds, including monetary assets, data units, and user identities. Smart contracts may be used to program token usage.
2) **Distribution**: Blockchain users connected on a distributed network operate nodes (computers) that execute a program that enforces the blockchain's business rules. Nodes also maintain a complete copy of the distributed ledger, which is automatically updated when new transactions occur.
3) **Decentralization**: No single entity controls the majority of nodes or defines the rules. A consensus process checks and authorizes transactions, removing the need for a centralized intermediary to control the network. The manner in which governance is implemented can be modified.
4) **Encryption**: To record data safely and semi-anonymously, blockchain employs technologies such as public and private keys.
5) **Immutability**: Transactions are cryptographically signed, timestamped, and appended sequentially to the ledger. Changes to the record are not permitted unless all participants concur.

Wallets allow users to engage with the blockchain. A digital wallet gives its owner a pair of keys (private and public). Every record is signed using the issuer's private key. If the record is valid, other users can read it using the issuer's public key and accept it. A genuine digital signature provides the reader with strong evidence that the record was created by its issuer (authenticity) and has not been altered (integrity). Digital signatures also include a non-repudiation property, which means that the signer cannot declare that the signature is false. Each wallet is paired with a group of tokens, the wallet contents. Using the private key, only the owner of the wallet has the authority to transfer the related tokens. The digital wallet tied to a smart contract can also function as the contract's identifier within the system. In essence, users pay for the contract's execution.

Blockchain also holds the potential of improved transactional performance. We used to rely on sluggish, expensive, analogue-based techniques to verify identity and legal standing in economic transactions. Blockchain eliminates these methods. Equally crucial is that blockchain enables faster and more diverse transactions, both in terms of type and size, than are achievable with conventional centralized systems. For years, businesses have relied on centralized infrastructures, such as payment systems, insurance, delivery and logistics services, and governments, in order to execute commercial transactions and manage risk. However, these systems were not meant to handle the complexity, volume, and scale of the machine-to-machine transactions enabled by digital platforms. Without employing an intermediary that collects data on each side and gets a share of the transaction's value, businesses need a new method for handling new digital assets and interactions. Blockchain offers a solution.

Once blockchain becomes a mature technology, its benefits will have a profound effect on many facets of society. In constantly changing societal, industrial, and economic interactions, direct exchanges between users based on self-established, evolving, transparent, and verifiable rules at any time might very well open up new markets and services while simultaneously improving the quality of the services and goods exchanged.

However, the genuine appeal or utility of blockchain must be demonstrated in real-world applications before it can be compared to other more proven or specialized alternatives. Many of the proof-of-concept projects created in recent years attempted to tackle problems with blockchain technology that were not necessarily the best fit for demonstrating its actual utility. Undoubtedly, the massive hype surrounding blockchain has prompted a lot of organizations to put up demos too rapidly, while others have questioned whether blockchain is now more of a solution than it is a problem in itself. All of this has contributed to an unfairly negative perception of these technologies.

Aside from crypto-currencies and speculative objects, there are other reasons why these technologies have not yet been able to fully display their true potential, including their inherent complexity and a number of challenges [62]. Blockchain technologies are still in their infancy, despite having to meet extremely stringent security requirements. To provide trust, blockchain must operate as a flawless system. Replacing a bank, a notary, or an auditor with a computer protocol running on a network necessitates that the blockchain meets similar security standards as important software systems. This is not the case yet.

Thus, blockchain solutions are still maturing and may eventually underlie new social and economic models. Indeed, blockchain technologies are showing gradual advances. In recent years, there has been a profusion of proofs of concept or trials in sectors such as energy, food, mobility, pollution control, administration, etc., although frequently with modest goals and modest investments.

Utilizing tokenized real-world assets and smart contracts to control them, enterprise applications ranging from aviation maintenance to food safety are beginning to reap the first benefits of blockchain. But despite the fact that blockchain solutions are currently available, the majority of enterprise initiatives only incorporate a subset of the five fundamental blockchain components. In addition, apart from the now famous and occasionally controversial cryptocurrency trading, we have yet to see "killer use cases" that would have to surpass present applications in terms of improving our lives.

Additionally, blockchains experience security issues. Since the "bad guys" experiment with new technologies far quicker and earlier than the "good guys", it takes time for the "good use cases" to catch up, and it takes much more time to apply fraud and security controls [61].

Hence before blockchain can take over the world, numerous significant scientific and technical problems must be addressed.

First, viable alternatives to proof of work must be identified in a sustainable society if blockchain is to make a breakthrough in the sector. Therefore, research naturally focuses on these alternatives. Emerging alternatives to proof of work have interesting design concepts, but they must still be validated. Specifically, the execution of the selection and renewal of validator committees is a significant security concern for these solutions today.

Blockchains must also be able to scale, i.e., to have light replication of information, but in a highly dynamic environment.

There is a need for more complex cryptographic approaches in order to verify the correctness of blockchain transactions without requiring access to their sensitive data, hence protecting privacy and trade secrets.

Another difficulty is the validation of the correctness of smart contracts. Smart contracts are software programs that are performed on a blockchain. Once a smart contract has been added to a block, it cannot be altered or deleted. As with any software, a poorly coded smart contract is susceptible to computer cyberattacks. Its immutability then becomes a problem. This was the situation with the well-known The DAO, which was attacked [63] causing

the loss of a significant amount of cryptocurrency. Numerous industrial applications would also be encoded in the form of smart contracts, making the issue of smart contract security all the more crucial. Fixing smart contracts is obviously a top priority, although it is crucial to remember that the majority of cyberattacks [64] have targeted bugs or weaknesses in the execution platform.

In addition, it is necessary to address the expressiveness of the languages in which smart contracts are written. There is a significant difference between a contract written or comprehended by an attorney and a smart contract. A smart contract is a piece of code written in a programming language (be it a general-purpose one, or a specialized one like Solidity [65]), which is largely unrelated to legal ideas. In the common conception, smart contracts are viewed as a potential method for automatically enforcing a contract in the legal sense. Nevertheless, from this perspective, smart contracts are no more powerful than a simple evidence log. Further study is required to develop smart contract languages that more accurately reflect actual legal contracts.

Scaling up blockchain is also a significant challenge that needs to be overcome, particularly in relation to energy usage. For Bitcoin for example, the consumption per transaction is predicted to range between 10 and 18 kWh [62]. However, exact consumption models are lacking and should be explored to provide an accurate appraisal of blockchain and its numerous incarnations.

The lack of interoperability between blockchains is another major challenge, as the blockchain industry is now fragmented over numerous incompatible platforms and protocols. This makes it extremely difficult to trade value between assets developed on different platforms, as it would require a centralized clearinghouse-operating intermediary. Also problematic is the exchange between apps produced on different platforms.

Monitoring or exploring blockchains remains an issue as well. Often referred to as explorers or scans, blockchain exploration tools (such as Etherscan and Ethereum blockchain explorer for Ethereum, Bitcoin Blockchain Explorer and Block explorer for BitCoin) typically display the blocks generated, the transactions within a block, and the miner who generated the block. However, these are still extremely basic tools. Exploration tools will be required to progress into far more advanced tools, similar to web search engines, with genuine monitoring or auditing capabilities. A good tool that satisfies the specified security standard will be the entry point to the underlying blockchain and will have a significant impact on its adoption. These mining tools will be the true guarantors of the auditability and transparency of blockchains.

The lack of proper software engineering tools at all stages of the design and development cycle of blockchains is a final but very significant problem for blockchain technology. At the design stage, formal methods and tools for the verification and certification of chains, as well as methods and tools for architecture and modularity, are required. The development stage needs methods and tools for multi-level simulation, testing, and benchmarking.

EU research is active in these fields, but it is dwarfed by e.g. American and Israeli competition. Major EU investments must thus be made to keep in the race for the future of blockchains.

Sovereignty depends on trustable and auditable hardware and software

A castle can have the deepest moat and the highest and strongest walls, but they are of no use if an adversary has the key to the secret tunnel or hidden postern. Similarly, if we don't control the parts of our IT systems that are crucial for cybersecurity, we can hardly guarantee it. If backdoors exist in the operating system or even in the hardware we buy, unknown to us, they can be exploited by attackers and it is very difficult to add extra elements of security to counter them. The same is true for development tools, like the compilers that generate the actual executables, or the software libraries used as building blocks to compose programs: being closed source and distributed as binaries, they could embed backdoors in the programs that are produced with them.

Currently, by basing most, if not all, of its activities on IT systems running on proprietary hardware and operating systems not made in the EU, the EU effectively entrusts the providers with the keys to its whole economy and all aspects of its (cyber)security. The situation is barely better for development tools, compilers, and libraries, in which EU production is quantitatively very limited. This is why it is crucial to retain the keys of the castle, which means to retain sovereignty and control over the important hardware and software components for at least the TCB, and hopefully the whole software stack.

The way for the EU to reclaim its IT **sovereignty** is thus to **base the TCB of EU IT systems either on open-source software and hardware, or on EU-made, trustable-because-audited, proprietary hardware or software.**

Liability and certification for IT systems

Liability and **certification** are crucial building blocks needed to mandate the consideration of non-functional security properties in IT systems [40]. Indeed, **adding the legal building block of liability**, introducing the threat of a potential penalty for non-secure systems, provides motivation for system builders and providers to commit effort, and hence money, to having secure systems. More and more voices call for such liability, including penal liability, for cybersecurity neglect [66], in similar way as liability exists for any other business (such as car manufacturers or plumbers). With liability, the extra time of specification and the extra step of verification would be worth taking.

Cybersecurity certification is the technical building block that makes the legal one of liability viable. With such certification, system builders can have their efforts for security quantified, priced and legally acknowledged; purchasers can mandate security based on an independent assessment; and regulatory bodies can outlaw low-security systems. Certification can be based on test suites that must be passed, on verification tools, on development practices that must be adhered to, etc. Cybersecurity certification schemes already exist in the world, like the CMMC (Cybersecurity Maturity Model Certification) in the USA [67] or the Cyber Essentials in the UK [68], and of course the revised EU directive on security of Network and Information Systems (NIS 2 Directive [69]).

With strong cybersecurity liability and certification, companies have the incentive to create secure software, or to keep finding, and mitigating or fixing vulnerabilities. Cybersecurity liability and certification must thus be put in force and reinforced as quickly as possible, at EU and national levels.

Education is a keystone for cybersecurity

Of course, there are not only the technical aspects to consider; education is involved as well in securing society.

First of all, we need to train more cybersecurity specialists. These are needed for researching and developing all kinds of defences, such as those mentioned in this article. As security is a kind of never-ending cat-and-mouse game between attackers and defenders, these cybersecurity specialists are also needed for continuing research into defining systems against the new and as yet unknown attacks malicious actors will come up with in the future.

Secondly, while not all developers need to become cybersecurity specialists, all developers should at least have some basic knowledge and understanding of cybersecurity. This is because while the cybersecurity specialists will develop new techniques and approaches to secure systems, these technologies still need to be integrated in and applied to systems. Application developers should furthermore be aware of the appropriate techniques to create secure systems by design, know how to apply these, and also know which techniques and, equally important, know which coding practices should be avoided when creating secure systems. This basic knowledge of cybersecurity will also enable better interaction between developers and cybersecurity specialists.

Finally, the public at large also should at least have some basic awareness of cybersecurity. Although this should not be necessary in an ideal world (where all systems are secure, and under no circumstances can these systems be abused or circumvented), we unfortunately do not yet live in this ideal world. As such, it is important that the general public is taught some basic cybersecurity concepts.

All this implies that we should invest in cybersecurity education, ranging from educating cybersecurity specialists, through giving all software developers an education into the basics of cybersecurity, to instilling an awareness in the public at large.

Conclusion

(Cyber)security is only as strong as its weakest link, which means that the level of security of an IT system is at the minimum of the level of security of its components. But "the physics of cyberspace are wholly different" to physical work, because of the **worldwide interconnection and instantaneousness of IT networks**, where an attack on a vulnerability in some obscure accounting software in Ukraine can spread to and **paralyse whole worldwide crucial supply chains**. The weakest link thus has to be seen on a worldwide basis.

For way too long security has been forgotten in the interests of better time-to-market and lower prices. However, increased awareness of the **cost of poor security**, coupled with the flurry of attacks in recent years, further increased during the COVID-19 pandemic, in a much more organized way than before, is now making it both *necessary and possible* to reverse this trend, and to take a path towards better security and resilience in IT systems, providing better resilience for economies, vital supply chains and people. Even when full protection cannot be achieved, simply increasing the cost for attackers to perform a cyberattack already has a positive effect, since cybercrime is a business that looks for return on investment, hence steers away from heavily defended targets.

The technical building blocks to this end are within sight in the research community and must be nurtured and pushed forward, so as to quickly mature and then irrigate the whole IT industry. Europe has a key role to play in terms of **IT systems that are more valuable because of their higher quality thanks to higher security and resilience**, by promoting **targeted research**, taking the appropriate **regulatory steps**, and taking the necessary steps to reclaim **sovereignty of the critical building blocks** of its own IT systems.

References

[1] TCB: Trusted computing base: https://en.wikipedia.org/wiki/Trusted_computing_base

[2] ISO/IEC TR 24772 Programming languages – Guidance to avoiding vulnerabilities in programming languages: https://committee.iso.org/sites/isoorg/contents/data/committee/04/52/45202/x/catalogue/p/1/u/1/w/0/d /0

[3] Frama-C: https://frama-c.com/features.html

[4] Ada SPARK's Discovery toolset: https://www.adacore.com/sparkpro

[5] Coq proof assistant: https://en.wikipedia.org/wiki/Coq

[6] CompCert compiler: http://compcert.inria.fr

[7] seL4 microkernel: https://sel4.systems

[8] Petya ransomware attacks: https://en.wikipedia.org/wiki/Petya_(malware)

[9] Wannacry ransomware attacks: https://en.wikipedia.org/wiki/WannaCry_ransomware_attack

[10] "Wargames" movie: https://www.imdb.com/title/tt0086567/

[11] INTERPOL report shows alarming rate of cyber attacks during COVID-19: https://www.interpol.int/News-and-Events/News/2020/INTERPOL-report-shows-alarming-rate-of-cyberattacks-during-COVID-19

[12] Precise and Scalable Detection of Double-Fetch Bugs in OS Kernels. Meng Xu, Chenxiong Qian, Kangjie Lu, Michael Backes, Taesoo Kim. IEEE Symposium on Security and Privacy, 2018

[13] Check It Again: Detecting Lacking-Recheck Bugs in OS Kernels. Wenwen Wang, Kangjie Lu, Pen-Chung Yew. ACM Conference on Computer and Communications Security, 2018.

[14] Control-flow integrity principles, implementations, and applications. Martín Abadi, Mihai Budiu, Úlfar Erlingsson, Jay Ligatti. ACM Trans. Inf. Syst. Secur. 2009

[15] Enhanced Operating System Security Through Efficient and Fine-grained Address Space Randomization. Cristiano Giuffrida, Anton Kuijsten, Andrew S. Tanenbaum. In USENIX Security Symposium, 2012

[16] SoK: Automated Software Diversity. Per Larsen, Andrei Homescu, Stefan Brunthaler, Michael Franz. IEEE Symposium on Security and Privacy, 2014

[17] Librando: transparent code randomization for just-in-time compilers. Andrei Homescu, Stefan Brunthaler, Per Larsen, Michael Franz In ACM Conference on Computer and Communications Security, 2013

[18] N-Variant Systems: A Secretless Framework for Security through Diversity. Benjamin Cox, David Evans. USENIX Security Symposium, 2006

[19] Cloning Your Gadgets: Complete ROP Attack Immunity with Multi-Variant Execution. Stijn Volckaert, Bart Coppens, Bjorn De Sutter. IEEE Trans. Dependable Secur. Comput. 2016

[20] Automatic Patch-Based Exploit Generation is Possible: Techniques and Implications. David Brumley, Pongsin Poosankam, Dawn Xiaodong Song, Jiang Zheng. IEEE Symposium on Security and Privacy. 2008

[21] Feedback-driven binary code diversification. Bart Coppens, Bjorn De Sutter, Jonas Maebe. ACM Trans. Archit. Code Optim. 2013

[22] The Jury Is In: Monolithic OS Design Is Flawed: Microkernel-based Designs Improve Security. Simon Biggs, Damon Lee, Gernot Heiser: APSys 2018: 16:1-16:7

[23] Olvid. https://olvid.io/technology/en/

[24] Signal. https://signal.org/docs/

[25] 81 Ransomware Statistics, Data, Trends and Facts for 2021. https://www.varonis.com/blog/ransomware-statistics-2021

[26] Ransomware As A Service (RaaS) explained. https://www.crowdstrike.com/cybersecurity-101/ransomware/ransomware-as-a-service-raas /

[27] Financial institutions; 90% of them have been targeted by ransomware. https://www.prdistribution.com/news/financial-institutions-90-of-them-have-been-targeted-by-ransomware/3279096
[28] 20 Ransomware Statistics You're Powerless to Resist Reading. https://www.thesslstore.com/blog/ransomware-statistics/
[29] Ransomware: the true cost to business. https://www.cybereason.com/hubfs/dam/collateral/ebooks/Cybereason_Ransomware_Research_2021.pdf
[30] Coveware Quarterly Ransomware Report Q4 2020. https://www.coveware.com/blog/ransomware-marketplace-report-q4-2020
[31] The State of Ransomware in the US: Report and Statistics 2019. https://blog.emsisoft.com/en/34822/the-state-of-ransomware-in-the-us-report-and-statistics-2019/
[32] More and More Companies Are Getting Hit with Ransomware. https://heimdalsecurity.com/blog/companies-affected-by-ransomware/
[33] Ransomware Trends 2021. https://www.hhs.gov/sites/default/files/ransomware-trends-2021.pdf
[34] September 2020 Healthcare Data Breach Report: 9.7 Million Records Compromised. https://www.hipaajournal.com/september-2020-healthcare-data-breach-report-9-7-million-records-compromised/
[35] Ransomware attacks on US healthcare organizations cost $20.8bn in 2020. https://www.comparitech.com/blog/information-security/ransomware-attacks-hospitals-data/
[36] Health Service Executive ransomware attack. https://en.wikipedia.org/wiki/Health_Service_Executive_ransomware_attack
[37] Java Logging Package RCE Vulnerability. https://media.cert.europa.eu/static/SecurityAdvisories/2021/CERT-EU-SA2021-067.pdf
[38] Log4j vulnerability - update from the CSIRTs Network. https://www.enisa.europa.eu/news/enisa-news/log4j-vulnerability-update-from-the-csirts-network
[39] 2020 United States federal government data breach. https://en.wikipedia.org/wiki/2020_United_States_federal_government_data_breach
[40] HiPEAC interview: Ingrid Verbauwhede on security by design, the internet of things and why European leadership is needed. https://youtu.be/2ZBfpwV2nx8
[41] The hybrid war in Ukraine https://blogs.microsoft.com/on-the-issues/2022/04/27/hybrid-war-ukraine-russia-cyberattacks/
[42] EU Action plan on military mobility 2.0.November 2022. https://defence-industry-space.ec.europa.eu/action-plan-military-mobility-20_en
[43] European Defence Agency (EDA) https://eda.europa.eu/
[44] Countering cognitive warfare: awareness and resilience. Cao et al. in NATO Review, 20 May 2021. https://www.nato.int/docu/review/articles/2021/05/20/countering-cognitive-warfare-awareness-and-resilience/index.html
[45] Hackers just found serious vulnerabilities in a U.S. military fighter jet. Joseph Marks, in The Cybersecurity 202 Newsletter. 14 August 2019. https://www.washingtonpost.com/news/powerpost/paloma/the-cybersecurity-202/2019/08/14/the-cybersecurity-202-hackers-just-found-serious-vulnerabilities-in-a-u-s-military-fighter-jet/5d53111988e0fa79e5481f68/
[46] European Parliament website hit by cyber attack. Le Monde. 23 November. 2022. https://www.lemonde.fr/en/europe/article/2022/11/23/european-parliament-website-hit-by-cyber-attack_6005369_143.html
[47] Inside a US military cyber team's defence of Ukraine. Gordon Corera. BBC. 30 October 2022. https://www.bbc.com/news/uk-63328398
[48] Aurora Generator Test. Wikipedia. https://en.wikipedia.org/wiki/Aurora_Generator_Test
[49] Inside the Cunning, Unprecedented Hack of Ukraine's Power Grid. Wired. 3 March 2016. https://www.wired.com/2016/03/inside-cunning-unprecedented-hack-ukraines-power-grid/
[50] Ukraine says it thwarted Russian cyberattack on electricity grid. James Pearson. Reuters. 12 April 2022. https://www.reuters.com/world/europe/russian-hackers-tried-sabotage-ukrainian-power-grid-officials-researchers-2022-04-12/
[51] Russian hackers targeting Dutch gas terminal: report. 25 The NL Times. 25 November 2022. https://nltimes.nl/2022/11/25/russian-hackers-targeting-dutch-gas-terminal-report
[52] Global surveillance disclosures (2013–present). Wikipedia. https://en.wikipedia.org/wiki/Global_surveillance_disclosures_(2013%E2%80%93present)
[53] What is WannaCry ransomware? Kasperky. https://www.kaspersky.com/resource-center/threats/ransomware-wannacry
[54] The Untold Story of NotPetya, the Most Devastating Cyberattack in History. Andy Greenberg. September 2018, Wired. https://www.wired.com/story/notpetya-cyberattack-ukraine-russia-code-crashed-the-world/
[55] The Hidden Costs of Cybercrime. McAfee. December 2020. https://www.mcafee.com/enterprise/en-us/assets/reports/rp-hidden-costs-of-cybercrime.pdf
[56] Actors behind PyPI supply chain attack have been active since late 2021. Dan Goodin. Ars Technica. September 1, 2022. https://arstechnica.com/information-technology/2022/09/actors-behind-pypi-supply-chain-attack-have-been-active-since-late-2021/
[57] Sabotage: Code added to popular NPM package wiped files in Russia and Belarus. Dan Goodin. Ars Technica. March 18, 2022. https://arstechnica.com/information-technology/2022/03/sabotage-code-added-to-popular-npm-package-wiped-files-in-russia-and-belarus/
[58] The SolarWinds Cyber-Attack: What You Need to Know. Center for Internet Security. March 2021. https://www.cisecurity.org/solarwinds
[59] What Is Blockchain? Gartner. March 2022. https://www.gartner.co.uk/en/articles/what-is-blockchain
[60] Gartner blockchain hype cycle 2021: where we are & what's next. Godfrey Benjamin. iMi Blockchain. June 2021. https://imiblockchain.com/gartner-blockchain-hype-cycle/
[61] Gartner Hype Cycle for Blockchain and Web3, 2022. Avivah Litan. Gartner. July 2022. https://blogs.gartner.com/avivah-litan/2022/07/22/gartner-hype-cycle-for-blockchain-and-web3-2022/
[62] Les verrous technologiques des blockchains. CEA, IMT, INRIA. April 2021. https://www.entreprises.gouv.fr/files/files/etudes-et-statistiques/rapport-final-blockchain.pdf
[63] What Was The DAO? Cryptopedia. March 2022. https://www.gemini.com/cryptopedia/the-dao-hack-makerdao
[64] N. Atzei, M. Bartoletti, and T. Cimoli. A survey of attacks on ethereum smart contracts sok. In Conference on Principles of Security and Trust 2017.
[65] Solidity language. https://soliditylang.org/
[66] Introduction to track "Supply chain cybersecurity" of the European Cyber Week. Admiral COUSTILLIERE. November 2022. https://www.european-cyber-week.eu/programme?lang=en
[67] Cybersecurity Maturity Model Certification. https://dodcio.defense.gov/CMMC/
[68] Cyber Essentials. https://www.ncsc.gov.uk/cyberessentials/overview
[69] The NIS 2 Directive. https://www.nis-2-directive.com/

Olivier Zendra is a tenured computer science researcher at Inria, Rennes, France.

Bart Coppens is a part-time assistant professor and a post-doctoral researcher in the electronics department of Ghent University, Ghent, Belgium.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.3, January 2023. Previous versions were published under the name "Cybersecurity must come to IT systems now".

Cite as: O. Zendra and B.Coppens. From cybercrime to cyberwarfare, nobody can overlook cybersecurity any more. In M. Duranton et al., editors, HiPEAC Vision 2023, pages 130-144, Jan 2023.

DOI: 10.5281/zenodo.7461910

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

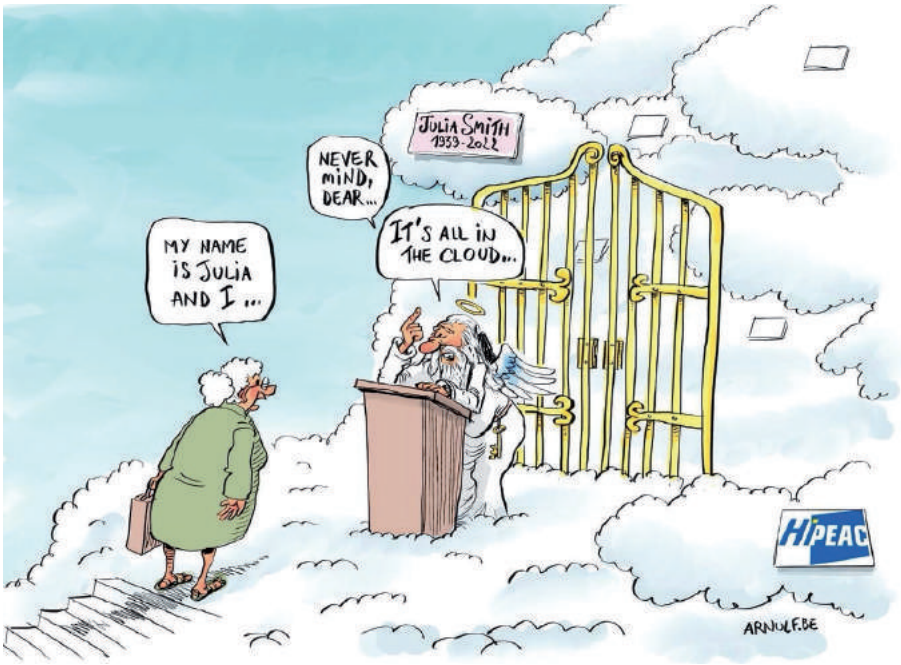
© HiPEAC 2023

There is growing awareness of the importance of privacy while, at the same time, we are sharing ever more private data with third parties. This creates an uneasy tension.

Is privacy possible in a digital world?

By BART COPPENS and OLIVIER ZENDRA

While privacy used to be a concern only for a limited number of people, in recent years awareness of it has been growing. This has been for a number of reasons, including the implementation of the General Data Protection Regulation (GDPR), the growing impact of data leaks, and data logging by governments and enterprises. In addition, there have been high-profile examples demonstrating how new legislation can turn seemingly innocuous private data against its owners. At the same time, the majority of us are consciously or unknowingly transmitting an increasing amount of private data to the cloud, which increases the likelihood of it being leaked or abused.



To try and reconcile these two opposing directions, both consumers and businesses should enhance their knowledge of privacy issues and their use of privacy-enhancing technologies, and enterprises should include privacy by design into their IT development processes.

Key insights

- Ever more data is being sent to and collected by governments and private companies alike.
- This data can then be analysed with big-data methods and artificial intelligence (AI), which has serious implications for privacy.
- The scope and volume of the data being collected and analysed is often not clear to consumers, who are sometimes even completely unaware that it is happening.
- However, there is increasing public awareness of privacy, not only of the fact that personal data is being collected, but also that it can be leaked, either on purpose or inadvertently.
- Technical solutions exist to improve privacy. The European Union (EU) has a role to play.

Key recommendations

- Promote research into technologies that enhance people's privacy and reduce the risks and impact of leaks of private data.
- Stand by EU principles of privacy for its citizens, requiring companies to actively adhere to the principles of privacy by design, and not allowing backdoors to be put into applications.
- To effectively enforce privacy, base information and communication technology (ICT) systems either on open-source software and hardware, or on EU-made, trustable because audited, proprietary hardware and software.
- Restrict how AI can be used to infer and reveal sensitive information from information which is seemingly less sensitive; or restrict the collection of such data.
- Invest in post-quantum cryptography so as to have EU-designed and EU-validated quantum-resistant encryption schemes.

We live in an era in which *almost everything we do is transmitted to servers beyond our control*. To give just a few examples, our private documents are stored in the cloud, while in some countries internet providers are legally obliged to keep track of which websites we visit [14]. Mobile service providers keep track of where our mobile phones make contact to their base stations and thus keep track of where we are; when we drive, our vehicle licence plates are captured by an increasing number of automatic number-plate recognition (ANPR) cameras which are placed for various purposes by governments and municipalities [15,16,17]. The list goes on. People even freely put microphone-based listening devices such as Amazon Echo [11], Apple Siri devices [12], Google Nest [13], etc. in their homes for purposes of convenience and comfort.

Sometimes, this sharing of information is quite intentional. When individuals share a document online, they expect that this information will be shared only with the intended recipients. What they do not usually realize is that others will also have access to this information. In fact, the document is stored on servers, which can be read by legitimate system administrators, teams that deal with abusive or illegal content that may be stored on that server, unrelated people through server misconfigurations, or hackers who gain access.

Moreover, as a result of the transition to cloud-based data storage and processing, data is often no longer held locally on the premises of the organizations to which we entrust our data, but is instead stored on infrastructure that is shared by several companies. Even if they deliberately choose to share information online, the majority of users are typically unaware of the extent to which it may be accessible to others. The amount of information gathered and stored which is *not* explicitly shared by users is orders of magnitude larger.

It is thus clear that privacy is an important topic that directly affects the lives of many people. In this article, we first discuss in more detail the kinds of **personal and private information** that nowadays are being generated, collected, and potentially leaked. We then describe some of the **technical and policy directions** that can help us to **protect** our **data and privacy** better.

Personal and private information

As a society, we are generating and storing an ever-increasing amount of (private) data. This includes (confidential) company information. Almost all of this data is delivered to and kept on cloud-based servers, regardless of its origin or whether this was done intentionally by a user. Although those of us who live in the EU should now be *informed about the fact that data is collected, most people are unclear about the scope of the increasing amount of data that is being collected, processed, and stored*. The scope of the data gathered takes on two dimensions: which data is actively gathered, and which *additional* data *can be inferred*. The trend towards a service-based cloud economy only exacerbates the scope to which this data is gathered and shared, and the risks to which this data is exposed.

All of this begs the question: what sort of personal and confidential data are we discussing? As previously indicated, there is the data that we create and upload ourselves. However, the data collected goes well beyond that, encompassing information pertaining to private conversations, gender, sexual orientation, race, date of birth, religion, political affiliation, medical data, and so on.

It is obvious that we do not want this type of information to be accessible to unrelated persons; yet this is precisely what happens. The online actions of individuals are tracked by advertising corporations such as Google and Meta in order to sell increasingly tailored advertisements [22,23]. On the basis of this information, data analysis and machine learning techniques are then used to develop a profile of personal interests, which is capable of deducing even sensitive information such as sexual orientation [28].

As part of so-called legitimate commercial objectives, this data is also traded and sold in the advertising industry [49]. Advertisers are interested in targeting extremely specific groups of individuals. Extreme instances include targeting individuals present at political demonstrations or individuals who routinely go to church [53]. This information can then be merged with other types of personal information, narrowing down the targeted population until it is possible to identify individuals. All of this information must be tracked and integrated in some way for such tailored advertising to be possible.

Some advertising companies go to quite some lengths to circumvent anti-tracking measures that are implemented by browsers [24]. Some totally unscrupulous companies even scrape the internet for people's pictures from different social media channels and other websites, in order to build and train very accurate facial recognition of these people without their consent [25]. These facial recognition engines can then be sold to governments and commercial organizations across the world [26,27]. This tracking of data has further ramifications with regards to sovereignty. US or Chinese companies keep track of the data of EU citizens, harvest their pictures and process this data abroad, outside EU laws and regulations; this can create problems that are hard to address and solve.

In theory, the GDPR requires explicit consent to be given when private information is kept of users. However, people quickly grow tired of cookie banners and information tracking pop-ups, and are trained to click on anything that allows them to consume their content faster. Furthermore, until recently, it was not clear that the same level of care needs to be taken when the tracked data in and of itself is not private, but when sensitive information can be deduced from it. For example, a recent case for the Court of Justice of the European Union tackled the following issue: keeping track of the name of someone's partner spouse allows one to deduce someone's *sexuality* [51]. And while the latter is clearly private information, it was not clear that the former was. However, the court ruled that, precisely because that inference is possible, the former is private information as well [51]. This is an important milestone, and should have far-reaching consequences.

How is it that all of this information, including sensitive information and information from which sensitive information might be deduced, is monitored in the first place? Obviously, there are countless sources. As previously indicated, people occasionally disclose information explicitly, for example to get access to content. However, much more common are inadvertent data disclosures. For instance, while we drive around, there are automatic number-plate recognition (ANPR) cameras everywhere, which aim to detect flagged vehicles or impose a maximum average speed limit along a road segment. In addition to these government-operated cameras, private residents are increasingly installing cameras as well, both security cameras and standard door cameras, such as those associated with Google Nest. As many of them are cloud-connected, not only do the owners of these cameras have access to the recordings, but also the corporations linked with these cameras, and they can potentially provide access to the police without a court order [55]. The latter mechanism, which may be necessary for efficient law enforcement, can be abused nonetheless by malevolent people to access information they should not have access to [59].

In the examples above, the cameras are highly visible. However, there are other examples where a device's role in compromising privacy is not so obvious. People are often astonished to learn that their listening gadgets not only send fragments of their private conversations to the device manufacturer, such as Amazon, but also to subcontractors who listen to them to improve the accuracy of the voice-recognition engines that power these devices [18]. The information regarding cell phone activities that is being tracked and stored is another example. This includes information about who is speaking with whom and for how long, as well as the location of individuals. This information can then be aggregated to get estimates of which locations are busier than others, and providers can sell this information [62].

Figure 1: A sign announcing camera surveillance at George Orwell Square in Barcelona [68].

Even when this information is *not* sold to others, it can still be problematic. Tesla vehicles, for instance, are fitted with a multitude of cameras and other sensors. Due to concerns that they could be used for espionage, Tesla vehicles have been barred from cities that host Chinese party leadership meetings and Chinese military installations [54,57]. In a recent lawsuit filed against Tesla by the father of a victim of a tragic car accident, Tesla supplied months of historical speed data in an attempt to transfer blame to the driver based on the fact that they had driven over the speed limit in the past [54]. Tesla maintains that it will not sell such information, but it may share it with third parties including business partners, service providers, and government agencies [58].

The ways in which companies track information can also be quite surprising. Information can even be deduced based on user behaviour. For example, when scrolling through a Facebook feed, Meta does not only keep track of which links a user clicks on, but also how long that user spends looking at certain items in their feed: do they scroll past them quickly, or do they linger on certain posts longer? This also allows Meta to deduce information about user likes and interests, even if these users don't explicitly provide them with this information [52].

There is thus a real issue surrounding privacy and awareness of privacy issues. While privacy has always had wide-reaching consequences for some groups of people and professions, ranging from LGBTQIA+ people to journalists who need to protect their sources [56], these consequences are also reaching ever more people.

One recent example is that of data related to abortions, in the wake of the US Supreme Court reversing its previous decision in the Roe v. Wade case and causing abortions to become illegal in several US states. What many people do not realize is that much data about abortions, again including data from which this can be *inferred*, is stored by a large number of (international) companies. Highly personal data such as text messages, emails, payment records, etc., which might merely *imply* that a user had an abortion is not only stored by companies, but can then also be accessed by authorities [43,45]. This includes data from apps which help women track their menstrual cycles [44,45], and if someone just *searches for* abortion pills online [45].

Companies tracking location data may find that this is suddenly a liability, as this information also includes visits to abortion clinics [46]. Some companies subsequently realized that storing as much data as possible about users might not have been the best idea, but still struggle to limit the scope of the captured data. For example, while they can try to simply remove (or not store) location data around abortion clinics, but this in and out of itself might not prove sufficient, as long as there enough location information is being kept from the neighbouring area. For example, in a related context, fitness tracking social networks allow users to enable endpoint privacy zones to hide the begin and end of tracks so as to hide the exact location of people's homes. However, creating schemes that prevent attackers from inferring these locations anyway turns out to be challenging and not necessarily user friendly [47]. Worse, even if people explicitly opt out from one kind of location data collection, there might be multiple systems independently collecting such data: Google recently settled a lawsuit over the fact that turning off the "location history" checkbox only stopped a Google Maps feature, not all Google location tracking [48].

In summary, *in addition to typical government-related activities, private companies are keeping track of increasingly more detailed information about people who are not even necessarily their users*. And even information which, at first glance, might not seem private in and of itself, can still lead to sensitive information being deduced from it.

Technical means to protect our data and privacy

Now, how can we effectively safeguard our data? The simplest answer would be to not have this data collected at all, and if it is collected to not share it at all: data that is not shared is really private data. However, the extent to which data is currently exchanged whenever we engage with today's predominantly digitized world makes this impractical unless people are willing to become (partial) digital and social recluses, so a middle ground needs to be found. This can be achieved at least in part by advocating and selecting technology that enhances our privacy, as opposed to ignoring it or, worse, deliberately trying to circumvent it.

In order to do so, we need to *design systems with privacy in mind from the ground up*. How a system handles private data and how it deals with privacy should be design requirements from the start. In 2010, the International Conference of Data Protection and Privacy Commissioners published a resolution encouraging recognition of the fact that privacy by design is an essential component of privacy protection, as well as the adoption of a set of foundational principles of privacy by design [29]. One of these principles is that privacy should be *embedded* in the design, and it should be an essential core of the functionality of the system [30]. These guiding principles were as true and necessary then as they are now.

For example, rather than sharing documents and messages with people where the shared data is stored in an unencrypted form on cloud servers, we can use solutions with true *end-to-end encryption*. Unlike Facebook Messenger, Signal [10] and the EU-based Olvid [9] encrypt the data such that intermediate servers cannot decipher the message[1]. While FaceBook Messenger supports end-to-end encrypted chats, this is not yet the default policy; users have to manually enable this function, and, if they don't, Meta still has access to the plain-text versions of the messages [60].

Where true end-to-end communications are not possible, or where people risk being tracked, it would still be beneficial to at least choose a *technology or solution which explicitly focuses on the privacy of its users*, like for example the EU-based Qwant search engine [31], or the Brave browser [32], that put an emphasis on protecting their users' privacy. Of course, in the case of a communications app, there needs to be a critical mass of people using it: there seems little point switching to an app for communicating with friends and family if they do not migrate. On the contrary: one is then stuck with two apps instead of one app: an app to communicate securely with a few people who have migrated, and an app to communicate with the rest. The EU should encourage more initiatives and further developments and investments into such privacy-aware technologies and companies, and encourage the adoption of these technologies, possibly mandating them in its own services, to help protect its citizens and its sovereignty over data.

Of course, a simple and effective way to have fewer problems with private data potentially being compromized is to not send it over the network at all. One way in which this can be solved is by ensuring that most or even all computations that normally take place in the cloud now take place locally, with the *processing being done at the edge*. This also has implications for industrial applications in the context of the internet of things (IoT) and cyber-physical systems (CPS): the more data is processed in those devices themselves, rather than having to be transmitted to cloud servers for processing, the less private data can be abused or leaked. A fog or federation of local devices sharing part of the global information in an encrypted way could solve the problem of accessing larger computing or storage resources in a more local manner.

If data does need to be transmitted or computed remotely, it is important to do this in a secure fashion that preserves as much security and privacy as possible. Most companies already try to *protect most sensitive data at rest and in transit with encryption*, for example with the Advanced Encryption Standard (AES) and Transport Layer Security (TLS). However, this data still needs to be processed, for which the data is currently still decrypted (and thus unprotected) on the systems that process it.

Furthermore, if this data processing involves the data being searchable or queryable in a database, many systems will still store this data in an unencrypted form. One way to mitigate this problem is to *do the data processing on encrypted data*, in such a way that the personally identifiable information (PII) is not known to the system performing the actual processing. Examples of such techniques are (fully) homomorphic encryption (FHE), which still requires research to decrease its computing resource requirements, and secure multi-party computation.

There are many fields in which homomorphic encryption would significantly increase the privacy of data in the context of cloud-based data processing. In the medical sector, users would be able to upload their ECG data and have a cloud provider monitor their health without actually sharing their data with that cloud provider [2]. Similarly, people would be able to have their genomes analysed by third parties without information being passed on about which genetic diseases they have or other PII such as gender, race, etc [3].

Modifying different cloud-based machine learning tasks to protect PII would also significantly reduce the risks associated with outsourcing the relevant data. For example, face verification or face recognition would no longer expose photographs of people [4], and performing optical character recognition would no longer leak the text being processed [5]. Furthermore, if the recognized text is from licence plates that need to be queried in a database of stolen and wanted vehicles, for exam-

1 In addition, Olvid does not require you to provide your phone number, since it does not rely on any central directory.

ple, one solution could be to prevent the processing of all licence plates from leaking information about *non-stolen* cars [6]. The EU should invest in more technologies such as these, so that if PII data does need to be processed, the amount of data that can be intentionally or inadvertently leaked is minimized as much as possible.

Given the urgency within today's business landscape to achieve more robust data privacy systems, we predict and advocate for an increase in the design and use of such homomorphic encryption and related techniques. Some start-ups already provide very specific applications of these techniques [7]. One limiting factor in applying FHE right now is its overhead. Both the time needed to process the data and the size of the messages that need to be exchanged with the cloud provider currently increase dramatically when FHE is applied. At present, this means that many of those techniques are unfortunately not yet usable in practice. In the meantime, some specific cases might not need to send the PII itself to third parties.

Another issue to take into account when protecting data by encrypting it is how resistant the encryption scheme is to the changing landscape of attackers' capabilities. One clear but constant change is the increase in the processing speed of computers. As one of the most obvious goals of an attacker is to recover the information, the question is how long information can remain private, and how this time decreases with an increase in processing speed, and by how much we then increase the strength of the encryption (for example, by increasing the key size) to compensate for this.

In the case of traditional computers, it is quite clear how these scaling laws work, and increases in computing power do not immediately threaten the security of data encrypted with traditional encryption schemes. However, when switching to the different computing paradigm of quantum computers, this is not necessarily the case, because certain algorithms are believed to run significantly faster on quantum computers than on traditional computers. With some algorithms, it is sufficient to choose larger key sizes to compensate for this. However, other algorithms could be completely broken with quantum computers. Such algorithms need to be replaced with algorithms that could withstand attacks from a quantum computer. This field is called post-quantum cryptography (which is sometimes abbreviated as PQC).

It is crucial that we begin defending not only against existing threats, but also against future threats. There are two compelling arguments for deploying post-quantum cryptography as soon as possible. First, the data we send or save now may remain confidential and valuable in the future, when potential adversaries may have access to quantum computers, so we should safeguard against future attacks.

Second, switching algorithms is not an instantaneous process; compatibility must be ensured across the board. While many devices may be upgraded quickly to support more algorithms, this is not always the case. Obviously, every new cryptographic algorithm must be extensively examined. One of the finalist algorithms in the NIST competition on post-quantum cryptography [66], for instance, was recently invalidated: researchers demonstrated that a standard laptop could crack it in less than an hour [64]. Therefore, it is clear that additional research into these algorithms is required. Additionally, switching algorithms should be done with caution. Google, for instance, recently opted to begin utilizing post-quantum cryptographic algorithms to secure its internal communications, but they ensured that this was an additional layer of security [65], rather than being the only one.

However, it is not sufficient to use state-of-the-art encryption algorithms to protect PII. Software that is not secure can still leak all kinds of confidential and private information to attackers, even if under normal circumstances this data is stored and transmitted securely. Some *security-related instruction set architecture extensions* have explicit implications for improving privacy. For example, one of the goals of Intel's Software Guard Extensions (SGX) is to protect the execution of certain code fragments from attackers that have control over the rest of the system, including the operating system itself. This can then be used to protect sensitive and private information even when the entire system is being attacked. However, the many recent attacks on SGX show that even this technology is clearly not yet mature enough to withstand such attacks in practice [19,20,21]. It may even be that the SGX model of allowing execution of code on private data, and general-purpose code execution by untrusted users, might not be feasible.

Protecting our data and privacy through policy

Thanks to *regulations such as the GDPR*, EU citizens should be better protected against at least some forms of unwanted

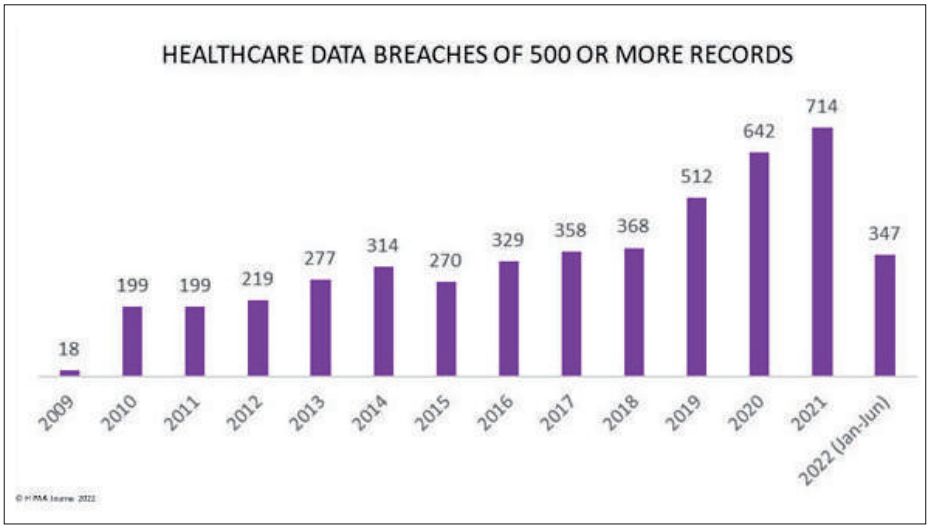


Figure 1: Reports of healthcare data breaches of 500 or more records reported to the US Department of Health and Human Services' Office for Civil Rights between 2009 and June 2022 [1].

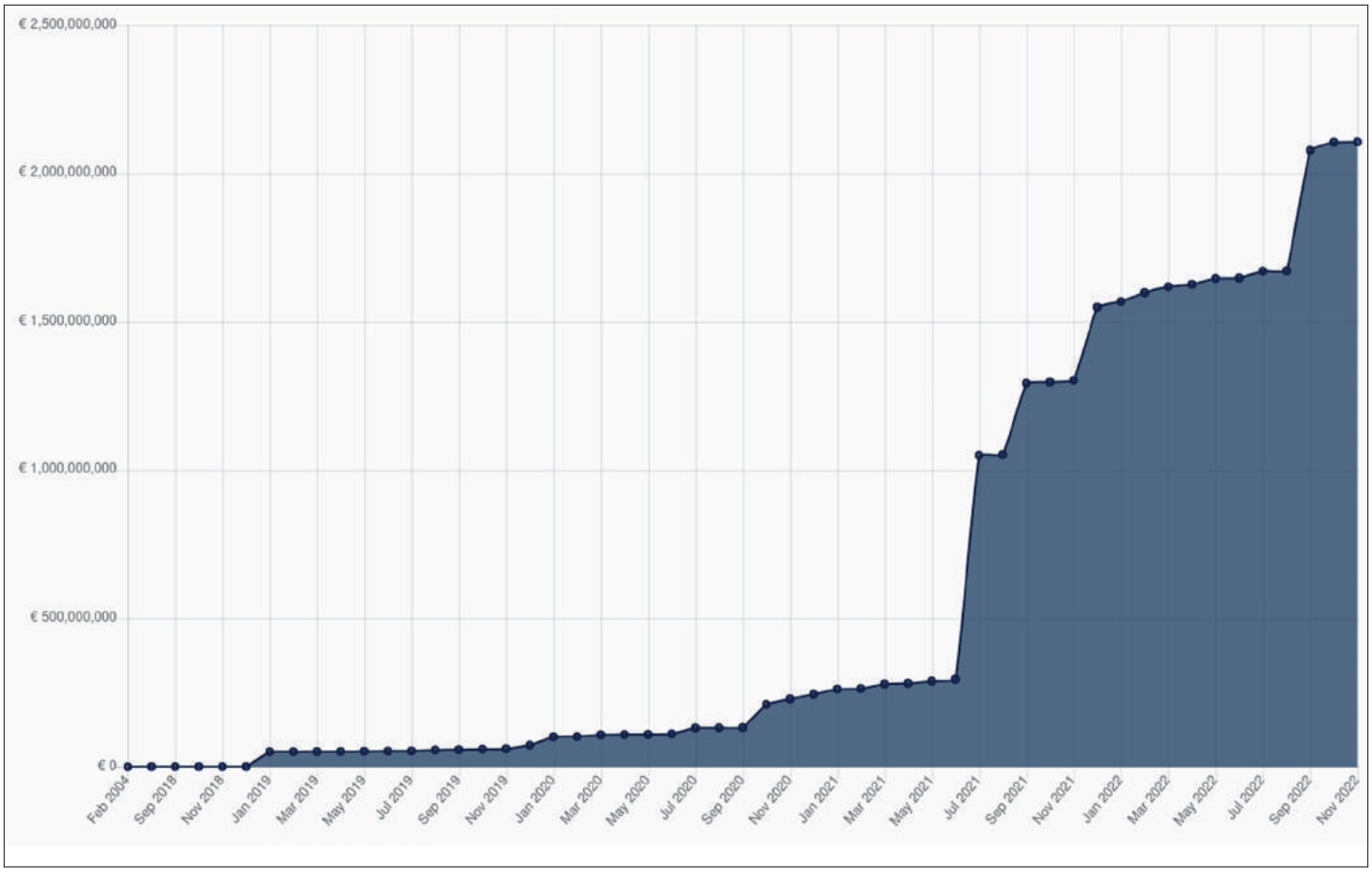


Figure 2: Cumulative sum of GDPR-related fines [63]

processing of private data, and they should now at least be informed when such data is leaked or mishandled. This represents huge progress in terms of the public being informed and aware of data protection issues and should be hailed as a very positive step. However, data leaks are still likely to be a problem. Figure 1 gives an idea of the scale of the problem, indicating healthcare data breaches of 500 or more records reported to the US Department of Health and Human Services' Office for Civil Rights between 2009 and June 2022. Unfortunately, rather than decreasing, the numbers are going up. Similarly, Figure 2 shows the cumulative sum of GDPR-related fines; again, the fines show no sign of decreasing.

In this context, it is important to stress the importance of the entire system being *secure and not weakened by backdoors*. Some governments have advocated for such backdoors in operating systems, telecommunications networks, and secure/encrypted communication platforms, so that only "they" may (legally) get access to systems and decode encrypted data. In 2020, the EU Council called for law enforcement to have access to encrypted private data [36], while in 2021, the Belgian government proposed sweeping new legislation mandating backdoors in encrypted communications networks [37]. This latter proposal was rescinded in response to public backlash and unfavourable feedback from experts and the Belgian Data Protection Authority [38].

Along these lines, Apple also proposed a new "feature" for iCloud in which an iPhone would check photographs for child abuse content. While this proposal appeared limited in scope, it met with a massive public outcry from both experts and civil rights organizations, who feared that it would open the door to massive government surveillance and transform mobile phones into permanent government spying devices, causing Apple to delay the release of this feature [43]. In exactly this context, though, the EU has again proposed to weaken their citizen's privacy and security [61], which we consider would be a highly insecure mistake.

A fundamental issue with these backdoors is that they reduce the security of the entire system [40,41,42]. There is in fact no guarantee that the law enforcement agents of the country will be the *only ones* with access to these backdoors: other countries and criminals might be able to use them too. For example, a pseudo-random number generator containing a weakness which had allegedly been introduced by the US National Security Agency (NSA) eventually found its way into firewalls, where it was exploited by unknown parties [34]. Some allege that Intel's closed-source management engine on chips not only supports well-intended remote management functions, but might also be used by other (malicious) parties to get remote access to machines [35].

Such backdoors decrease the security of individuals and their data. They also undermine the EU's efforts to protect the privacy and security of its residents by diminishing people's trust in computers and telecommunications infrastructure. Furthermore, such measures would affect the confiden-

tiality of company data, which would then be susceptible to leakage through these backdoors. Thus, to protect the privacy of its inhabitants and the secret information of its businesses, the EU should not even consider legalizing such backdoors.

However, while an insecure system can lead to information leaks, the converse is not necessarily true. A secure system cannot distinguish between purposeful leaks of information (for example, a user who wants to print his/her own bank statements), versus inadvertent leaks of information (for example, these bank statements being stored unencrypted on disk). One possible solution here is language-based information-flow security that allows programmers to explicitly define which flows of information are allowed, and to define properties on these flows [8].

A final source of leaking private information is the users themselves: frequently, they are unaware of the actual private information that can be extracted from the data being shared: posting pictures of someone in a bar or nightclub could be interpreted by an insurance company as evidence that the individual is a health risk because they drink alcohol. There are systems based on artificial intelligence that can analyse such public content and warn users about such "social" information [33].

In light of this, it would be useful to consider restricting which information can be collected and stored, and how data that can be inferred from it should be treated.

Conclusion

Public awareness of privacy issues is slowly increasing thanks to initiatives such as the enactment of the GDPR, which is widely considered as an outstanding and pioneering move on the part of the EU, but also with people seeing the impact of how seemingly innocent data collection can make them vulnerable to prosecution. These will hopefully be a trigger for people to think more about where and how their private data is being collected, stored, and used, which in many cases is anyplace, anytime, by most of the tools and applications. This will hopefully encourage individuals to be more vigilant about protecting their own privacy, as well as demanding more robust privacy protections from ICT companies.

The EU has a role in regulating, promoting, and supporting EU-based solutions that protect privacy and sovereignty. It should build on its position as a leader in privacy legislation, which has already had an influence on laws in other parts of the world [67]. If companies wish to sell ICT products and services in the EU, they should meet stringent EU-defined privacy requirements. Lastly, companies should be obliged to ensure privacy by design in their products and services.

References

[1] Healthcare Data Breach Statistics, HIPAA Journal, online, accessed November 22, 2022. https://www.hipaajournal.com/healthcare-data-breach-statistics/

[2] Kocabas, Ovunc, et al. "Assessment of cloud-based health monitoring using homomorphic encryption." 2013 IEEE 31st International Conference on Computer Design (ICCD). IEEE, 2013

[3] Miran Kim, Kristin Lauter, "Private genome analysis through homomorphic encryption", BMC Med Inform Decis Mak. 2015; 15(Suppl 5): S3.

[4] J.R. Troncoso-Pastoriza, D. González-Jiménez, F. Pérez-González, 2013. Fully private non-interactive face verification". IEEE Transactions on Information Forensics and Security, 8(7), pp.1101-1114

[5] Nathan Dowlin, Ran Gilad-Bachrach, Kim Laine, Kristin Lauter, Michael Naehrig, and John Wernsing. 2016. "CryptoNets: applying neural networks to encrypted data with high throughput and accuracy. In Proceedings of the 33rd International Conference on International Conference on Machine Learning – Volume 48 (ICML'16), Maria Florina Balcan and Kilian Q. Weinberger (Eds.), Vol. 48. JMLR.org 201-210.

[6] Sunil, Archana Bindu, Zekeriya Erkin, and Thijs Veugen. "Secure matching of Dutch car license plates." Signal Processing Conference (EUSIPCO), 2016 24th European. IEEE, 2016

[7] https://www.privatebiometrics.com/, accessed December 4, 2020

[8] A. Sabelfeld and A. C. Myers, "Language-based information-flow security", in IEEE Journal on Selected Areas in Communications, vol. 21, no. 1, pp. 5-19, Jan. 2003.

[9] Olvid. https://olvid.io/technology/en/

[10] Signal. https://signal.org/docs/

[11] Amazon Echo. https://en.wikipedia.org/wiki/Amazon_Echo

[12] Apple Siri. https://en.wikipedia.org/wiki/Siri

[13] Google Nest. https://en.wikipedia.org/wiki/Google_Nest_(smart_speakers)

[14] Judgments in Case C-623/17, Privacy International, and in Joined Cases C-511/18, La Quadrature du Net and Others, C-512/18, French Data Network and Others, and C-520/18, Ordre des barreaux francophones et germanophone and Others. Press release 123/20, Court of Justice of the European Union, 6 October 2020

[15] Automatic Number Plate Recognition, Police.uk https://www.police.uk/information-and-advice/automatic-number-plate-recognition/ accessed December 2020

[16] Denmark: Targeted ANPR data retention turned into mass surveillance EDRi, September 6, 2017, https://edri.org/our-work/denmark-targeted-anpr-data-retention-turned-into-mass-surveillance/

[17] Automatic number-plate recognition - Usage, Wikipedia, https://en.wikipedia.org/wiki/Automatic_number-plate_recognition#Usage Accessed December 2020

[18] Apple contractors 'regularly hear confidential details' on Siri recordings, The Guardian, July 26, 2019.

[19] Foreshadow - Extracting the Keys to the Intel SGX Kingdom with Transient Out-of-Order Execution. USENIX Security Symposium 2018.

[20] Plundervolt: Software-based Fault Injection Attacks against Intel SGX. Murdock et al. IEEE Symposium on Security and Privacy 2020

[21] CrossTalk: Speculative Data Leaks Across Cores Are Real. Ragab et al. Accepted in the IEEE Symposium on Security and Privacy, 2021.

[22] Why targeted ads are the most brutal owns. Vox, September 25, 2018. https://www.vox.com/the-goods/2018/9/25/17887796/facebook-ad-targeted-algorithm

[23] Google's ad tracking is as creepy as Facebook's. Here's how to disable it. The Guardian, October 21, 2016. https://www.theguardian.com/technology/2016/oct/21/how-to-disable-google-ad-tracking-gmail-youtube-browser-history

[24] Ad Tech Surveillance on the Public Sector Web, Cookiebot Report, version July 14, 2020. https://www.cookiebot.com/media/1136/cookiebot-report-2019-ad-tech-surveillance-2.pdf

[25] Scraping the Web Is a Powerful Tool. Clearview AI Abused It. Wired, January 25, 2020. https://www.wired.com/story/clearview-ai-scraping-web/

[26] Clearview's Facial Recognition App Has Been Used By The Justice Department, ICE, Macy's, Walmart, And The NBA. Buzzfeed, February 27, 2020. https://www.buzzfeednews.com/article/ryanmac/clearview-ai-fbi-ice-global-law-enforcement

[27] Secret Users Of Clearview AI's Facial Recognition Dragnet Included A Former Trump Staffer, A Troll, And Conservative Think Tanks. Buzzfeed, March 11, 2020. https://www.buzzfeednews.com/article/ryanmac/clearview-ai-trump-investors-friend-facial-recognition

[28] Researchers Claim Facebook Ads Could Out LGBTQ+ Users. Out, August 30, 2019. https://www.out.com/tech/2019/8/30/researchers-claim-facebook-ads-could-out-lgbtq-users

[29] Resolution on Privacy by Design. 32nd International Conference of Data Protection and Privacy Commissioners. Jerusalem, Israel 27-29 October, 2010. https://edps.europa.eu/sites/edp/files/publication/10-10-27_jerusalem_resolutionon_privacybydesign_en.pdf

[30] Privacy by Design: The 7 Foundational Principles. Ann Cavoukian, Ph.D., Information & Privacy Commissioner of Ontario, Canada. https://www.ipc.on.ca/wp-content/uploads/Resources/7foundationalprinciples.pdf

[31] Qwant. Accessed December 2020. https://www.qwant.com/?l=en

[32] Brave. Accessed December 2020. https://brave.com/

[33] https://www.researchgate.net/publication/301221061_Personalized_Privacy-aware_Image_Classification

[34] Researchers Solve Juniper Backdoor Mystery; Signs Point to NSA, Wired, December 22, 2015. https://www.wired.com/2015/12/researchers-solve-the-juniper-mystery-and-they-say-its-partially-the-nsas-fault/

[35] Is the Intel Management Engine a backdoor? TechRepublic, July 1, 2016. https://www.techrepublic.com/article/is-the-intel-management-engine-a-backdoor/

[36] Encryption: Council adopts resolution on security through encryption and security despite encryption. Council of the EU Press release, December 14, 2020. https://www.consilium.europa.eu/en/press/press-releases/2020/12/14/encryption-council-adopts-resolution-on-security-through-encryption-and-security-despite-encryption/

[37] Belgian Government Wants To Add Encryption Backdoors To Its Already-Terrible Data Retention Law. Techdirt, October 6, 2021. https://www.techdirt.com/articles/20211003/17223447690/belgian-government-wants-to-add-encryption-backdoors-to-already-terrible-data-retention-law.shtml

[38] Victory! Belgium Scraps Proposed Law to Backdoor End-to-End Encryption. Center for Democracy & Technology, December 22, 2021. https://cdt.org/insights/victory-belgium-scraps-proposed-law-to-backdoor-end-to-end-encryption/

[39] If You Build It, They Will Come: Apple Has Opened the Backdoor to Increased Surveillance and Censorship Around the World, EFF. August 11, 2021. https://www.eff.org/deeplinks/2021/08/if-you-build-it-they-will-come-apple-has-opened-backdoor-increased-surveillance

[40] Keys Under Doormats: mandating insecurity by requiring government access to all data and communications. Abelson et al. MIT CSAIL Technical Report MIT-CSAIL-TR-2015-026. July 6, 2015

[41] End-to-end encryption protects children, says UK information watchdog. The Guardian, January 21, 2022. https://www.theguardian.com/technology/2022/jan/21/end-to-end-encryption-protects-children-says-uk-information-watchdog

[42] Bugs in our Pockets: The Risks of Client-Side Scanning. Abelson et al. CoRR abs/2110.07450. 2021

[43] Apple delays controversial child protection features after privacy outcry. The Verge, September 3, 2021. https://www.theverge.com/2021/9/3/22655644/apple-delays-controversial-child-protection-features-csam-privacy

[43] How overturning Roe v Wade has eroded privacy of personal data. David Cox. BMJ 2022;378:o2075 https://www.bmj.com/content/378/bmj.o2075

[44] How period tracking apps and data privacy fit into a post-Roe v. Wade climate. NPR, June 24, 2022. https://www.npr.org/2022/05/10/1097482967/roe-v-wade-supreme-court-abortion-period-apps

[45] The post-Roe data privacy nightmare is way bigger than period tracking apps. K. Bell. Engadget, June 29, 2022. https://www.engadget.com/data-privacy-period-tracking-apps-130054404.html

[46] Google Says It Will Delete Location Data When Users Visit Abortion Clinics. New York Times, July 1, 2022. https://www.nytimes.com/2022/07/01/technology/google-abortion-location-data.html

[47] A Run a Day Won't Keep the Hacker Away: Inference Attacks on Endpoint Privacy Zones in Fitness Tracking Social Networks. K. Dhondt et al. ACM CCS 2022.

[48] Google settles "Location History" lawsuit with 40 states, will pay $392 million. R. Amadeo, Ars Technica. November 14, 2022. https://arstechnica.com/gadgets/2022/11/google-settles-location-history-lawsuit-with-40-states-will-pay-392-million/

[49] There's a Multibillion-Dollar Market for Your Phone's Location Data. J. Keegan and A. Ng, The Markup. September 30, 2021. https://themarkup.org/privacy/2021/09/30/theres-a-multibillion-dollar-market-for-your-phones-location-data

[50] Top U.S. Catholic Church official resigns after cellphone data used to track him on Grindr and to gay bars. M. Boorstein et al. New York Times. July 21, 2021. https://www.washingtonpost.com/religion/2021/07/20/bishop-misconduct-resign-burrill/

[51] Sensitive data ruling by Europe's top court could force broad privacy reboot. N. Lomas. TechCrunch. August 2, 2022. https://techcrunch.com/2022/08/02/cjeu-sensitive-data-case/

[52] Facebook Now Cares About How Long You Look At Stuff In Your News Feed. G. Kumparak. TechCrunch. June 12, 2015. https://techcrunch.com/2015/06/12/facebook-now-cares-about-how-long-you-look-at-stuff-in-your-news-feed

[53] Political Campaigns Know Where You've Been. They're Tracking Your Phone. S. Schechner, et al. WSJ. October 10, 2019. https://www.wsj.com/articles/political-campaigns-track-cellphones-to-identify-and-target-individual-voters-11570718889

[54] The Radical Scope of Tesla's Data Hoard. M. Harris. IEEE Spectrum. August 3, 2022. https://spectrum.ieee.org/tesla-autopilot-data-scope

[55] Google feels it is 'important' to be able to share Nest Cam recording with police, but never has. B. Schoon. 9to5Google. July 27, 2022. https://9to5google.com/2022/07/27/google-nest-cameras-police/

[56] Digital and Physical Safety: Protecting Confidential Sources. Committee to Protect Journalists. November 22, 2021. https://cpj.org/2021/11/digital-physical-safety-protecting-confidential-sources/

[57] Tesla cars barred for 2 months in Beidaihe, site of China leadership meet. Reuters. June 20, 2022. https://www.reuters.com/business/autos-transportation/chinas-beidaihe-district-bar-tesla-cars-driving-july-local-police-2022-06-20/

[58] Who Actually Owns Tesla's Data? M. Harris. IEEE Spectrum. August 5, 2022. https://spectrum.ieee.org/tesla-autopilot-data-ownership

[59] Hackers Gaining Power of Subpoena Via Fake "Emergency Data Requests", KrebOnSecurity, March 29, 2022. https://krebsonsecurity.com/2022/03/hackers-gaining-power-of-subpoena-via-fake-emergency-data-requests/

[60] How to encrypt Facebook Messenger chats B. Krasnoff. The Verge. August 17, 2022. https://www.theverge.com/23307995/facebook-messenger-end-to-end-encryption-how-to

[61] Re: Protecting digital rights and freedoms in the Legislation to effectively tackle child abuse. Open letter by European Digital Rights (EDRi) et al. March 17, 2022. https://edri.org/wp-content/uploads/2022/03/Civil-society-open-letter-Protecting-rights-and-freedoms-in-the-upcoming-legislation-to-effectively-tackle-child-abuse.pdf

[62] Proximus Analytics - Big Data - Location Data. Accessed November 22, 2022. https://www.proximus.be/en/id_cl_analytics/companies-and-public-sector/it-services/iot/proximus-analytics.html

[63] GDPR Enforcement Tracker - list of GDPR fines. Accessed November 22, 2022 https://www.enforcementtracker.com/?insights

[64] Post-quantum encryption contender is taken out by single-core PC and 1 hour. D. Goodin. Ars Technica. August 2, 2022. https://arstechnica.com/information-technology/2022/08/sike-once-a-post-quantum-encryption-contender-is-koed-in-nist-smackdown/

[65] Securing tomorrow today: Why Google now protects its internal communications from quantum threats. S. Kölbl et al. Google Cloud Blog. November 19, 2022. https://cloud.google.com/blog/products/identity-security/why-google-now-uses-post-quantum-cryptography-for-internal-comms

[66] Post-Quantum Cryptography Standardization | CSRC. NIST. Accessed November 23, 2022. https://csrc.nist.gov/Projects/post-quantum-cryptography/post-quantum-cryptography-standardization

[67] GDPR: An impact around the world.. D. Ruiz, Malwarebytes Labs. April 1, 2020. https://www.malwarebytes.com/blog/news/2020/04/gdpr-an-impact-around-the-world

[68] george_orwell_bcn. by fibercool. Taken on July 5, 2007. From https://www.flickr.com/photos/76499396@N00/728743297/ Licensed under CC BY-SA 2.0 https://creativecommons.org/licenses/by-sa/2.0/

Bart Coppens is a part-time assistant professor and a post-doctoral researcher in the electronics department of Ghent University, Ghent, Belgium.

Olivier Zendra is a tenured computer science researcher at Inria, Rennes, France.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

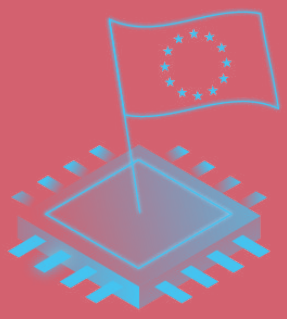
This is release v.3, January 2023. Previous versions were published under the name "Privacy – whether you're aware of it or not, it does matter!"

Cite as: B.Coppens and O. Zendra. Is privacy possible in a digital world? In M. Duranton et al., editors, HiPEAC Vision 2023, pages 130-152, Jan 2023.

DOI: 10.5281/zenodo.7461921

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

© HiPEAC 2023



HiPEAC Vision 2023

THE RACE FOR SOVEREIGNITY

As geopolitical lines are shifting, world regions want to become less dependent on the rest of the world.

The race for sovereignty

By KOEN DE BOSSCHERE

After decades of globalization, the world has discovered that a globalized economy has its weaknesses too, especially if supply chains pass through regions that face production issues (due to a pandemic, a local conflict, natural disaster, social unrest, energy or water shortage, geopolitical tensions, sanctions, …). Local governments are making supply chains more resilient by shortening them, especially for products that are considered of strategic value (like chips, drugs, food, energy, …).

But shortening supply chains brings its own challenges and is anything but self-evident.

- It might make goods more expensive due to the higher labour costs in Europe
- It means that Europe has to import raw materials instead of products, and that supply chains will have to be rerouted. There will be fierce international competition for scarce raw materials. It also means that Europe will also need more energy and water (process water, cooling, irrigation for local food production, …) in order to support industrial activities
- Digital sovereignty means that Europe also needs to have control over strategically important software stacks and platforms. Essential European software infrastructure should not break if a major information and communication technology (ICT) company fails, or if a foreign government decides to limit the use of a platform in Europe
- The European labour market is losing one million workers per year, that is, almost 0.5% of the EU work force. Reshoring manufacturing will require workers that are also needed in the health sector (ageing population), in education, and in the renewable energy sector (which is considered Europe's growth strategy). A productive local work force is essential for sovereignty.
- Some production facilities like semiconductor fabs require highly specialized workers. Since higher education needs 10 years (engineering degree + PhD) to produce such profiles, reskilling, upskilling or immigration will be needed in the short term. The European Pact for Skills provides support for reskilling and upskilling for the green and digital transformations.

This chapter contains four contributions.

- *"The position of Europe in the world"*
  This article contains a strengths, weaknesses, opportunities and threats analysis (SWOT) of Europe. The biggest strength of Europe is its research, and its biggest weakness is its inability to bring the results of this excellent research to the market. Its biggest threat is its rapidly shrinking work force, and the biggest opportunity is that the future growth in digital technology will be in automotive, industrial automation and cybersecurity – three of the strengths of Europe.
- *"Open source is the enabler for innovation and collaboration"*
  Open source has become a global framework, but can only work well if there is enough support to help the stakeholders to do open source the right way. It is an opportunity for Europe to promote collaboration in major application domains and between academia and industry. Successful European open-source projects will not only contribute to Europe's competitiveness, but also to Europe's soft power.
- *"Europe's need for digital essentials, individual sovereignty and consumer protection"*
  This article focuses on software sovereignty. It states that we should not only try to become technologically independent of countries like Russia, or China, but also of the US, and not only for hardware, but also (and especially) for software. For that to happen, the article advocates "digital essentials" that would ensure that digital functionality and services keep working no matter what happens at the geopolitical level. It states that working software is as important as food and pharmaceutical drugs for society.
- *"Rethinking education"*
  The European educational system is very old and of excellent quality, but it might no longer be adapted to the needs of the 21st century. In order to be ready for the job market, graduates should have a T-shaped profile, solid technical training in a discipline supplemented with a broad set of soft skills, and a strong attitude towards lifelong learning.

Key insights

- Europe's global economic impact is dwindling because other regions are growing faster due to their demographics, rapid economic development, or abundance of natural resources.
- A major long-term challenge for Europe is to sustain economic growth with a shrinking active population (up to 0.5% per year) and growing costs for social security. Maintaining the current standard of living will require a highly educated and productive workforce.
- Bringing back critical value chains to Europe is a huge challenge: supply chains have to be rerouted, and a qualified work force needs to be trained. This is a challenge in a shrinking labour market.
- European higher education is of high quality, widely accessible, and affordable.
- Europe is a scientific powerhouse, but it fails to bring its research results to the market due to the lack of entrepreneurial talent, venture capital and a large enough IT workforce.
- Europe currently lacks digital sovereignty, both in hardware and in software. Modern IT systems are mostly made of components that are produced and controlled outside the EU.
- Open source is a global framework for collaboration and innovation that can help Europe to better collaborate and innovate, and to build soft power.
- The European Chips Act wants to strengthen chip manufacturing in Europe, and get 20% of the global chip market by 2030. This is a very ambitious goal under the constraints just mentioned.
- Digital sovereignty requires that Europe's essential software infrastructure keeps working no matter what happens outside Europe (e.g. the bankruptcy of a major IT company or interference by a foreign government).

Key recommendations

- The creation of well-funded European competence centres will help to retain and attract top talent, and to stay at the forefront of new digital technologies.
- Europe should continue to invest heavily in research and innovation, and in a more entrepreneurial Europe that generates lots of start-ups, scale-ups and global companies. Europe needs more venture capital to support the growth of scale-ups. The European Innovation Council (EIC) could be instrumental in this area.
- Europe should invest in future-proof areas like the silver economy, health, mobility, energy, automation and sustainability, and secure a global leading position in these areas.
- Europe should strive for digital sovereignty, both in hardware and in software because digital technology is the cornerstone of modern society. The goal is not to strive for full digital independence, but for at least digital interdependence with the other global regions.
- Europe should also work on "digital essentials" that can guarantee that Europe's essential software infrastructure keeps working uninterruptedly no matter what happens outside Europe.
- Europe should promote open source and create an infrastructure to support the open-source software and hardware supply chain in Europe. Open source helps collaboration and innovation, as well as building soft power.
- Higher education should pick up two additional important roles: lifelong learning and supporting regional entrepreneurial ecosystems.

Koen De Bosschere is a professor in the electronics department of Ghent University, Ghent, Belgium.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2023.

Cite as: K. De Bosschere. The race for sovereignty. In M. Duranton et al., editors, HiPEAC Vision 2023, pages 154-155, Jan 2023.

DOI: 10.5281/zenodo.7461929

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

© HiPEAC 2023

Europe is currently no superpower in computing, but it is never too late to become one.

The position of Europe in the world

By KOEN DE BOSSCHERE

In order to set out a strategy for the future, it is important to know one's strengths and weaknesses. It also helps to identify opportunities, and to prepare for threats. Although Europe is currently no superpower in computing or in the business-to-consumer information and communication technology (B2C ICT) market, it is never too late to become one. For that to happen, it is important to understand the obstacles that make it difficult to grow global companies, and to develop policy to remove such barrriers.

Key insights

- Europe's global economic impact is dwindling because other regions are growing faster due to their demographics, rapid economic development, or abundance of natural resources. The biggest long-term challenge for Europe is to sustain economic growth with a shrinking active population and growing costs for social security. Maintaining the current standard of living will require a highly educated and productive workforce.
- It is important for Europe to realize that computing is a key enabling technology of strategic importance because it is at the basis of all modern smart products and services. Europe should invest in the capacity to build its own computing solutions.
- Europe is a scientific powerhouse, but it fails to monetize some of its research results due to the lack of an entrepreneurial culture, venture capital and a sufficiently large ICT workforce.

Key recommendations

- The creation of well-funded international competence centres will help to retain and attract top talent, and to stay at the forefront of new digital technologies.
- Europe should continue to invest heavily in research and innovation, and in a more entrepreneurial Europe that generates lots of start-up, scale-up and global companies. Europe needs more venture capital to support the growth of scale-ups….
- Europe should invest in future-proof areas, such as the silver economy to support the ageing population (home automation, health, entertainment, …) and technologies for sustainability (low power, the circular economy, …). In terms of technology, hardware accelerators and artificial intelligence are key elements of future computing systems.

In this article, we present a SWOT (strengths, weaknesses, opportunities, threats) analysis of the European computing systems ecosystem. We make a distinction between three stakeholders: (i) publicly funded universities and research institutions ("Science and Technology"), (ii) the computing industry and its market, and (iii) the local and European governments responsible for creating an environment in which research, innovation and commercialization can take place ("Policy and Government"). Most data in this article comes from the European Commission resport *Science, research and innovation performance of the EU 2022* [1].

Strengths and weaknesses

	Strengths	Weaknesses
Science and Technology	• High-quality education • Excellent research • Growing academia-industry link • World leader in lithography for semiconductor manufacturing	• Strong in research, but not in commercialization
Industry and Market	• Third largest market in the world • Stronger in systems than in components	• EU ICT contributes less to GDP than in other advanced countries • Lack of venture capital culture • Lack of advanced foundries
Policy and Government	• Common market • Decent public funding level of research	• Lack of ICT workers • Fragmentation of funding

High-quality education

Europe has an excellent and affordable educational system from preschool to university. According to the 2023 *Times Higher Education World University Rankings* [2], more than one third of the top 100 universities are located in Europe. The US dominates the top 10 and the top 100, but all international rankings put Europe as the dominant continent in the top 500. This shows that Europe has a very solid higher education system.

Participation in higher education is growing in Europe, and there is still a lot of room for growth in some parts of Europe, see Figure 1.

A positive evolution is that the share of graduates of higher education in the science and technology sector is rapidly increasing too; see Figure 2.

Given demographic evolutions, Europe will not be able to match the number of higher education graduates of China in the future. There are only three ways to increase the number of graduates.

- Try to further increase participation, although there is not an unlimited number of students that are qualified to undertake higher education (Europe sets 40% as a target; some countries have already reached this level).
- Try to increase the number of graduates via lifelong learning. That means that workers are (re)trained while working, or between jobs.
- A final option would be to try to attract more foreign students/graduates, especially those who have plans to stay in Europe after graduation. Given the fact that almost every country in the world is trying to stop brain drain, and has created incentives to bring successful expats back to their home country, the impact of recruiting overseas students is also limited, and the numbers will always be lower than the number of local students.

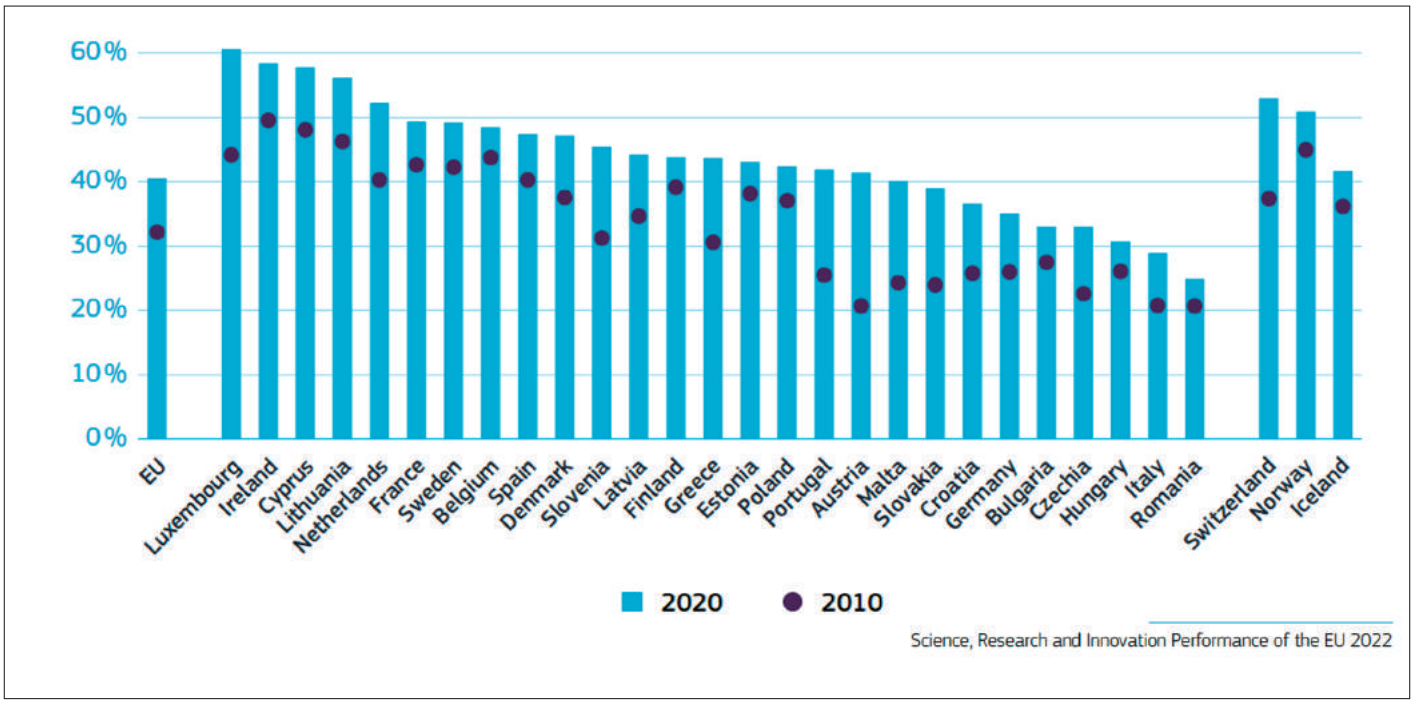


Figure 1: Share of population aged 25-34 who have successfully completed tertiary studies

European universities generally do not have access to the huge endowments of some US universities, see Figure 3.

In the European Union (EU), the students pay on average about 20% of the real costs as tuition fees while in the United States and the UK it can be up to 66%. That makes higher education affordable for most young people in the EU. Especially during a recession, it also allows freedom of choice. Students do not have to worry how they are going to pay back student loans. This is important for study programmes that do not lead to highly paid jobs, but are still societally very important (nursing, education, childcare, …). Most EU students do not start their career in debt.

Excellent research

European universities produce significantly more PhD graduates per 1,000 of the population than American or Chinese universities, see Figure 4. The majority of European countries perform better than the United States, even in science and technology.

During the last 20 years, Europe has maintained its global share of scientific publications, while the United States has seen a steady decline, and China has shown spectacular growth (Figure 5). Europe has not only retained its share of scientific publications, but also its share of the 10% most highly cited publications. The US is

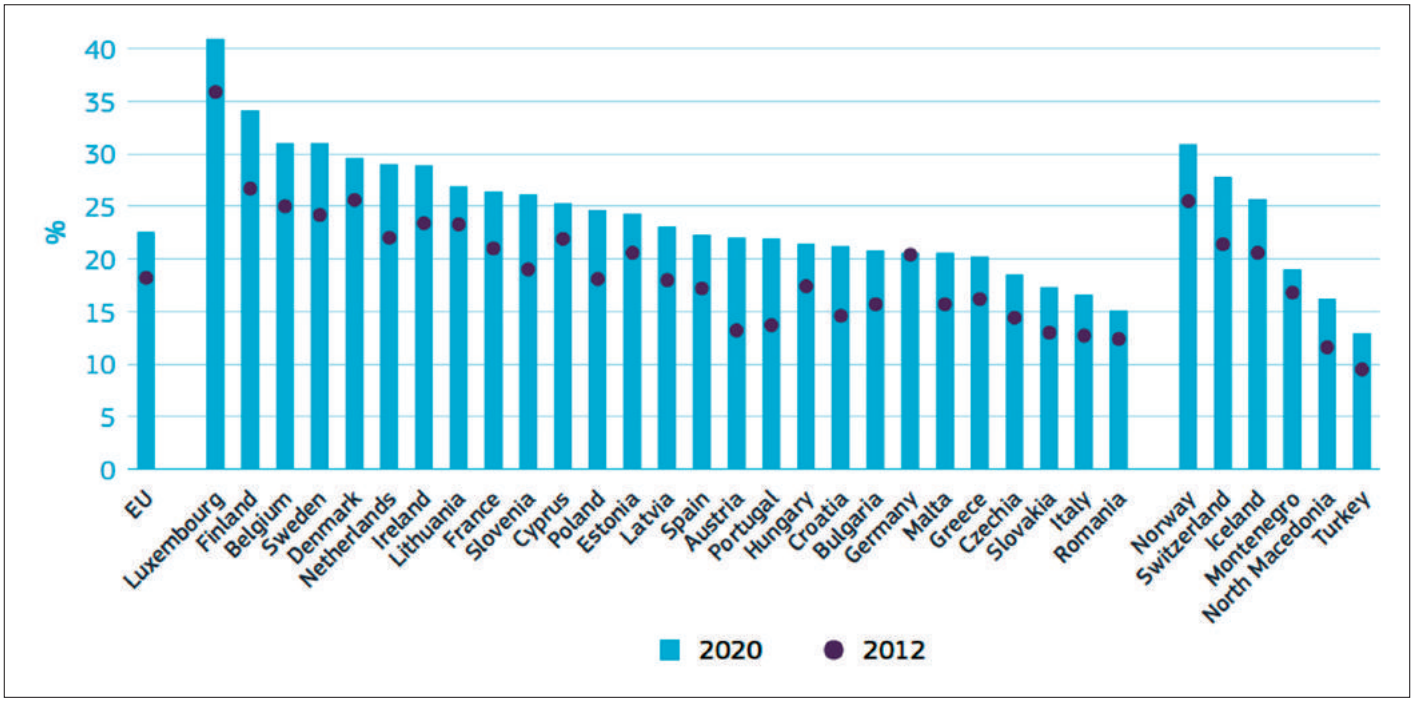


Figure 2: Share of higher education graduates employed in science and technology

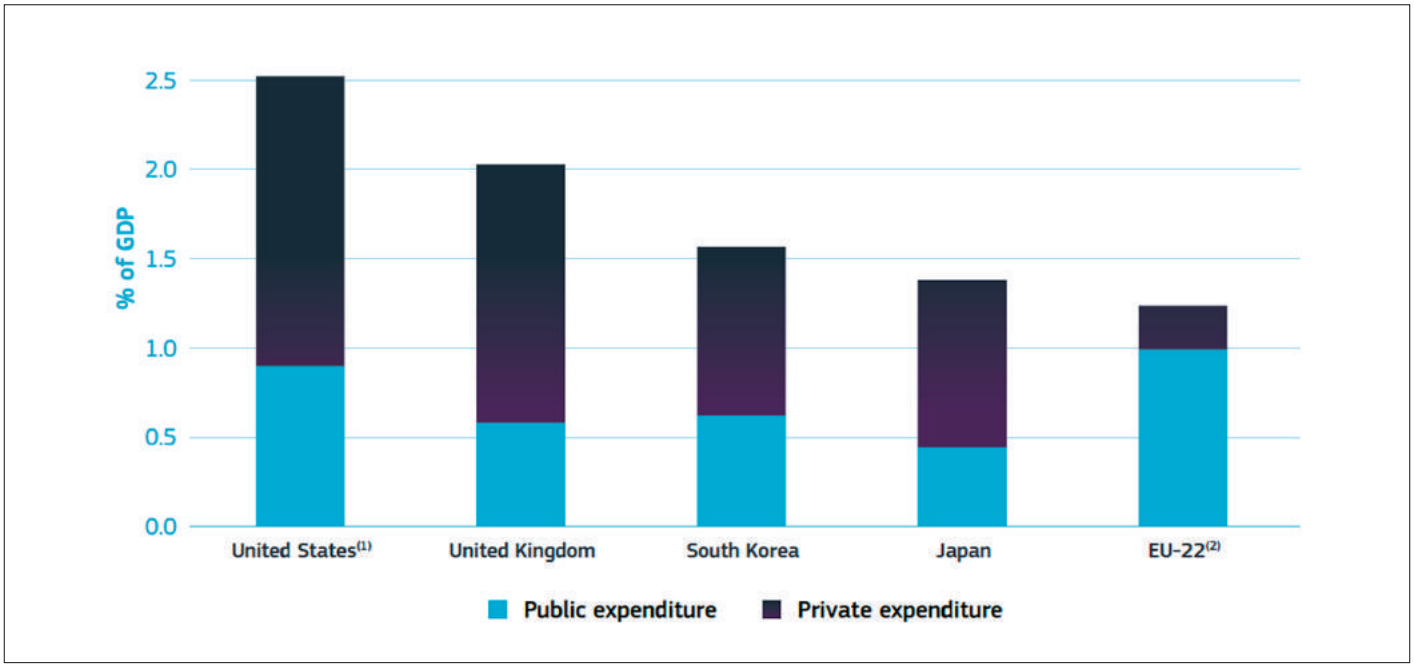


Figure 3: Expenditure in tertiary education, 2018

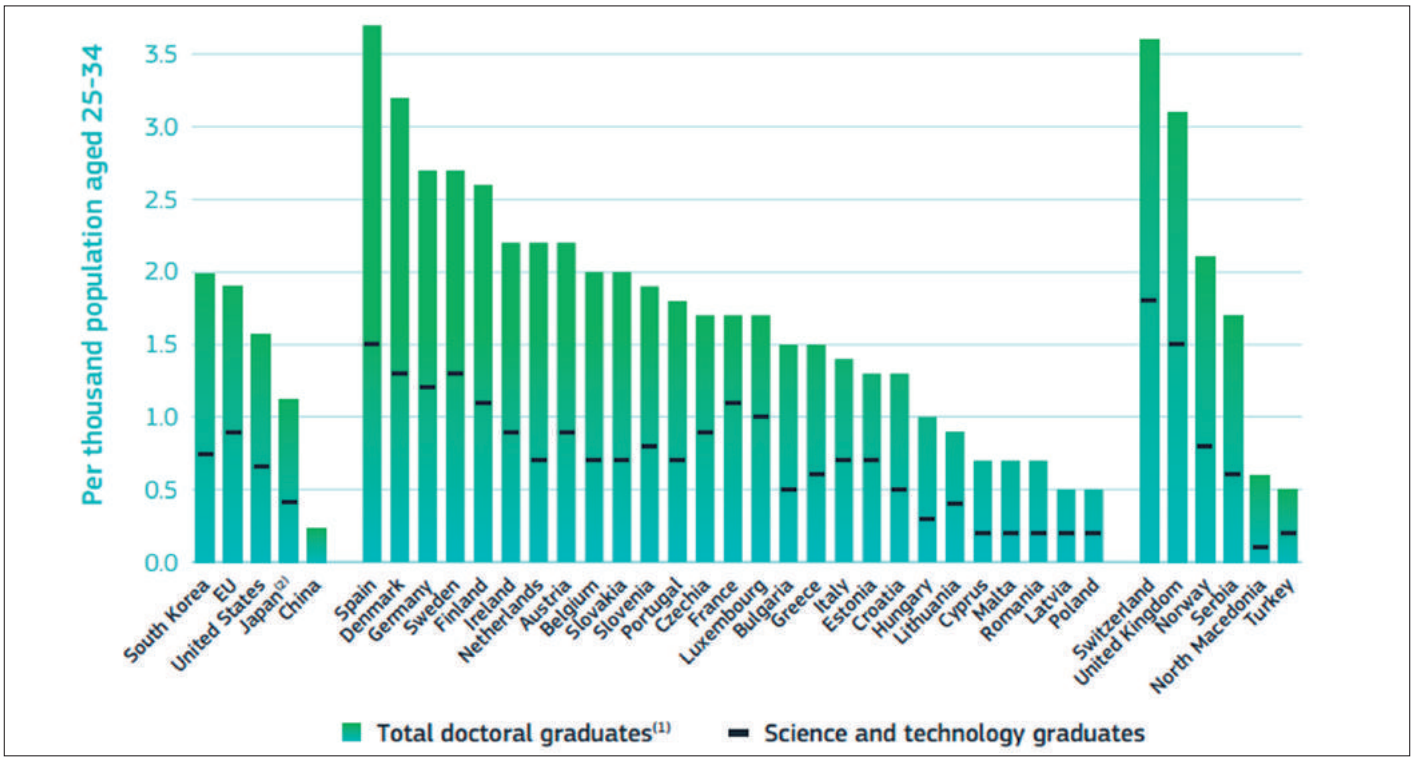


Figure 4: New doctoral graduates per 1000 population aged 25-34, 2017

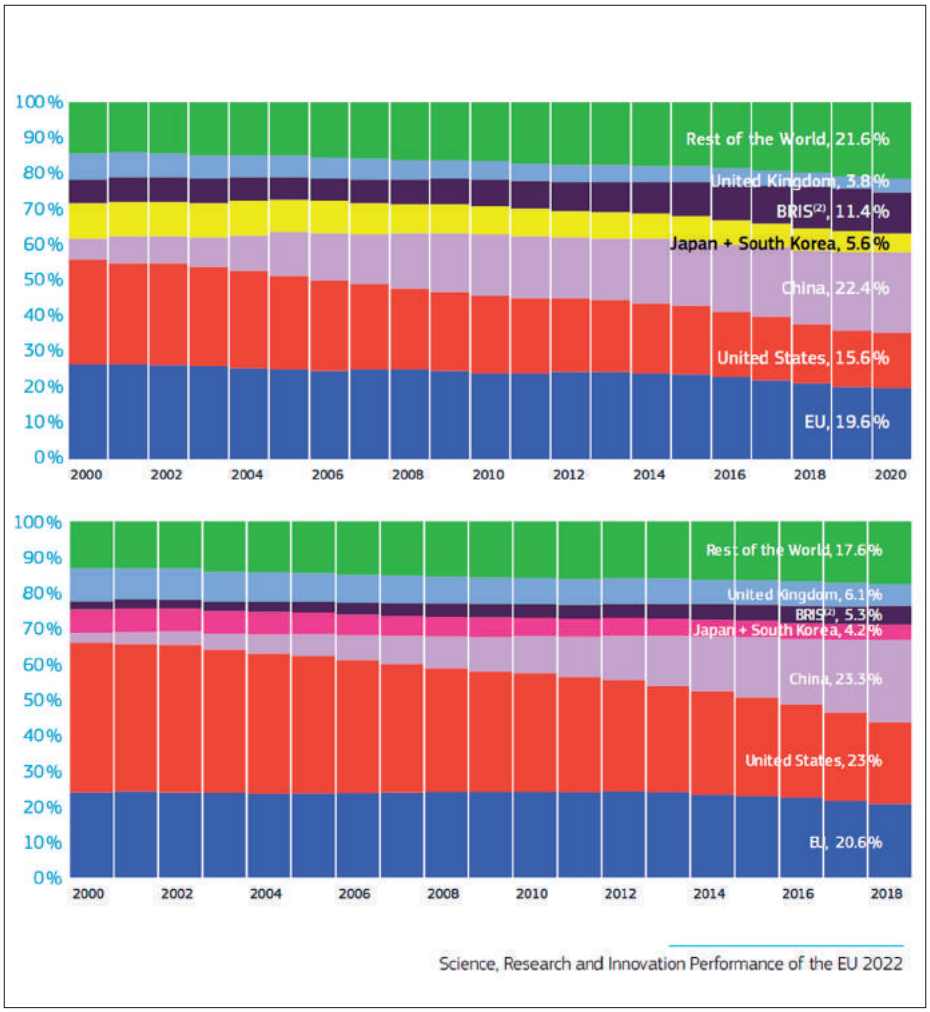


Figure 5: World share in scientific publications and 10% highly cited scientific publications

also losing its share of highly cited publications. The same trend is visible for the top 1% highly cited publications.

It is remarkable that the sum of highly cited publications from the United States and China has been almost constant since 2000, which seems to indicate that Chinese presence is growing at the expense of the US. The explanation is that the US and China are competing in the same domain (technology), while the EU publications are focussing on different domains (arts, humanities, …). If this trend continues, China will become the leader in publications and citations by the end of this decade. Hopefully, Europe will be able to defend its position in second place. As there is a correlation between the amount of public research funding and the number of highly cited publications, it seems crucial not to cut down on public research funding. The US and Japan have been cutting public research funding in the last decade, and their number of publications and citations have followed the same path.

Growing academia-industry link

Compared to the United States, Europe has now more public-private co-authored publications (Figure 6). This might be

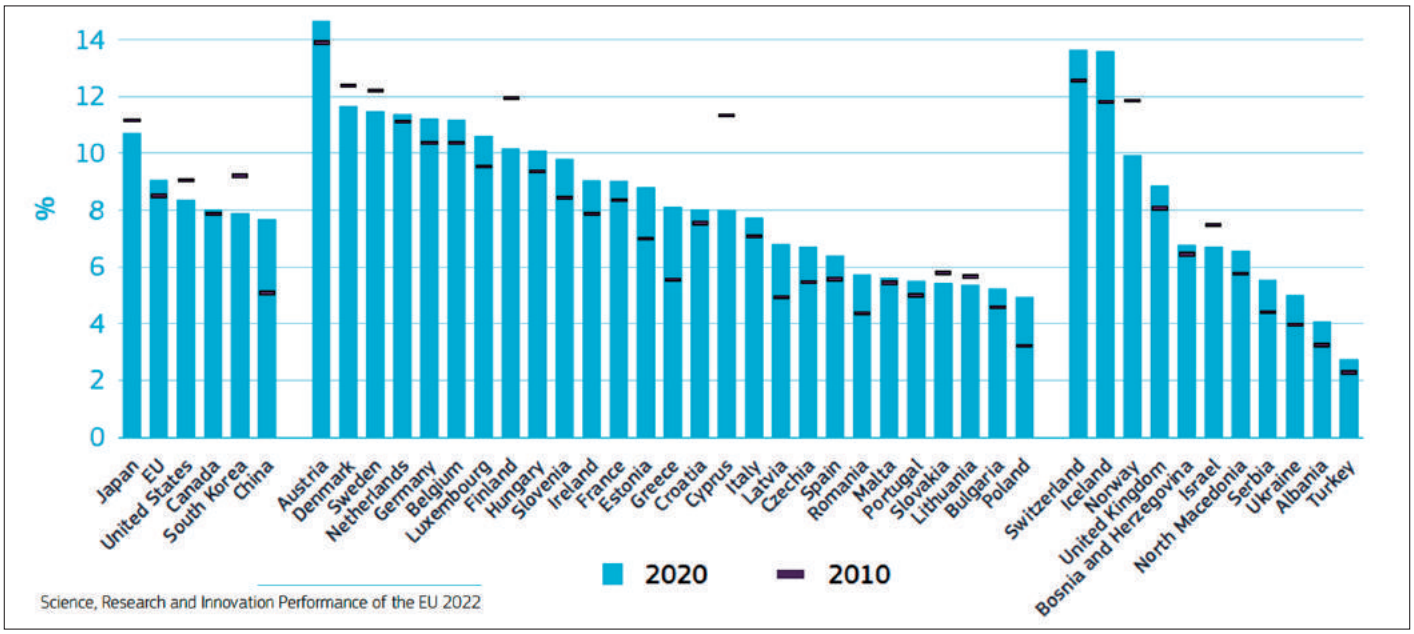


Figure 6: Public-private co-authored scientific publications in total scientific publications

a result of the EU funding instruments in which universities and businesses are encouraged to collaborate. It is definitely a positive evolution.

World leader in lithography for semiconductor manufacturing

Europe has several research institutes and companies that are key players in semiconductor technology development (including ASML for making advanced EUV lithography machines, CEA, imec and Fraunhofer). They are Europe's biggest asset when it comes to the further development of CMOS-technology, and their expertise might be crucial to the development of post-CMOS technology. This asset is very important because it means that Europe is at the table when important technology deals are being made.

The US and Europe want to become less dependent on foreign supply chains for their digital technology and want to create more domestic fabs. TSMC is investing in fabs all over the world, but so far not in Europe. Intel has already announced plans to build a fab in Europe.

Third largest market in the world

According to the International Monetary Fund, Europe (EU-27) has the third largest economy in the world (it lost its second place to China in 2021; see Figure 7). Obviously, Brexit has caused a serious reduction in the GDP of the EU, but the economy of China is growing much faster than Europe's.

Still, European businesses have access to a large internal market, with significant potential for growth in the new member states. Having access to a large internal market is an important advantage in times of troubled international trade relations. However, the European market is very fragmented due the diversity of regulations and languages across countries.

Stronger in systems than in components

In major embedded application domains, Europe is a global leader. According to the *Strategic Research Agenda for Electronic Components and Systems 2020* [3], Europe produces 25-30% of global annual value at the system level, compared to 13% at the component level (Figure 8, Figure 9). The system level is expected to grow tenfold between 2016 and 2025 while the component level will only double. The system level is a clear strength in Europe, and a strength that we should exploit. The

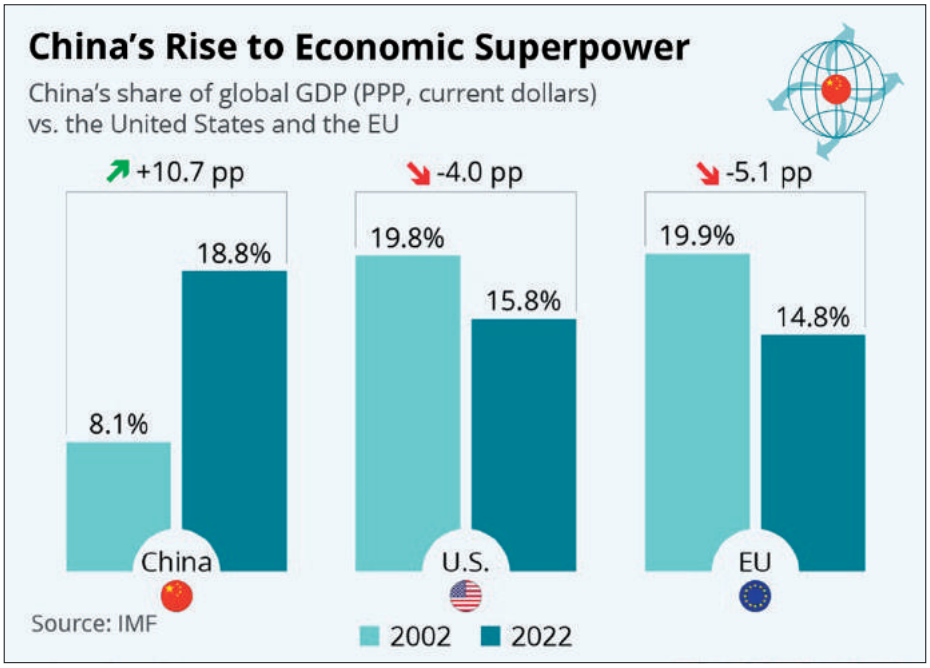


Figure 7: Top 3 biggest global economies

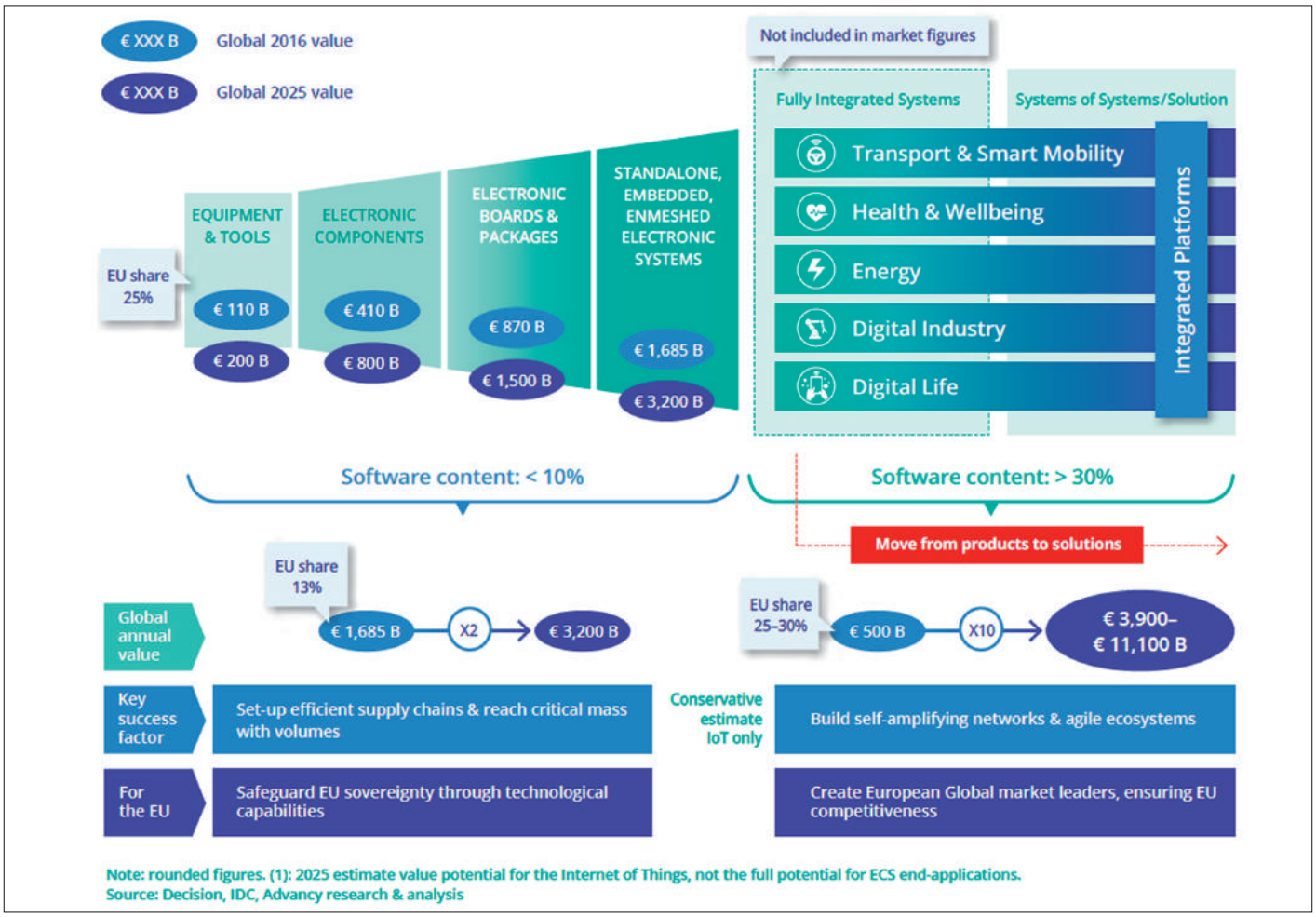


Figure 8: Embedded application domain markets. (Source: ECS)

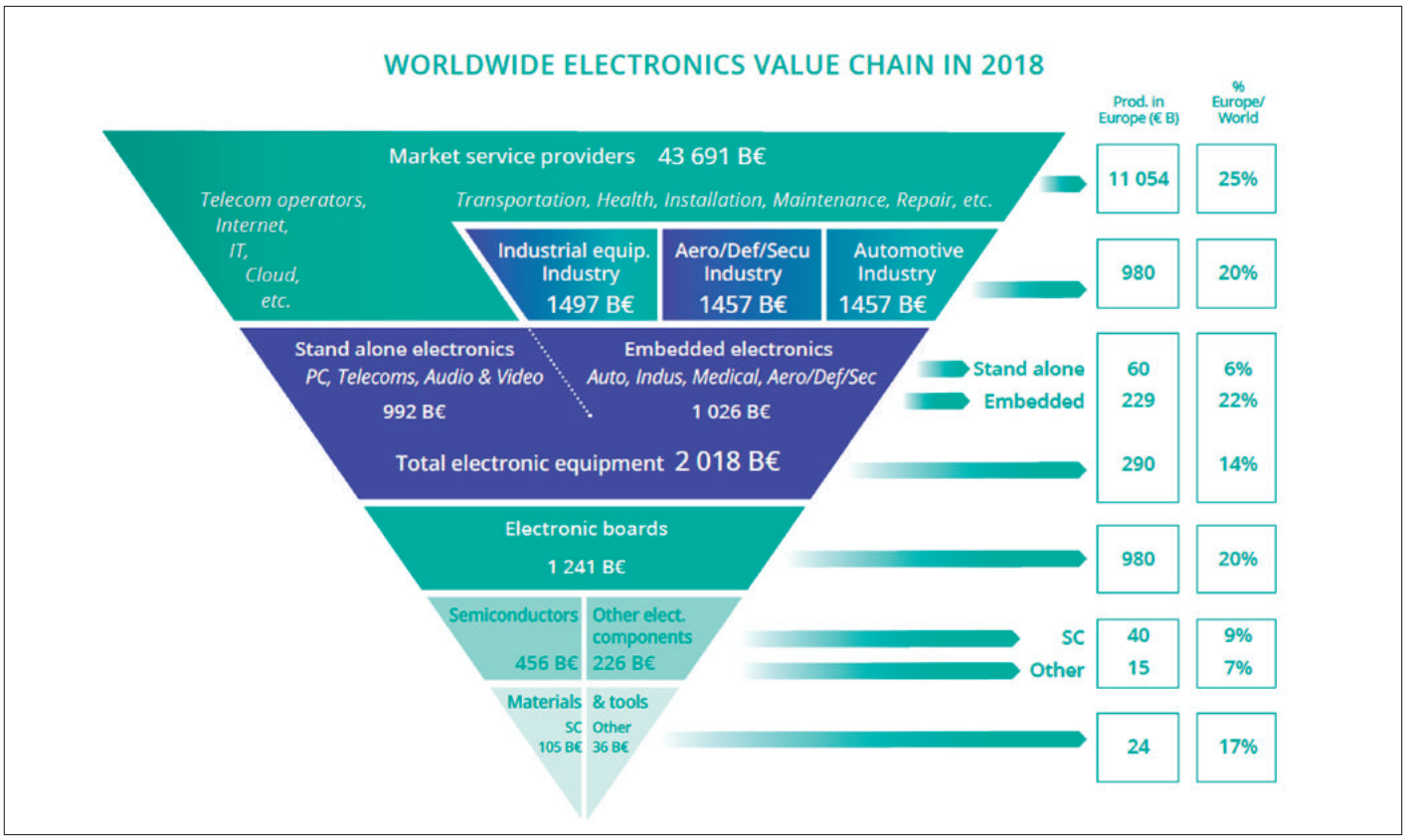


Figure 9: Worldwide electronics value chain in 2018. (Source: ECS)

sector that is particularly strong is transportation (automotive, air, rail).

Common market

At the policy level, one of the strengths is the common market, and the fact that Europe can act as one economic bloc in global trade negotiations. Individual countries do not have to negotiate individual agreements. However, there is still a long way to go before Europe becomes a fully integrated market with one set of laws, one currency and one tax system. The difference in minimum wages across Europe shows how pronounced the difference between countries is (Figure 10).

Strong public funding

Europe has a variety of research funding instruments, complementing national funding instruments. The research and innovation programmes of the European Commission help to stimulate research collaboration. European Research Council instruments support research excellence, the flagship programmes aim to create critical mass in key research areas, the European Institute of Technology aims to stimulate research and innovation, and joint undertakings (JUs) like the Key Digital Technologies (KDT) JU, soon to become the Chips JU, aim to pool local and European funding to encourage cross-border research and innovation.

The total amount of public research and development (R&D) funding available makes Europe a good place to carry out R&D (at 0.73% of GDP). Worldwide, Europe is in second place after South Korea (Figure 11).

However, the relatively high amount of public funding across the EU does not compensate for the low R&D investments by industry (see weaknesses). When considered as a whole, Europe is dramatically lagging behind other regions of the world. The aim for Europe is to spend 3% of GDP on R&D, but it is still far away from that target (Figure 12).

The intensity of R&D translates into the number of researchers employed. Although Europe produces a higher number of PhD graduates per 1,000 of the population than any other world region, this has not resulted in more researchers in employment. In addition, almost half of the researchers working in the EU are employed by the public sector (Figure 13).

Furthermore, there is great variation between European countries (Figure 14). There are frontrunners (first quadrant), and laggers (third quadrant).

Weaknesses

Strong in research, but not in commercialization

Europe is lagging behind Japan and at a comparable level with the US with respect

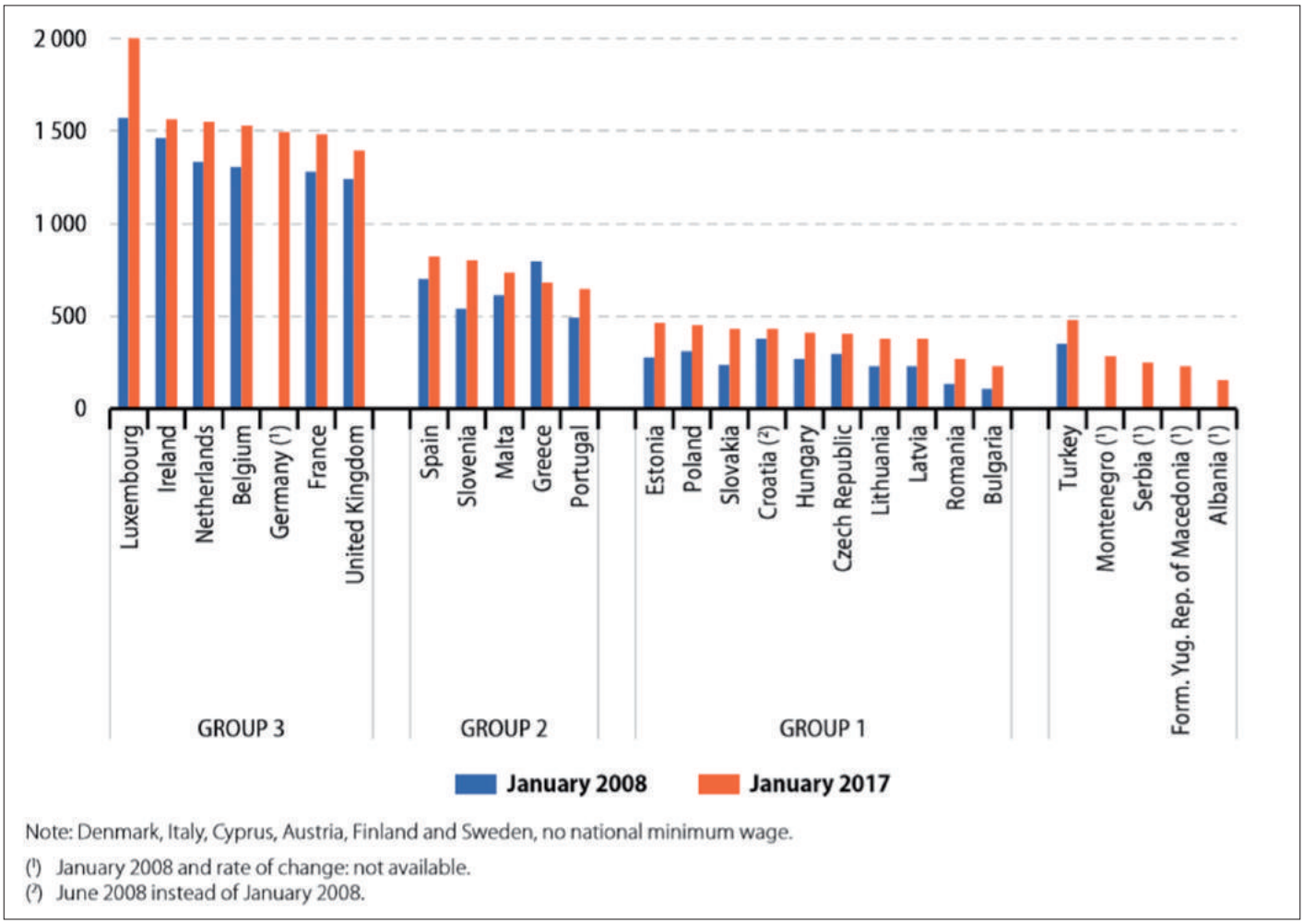


Figure 10: Minimum wages, January 2008 and 2017 (EUR per month) (Source: Eurostat)

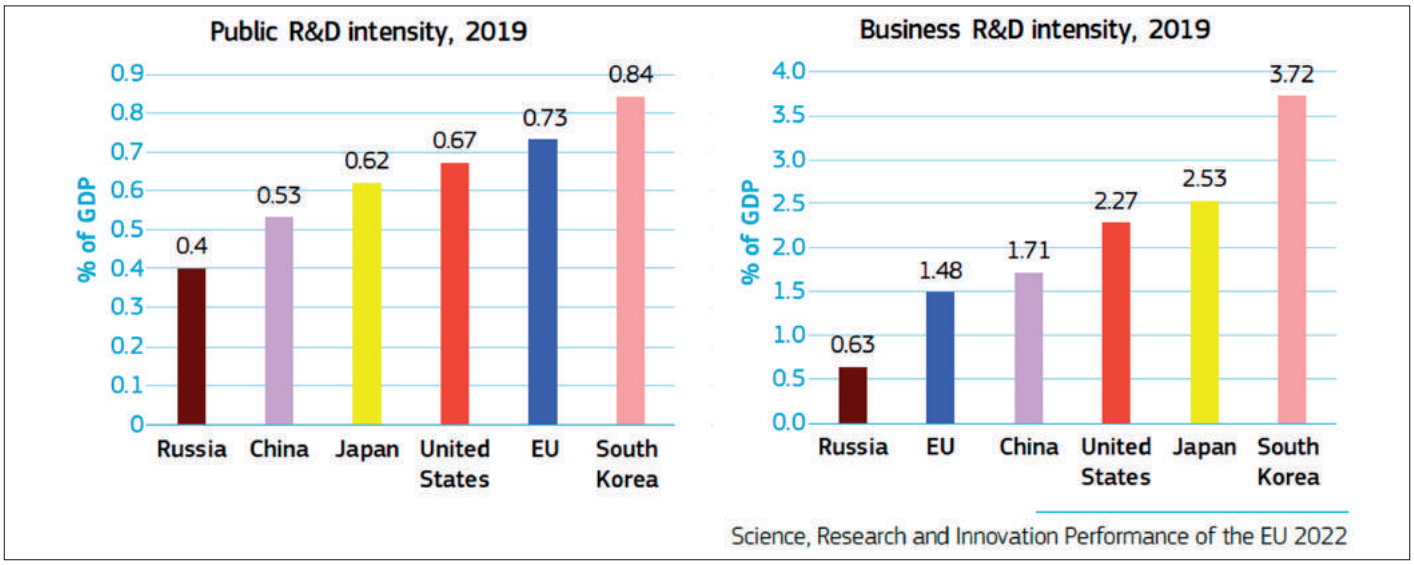


Figure 11: Evolution of public and Business R&D intensity 2019 (Source: DG Research and Innovation)

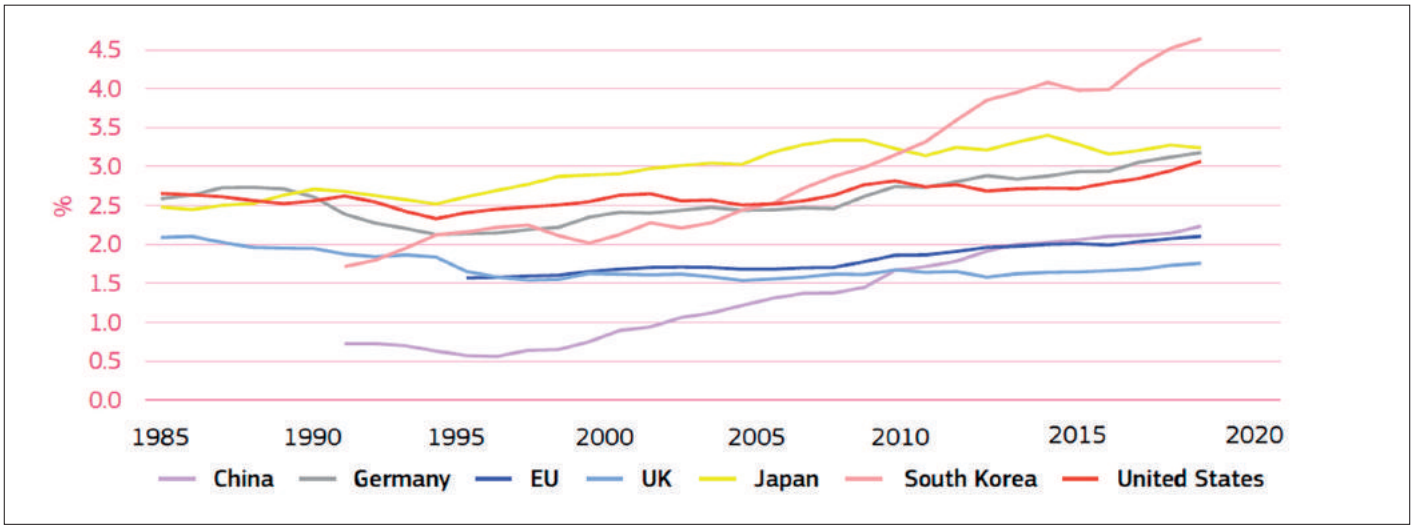


Figure 12: Evolution of total R&D intensity 1985-2019 (Source: DG Research and Innovation)

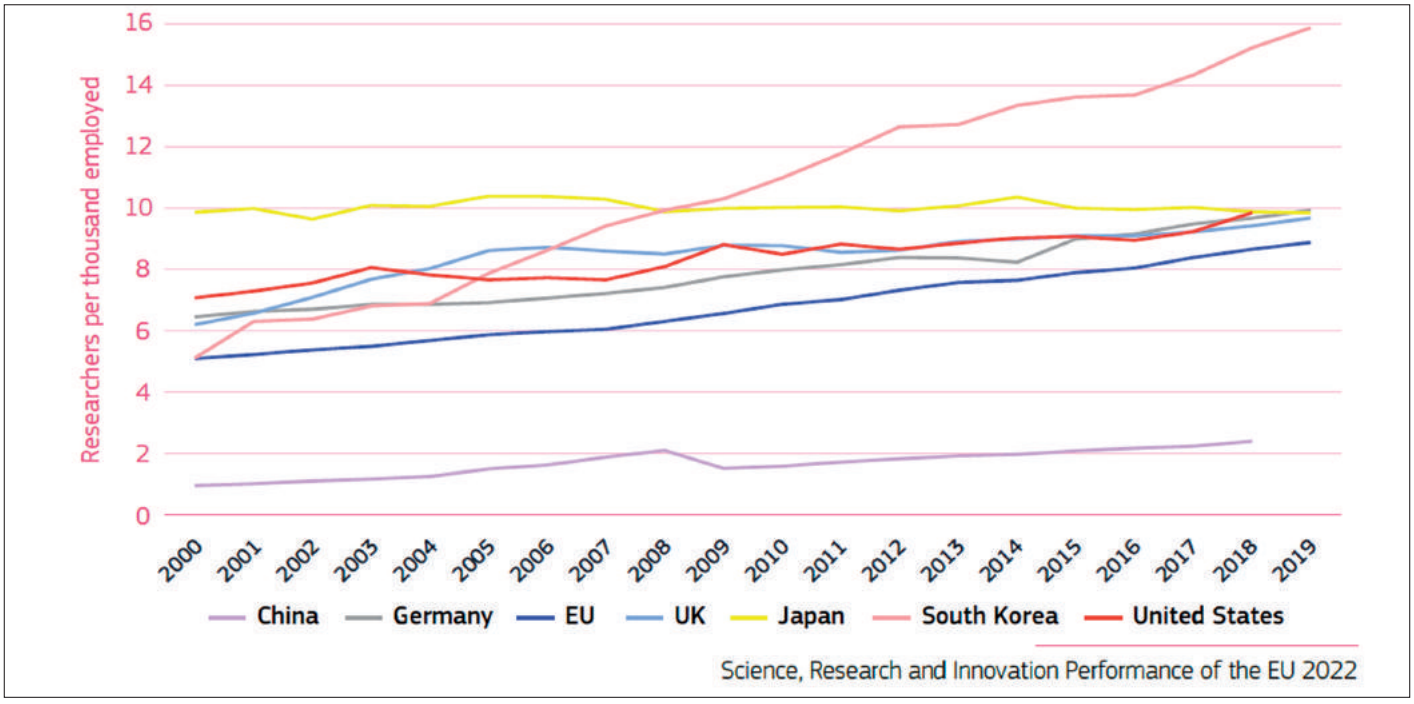


Figure 13: Total researchers (FTE) as % of total employment 2000-2018 (Source: DG Research and Innovation)

to the innovation output indicator (based on four components: patents, employment in knowledge-intensive activities, trade in knowledge-based goods and services, and the innovativeness of high-growth enterprises), see Figure 15.

On the other hand, in recent years several European cities have emerged as start-up ecosystems. Regional and European authorities are currently investing heavily in the creation and the support of start-up ecosystems in all European urban areas. Globally, 37% of all newly created ecosystems are based in Europe (Figure 16).

ICT contributes less to GDP in the EU than in other advanced economies

The European ICT industry contributes around 4% to GDP, compared to more than

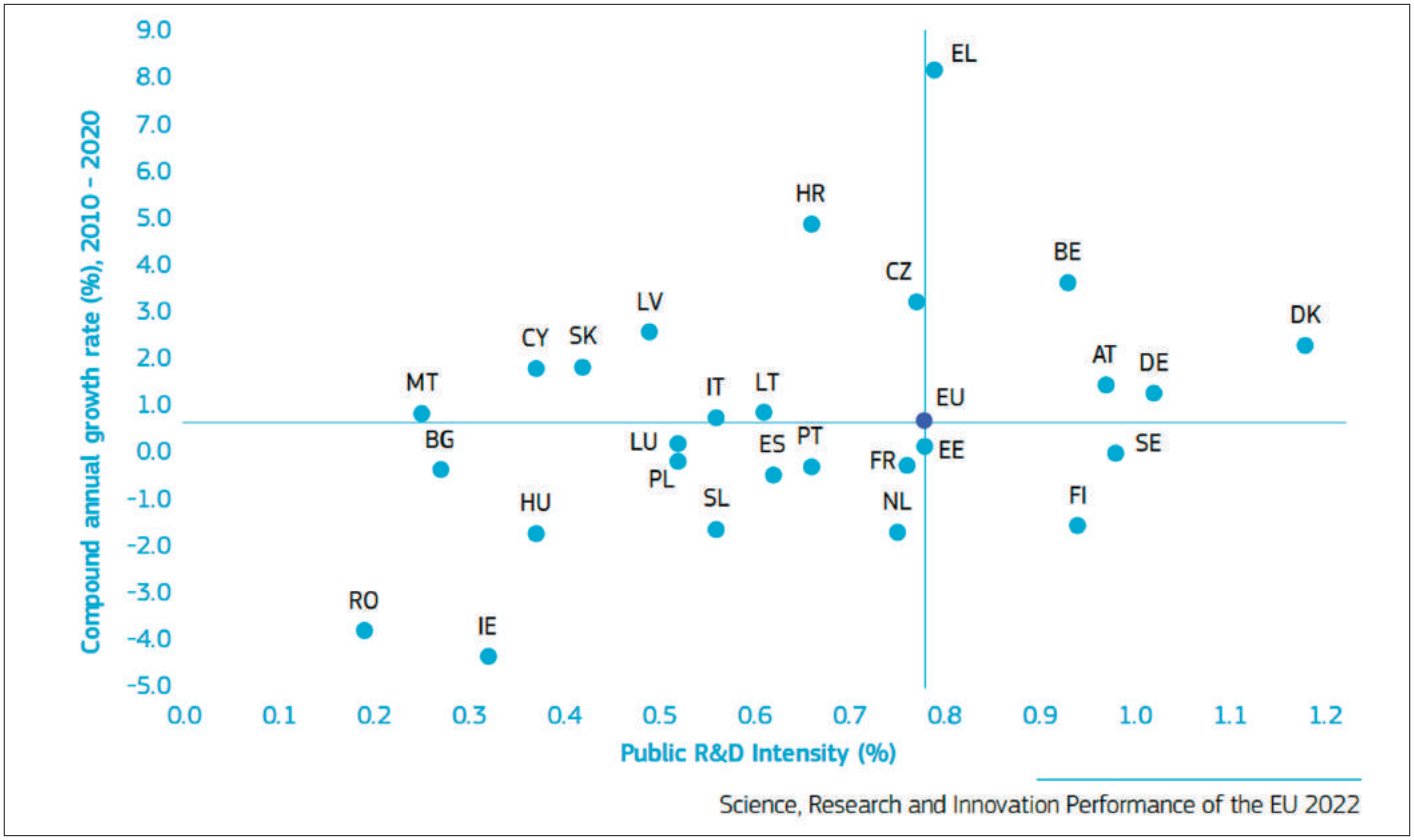


Figure 14: Public R&D intensity, 2018 and compound annual growth (%) 2007-2018 (Source: DG Research and Innovation)

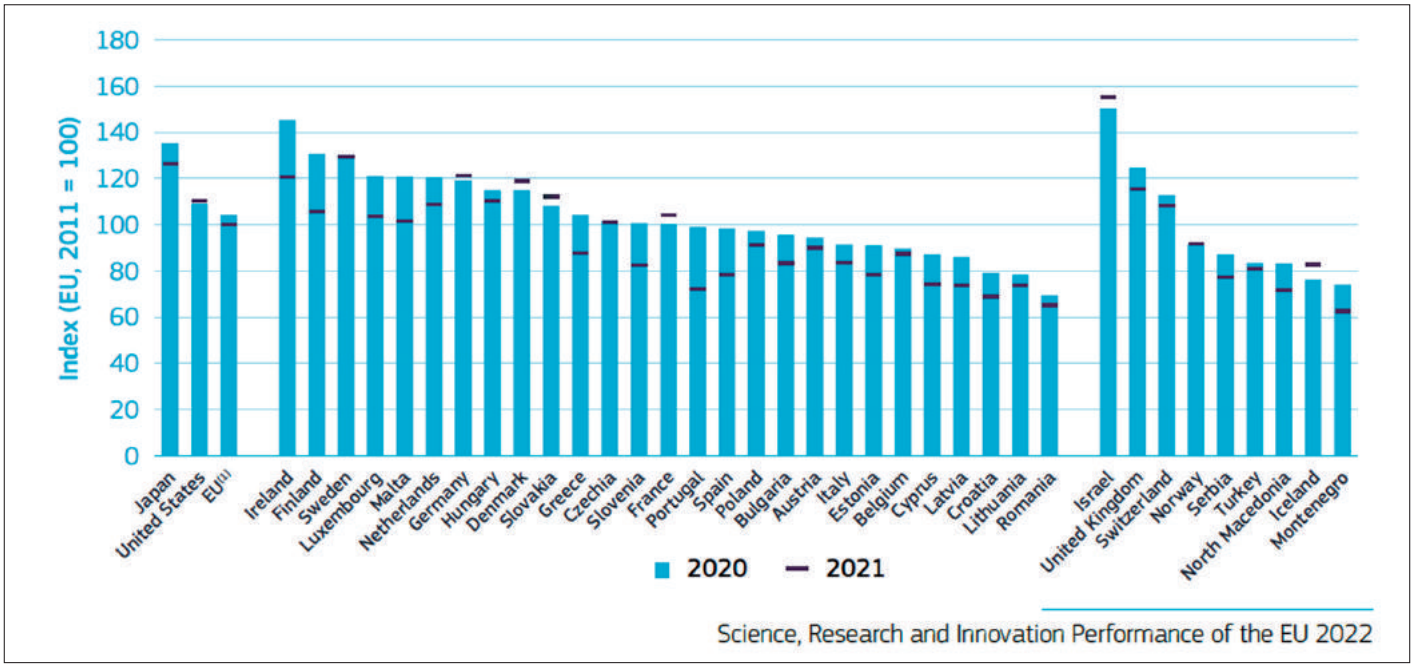


Figure 15: Innovation output indicator (EU, 2011=100), 2020, 2021 (Source: DG Research and Innovation)

6% or more in competing world regions (Figure 17). One explanation is that Europe lacks major technology companies like Alphabet, Apple, Amazon, Meta or Microsoft (in the US) or Baidu, Alibaba, Tencent and Xiaomi (in China), as well as other major ICT companies like HP, Dell and IBM, and the ecosystem supporting them.

The lack of such large corporations can be explained by the lack of venture capital (VC) culture in Europe. In order for companies to grow to US$50-100 million, they have to enter non-European capital markets like the United States or China. The US market is very competitive and sophisticated, and Asian markets are even more challenging. Even growing within Europe has its challenges, because Europe is not a single entity, it is composed of a plurality of markets, languages, laws, cultures and so on. Therefore, it is difficult for a company to address the whole of Europe without extra work to adapt to each country. As an example, voice assistants appear later in non-English speaking countries due to the additional effort required to adapt them to different languages. Neither US or Chinese companies face such challenges.

The fact that Europe lacks major ICT companies has far-reaching consequences: it also means that venture capitalists are less eager to invest in European start-ups and scale-ups because there are fewer companies that might be able to acquire them. Companies that do grow significantly are often acquired by non-European companies: Nokia was acquired by Microsoft, Movidius and Yogitech by Intel, Silexica by Xilinx, for example. Fortunately, there are also counterexamples like Sysgo, which was acquired by Thales.

Non-European business leaders often articulate a clear vision of the future and tend to generate considerably more publicity than their European counterparts.

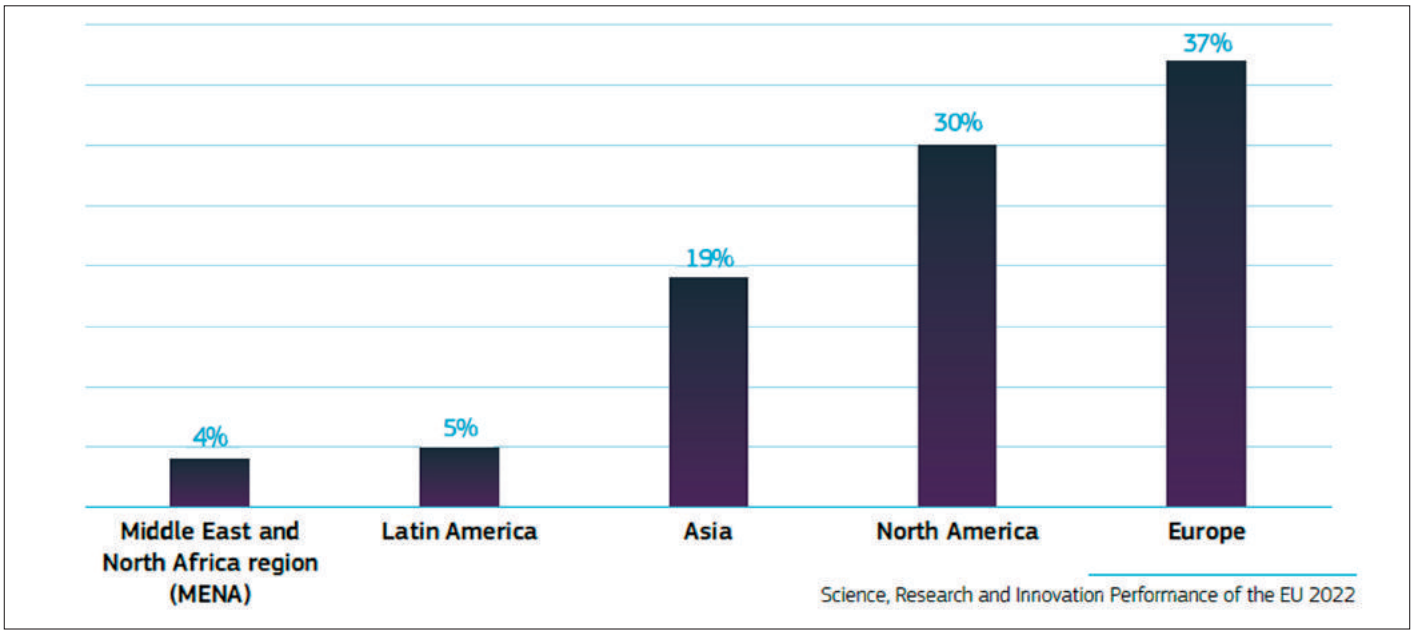


Figure 16: Share of emerging start-up ecosystems per world region (Source: DG Research and Innovation)

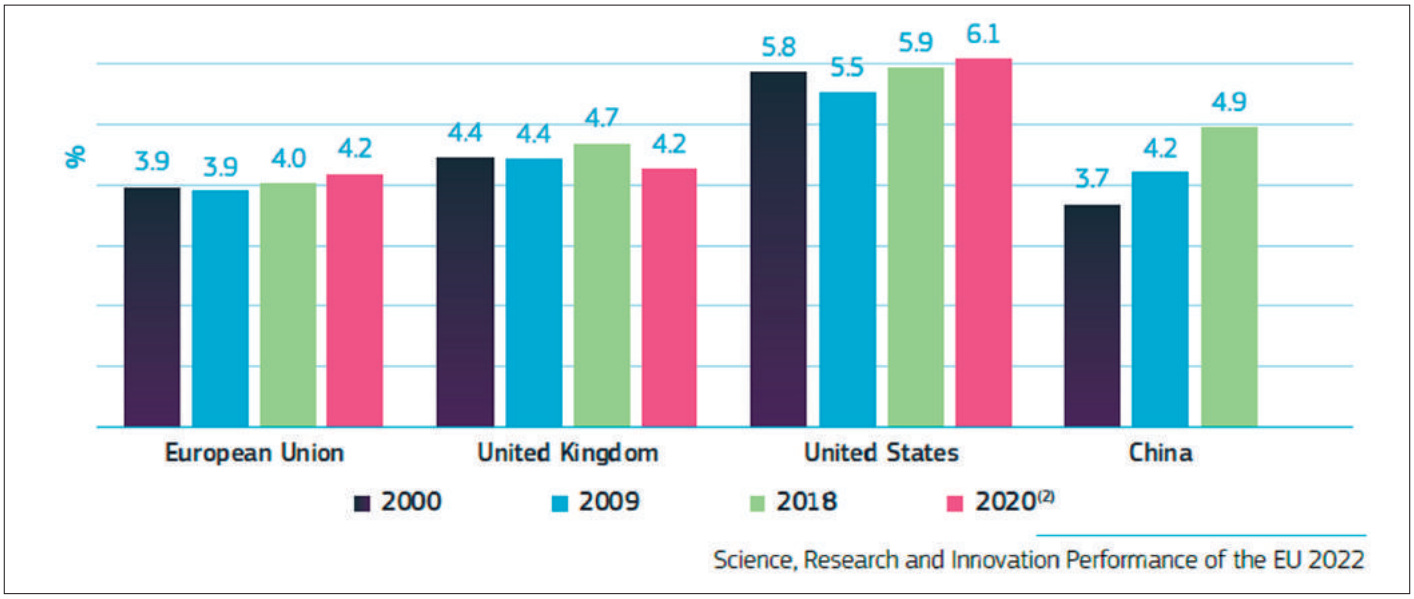


Figure 17: ICT sector value added share of GDP

While the personalities in question may vary, from the apparent instability of leaders like Elon Musk to the steady guidance of others like Lisa Su, what they have in common is frequent media coverage. In contrast, very few people know the CEOs of major European computing companies like Infineon, Ericsson and STMicroelectronics, who lack the glamour (or, in some cases, notoriety) associated with their international equivalents.

Lack of VC culture

More generally, Europe lacks a VC culture and, in this metric, the gap between the United States and Europe could not be bigger (Figure 18). This observation, in combination with the large number of young start-up companies, is problematic. It means that start-ups have to fight hard to get the funding to become a scale-up company. Some of them turn to the US to get VC investments.

Fortunately, the majority of the investments that do take place are in the ICT sector (Figure 19).

Lack of advanced foundries

There used to be foundries in Europe, but they were acquired by non-European companies and disappeared. The fact that Europe depends on foreign foundries means that it has to import most of its semiconductors. The leading foundries are not located in low-wage countries, meaning that they did not leave Europe due to labour costs. Given the fact that Europe is a world leader in the development of the technology used in foundries (CEA, imec, Fraunhofer, ASML), it is surprising that no large foundries are left in Europe and that Global Foundries some years ago decided to stop the development of 7nm technology and instead make its 14/12 nm FinFET platform more relevant to its customers.

One explanation is that European countries did not aggressively invest in new foundries (as was the case in South Korea and in Taiwan), and that European VCs are not interested in foundries (while they are in the United States). Another is that the European customers of the foundries like STM, NXP and Infineon are making products that do not require advanced processes because their market is micro-

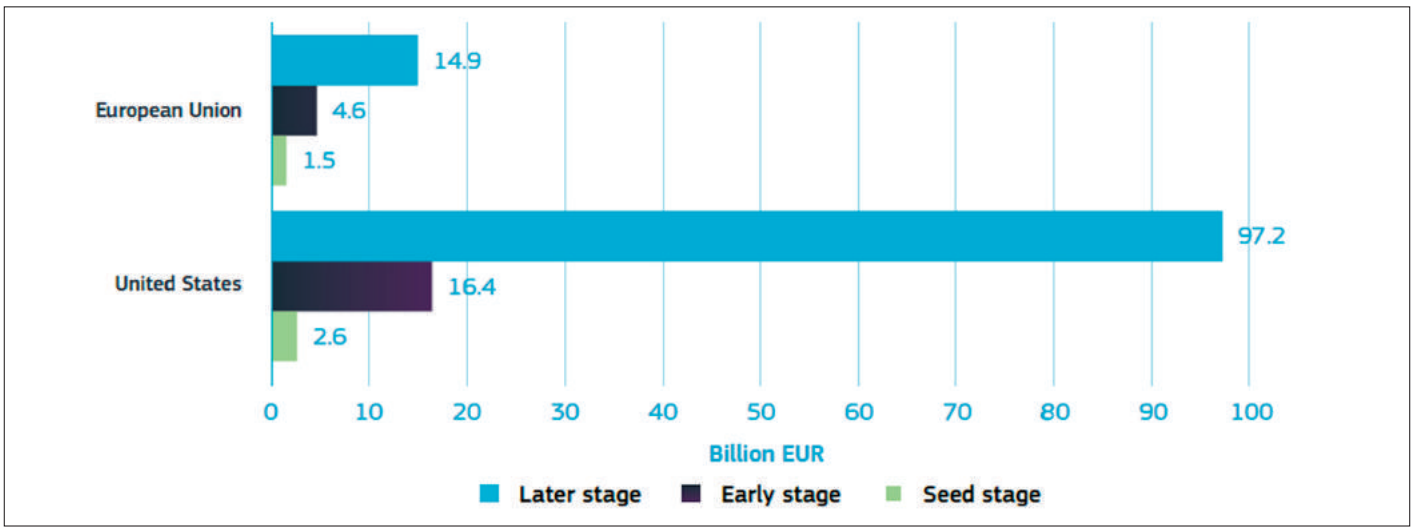


Figure 18: Venture capitalist investments in the EU vs the US by development stage in 2020 (Source: DG Research and Innovation)

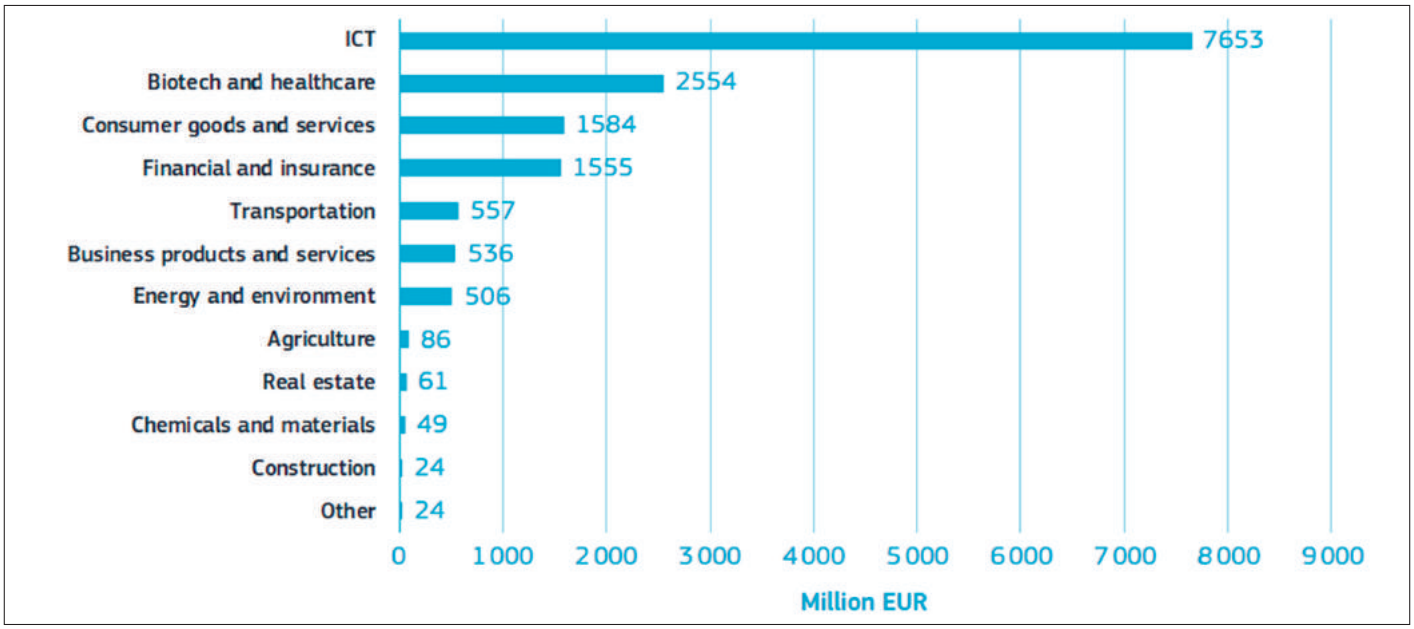


Figure 19: Venture capital investments in the EU per sector, 2021 (Source: DG Research and Innovation)

controllers instead of microprocessors, analogue devices versus memories.

To address this issue, the EU aims to create production capacity of 20% of the global chip market in Europe by 2030 to secure digital sovereignty, as part of the Chips Act [17]. Recently, Intel announced plans to build a state-of-the art foundry in Germany. TSMC might make a similar decision in the near future.

EU Chips Act and international public initiatives for semiconductors

On the 9 March 2021 the EU President von der Leyen announced the "European Digital Decade" and the associated "Digital Compass" [4, 5] stating the ambitions and targets of the EU policy for the digital transformation of our society and economy by 2030. The pandemic highlighted the central role of digital technologies in building and maintaining a successful and sustainable future. It also exposed the divide that still exists between areas in EU, as well as strategic weaknesses compared to some of the other world areas.

Among the objectives set, one concerns the semiconductor industry and aims at reaching a 20% market share of the semiconductor world market in value by 2030. As the current share is below 10% and the prediction for the market size is to double over the period, this implies at least a five-fold increase in production capabilities in EU. Moreover, the objectives include leading-edge technologies (below 5nm) where the current share of EU is 0%. Such ambitious targets cannot be attained without massive investments by EU industry and, given the current size and number of EU players, without large companies from outside EU investing here.

To support such massive investments, the EU Commission has proposed a new piece of legislation submitted for approval to the EU Parliament and the EU Council establishing a wide range of actions. This "European Chips Act" [6] proposal was presented on 8 February 2022 and is composed by a series of documents including a 36-article legislative text that was complemented, on 12 May, by a "Staff Working Document" [7] which provides a large number of details on the expected implementation.

The Chips Act is structured around three "Pillars". Pillar 1 addresses the R&D effort necessary to maintain and improve technological capability in Europe and aims at creating "Pilot Lines" and networked "Design Competence Centres" that could support the EU stakeholders in moving faster towards leading-edge technologies and inject innovations in the industry. Pillar 2 addresses the need for support to increase capacity in EU by establishing a fast track for the validation of national funding plans to "first of a kind" new fabrication facilities taking the EU beyond existing capabilities. Pillar 3 aims at creating an early warning and crisis management system to avoid industry shortages affecting critical EU sectors.

Currently, the outcome is still not clear (e.g. the ITRE Committee of the EU Parliament has recently published a list of over 100 proposed amendments [8] to the Chips Act) but the reactions are mixed. The accent on leading-edge technologies is questioned as it is seen as unreachable and outside the scope of the EU semiconductor industry, although EU industrial semiconductor tools suppliers see it as fundamental to maintain their leadership positions in some market segments (lithography, atomic layer deposition, ….). Bringing the fabrication facilities of non-EU players to Europe is seen by some as a drain on already limited public investment support and skilled labour force that would penalize EU players even if not in direct product competition.

Some member states do not see the benefit how the initiative would benefit them and are against the principle of public support for the semiconductor sector. At the other end of the scale, some stakeholders lament the low level of EU support (about €4 billion for Pillar 1) compared to similar actions in US, China and generally in Asia. The reliance on national funding for Pillar 2, which is likely to privilege national interests compared to more EU-wide needs of technology sovereignty and secure access to supply of chips, has also been raised as a concern.

In any case, just the announcement of the initiative has spurred a series of announcement of new facilities. Intel has announced plans to massively invest in Germany and Italy respectively on an advanced foundry facility (about €15 billion initially and up to €68 billion over 10 years) and advanced 3D and packaging (about €1.5 billion and up to €5 billion). These investments will be complemented by R&D, design and customer support investments in France, Spain and Poland. STMicroelectronics has announced an investment of nearly €700 million in Catania (Italy) on SiC production and the EU has already cleared €270 million of state aid for the endeavour. STMicroelectronics, Soitec and GlobalFoundries have announced a joint effort of about €1.5 billion in Crolles (France) on the FD-SOI ecosystem with both dedicated substrate production and a fab capable of producing up to one million 300mm wafers per year.

As a comparison, the US in their Chips Act have dedicated US$53 billion of federal budget and US$24 billion of tax credits to the semiconductor sector on top of the individual states' initiatives. Of the US$53 billion, US$15 billion are dedicated to R&D initiatives very similar to the EU "Pilot Lines" and the existing 300mm infrastructures of imec, CEA and Fraunhofer in a clear attempt to repatriate some of the R&D currently done in EU by US industrial actors.

China published plans are of the order of US$159 billion over 10 years with some of the larger regions (Shanghai, Beijing, Chengdu, …) investing another US$40 billion.

South Korea has launched a tax incentive to support US$450 billion of industrial investments in R&D and manufacturing until 2030, while Japan only in the last year has allocated US$5 billion to the sector and has signed contracts to support the installation of TSMC (one fab and one R&D centre) domestically.

Lack of ICT workers

Europe lacks hundreds of thousands of ICT workers. Fortunately, there is significant growth in the number of graduates (Figure 20), but the number of graduates does not yet match demand. It seems that Europe is not succeeding in convincing large enough numbers of high school students to start a career in the ICT sector. This is unfortunate because the competitiveness of this sector in Europe will depend on a sizeable workforce in order to innovate in big data analytics, artificial intelligence, robotics and so on.

Encouraging well-trained foreign workers to relocate to Europe en masse to help mitigate the shortage is not an effective solution. First, Europe needs more than one million ICT workers in the next decade. Secondly, the countries of origin try hard to keep local talent in their own countries. Finally, Europe has become less inviting to immigrants during the last decade. To complicate things further, foreign ICT workers will be attracted by well-paid jobs in the major global innovation hubs, and it will be more difficult to convince them to accept a job in smaller cities, or in poorer countries.

The only long-term and sustainable solution is to invest heavily in the technical education of local people. In parallel, a set of highly visible European competence centres in strategic digital areas could (i) stop the brain drain of high-potentials for other world regions, and (ii) attract high-potentials from outside Europe.

Fragmentation of funding

The public funding system in Europe is highly fragmented. There are national funds, regional funds and European funds. There are funding instruments for applied research, for innovation, and for fundamental research. There are individual grants and collaborative research grants. A particular research proposal could fit multiple funding instruments and calls. Sometimes a research proposal can only be funded if different agencies agree to each fund a part of the proposal. On top of all this, the success rate for research proposals is sometimes lower than 10%.

Within a funding agency, different committees deal with particular topic areas, which makes it very hard to secure funding for multidisciplinary project proposals because committees tend to give priority to the proposals that belong to the core of a topic area, leading to lower acceptance rates for interdisciplinary projects. It is therefore very hard for technologies that are common to several application domains, such as research in computing hardware and software, to be funded for their own intrinsic development. Instead, they must piggy-back on more domain-specific project calls. The organizational structure of the funding agency thus ends up constraining the research work that can be proposed in one single project. The fact that European Regional Development Fund has also started to be used to fund research only adds to the complexity.

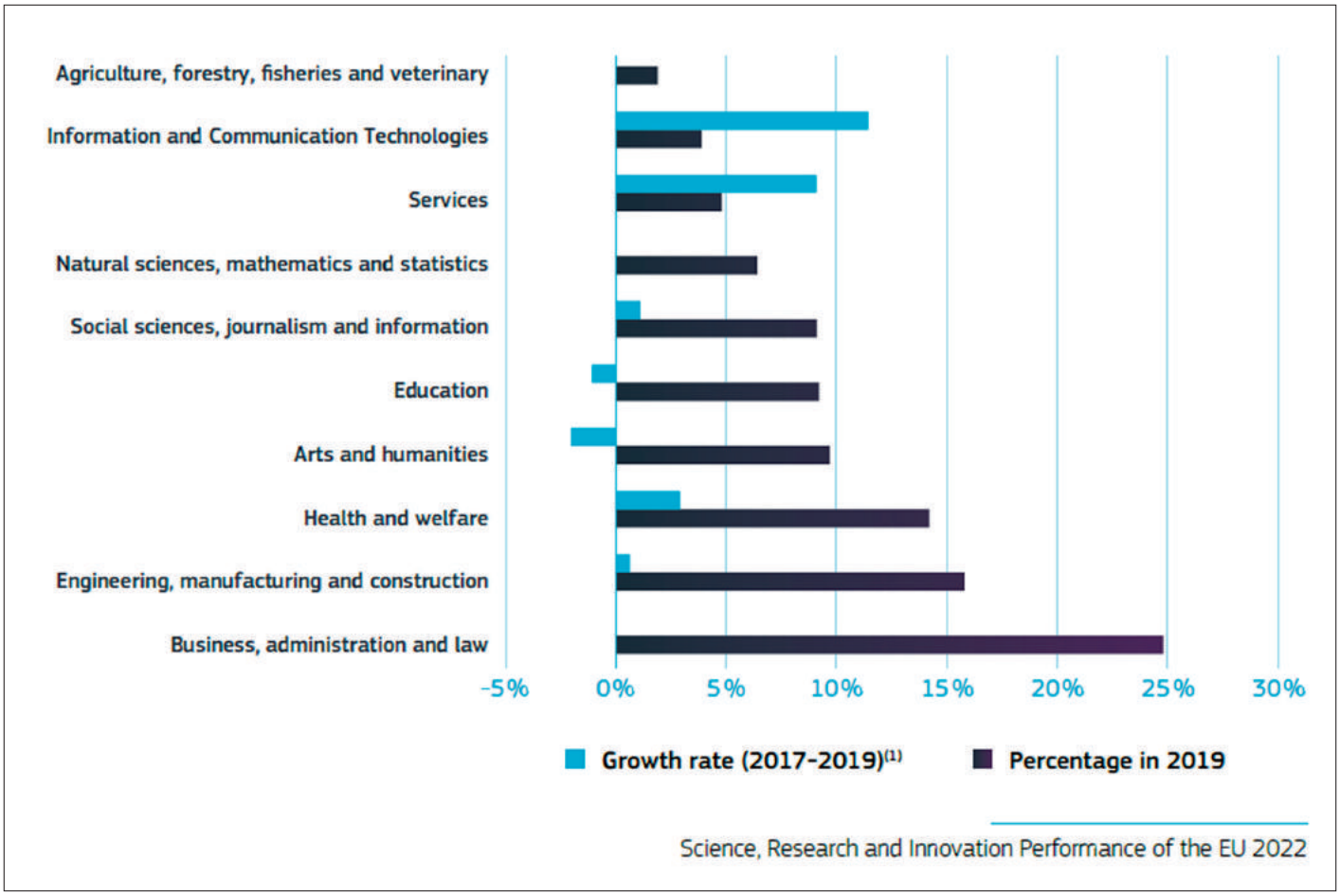


Figure 20: Share and growth of tertiary graduates by field of study in the EU

Opportunities

Fortunately, there are also opportunities.

	Opportunities	Threats
Science and Technology	• The end of Moore's law	• Economic stagnation and shrinking work force • Brain drain
Market and Industry -	• Embedded systems, IoT, CPS, edge intelligence • Open source	• Saturating markets • Computing initiatives in countries such as China, Russia and Japan
Policy and Government	• Solutions for societal challenges	• Geopolitical instability

The end of Moore's law

The increase of sequential performance at the pace of Moore's law already ended a decade ago (end of Dennard's scaling); parallelism kicked in to keep performance increasing in lockstep with number of transistors and cores, and now accelerators are the preferred technique to further improve performance, but the parallelism and the heterogeneity add a lot more complexity for software developers.

The design of accelerators marks a new era of architectural research to devise clever solutions to improve the performance per Watt in particular application area (graphics, cryptography, machine learning, …).

There is however room (and also a need) for more disruptive solutions, possibly replacing the (rather inefficient) von Neumann architecture with other computing paradigms, but only if it is done with a high level of efficiency and in a short period of time. Progress in artificial intelligence could help in the efficient design of new systems.

Designing a new accelerator or launching a new technology will always be a daunting task as the proposed solution has to be better than current solutions, and also needs to have a roadmap in order to keep the lead.

Embedded systems, IoT, CPS, edge intelligence

The number one market opportunity in computing systems is the strongly growing market of secure edge computing (which encompasses the internet of things (IoT), cyber-physical systems (CPS), edge intelligence, automotive and industrial automation). Europe has the third largest economy in the world, it has a number of world-class players producing the key enabling technologies for advanced embedded systems, and it has strong automotive, health and industrial automation industries.

Furthermore, there are no non-European dominant companies like Alphabet, Amazon, Apple, Meta or Microsoft in this space yet and Europe is quite strong in . The stars in this domain will probably not be the same as those of the internet era (which are different from those of the mainframe era). Could the company dominating computing in 2030 be European? The only way to win this race is to invest in Europe's strengths, support them to scale up, and hope that they will become world leaders.

Open source

Open source can help promote digital autonomy by removing the reliance on (mostly non-European) proprietary ICT solutions. Instead, open source relies on skills, which gives Europe, with its excellent higher education system, an advantage.

Moreover, as a key driver of innovation, open source may allow researchers to help European companies leapfrog over the incumbent players to become leaders in a specific field. As an example, recently there has been significant investment in the RISC-V open-source instruction set architecture at both the EU and member-state level. RISC-V has the potential of allowing Europe to compete in markets which for years have been dominated by non-EU companies, such as high-performance computing.

Given the lack of major vertical technology companies in Europe, the way for European companies to survive is usually to collaborate. Open source allows them to do this and, eventually, can give their technology a dominant position on the market; see for example Bosch's AUTOSAR project in the automotive sphere. See [18] for a full discussion of the potential of open source to give Europe a competitive edge.

Solutions for societal challenges

Societal challenges form a huge opportunity for the European computing industry. Europe is the region with the highest number of people aged 60 or older [14]. The old age dependency ratio (OADR) is the number of 65+ people per 100 people of working age (24-64). That ratio will increase dramatically over the next 30 years (Figure 21). That means that, globally, Europe will have to search for solutions for the ageing population first. Since the rest of the world will face the same challenges in the future, Europe has an opportunity to develop and commercialize services and products for the silver economy first and sell them to the rest of the world.

The same reasoning holds for the environment. The European population (together with the US) has one of the largest ecological footprints in the world. Reducing the ecological footprint will become one of most important global challenges of the rest of the century. The European Green Deal will help European scientists and industry to find solutions for footprint reduction that can be used and applied across the world. This is a once in a lifetime opportunity.

At this point, Europe is the global leader in publications and patents in the domains of climate action, environment, and renewable energy. The opportunity (and challenge) for Europe is to exploit and mone-

tize all this technological knowledge, and build a handful of global companies in the green tech domain. Europe has the potential to lead in this area, but it should make use of this window of opportunity, and not wait until it closes.

Overall, Europe could position itself as a provider of "quality" ICT which does not merely take monetary imperatives into account, but also meets "European" standards. Combined with regulation of access to the EU single market, this could prove a powerful way to boost European strengths. As an example, Europe could build on its strengths in cybersecurity to create a "kite mark" standard for security in electronic devices and systems, and ensure that only products meeting the standards can be sold within the EU, as suggested by [19].

Threats

Economic stagnation and shrinking work force

Europe has been characterized by low economic growth over the past decade (Figure 22).

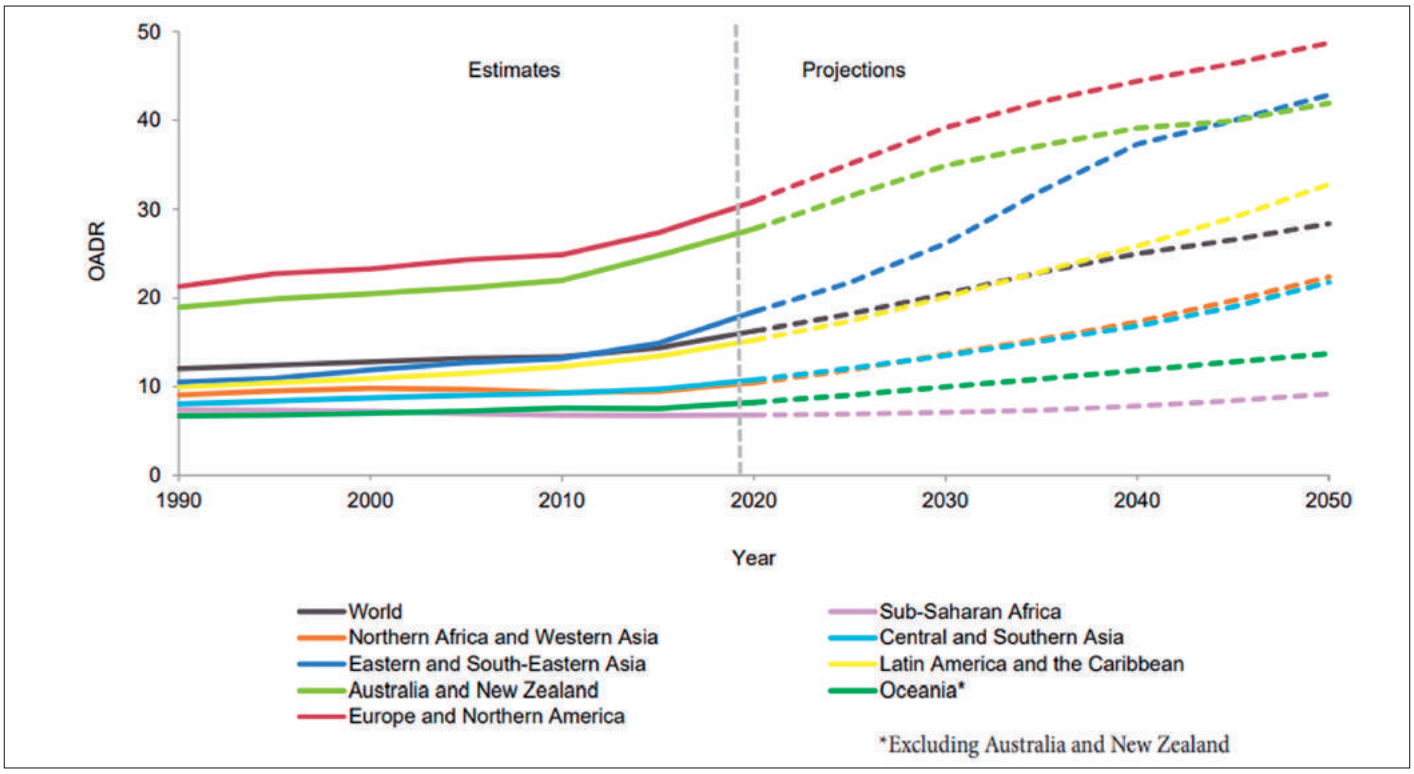


Figure 21: Percentage of 65+ people depending on 100 people of working age (old age dependency ratio) (Source: United Nations (2019)

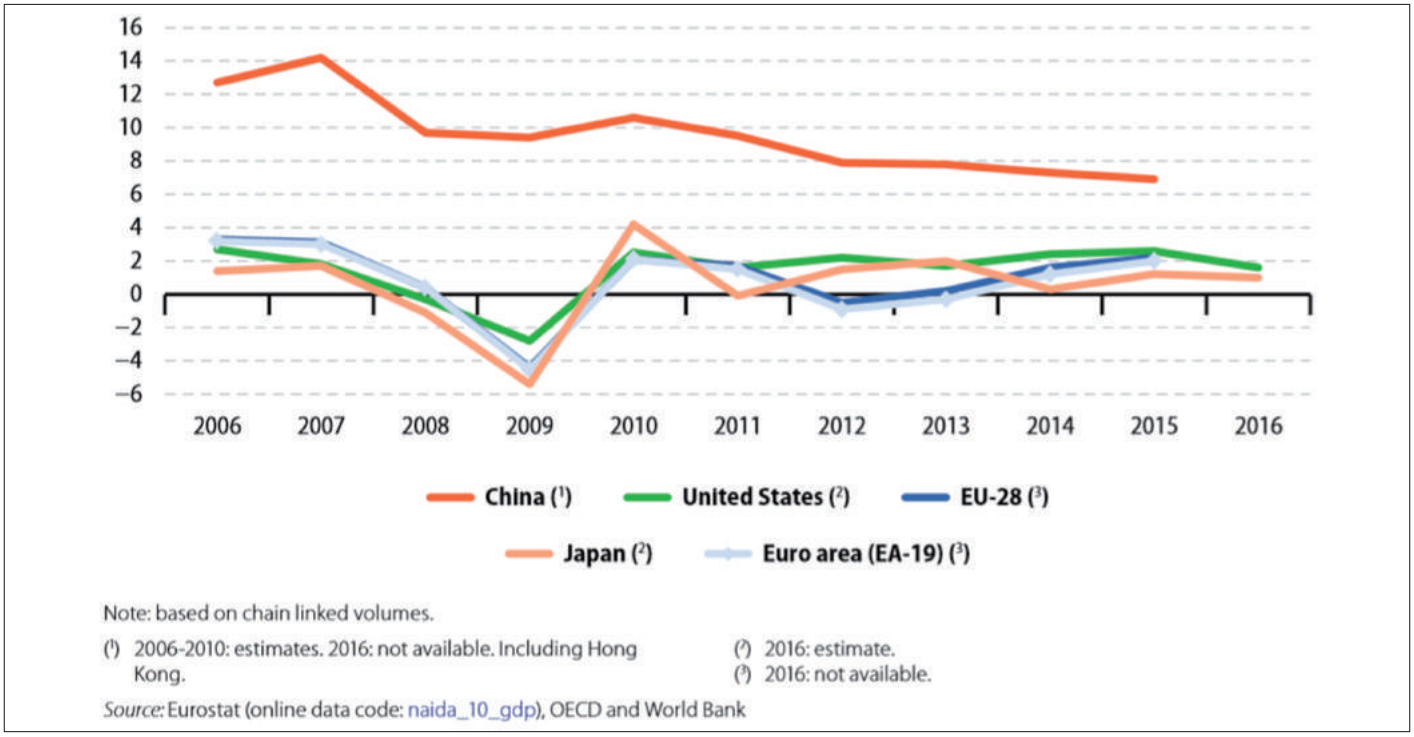


Figure 22: Real GDP growth 2006-2016 (% change compared with previous years) (Source: Eurostat)

This situation has recently been aggravated by the economic impact of the COVID-19 pandemic. Then came supply chain problems in many sectors, the effects of the Brexit, and the energy crisis due to the Ukrainian war. In the background, there is the increasing cost of supporting the ageing population. The cost of pensions will continue to grow until 2040 (when "baby boomers" will have reached their life expectancy).

An often forgotten corollary of the ageing population is the shrinking workforce. Figure 23 shows the number of people entering the job market per year (aged 21), and leaving the job market (aged 65). Until 2018, there were more people entering the job market (as many as five million in 2008, which is only 15 years ago). From now until 2040, Europe will lose one million active professionals per year, and this will continue for the next 20 years. After 2040, it will even increase. Europe has a labour force of about 200 million (70% of the working age population aged 20 – 64 (which is 57% of total population)). A reduction of one million per year means that 0.5% of the workforce will leave the job market per year. Per 200 seasoned doctors, teachers, researchers, nurses, only 199 can be replaced by junior profiles. This leads to a war for talent, and increasing salary costs. We are seeing the first signs all

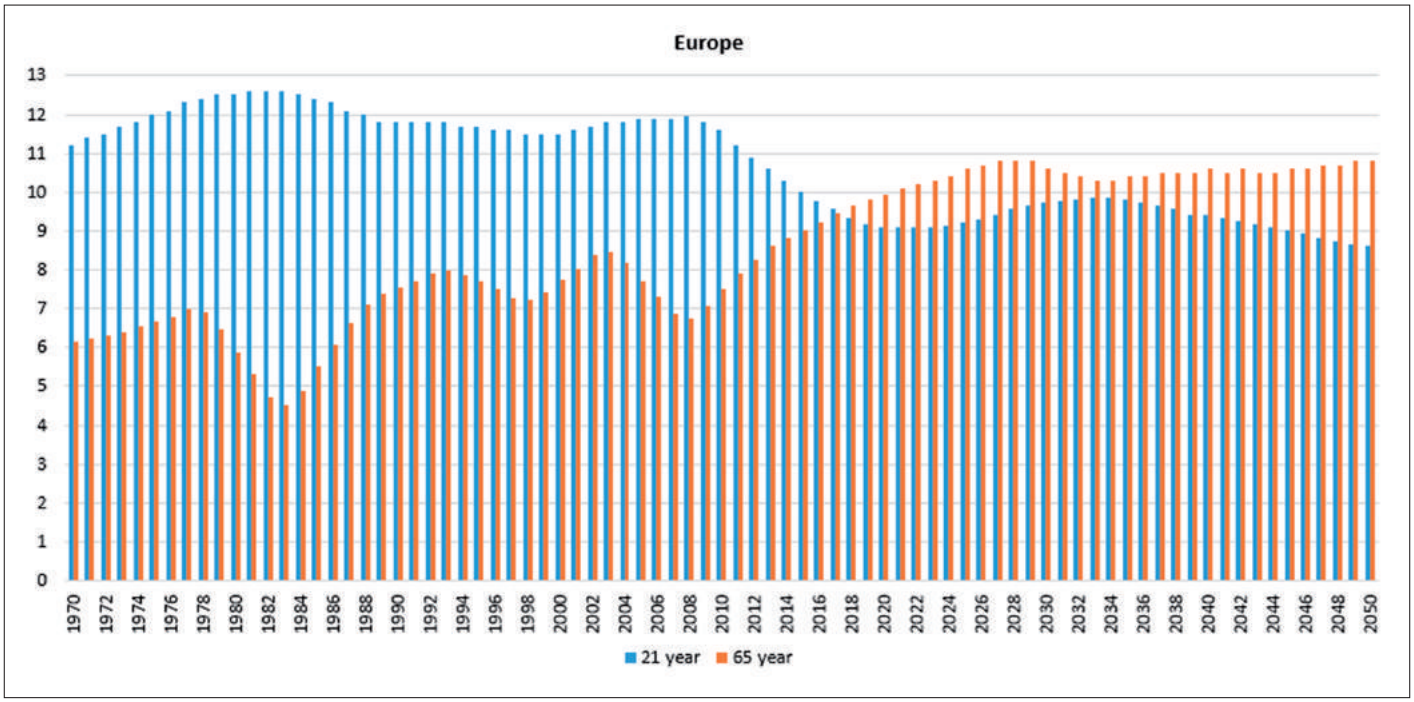


Figure 23: Evolution of the size of the European job market

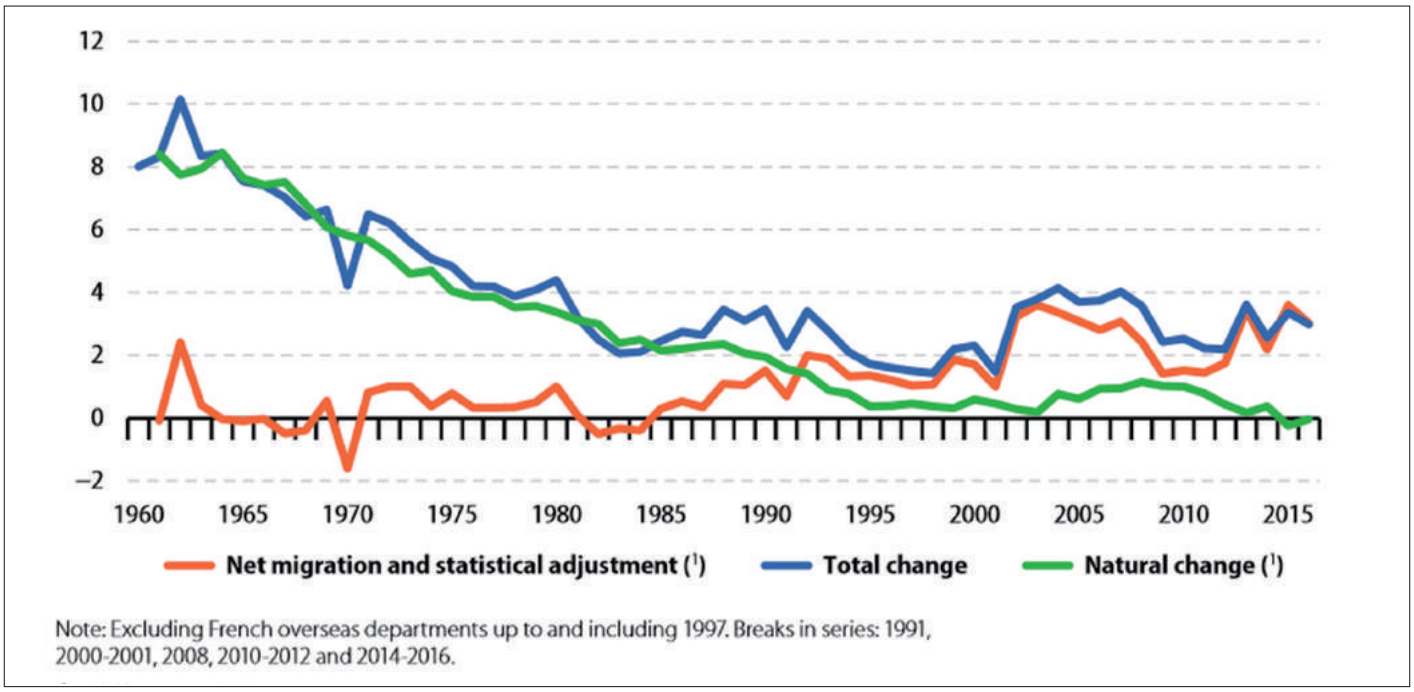


Figure 24: Population change by component (annual crude rates), EU-28, 1960-2016 (per 1.000 people)

over Europe. However, it is also important to note that 0.5% is the average for Europe. In some countries in Eastern Europe, the situation is much more dramatic in that for every 200 doctors, teachers, researcher, only 150 can be replaced. For these countries it is almost impossible to grow the economy due to the lack of workers.

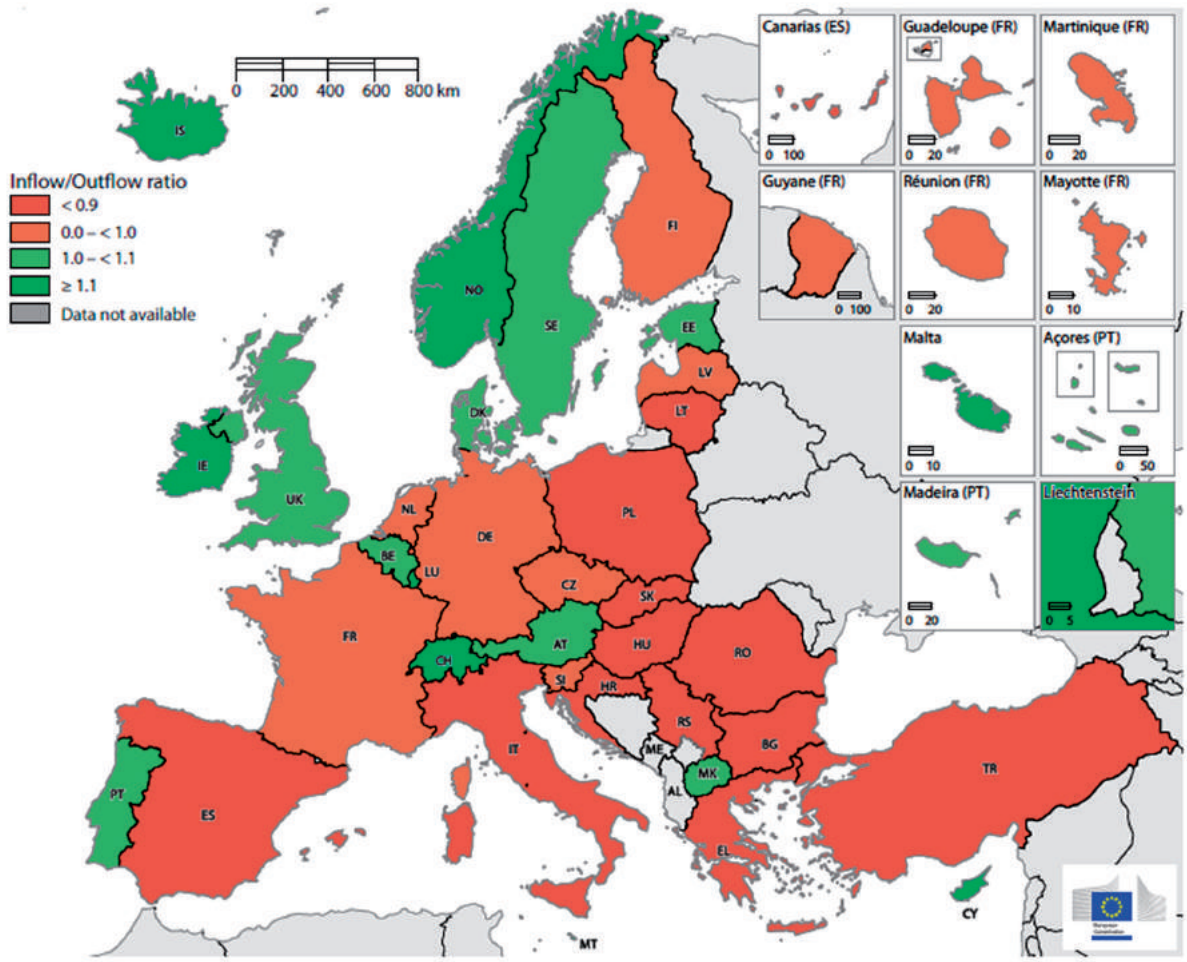
Brain drain

There is a lot of public attention given to the topic of immigration in Europe. It is indeed the case that immigration has increased since the fall of the Berlin Wall in 1989 and is now a major source of population growth in Europe (Figure 24).

This graph, however, masks the fact that net immigration is the difference between immigration and emigration. Emigration usually takes place from economically weaker countries toward economically stronger countries: from the Middle East and North Africa to Europe, from Eastern and Southern Europe to North-Western Europe, but also from North-Western Europe to the United States and other rich countries in the world.

In the research domain, there are countries who have succeeded in reversing the brain drain (the green countries in Figure 25). These countries are attractive because they offer good career perspectives for researchers who want to move to Europe (funding, but also the ecosystem in which they will operate). Southern and Eastern Europe has not been successful in creating the right environment.

In computing, there seems to be a brain drain from Europe to the US. Top researchers and ambitious entrepreneurs are attracted by the merit-based American society and top salaries for high-potential candidates in both academia and industry. Large multinational ICT companies are attractive employers for young European talent eager to travel the world and make a fast career. Even if they do not want to move, US-based companies often acquire European companies in order to have access to their talent or to open subsidiaries in Europe. This is a less visible form of brain drain. Particularly in machine learning, there has been a very strong pull on the top talent in Europe by companies like Meta and Alphabet.



Science, Research and Innovation Performance of the EU 2022

Source: DG Research and Innovation – Common R&I Strategy and Foresight Service – Chief Economist Unit based on Science-Metrix using Scopus database

Figure 25: Map of inflow and outflow or reserachers

Europe should create attractive research ecosystems in order to attract top talent from outside Europe. It could create large and well-funded competence centres to retain European talent, and to attract excellent workers from abroad. CERN is a good example of such a competence centre, attracting talent from all over the world. The proposals for pan-European centres in artificial intelligence [11] and cybersecurity [12] launched recently will hopefully help fulfil this need.

Saturating markets

The market for desktop computers and laptops is shrinking, and the market for smartphones and tablets is shrinking too (after having cannibalized the markets of other devices like navigation systems, cameras, music and video players). The reason is that we have reached human scale: most people in the Western world already have all the devices they need, and the features of most devices have stabilized, eliminating the need to replace devices to get access to more features. The COVID-19 pandemic and the requirement to work and study from home might have created a short peak in the demand for devices needed to telework (mobile devices, headsets, webcams), but this is not a long-term trend.

A greater consciousness of the need to live more sustainable lifestyles is also encouraging people to use devices until the end of their lifetimes, or have them repaired if they are still usable. This will further reduce demand in the longer term. Fewer sales mean fewer resources to spend on the development of new devices and new features. This is not good news for major global B2C companies, but it could create an opportunity for Europe, which in any case is not strong in the B2C sector.

At this point, it seems that investments in energy savings have the best return-on-investment (e-mobility, photovoltaic systems, heat pumps, …) and it is to be expected that these markets will be booming in the coming decade. They are full of digital technology, and Europe could focus on producing components and systems for these applications.

Computing initiatives in countries such as China, Russia and Japan

A threat to the European computing industry is the rapid development of the computing industry in other world regions. Many countries understand that computing is a key enabling technology of strategic importance, and are investing in their own research, products and companies. If Europe fails to do the same, it might eventually become dependent on technology that is designed, developed, produced and controlled outside Europe. The same holds for cybersecurity solutions.

The fastest growing country of the moment is China. There are several sectors where it has the ambition to become a world leader (artificial intelligence [16] and renewable energy being just two examples). This is evident from the quickly growing number of patent applications by Chinese companies, see Figure 26.

The ambition of China to become the frontrunner in artificial intelligence was made very clear in 2017 in their Next Generation Artificial Intelligence Development Plan [13]. It states: "*… by 2030, China's AI theories, technologies, and applications should achieve world-leading levels, making China the world's primary AI innovation centre, achieving visible results in intelligent economy and intelligent society applications, and laying an important foundation for becoming a leading innovation-style nation and an economic power*". China created a five-year AI talent training programme, and invested more than US$2 billion in a huge AI industrial park in the suburbs of Beijing. The presence of Baidu, Alibaba and Tencent is an asset in developing advanced AI applications. The fact that US administrations have put restrictions on Chinese companies will not change this.

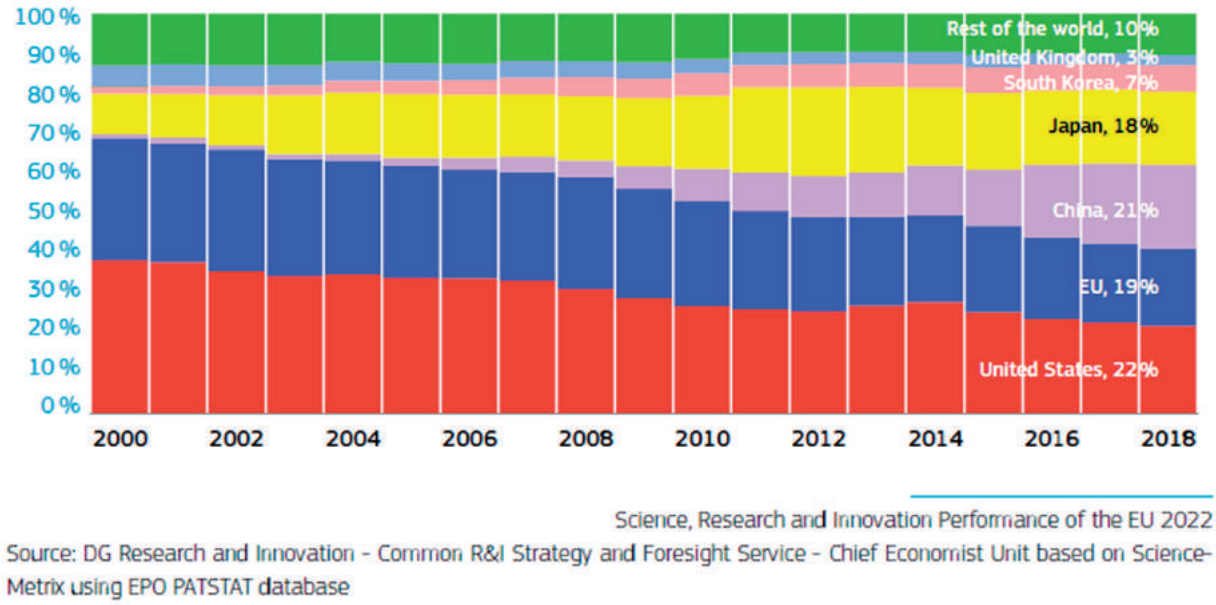


Figure 26: World share (%) of PCT patent applications 2000-2018 Source: OECD ((Patents by technology))

Geopolitical instability

Another threat is the geopolitical instability that Europe and the rest of the world are currently experiencing. After Brexit came the war in Ukraine, and the energy crisis in its wake. The US is tightening its grip on the Chinese tech industry. Meanwhile, the Chinese leadership are showing a renewed interest in Taiwan. According to Bloomberg, a Chinese invasion in Taiwan and the loss of the TSMC factories would lead to a global recession of US$1 trillion [4]. Most countries are therefore trying to move part of the production of the chips they need closer to home [10].

Conclusion

Europe has traditionally struggled to transform its well-established, advanced research capacity into significant economic growth. However, with a potentially global shift to a mindset where the sky is no longer the limit, Europe can capitalize upon its strengths by prioritizing quality electronics for sustainable development, promoting company collaboration to pool resources and focusing on key verticals in which the continent is strong.

References

[1] European Commission, "Science, research and innovation performance in the EU 2022", https://research-and-innovation.ec.europa.eu/knowledge-publications-tools-and-data/publications/all-publications/science-research-and-innovation-performance-eu-2022-report_en

[2] Times Higher Education World University Rankings 2023, https://www.timeshighereducation.com/world-university-rankings/2023/world-ranking

[3] Electronics Components and Systems Strategic Research Agenda 2020, https://artemis-ia.eu/documents.html

[4] European Commission, "Europe's Digital Decade: digital targets for 2030" , https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/europes-digital-decade-digital-targets-2030_en

[5] European Commission, 2030 Digital Compass: the European way for the Digital Decade, https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX%3A52021DC0118

[6] European Commission, "European Chips Act: Communication, Regulation, Joint Undertaking and Recommendation" https://digital-strategy.ec.europa.eu/en/library/european-chips-act-communication-regulation-joint-undertaking-and-recommendation

7] European Commission, "European Chips Act: Staff Working document", https://digital-strategy.ec.europa.eu/en/library/european-chips-act-staff-working-document

[8] European Parliament Committee on Industry, Research and Energy, Draft report on the proposal for a regulation of the European Parliament and of the Council. Establishing a framework of measures for strengthening Europe's semiconductor ecosystem (Chips Act), https://www.europarl.europa.eu/doceo/document/ITRE-PR-731655_EN.pdf

[9] Alan Crawford, "The US Has a $1 Trillion Taiwan Chip Dilemma", https://www.bloomberg.com/news/newsletters/2022-10-07/the-us-has-a-1-trillion-taiwan-chip-dilemma

[10] J. Salmon, "Great pensions divide: Private sector staff must put in a third of their pay to match state worker benefits", https://www.dailymail.co.uk/news/article-2009669/Great-pensions-divide-Private-sector-staff-pay-match-state-worker-benefits.html

[11] Claire, "Confederation of Laboratories for Artificial Intelligence Research in Europe", accessed December 2022, https://claire-ai.org

[12] ECSO, European Cyber Security Organisation, accessed December 2018, https://ecs-org.eu

[13] R. Creemers, "A Next Generation Artificial Intelligence Development Plan, China Copyright and Media", 1 Aug 2017, https://chinacopyrightandmedia.wordpress.com/2017/07/20/a-next-generation-artificial-intelligence-development-plan

[14] UN, World Population Ageing 2019 https://www.un.org/en/development/desa/population/publications/pdf/ageing/WorldPopulationAgeing2019-Highlights.pdf

[15] S. Pham, "Taiwan chip maker TSMC's $12 billion Arizona factory could give the US an edge in manufacturing", https://edition.cnn.com/2020/05/15/tech/tsmc-arizona-chip-factory-intl-hnk/index.html

[16] F. Westerheide, "China – The First Artificial Intelligence Superpower", https://www.forbes.com/sites/cognitiveworld/2020/01/14/china-artificial-intelligence-superpower/?sh=5c6dc28f2f05 +

[17] European Commission, "Digital sovereignty: Commission proposes Chips Act to confront semiconductor shortages and strengthen Europe's technological leadership", https://ec.europa.eu/commission/presscorner/detail/en/IP_22_729

[18] Paraphrasing Gaël Blondelle, Philippe Krief, Mateo Valero and John Davis in "Breaking out of the black box: Open source gets serious in Europe", HiPEACinfo 67, June 2022, https://www.hipeac.net/magazine/7162.pdf#page=18

[19] HiPEAC video, "Ingrid Verbauwhede on why hardware needs to be secure by design", https://youtu.be/2ZBfpwV2nx8

Koen De Bosschere is a professor in the electronics department of Ghent University, Ghent, Belgium.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2023.

Cite as: K. De Bosschere. The position of Europe in the world. In M. Duranton et al., editors, HiPEAC Vision 2023, pages 156-174, Jan 2023.

10.5281/zenodo.7461934

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

© HiPEAC 2023

Over the last decade, open source has proven to be the best way to collaborate on technology platforms and promote the adoption of innovation.

Open source is the enabler for innovation and collaboration

By GAEL BLONDELLE, with contributions from PHILIPPE KRIEF

Open source and open hardware have become more and more popular over the last two decades, in information technology (IT) circles, in the research and academia ecosystem, and even in industry. However, there are different types of open source project, from individual-driven projects to more corporate-driven and even collaborative projects, and every user or contributor should be conscious about what they get from the open source ecosystem, how they should contribute, and what they should put in place to benefit from the open source.

In this article, we cover how open source provides a global legal framework for collaboration (including between research and industry players), how open source foundations play a key role in helping organizations do open source properly, and how this can enable European leadership for upcoming technology trends.

Key insights

- Open source is a global framework for collaboration and innovation.
- Permissive licences and weak copyleft licences lower the barrier to usage and participation for industry players.
- Open source foundations like the Eclipse Foundation are there to support research and industry players to do open source the right way.
- European research and industry actors can thrive with open source by promoting global adoption of their innovations.

Key recommendations

- Use of open source to promote collaboration between academia and industry.
- Adopt permissive licences to lower barriers to collaboration with industry.
- Leverage the experience, governance, and processes of existing open source foundations established in Europe.

When I started working on open source in 2002, I was a technical architect at a French telecommunications operator, and our goal was to reduce the cost of our platforms by moving from proprietary Unix systems to Linux running on Intel hardware. At the time, open source was mostly a commoditization approach, and the main goal was to challenge monopolies and to reduce the cost of platforms by building and using open source alternatives to existing platforms: That's when the Apache HTTP Server, MySQL and to some extent PostgreSQL or the Eclipse development tools bloomed.

That's also when Linux won on the server side, while later the Linux kernel achieved more adoption in embedded systems.

However, as different stakeholders increasingly began to understand the open source model, it became clear that this was also a good way of getting adoption for innovative technologies. For the last 10 years, the platforms for big data, cloud and artificial intelligence (AI) technologies have been open source.

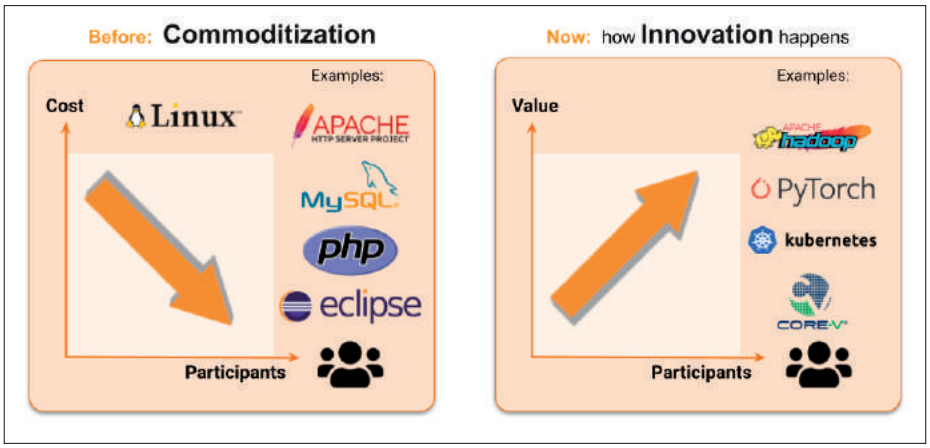


Figure 1: The shifting impact of open source

Open source superpowers have been used for more than 10 years to enable global collaboration in key technologies. This is because, as discussed below, 1) open source provides a proven global legal framework for collaboration and 2) open source provides a platform for most applications.

Open source as a global legal framework

The Free Software Foundation was founded in 1985 by Richard Stallman, from MIT, to support the free software movement. The foundation's "Free Software Definition" relies on the four essential freedoms [1] that allow anybody to run, study, modify and redistribute "free software".

In 1998, after the announcement of the release of the Netscape source code, the term "open source" was suggested by Christine Peterson to promote free software without the additional connotations of "free" (i.e. "free of cost") in the English language. The Open Source Initiative (OSI) was founded in February 1998 by Eric Raymond and Bruce Perens. The OSI published the "Open Source Definition" [2] that builds on top of the four freedoms and expresses with 10 clear points how a licence can comply with the definition. Since its creation, the OSI has been maintaining a list of approved licences [3].

Developers sometimes lose sight of the fact that both the free software and the open source definitions have enabled a global legal framework for software:

1. When an individual or an organization publishes software under an open source licence, they select the legal clauses of this specific licence to publish their intellectual property,
2. When an organization decides to use an open source component, they accept the use of it under the terms of the specific licence,
3. Open source licences are short (five pages long for the Apache Software Licenses V2 [4] or the Eclipse Public License V2 [5] or even shorter for the MIT License that is fewer than 200 words long [6]) and well understood globally).

Open source grows as it empowers application developers

Recent statistics show that 80-90% of applications use open source components [7]. This means that organizations can build applications by putting together existing open source components and focus on the 10-20% code that forms their unique value proposition.

This is possible thanks to the four fundamental freedoms. A good example of how open source enables modern applications is that, just as some proprietary licences explicitly restrict the usage of the software to non-life-critical functions, the Open Source Definition has the following "No Discrimination Against Fields of Endeavor" clause: "The license must not restrict anyone from making use of the program in a specific field of endeavor. For example, it may not restrict the program from being used in a business, or from being used for genetic research."

Any organization is empowered to use an open source kernel, platform, or component in their application without asking the authors of such components for permission.

All open source licences comply with the Open Source Definition, they all come with rights (at least the four freedoms, and often more), but they also come with duties and restrictions. The "JoinUp Licensing Assistant" describes the characteristics of different licences and enables comparison of different licences [8].

Over the years, users have identified that open source is not exempt from licence compatibility issues, and most companies have created an "open source compliance" process to ensure that they comply with the licences of components they redistribute in their products.

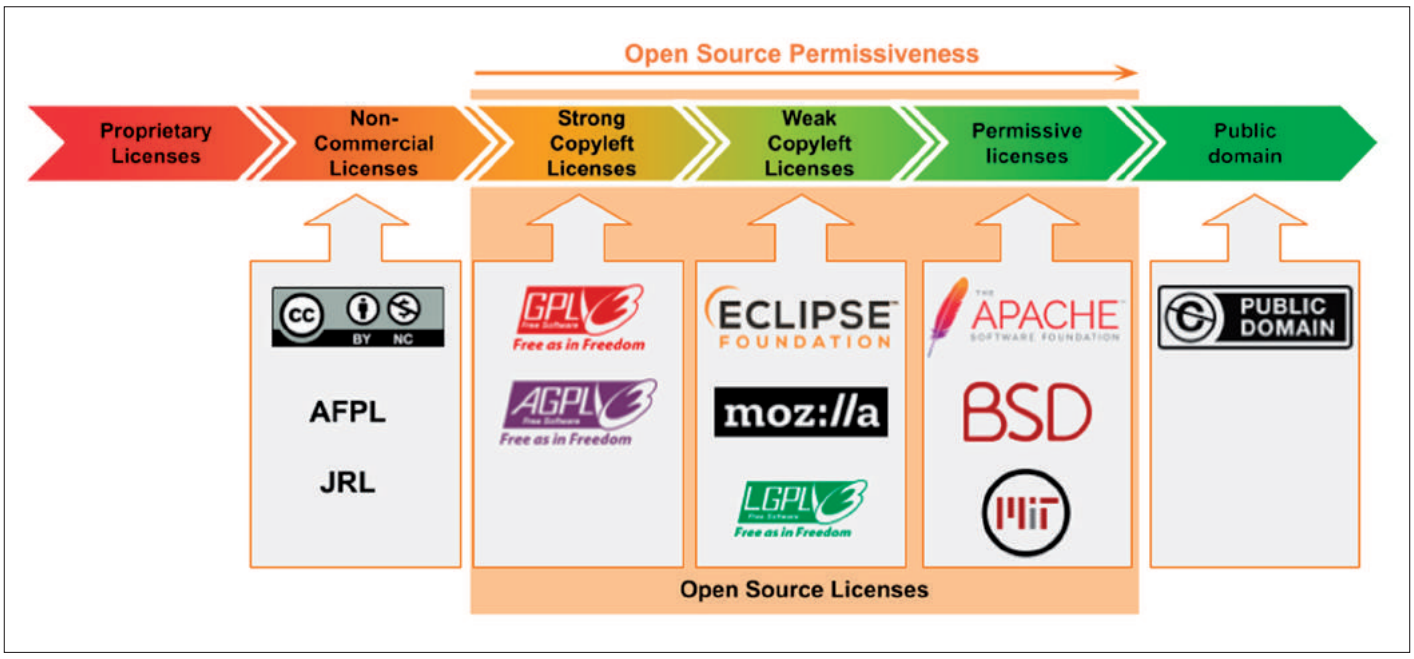


Figure 2: Open Source licences spectrum

To make things simple, developers and users of open source have widely adopted "weak copyleft" and "permissive licences" to facilitate collaboration and reuse of open source components in a product, without imposing constraints on the licence of the final product. That's the main reason why the Apache Software License became the most popular open source licence [9].

Beyond making open source compliance easier, the advantage of weak copyleft and permissive licences is that they enable collaboration between industry partners independently of how they will exploit the open source software: professional services, training, support, or even integrating the open source components in a proprietary software product. With these licences, as they limit the duties on users, the focus is on collaboration and, in most companies, components under licences like the Apache Software License, the BSD family of licences, the MIT License, or the Eclipse Public License can be used without any additional checks from the legal team.

Not all open source projects are "equal"

To simplify, one can consider that there are three main categories of open source projects: individual-driven open source, single-vendor open source, and open source hosted by foundations.

Some projects, started by individuals, are casually hosted on public platforms like GitHub or GitLab, and organically grow in popularity. It is more likely that an open source project driven by a single individual may struggle at some point to address different issues. One example is that the more the project becomes successful, the more it may get external contributions, and the more pressure the initial contributor may receive. Additionally, the individual may not have the skills or interest to address non-functional topics like security, performance or scalability. Finally, an individual, or a small group of individuals can be overwhelmed, shift interest to a new project, or even find a new job thanks to their new visibility acquired through open source that could get them away from the project. As we will see later, it is important to consider that when you use an open source component, some of its dependencies can be such a project, or even in some cases, an "orphan project" with barely any active maintainers.

Other projects that we call "single-vendor open source projects" [10], are mainly developed and supported by a single company. Most of the time, such a company requires contributors to sign a "Contributor License Agreement" to bring together all the intellectual property rights of the project. That enables different business models which likely make the project sustainable, but recent history has shown that single-vendor open source may sometimes stop being open source at some point [11]. This can be an issue for users who then need to find a plan B like becoming a customer of a product which is no longer open source, or fork the open source project...

The third main option is open source projects hosted by a well-established open source foundation.

Open source foundations as catalysers of best practices

Beyond code and documentation, an important aspect of open source projects is their governance. Open source governance is the set of rules that define who takes decisions in an open source project, how to engage with an open source project, how contributions are accepted, how a developer can gain write access to the project repository by becoming a committer, the rules that ensure the vendor neutrality of the ecosystem, …

Most successful projects have well-defined open source governance, as this provides a safe environment for users and contributors of a project. The goals of open source governance are to provide vendor-neutrality, to enable a multi-vendor ecosystem, to sustain a diverse community around the project, and to protect the investments of contributors and users of the project.

For more than 20 years, open source foundations have provided the governance framework for fruitful open source collaboration. Let's take the example of the governance framework of the Eclipse Foundation. As a non-profit organization established in Europe, the Eclipse Foundation has a mission to provide governance and sustainability for the projects that are hosted by the foundation.

In addition to that, the Eclipse Foundation provides a "community of practice" that enables companies to adopt open source best practices and fosters values of transparency, openness, meritocracy, and vendor neutrality.

Open source as a catalyser for European technology transfer

European research teams constantly make important advances in their areas of study, but they sometimes fail to provide a clear path to build on the results of their research in a way that creates value for industry and society. For almost ten years, the Eclipse Foundation has been working with European researchers and

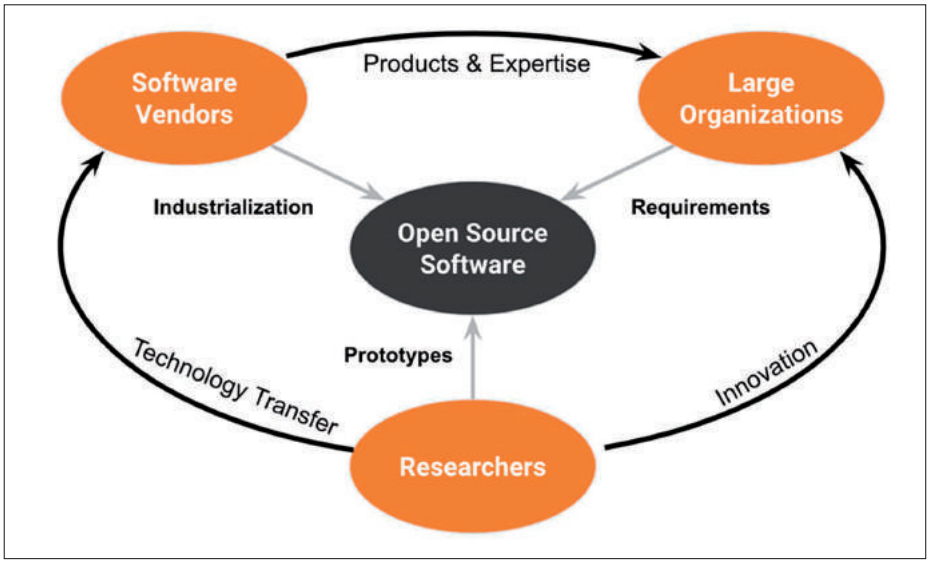


Figure 3: Open Source software as a catalyst for technology transfer and innovation

academics to help them create, evolve, and sustain open source projects that capture the results of their research so they can be disseminated and commercialized to solve real-world challenges

Researchers can leverage open source to build an open ecosystem that enables industry players and technology adopters to collaborate and build on their research. Everyone interested in using and contributing to further the research results can collaborate without the significant complexity and costs of setting up multiple contracts and licence agreements.

This is possible because the open source software acts as a catalyst for technology:

- industry players bring their requirements,
- academics and researchers innovate to develop solution prototypes that meet industry needs,
- software vendors industrialize these innovations to deliver products.

This kind of collaboration is greatly facilitated by open source best practices and business models.

As we argued in [12], thanks to its role in driving innovation, open source can allow researchers to challenge the status quo and become leaders in a specific field: look at the RISC-V initiative and OpenHW Group as examples. This is particularly important for Europe, which lacks the major vertical companies dominant in other parts of the world but which has a strong research infrastructure. Open source enables digital autonomy because it replaces dependency on proprietary intellectual property mainly owned by non-European companies with the need to invest in people and skills. European countries, with their well-established universities and robust education systems, are a great source of skilled researchers and engineers.

If European companies play alone against the major technology giants, at best they will be bought out, at worst put out of business. So if a European company wants a competitive edge, it needs to build a network of partners to pool resources. One example of this strategy is Robert Bosch: executives at Bosch realized that if each platform promoted their technology to the detriment of others, they would eventually collapse, resulting in a dependency on technologies by big tech companies. Instead, they decided to put their IoT technologies developed internally into the open arena, most of it being contributed to the Eclipse IoT Working Group [13], in order to benefit from a network effect leading to wider take up of their technologies on one hand, and the continued improvement of these technologies on the other.

Creating a new open source foundation is a long and expensive process. Some groups think that they should create a new association to develop an ecosystem around the results they publish in open source. Creating such an organization is certainly a very exciting process, but it is very likely more efficient to leverage the proven collaboration models provided by existing foundations:

- In many cases, one or several open source projects can provide enough of a meeting point to build a successful community.
- Several open source foundations support a "foundation in a box" approach. For example, the Linux Foundation provides a toolbox to organizations that want to create their own "foundation" managed by the Linux Foundation, and the Eclipse Foundation provides processes to create Interest Groups (lightweight) or Working Groups (full featured) that enable collaborations, and for example the development of open specifications.

As a reference, the OpenHW Group [14] created an Eclipse Working Group, OpenHW Group Europe, to specifically support its European ecosystem around Core-V applications.

We estimate that creating an association from scratch that could provide the right level of open source governance would take more than 18 months before it would be fully operational and would require a legal budget of hundreds of thousands, if not millions, of euros. Teaming up with an established foundation makes the process much faster and an order of magnitude less expensive.

Non-functional aspects of open source are important too

Having looked at code, community, and the ecosystem, let's approach two challenges that open source has to face in the coming years: supply-chain security and certification.

For a long time, it was clear that even if the OSS movement was not designed with security in mind, the Linus' law in open source states that: "given enough eyeballs, all bugs are shallow." [15].

It is obviously the case that major security vulnerabilities impact proprietary software as much as open source software. The SolarWinds Hack [16] is an example of a security breach that has impacted a proprietary product without any relations to open source.

But in the last few years, the conjunction of a few vulnerabilities in critical open source software like OpenSSL [17] – and the fact that open source has become pervasive and is used in almost every piece of software [18] – makes open source security and more specifically supply-chain security a major topic. Recently, the Log4Shell [19] vulnerability has been a wakeup call for governments, communities and companies to put supply-chain security on top of their priority list.

As a result, several open source foundations are addressing the challenge of open source supply chain. The Open Source Security Foundation [20] was created by the Linux Foundation and is hosting the Alpha-Omega project, funded by Google and Microsoft, whose mission is to "Protect society by improving the security of open source software through direct maintainer engagement and expert analysis". The Eclipse Foundation is also creating its own security team, partly supported by funds from the Alpha-Omega project, to help the Eclipse community at large address the supply-chain security of the more than 400 projects hosted by the Eclipse Foundation. This includes developing new processes and deploying new tools to support projects so we can guarantee the traceability of software that users consume from the Eclipse Foundation,

while also ensuring that vulnerabilities are found early and are fixed in a predictable way.

Finally, as we see new open source initiatives being created like the Eclipse Software Defined Vehicle (SDV) Working Group [21] where automotive companies collaborate not only on development tools but also on software that will be embedded in future generations of cars, it is important to mention the topic of certification. For a long time, lots of people considered that open source components could not be certified, but, as in the case of security, the open source collaboration model and the approach to certification are orthogonal topics and there are several cases where open source components have been reused in certified environments. The point is that either an organization contributing to open source projects executes the certification, or the users of the open source component will have to execute it in their context. As an example, Eclipse iceoryx is striving for being able to safety-certify the code base up to ISO 26262 ASIL D that is essential to the automotive industry [22].

Conclusion

As it enables scalability, collaboration and adoption, forthcoming technology trends like edge computing will largely leverage open source. This is a unique opportunity for the European ecosystem if European organizations move quickly from using and contributing to open source to a more strategic approach of open source.

Open Source has been used successfully to grow the adoption of some of the platforms we all use today, like Kubernetes. As we develop exciting technologies in research and industry, let's make sure that we use the superpowers of open source to foster global adoption of these technologies.

There is good news: the European Commission is pushing for open source, the industry is increasingly embracing open source, and open source foundations are ready to help.

So, everything is aligned to show strong leadership for the global ecosystem to collaborate on open source technologies with strong roots in Europe.

References

[1] Free Software Foundation, "The four essential freedoms", https://www.gnu.org/philosophy/free-sw.html#four-freedoms

[2] The Open Source Definition https://opensource.org/docs/osd and the annotated version https://opensource.org/osd-annotated

[3] List of licenses approved by the Open Source Initiative: https://opensource.org/licenses

[4] Apache Software License V2.0: https://www.apache.org/licenses/LICENSE-2.0

[5] Eclipse Public License V2.0: https://www.eclipse.org/legal/epl-2.0/

[6] MIT License: https://en.wikipedia.org/wiki/MIT_License

[7] Open Source Software are everywhere: https://www.digitalinformationworld.com/2022/11/open-source-software-are-everywhere.html

[8] Joinup License Assistant – Comparison of some popular licenses: https://joinup.ec.europa.eu/licence/compare/EPL-2.0;Apache-2.0;MIT;GPL-3.0-only;GPL-2.0-only

[9] Permissive Open Source Licenses Continue to Trend: https://www.mend.io/resources/blog/open-source-licenses-trends-and-predictions/

[10] Single Vendor Open Source Firms and their intellectual property strategies – Dirk Riehle: https://dirkriehle.com/publications/2020-selected/single-vendor-open-source-firms-and-their-intellectual-property-strategies/

[11] The SSPL is Not an Open Source License: https://opensource.org/sspl-not-open-source

[12] "Breaking out of the black box: Open source gets serious in Europe", HiPEACinfo 67 pp.18-23

[13] Eclipse IoT Working Group: https://iot.eclipse.org/

[14] Open HW Group: https://www.openhwgroup.org/

[15] Considerations About Open Source and Security: https://www.eclipse.org/opencert/resources/security/

[16] SolarWinds hack explained: Everything you need to know: https://www.techtarget.com/whatis/feature/SolarWinds-hack-explained-Everything-you-need-to-know

[17] Heartbleed: https://heartbleed.com/

[18] The State of open source software: https://octoverse.github.com/

[19] Log4Shell: https://en.wikipedia.org/wiki/Log4Shell

[20] Open SSF: https://openssf.org/

[21] Eclipse Software Defined Vehicle: https://sdv.eclipse.org/ The Eclipse Foundation Announces the Release of Eclipse iceoryx Almond (1.0): https://newsroom.eclipse.org/news/announcements/eclipse-foundation-announces-release-eclipse-iceoryx-almond-10

Gaël Blondelle is chief membership officer at the Eclipse Foundation AISBL. He has been working in open source since 2004 and is helping organizations collaborate in open source since he joined the Eclipse Foundation in 2013.

Philippe Krief is the director of research relations at the Eclipse Foundation. Since 2016, he has been working with European research projects to implement their open source publication and dissemination strategies.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2023.

Cite as: G. Blondelle, with P. Krief. Open source is the enabler for innovation and collaboration. In M. Duranton et al., editors, HiPEAC Vision 2023, pages 175-179, Jan 2023.

DOI: 10.5281/zenodo.7461971

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

© HiPEAC 2023

Digitalization made IT essential for a working society. Europe needs to define and defend its digital essentials for sovereignty.

Europe's need for digital essentials, individual sovereignty and consumer protection

By THOMAS HOBERG

Information technology (IT) is now critical to a functional society and should be treated as such. The spread of digitalization across the globe has created new power structures, ruled over by technology monopolies. Perhaps more profoundly, the digitalization of global interaction has brought humanity into a contact that is far too close, immediate and unfiltered for comfort. Without the support of a spatiality[1] which allows us to reuse the social code of civilization and diplomacy learned through centuries, very much based on distance and visibility, we enter a phase where basic civility needs to be relearned, while our ability to inflict annoyance, anger or outright harm has been greatly enhanced.

This article explores some of the trends towards a digital chasm. It outlines some of the key issues in terms of platform power, vendor lock-in, and digital essentials, and suggests ways in which Europe can reclaim sovereignty over its participation in the digital sphere.

Key insights

- Digitalization has made IT become globally essential and critical, causing the underlying rifts in political values to widen and split. Political battle lines are currently drawn across supply chains and trenches spring up in software, and there is no sign of reversal of this new digital chasm.
- National governments, legislation, courts and plebiscites, who intrinsically have only a local mandate, can suddenly have a global impact, cause disruption or even completely cut off other nations and regions, merely because of where an essential piece of software was created or a where a critical service is operated. This risk is hardly even evaluated, let alone managed.
- The mere existence of open-source solutions or alternatives is not enough to compensate that risk, because they need continued maintenance and evolution.
- Peddlers of some niche products[2] in quirky emerging markets have turned into global cloud giants that mediate nearly every significant transaction between people and businesses, and try to shoehorn their terms-of-purchase into constitutions of a platform or metaverse they believe they own, selling often only an illusion of ownership in return for digital indenture.
- Digital platforms have grown from shopping malls into omnidirectionally integrated digital giants that challenge sovereign nations simply by force of economic gravity and transactional reach. Peaceful coexistence is impossible beyond a certain size; sovereigns need to act decisively.

1 See the "Gaming, content and the metaverse" article in this HiPEAC Vision.

2 A mobile digital music player using digital rights management (DRM) to control content distribution without physical media – the Apple iPod that morphed into the iPhone later.

Key recommendations

- The ability to survive in a strongly partitioned internet depends on the ability to take control, create a local fork and maintain software digital essentials at the required quality. It means that the EU cannot wait for such software access partitioning to happen; it needs to sponsor a fully capable developer community within Europe for all digital essentials today, ready to fork in case of a partition, but also ready to rejoin once conflicts has been resolved.
- Europe needs to foster the creation of new digital essentials, especially in open source and around enabling smart digital assistants that are loyal, support multi-compliance and enhance privacy, which can be used as division busters and bargaining chips for re-merging. Most attractive may be alternatives to proprietary (e.g. Office) or control-locked (Android) offerings.
- Platform power is generated by retaining control beyond "purchase" and trying to force consumers into subscriptions. Any control beyond purchase[3] must be broken via regulation and a full purchase with or without support must be made available at a reasonable price.
- The roles of a device vendor, software vendor, mall operator, gaming server provider, digital asset notary etc. must be segregated cleanly and completely to ensure that consumers cannot be blackmailed, and independent merchants be abused. That segregation may not just have to be functionally, but also geographically to ensure that functionality and services can be maintained in the EU, when a new digital division opens up and even if such a segregation isn't required or even prohibited in a merchant's/publisher's home market.
- Mandate very strong protections for individual sovereignty of personal computing device owners, right-to-repair and the de-coupling of devices from vendors without losing current functionality. Break vendor lock-in on software shops, mandate software quality control and insurance (against malware, viruses), potentially at distinct levels on shop operators much like food safety[4].

Global digital divisions

While the internet might have taken the world by storm, digitalization today also storms the internet. Yes, the internet may have lowered, even levelled, some barriers, just as travel, the printing press, radio and television did. But when the majority of all interpersonal business and communication is transacted on the internet, all the remaining cultural and political divisions will follow and mirror the complexities of social and political graphs in cyberspace.

The tale of a truly open and neutral cyberspace is at best a pipe dream, more

The Berlin Wall (source: Andreas Prott on Adobe Stock). Today, geopolitical divisions are often manifested in the "digital chasm".

3 DRM requirements may require enforcement support.

4 Which is certainly not failsafe, but probably better than having none.

likely an attempt to exhaust a technology lead before regulation catches up.

The internet was about objective science and technology at its beginning, where opinions and beliefs seem to matter less. It is mostly about content today, where values and culture are very subjective and local.

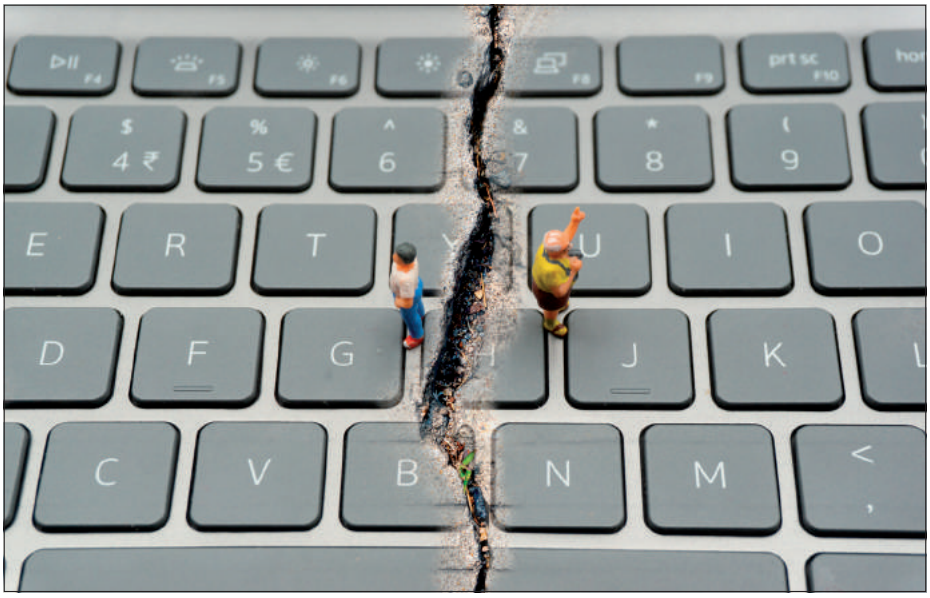
Trade has often resolved conflicts of interest very in human history because war is terribly expensive. But when interests can no longer be balanced via exchanges, trenches spring up. On the internet this may still seem recent, even if the Great Firewall was built only three years after the internet arrived in China 1994[1].

We still think that everything on the internet is immediately available, and disturbances a technical problem that will be resolved quickly. When the world is conflicted and when even within Europe fragmentation seems on the rise, we need to plan and prepare not just for cyberwar, which already has become as ubiquitous as cybercrime, but for a highly variable multitude of much deeper divisions, some of which may last a long time.

The supply-chain issues which have gained significant attention over the last years are only a hint of what is to follow in terms of internet content and services, including some of the very basic and early ones, which seem rather technical and should therefore provoke little debate.

While there has been some attention on supply chains for raw materials, foodstuff, and hardware, we may be much too careless about the software and services, especially where open-source code, free services or content are concerned.

We came very close to a coup d'état in the United States on 6 January 2021, where a person was grasping for power that likes to dip its fingers directly into the dearest secrets of any person or nation state from his private estate. We also live in a world where a few people on the US Supreme Court feel no compunction about holding the bodies of half their own population hostage, so it is safe to assume they feel



even fewer compunctions to impose their whim on everyone else on the planet.

The dominance of US hyper-scalers over the internet in its current form means that US presidents and courts can seize, stop, hijack, infiltrate, abuse pretty near all of it, at will or the drop of a gavel. Even if Microsoft, Amazon, Google, Meta, and the various other internet giants consider themselves truly global and maintain the physical assets of their clouds very close to their global client base, they can't but bow to the ultimate authorities back home, wherever they operate. In that regard they are no different to Chinese or Russian companies; nor in fact could a European company free itself from EU regulation just because parts of it operate elsewhere.

In October 2022, the so-called "Bureau of Industry and Security", an office inside the US Department of Commerce, unilaterally decided to cut off China's semiconductor industry from nearly every relevant supplier and customer on the planet. They did not ask or consult the European Union, Japan, South Korea, Israel or any other nation heavily invested in the international chip industry, nor the global customer base. The UNO was not involved, neither was the World Trade Organization or any lesser international body[2]. An undersecretary of the US government assumed control over the global chips industry because the US believe their historical lead in that industry not only entitles them to this power, but that this lead should also be maintained at a point when China is at the verge of overtaking. They did so secure in the knowledge that without US-controlled technology a full range and stack of IT products cannot be created. But they also chose to ignore that without many global players outside the US, it is similarly handicapped itself.

If Europe could rest assured that the United States would treat it with "near domestic" privileges and would remain its staunch ally in the pursuit of democracy and human rights against totalitarian regimes in mainland China, Russia and North Korea, such initiatives eventually might even find political support, especially when asked politely. But blind faith into the United States as a pillar of democracy and an intimate friend of Europe has obviously become misplaced. The United States have declared the equivalent of a digital cold war on China, but as a hardly noticed collateral they have filed for a digital divorce with Europe, choosing "America First". Instead of negotiating, the US government seems to have adopted the "disruption religion", which according to Peter Thiel[3], one of its main proponents, regards democracy as incompatible with freedom.

No one outside the United States, no person, no enterprise, and no government may ignore that they are at the whim of

5 Including Microsoft Edge.

powers completely outside their control. To do so, after the events of the last years, can't be anything but grossly negligent. This must be changed quickly, decisively, and with regulatory support, because otherwise in many cases such sudden and drastic changes would violate contract law.

Digital essentials for Europe

Few things are as important to the IT industry as economy of scale. Whether it's software code or a hardware description language, creating truly original lines of code is extraordinarily expensive and those high costs can only be offset by maximum reuse and deployment at maximum scale, preferably global. This naturally leads to code monoculture, because maintaining more than one code base for the same functionality adds cost without apparent benefit.

Industry giants use this effect to create effective monopolies and lock-in effects even if there is no purchase price involved by sponsoring open-source code projects or free services. Perhaps the best-known example currently is Google's Chrome browser, which together with all its derivatives dominates the browser market everywhere outside the Apple universe. A broken or unavailable Chrome nearly breaks the Internet, which is why the EU must guarantee the free availability of such digital essentials, independent of any other political bloc, in case of a political rupture, a cyberwar or physical conflict that could last years.

It is a move that has been similarly anticipated and undertaken by political blocs which are further away from the US than Europe e.g. North Korea, Russia and China for the same purpose of ensuring a minimal degree of digital autonomy.

While it is easier to. maintain software digital essentials than to produce a full stack of hardware locally, it is not enough to just mandate the use of open-source code. **The ability to survive in a strongly partitioned internet depends on the ability to take control, create a local fork and maintain software digital essentials at the required quality. It means that the EU cannot wait for such software access partition to happen; it needs to sponsor a fully capable developer community within Europe for all digital essentials today, ready to fork in case of a partition, but also ready to rejoin once a conflict has been resolved.**

Compared to the number of public employees currently involved in the maintenance and evolution of our societies cultural and legal code (e.g. teachers and all branches of government), those required to maintain our societies software base would be miniscule, yet their importance is difficult to overrate. Just imagine Europe were to lose some or all of the digital essentials listed in the box.[6]

1. a Linux distribution for a) PCs, b) servers, c) mobiles, d) IoT/embedded (e.g. Debian[4])
2. an Android fork[5] for tablets and phones, a Chromium fork for "Chromebooks"
3. a hypervisor (unless included in 1.[6])
4. a browser (e.g. Firefox[7])
5. an office suite (e.g. Libreoffice[8])
6. a team collaboration suite including conferencing (e.g. NextCloud[9])
7. an e-reader (PDF & book formats e.g. SumatraPDF[10])
8. a multimedia player (e.g. VLC[11])
9. mail clients/server unless included in 4[12].
10. a firewall/content-filter appliance (not necessarily as a separate image but included in 1. or in a container) e.g. OpenSense[13]
11. a messaging server and client platform (e.g. Signal[14])
12. thin-client/remote access tools[15]
13. VPN tools[16]
14. a compiler (e.g. GCC/LLVM)
15. an IDE
16. code repository tools (e.g. GitLab)
17. archiver (zip)
18. a Wikipedia
19. a search engine
20. DNS/PKI/CERT

How the EU ensures these assets and capabilities will remain available without restrictions in case of a partition from a major part of these software communities needs to be worked out and the approach may differ and change between them. Prescribing the use of such open-source essentials over proprietary solutions from outside vendors wherever the government is in direct control (that includes education), would be a much-needed start, especially since companies like Microsoft tend to sponsor deployments of Office 365 in those very same populations to deprive open-source alternatives like Libreoffice[17] and Nextcloud[18] of growth opportunities[19]. EU sovereignty can't bow to the lowest bidder, especially one that makes EU citizen private data available for corporate and US government abuse by design.

The economic threat from the cloud giants

Cloud computing has generated huge scale-benefit profits, which hyper-scalers have invested into bespoke solutions for all use cases that aggregated into sufficient mass.

For those solutions they have transitioned over the last ten years from custom software on commodity hardware to tailor-made everything, all parts of data centre infrastructure to proprietary chips, chassis, fabrics and software stacks, none of which are typically available to competitors, with the exception of the Open Compute Project (OCP)[20].

Corporate IT and smaller IT service providers can no longer compete, because they cannot purchase the ingredients that would level the field. Facebook's OCP was an attempt to enable a level field, because Facebook itself did not want to be forced into a competitor's cloud. But they are also losing that battle, and are becoming too small against what Amazon can do while facing rough headwinds from all directions.

Today only China has tried to out-scale the US technology giants, but the US technology sanctions currently seem to steer

6 Please note that curating this list is the first action to take.

China towards profiting from its greatest asset: the size and growth perspective of its domestic market protected from US competition.

IT is naturally global. Without restrictions, every service, even the most popular or critical, consolidates to a minimal number of providers, even a single one.

Europe needs such essential services, either some that are uniquely European or others protected against global competition while they incubate. Privacy and democracy are currently the only ones that seem scarce in the US, China, Russia, and other large parts of the world with less disposable income. If Europe can't utilize this or find others, it can only become a niche presence in global IT.

The personal computing space

We tend to regard (personal) computing devices as a single "thing", and apply a single set of rules to them, including all their external communications.

That is no longer how we use them, and it creates unsolvable problems both for sovereignty of individuals or civil collectives, and for governments or other regulators who need to be able to wield executive power to remain sovereign.

The current executive overreach into encrypted messengers and on-device scanning for "harmful content" risk criminal and executive abuse and expose device owners to harm. Yet the demand for total "privacy" for both these devices and all their communication effectively eliminates executive power in a fully digitalized world and results in anarchy.

We need to create differentiations and discriminations which currently do not exist, mostly because the device vendors and internet giants are profiting from an effective executive monopoly, acting as pseudo governments, or creating metaverses as regulatory exclaves from our ground universe.

These vendor "dictatorships" have been going on for more than a decade now, so they have fully exploited any "move fast and break things"-innovation sprints one might grant to juveniles; there is no excuse to further delay taking the reins of absolute power from mature industries big enough to cower smaller nations. China and other totalitarian regimes are completely right to understand the necessity; the EU and other democracies urgently need to develop matching frameworks, legal and technical.

We need to distinguish between **activities/apps**, who often act as agents of proxies not just of the owner, but for others (e-government/health, employers, even service providers), distinct **pools of data** these may need to maintain separately and securely and then **communications**, both external but also internally within the device.

Recommendations to support individual sovereignty

1. A personal computer (which includes all smartphones, tablets, chromebooks etc.) is *personal*, the owner the only true sovereign. A manufacturer has no right to restrict an owner's rights as to the software and services he runs on them. Any software that acts against the interest of the owner is illegal by default, especially when it does so unbeknownst to the computer owner, grossly so after the interest against it has been expressed[7].
2. Devices, software, and services sold in the EU need to comply with EU law and require an EU legal entity that can be held responsible, penalized for incompliance, and sued for damages from within the EU. Bigger, more critical services require the same at country level; they also need a regional failback service to ensure continuity and easy consumer protection[8].
3. Consumers must be given a choice to manage their computing devices on their own ("rooting"), or delegate management authority to a provider of their choice (not just the vendors like Apple, Google, and Microsoft).
4. Computing devices need to support enclaves, software partitions or embassies in such a way that personal, employer, e-health, e-government, and service provider applications can run in an environment that protects these apps and their data from any other, including the device owner (to enable digital rights management), who retains "traffic control" (not content control) over apps and data. That traffic control must include the ability to freeze, back up, remove and restore (encrypted) enclaves to manage contractual changes or international travel.
5. **We recommend that the EU fund research in proper international enclave management, as this topic is not sufficiently explored, to put in full practice with consumers travelling across bigger borders.**
6. Personal computers must have either an open-source or open application programming interface (API) base firmware, so owners can choose or write their own operating systems.
7. Any operating system sold with the device needs to continue to work, and do so with support from within the EU, if a political or environmental crisis erupts at or beyond EU borders.
8. Firmware that is no longer maintained by the vendor must be verified to contain no exploitable vulnerabilities from their APIs or open sourced for a minimum of ten years after stop of sales. Vendors must set up matching escrow services for devices sold in Europe.
9. There must be no apps already preinstalled on computers. Install-ready images can be provided on the device but need to be removable easily. No vendor may prescribe the use of a particular variant of a generic app for generic service or document type e.g. Edge- or Safari-browsers, PDFs or office documents, video or audio files, conferencing or remote connections.
10. **Service providers and software vendors may need to be regulated similar to food safety or building codes.** Yes, the safety of software is a bit more difficult to assess than testing the stability of

7 When internet users send "do not track", that is a very clear declaration of will, which may be ignored without penalty.

8 Really free and open-source software will require special attention, software that's only free to start and then commercialized via subscriptions or in-app purchases needs to be regulated similar to a purchase.

concrete, but the industry is developing methods and frameworks, a state of the art that is evolving much like building codes or food-safety regulations are not static. And like those they need not be entirely government operated, while they certainly need to be strictly split between vendors, underwriters, the creators and the monitors of the processes. The effort manufacturers are currently investing can certainly be replicated and perhaps improved within the EU and in accordance with the privileges applications demand and the context or enclave in which they run: there is less need to check for GDPR compliance in a game that runs in an enclave securely kept free from any personal data. The IT software industry and even PCs, mobile devices, the IoT and the cloud have come of age and grown far too big to be treated like kids: the gloves need to come off. Safety, security and compliance are in the public interest and a sovereign mandate in a digital society. Note that unfettered "innovation" in steam boilers also had to be reined in, once they spread too widely to ignore their potential harm. And just like you cannot rely on used-cars salesmen to run road safety inspections, the major technology companies cannot remain in sole charge for developing app safety guidelines and their execution on the devices they sell.

11. No firmware, operating system or app may send telemetry, usage, or personal data without the owner's informed consent; the initial setting must always be deactivated. The only exception is digital rights management at reasonable granularity (e.g. once every six months). That data format must be documented and inspectable. "Embedded" browsers or viewers must follow the user preference determined before first use.
12. "Do not track" must be respected; ignoring it must be made a criminal offence. Research into creating a variant that works should be funded by the EU.
13. Code injection by apps into websites must be illegal, selective code execution (and content filtering e.g. for ad protection) from websites an unalienable right: copyright on code may not beat ownership of the execution environment.
14. Content filters searching e.g. for child pornography, blasphemy or content judged politically incorrect may not be performed on data at rest within the private enclave without previous consent. The owner must be given the ability to delete or quarantine incompliant data without triggering outside communication ("spying on the owner"). Local legal, corporate or vendor ethical rules should only be applied to data that is being moved between enclaves or transmitted on public networks (which really is just another enclave).
15. We need to support a facility by which e.g. family members can exchange data privately between themselves, that might be judged incompliant for more public transmission, e.g. parents exchanging pictures or videos of their babies bathing.
16. Disconnecting digital service for incompliance must be contestable in a local court [21].
17. No computer must require an internet connection or the registration of an owner to operate as purchased, with the installed operating system or provided applications.
18. No app or service may be withheld from an individuum that does not want to be identified if it's not necessary to function for the consumer. A vendor's interest in "improving our services" is not a necessity.

Consumer protection

In a mostly digital industry with few physical assets, horizontal expansion is at the heart of the software revolution and cloudification. But verticalization also has both a much lower investment hurdle and higher incentives for control, speed to market and cost savings. It helps keep out the competition as well, to the point where every content provider, from device vendors like Apple or Sony to social network providers like Meta and game publishers like Tencent is aiming to create their own walled garden metaverse, using the Amazon approach of providing a business platform that "compekills" its more successful customers via the upper hand built into the base.

These big players then seek to protect their near total control of these captive markets by claiming that governmental regulation would stifle innovation and their ability to create "disruptive new products"' But after their initial transformative idea, all their remaining innovation is the smokescreens thrown up to keep the regulators out. Underneath are business practices similar to the ones that J.D. Rockefeller employed 150 years ago, and the very reason anti-trust legislation was invented. These metaverses are little more than the attempt to invent a private ethos and evade ground rules, a rather ancient technique also employed by the mafias and triads for centuries.

They rely on rapid addition of features and services to keep ahead of regulation and control, which today tends to wait for the first complaints to start investigating. By the time they might have won an anti-trust case, starved-off competitors have fallen to the wayside, because the legal system operates far too slowly for a digital society.

That is why regulation must anticipate the vertical and horizontal integration and walling of metaverse gardens and reverse the liability for proving competitive good behaviour. Vendors and service providers need to prove how the market remains open to competition, while breakup along verticals, horizontals and perhaps across value system borders need to be automatic once a threshold of significance is achieved.

This reversal of proof is also necessary and important in other areas cited below.

Malls vs merchants

For a couple of decades towards the end of the last century, large shopping malls dominated retail. Much like an apartment building, malls provided central services like parking, security, power and water supplies, ventilation/heating/air-conditioning, waste disposal, perhaps even delivery services to shop owners who could rent out space for their business.

Mall owners/operators and merchants – in their role as mall consumers – were separate and would not compete with one another, as that would give any shop owned or operated by a mall owner an unfair advantage over other merchants.

Amazon, Apple, Google, Microsoft, Steam today operate digital malls, but they also compete with merchants as their customers within those malls. And that competition is rather lopsided and far too easy to exploit e.g. by Amazon, who have an extraordinary amount of insight into the business of any merchant using their platform and can use that knowledge to replace them.

Steam[22] is mostly seen as a game distribution platform today but started as an internal patch and update service for Valve as a publisher. It gained quick popularity and dominates the PC games market because it eliminated the copy protection worries of publishers and de-materialized distribution e.g. via CD/DVD media with the help of a reliable digital rights management (DRM) platform. Consumers let go of purchasing physical media with the understanding that Steam would safeguard what they bought and enable them to re-install or move any game they owned to any device they operated. For that combination of a working DRM and consumer convenience Steam charges 30% "service tax", generating revenue that far outstrips Valve's earning as a publisher.

But it has also sparked the envy of the largest game publishers, as well as those companies who believe they retain ownership over the platform they sell devices or software for like Apple or Microsoft.

The roles of a device vendor, software vendor, mall operator, gaming server provider, digital asset notary etc. must be segregated cleanly and completely to ensure that consumers cannot be blackmailed, and independent merchants be abused. That segregation may not just have to be functionally, but also geographically to ensure that functionality and services can be maintained in the EU, when a new digital divide opens up and even if such a segregation isn't required or even prohibited in a merchant's/publisher's home market.

The game server and digital asset notary functions also need to be segregated out to ensure that the purchased digital assets remain functional even if the original publisher has ceased operations.

In short: **dematerialization enforced by the publishers of any digital content may not result in compromised ownership.**

Purchase rights and ownership of digital content

1. A purchase transfers ownership over a digital copy. A limited use "purchase" or revocable content may not be labelled as "purchase".
2. The copy may be watermarked or branded to reflect the purchaser and the copyright expiry and DRM can be used with accordance to restrictions (e.g. multiple concurrent devices) clearly stated at the time of purchase.
3. Any content without copyright, even if purchased, must not be distributed with DRM.
4. Any purchased content that runs out of copyright must be de-protectable by the owner. A service or tool must be provided for ten years after purchase.
5. Any purchased content under copyright needs to be kept available for download for as long as specified during sale.

Addiction

Computer and console gaming originally depended on selling physical media (ROM cartridges, CDs, DVDs) for monetization, which was tailored to affluent Western markets where consumers could afford to own computers and game consoles.

In Asia, most gamers could not afford their own computers but went to internet cafés instead, where for an hourly fee they could play games that were either pirated outright or re-implemented as clones.

Internet cafés enable new business models based on free multi-user online games with player IDs and a steady flow of tiny amounts for skins, weapons, extra time etc. using psychological techniques to keep players playing and paying.

It turns games into gambling and the near automatic result is addiction. Today mobile games measure an *individual* gamer's thresholds for loss aversion, endowment/envy/embarrassment effects, susceptibility to scarcity perception live. They use this information to personalize the game, adjusting offers and prices accordingly, e.g. via ironSource technology purchased by Unity.

Vendors will argue that data collection and game individualization is necessary and their business model depends on it to make their services "better".

They have started to turn the data collection not just into a business model for themselves, but also into a business model for the player-consumer having them gain staked digital assets from competing gamers via distributed ledger technologies (DLTs) ("play-to-earn"), blurring gaming and gambling, encouraging addiction.

This personalized content loop will spread from mobile games to wherever else it fits. We know the temptation to create addictive content is irresistible. So we can and must regulate before it happens.

- Measurement-based individual personalization needs to be heavily regulated or outright stopped via anti-addiction regulation, where service providers need to

prove their offers do not create addition, perhaps by proving that their revenues do not profit from addiction.

- These certifications need to be made externally and independently, similar to the food safety/building code approach.
- The residual risks and consequences need to be compensated with via regulated underwriting and insurance.
- Pay-to-win, play-to-earn schemes and loot boxes may need to be illegal from the outset, or at least have age restrictions or the financial exposure limited to fixed total amounts per individual and time period.

Summary

Digitalization is no longer new; the "disruptive moment" is far behind us. Governments and companies failing to align ground and cloud rules aren't "supporting innovation", they are simply negligent.

With digitalization reaching every aspect of citizens' personal, professional, and political life, the clashes in values between geographies need to be anticipated and prepared for.

Events which from the EU look very remote and should only affect the poor residents of a US state with a dwindling population, can very easily have a global impact: righteous US judges and juries have no qualms coercing global cloud giants with a US base or connection into enforcing an ethos as controversial as refusing abortion to teenage rape victims on the planet. So far, we feel less pressure from Malta, an EU member with a similar stance, while France aims to make abortion a constitutional right.

China has long ignored criticism on its most egregious human rights transgressions within, and likewise refrained from imposing its doctrines to the general public outside its borders. But that might change about something as seemingly insignificant as a candy bar campaign mentioning "Taiwan" in the context of "country"[23]. This trend of having a very localized ethos overreaching far beyond "natural" borders via internet mechanics completely lacks compensating controls.

The digitalization of global interaction brings humanity into a contact that is far too close, immediate and unfiltered for comfort. Without the support of a spatiality[9], which allows us to reuse the social code of civilization and diplomacy learned through centuries, millennia or even along our path to humanity and very much based on distance and visibility, we enter a phase where basic civility needs to be relearned, while our ability to inflict annoyance, anger or outright harm has been greatly enhanced.

The digital chasm is not just a technical problem or a short-term issue. It is a reboot of civilization with superpowers humans have not culturally evolved to manage properly.

That's why we urgently need to deal with it with an urgency that fits the head-on collision we are trying to avoid.

For a start it requires essential international IT services like mobile play stores, games platforms, social media, payment services, security services must be split or pre-split at political borders that could foreseeably cause a disruption. We can't allow a situation where domestic operations of a service are cut off because it's delivered by a company that suddenly sits behind a new political wall. Businesses, agencies and consumers must know before they buy or contract, if a service can continue autonomously in a country or region like the EU. For critical services a local failback must be provided and tested regularly.

9 See the "Gaming, content and the metaverse" article in this HiPEAC Vision

References

[1] https://en.wikipedia.org/wiki/Great_Firewall
[2] https://www.semianalysis.com/p/china-and-usa-are-officially-at-economic
[3] https://www.cato-unbound.org/2009/04/13/peter-thiel/education-libertarian/
[4] https://www.debian.org/
[5] https://lineageos.org/
[6] https://www.linux-kvm.org/
[7] https://www.mozilla.org/
[8] https://www.libreoffice.org/
[9] https://nextcloud.com/
[10] https://www.sumatrapdfreader.org/
[11] http://www.videolan.org/
[12] https://www.thunderbird.net/
[13] https://opnsense.org/
[14] https://github.com/signalapp
[15] https://virt-manager.org/
[16] https://openvpn.net/
[17] https://www.libreoffice.org/
[18] https://nextcloud.com/
[19] https://www.theregister.com/2022/11/22/france_no_windows_google/
[20] https://www.opencompute.org/
[21] https://www.heise.de/news/Nacktscanner-Unbedachte-Fotos-vom-Kind-fuer-den-Arzt-Google-Dienste-gesperrt-7238900.html
[22] https://en.wikipedia.org/wiki/Steam_(service)
[23] https://www.abc.net.au/news/2022-08-06/snickers-apologises-to-china-for-calling-taiwan-a-country/101308044

Thomas Hoberg is the technical director at Worldline Labs, Frankfurt, Germany.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2023.

Cite as: T. Hoberg. Europe's needs for digital essentials and individual sovereignty. In M. Duranton et al., editors, HiPEAC Vision 2023, pages 180-187, Jan 2023.

DOI: 10.5281/zenodo.7461962

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

© HiPEAC 2023

The current education system was designed in the 19th century to support the first industrial revolution. Does it still meet the needs of the 21st century?

Rethinking education

By KOEN DE BOSSCHERE and TULLIO VARDANEGA

The education system as we know it was designed in the 19th century to prepare the workforce for the needs of industry and government, as understood at the time. The basic concept has not fundamentally changed in the last 150 years. It worked well, provided that (i) people had lifelong jobs for which they could be trained at a young age, and (ii) knowledge was mostly shared via specialized books and libraries and changed slowly.

Today, (i) schools can no longer prepare the students for a lifelong career, and (ii) the ease of internet access and use has shattered schools' strict monopoly on knowledge dissemination. This has profound consequences for the future of education, which should be focused less on delivering static degrees of competence, and more on flexibility, self-directed learning, entrepreneurship and innovation. At the same time, while much younger in age, computer science and computing engineering education also have to change in order to encompass more profound awareness of the changes that informatics and digitalization are having on the very fabric of our society.

Key insights

- Schools have lost their monopoly on teaching; the internet has become one big international school: affordable and open to everybody, 24/7.
- The current education system was designed to serve 19th century industry and society, and it has to adapt to serve the needs of the 21st century.
- Schools can no longer educate people for a whole career; they can only educate people for their first job. Lifelong learning is essential for individuals to stay relevant in the job market, and to keep Europe competitive.
- High-quality education is a crucial element of sovereignty. If the European Union (EU) cannot produce enough local graduates with the right skills, it will not reach full sovereignty but will have to depend on foreign resources. In order to reach digital sovereignty, the EU needs enough information and communication technology (ICT) workers.

Key recommendations

- Modern education should be a mix of formal and informal learning, coaching students to become T-shaped professionals who serve the economy and society in the 21st century. All graduates should have basic digital skills.
- The EU should work to ensure a sufficient supply of world-class computer science and computer engineering courses, as well as ways to retain talent trained in Europe. Unconventional routes to technology careers should also be considered to broaden the talent pool.
- Universities should take on two important additional roles: lifelong learning and supporting regional entrepreneurial ecosystems.
- It is time to develop digital ethics as a separate discipline and to integrate it into higher education curricula.

The evolution of compulsory education

Compulsory education has a long history. In Europe, it started in the 16th century with the Protestant Reformation in Germany, where Martin Luther called for compulsory schooling in order to make sure that Protestants could read the German translation of the Bible by themselves. In the 17th century, several Protestant territories in Europe and the North American colonies implemented compulsory education for both boys and girls. The first state-wide compulsory education system was installed in Prussia in 1763 (for all children aged five to 13).

Around the same time, Prussia also invested in the creation of modern universities that used German as the language of instruction, chose rationalism over religious orthodoxy, applied new modes of teaching, and gave significant freedom to the professors (academic staff) who could also spend part of their time on research. The universities became the centre of the German Enlightenment in the 17th and 18th centuries [1].

Unleashing the intellectual power of the smartest people led to an unprecedented series of leaders in all areas of human cognition in the German-speaking world: Bach, Mozart, Schubert, Mendel, Freud, Engels, Marx, Kant, Nietzsche, Bonhoeffer, Ratzinger, Euler, Gödel, Gauss, von Liebig, Kekulé, Koch, Clausius, Boltzmann, Hertz, von Helmholtz, Röntgen, Planck, Einstein, Goethe, Schlegel, Hegel, van Beethoven,

von Humbolt, Schiller... This intellectual development also led to many technological innovations [2].

In the 19th century, compulsory education became mainstream in most Western countries and was universal by the time of the Second World War. Compulsory education was no longer inspired by religion, but by the needs of the industrial revolution: the availability of lots of workers with the right skillset. This explains why the modern compulsory education system was organized like a factory assembly line: children all entered at the same age and, year by year, they learned a standard set of facts, skills, competences and attitudes. Children and adolescents were also disciplined as if they worked in a factory: schools were run according to a strict daily and weekly schedule; arriving late at school was not permitted. There was also quality control: if they failed the test for a particular year, they had to retake the year, or switch to a different track. Some countries established standardized tests at crucial transition points in the name of "guaranteeing" quality. Out of such schools came a steady stream of standardized and government-approved workers: doctors, lawyers, teachers, nurses, construction workers, mechanics, ICT workers, …

In every country, public education is a very important department of government because it is not only large scale but also politically sensitive, as this is the place where future generations are being formed, and where the future of society is shaped. The focus often depends on the government that is in power: there may be more or less focus on nation building, on integration of under-represented groups, on STEM education, on religious or cultural studies, on entrepreneurship, on excellence, …. Furthermore, many countries honour freedom of education, which means that parents have the right to have their children educated in line with their personal views (political, religious, social, language, …) and without intervention of the nation state (i.e. via private schools or home schooling). In the 20th century, a number of governments experienced difficulties because they could not agree on education policy [14].

At the transition between the 20th and 21st centuries, the education model has changed minimally: (i) compulsory education has been extended in most countries (lowering the age of entry and raising the age of completion to 18), and (ii) the body of knowledge designed for learning has been continuously updated. In an evolution that started after the Second World War, many students started higher education because secondary education was no longer considered a sufficient basis for a well-paid career. Today, for many young adults, their formal schooling ends when they obtain a bachelor's or master's degree. The number of PhD students has also increased over the last 50 years, leading to increasing numbers of doctoral degrees. This model has served society and industry very well over the last 150 years, allowing countries to exploit the intellectual potential of the population.

Schools are no longer the only option for education

Schools focus on formal learning: the teachers explain the material, and the pupils prove that they understand it via testing. Eventually, pupils obtain a diploma that proves that they completed a particular study programme. Schools have a historical monopoly on formal learning but, with the rise of the internet, they now have strong competition [8,9].

The under-18s of today do a lot of non-formal (structured learning outside school, e.g. learning to play a musical instrument) and informal learning (in daily life) about their interests, hobbies and the world around them. A significant amount of non-formal and informal learning happens on the internet, which offers many opportunities for self-directed or incidental learning. The material is often presented in a way that is both very attractive (short movies, animations, games, demos, …) and fun, and does not feel like learning.

In Europe, the complete educational track from preschool to final graduation is funded by the government (at least for the public schools). Governments often start funding educational places for children aged two to three years, when they enter preschool, until they obtain their highest degree: up to 20 years in the case of a master's degree. Such a period equates to almost 50% of the span of a full career; it hence represents a huge societal investment in every single newborn.

Some countries allow universities to charge tuition fees to cover part of their costs. In countries with high tuition fees, it is common to finance a higher education degree with a student loan that has to be repaid. The offering on the internet is available 24/7, and is very cheap compared to the cost of schools. It is an attractive option for people who cannot afford the classic option. The offering is overwhelming in its breadth and range, and is growing daily.

Future education will be a mix of formal, non-formal and informal learning

Change is accelerating; there are no longer lifelong guarantees. Today's globalized world has been described as VUCA: volatile, uncertain, complex and ambiguous [13]. Unfortunately, schools are all but VUCA: they are usually structured and predictable, they often simplify things and avoid ambiguity, they offer a protected environment for students. This is quite different from the real world.

Despite this, schools do not work equally well for all pupils. Teaching may progress too fast for some, too slow for others. Some have problems with strict daily and weekly routines, and some encounter difficulties when they hit puberty. The one-size-fits-all model is not the best choice for all children. Thanks to freedom of education, parents who can afford to do so may decide to send their children to a private school, which may be able to offer a more personalized education.

In the real world, the evolution of science and technology goes at breakneck speed, and the knowledge base delivered by school education is rapidly and increasingly proving insufficient. What is certain is that it will not serve a complete career. Today, the half-life of knowledge in some disciplines is less than 10 years [3]. Much of the information that people learned in school 20 years ago has been refuted by new scientific insights. As a result, learning cannot stop with graduation. Adults will

have to further develop their knowledge and competences throughout their lifetime.

Obviously, we cannot send adults back to school to update their knowledge after a number of years. Instead, we must work with non-formal learning and informal learning. But then the question is: why wait until graduation to start lifelong learning? And why not start much sooner with non-formal and informal learning? In other words: is 20 years of uninterrupted formal education the best preparation for a career in the 21st century? If it isn't, which things should be part of formal education, and which could be part of the non-formal and informal education of children?

Formal education might be the best option to learn the basics of the established disciplines: mathematics, physics, biology, chemistry, languages, history, culture, economics, …. When studying the basics, study programmes should not focus too much on teaching solutions (which are by definition changing), but instead focus on reasoning and on the fundamental principles of the discipline, which have a much longer half-life. Furthermore, it is the fundamental principles that are needed to develop outside-the-box solutions in the future. Learning about the solutions can be done more easily in a non-formal or informal learning setting by reading, watching documentaries, and undertaking internships and voluntary work, often on a need-to-know basis in the context of project work. Making students partially responsible for their own education will lead to them being responsible for it after graduation too.

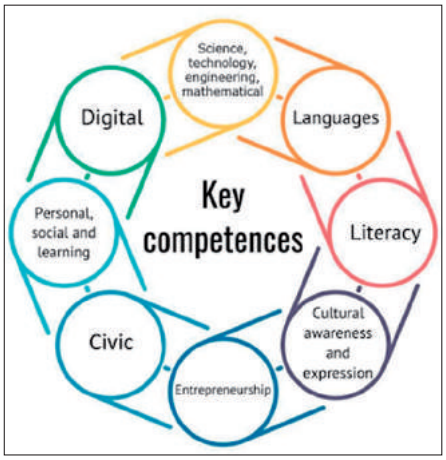


Figure 1: Key competences for lifelong learning [4].

At the competence level, study programmes should focus on the eight key competences for lifelong learning as adopted by the European Parliament in 2018 [4].

Notable in that recommendation is the focus on science, technology, engineering and mathematical competences, combined with digital skills as key competences for all citizens in Europe. The focus on entrepreneurship in combination with the soft skills of personal, social and learning competences is intended to make Europe more competitive. The combination of the "cultural awareness and expression" competence with civic competences should provide all Europeans with a common framework for values, democracy, globalization and multi-culturalism. Finally, literacy and (foreign) languages and culture are important as a means to learn, listen and express ideas.

These eight key competences are fundamental for personal fulfilment and development, employment, social inclusion and active citizenship. They break with two legacy traditions that have burdened formal education worldwide since the 20th century: the dichotomy between the humanities and the sciences, and the dichotomy between pure and applied training [5].

Towards a new role for universities

There has been a constant evolution in the role of higher education. Originally, universities were pure teaching (viz. transmissive) institutions in which professors passed on to the next generation the accumulated knowledge of the previous generations.

Later, professors were encouraged also to do personal research, and to create new knowledge. With the appearance of research universities, all professors were required to be active researchers. This requirement went hand in hand with the development of a number of research degrees (master's and doctorate degrees). Research universities have contributed enormously to the development of the modern world. Many of the things that we take for granted today have been developed at (or in collaboration with) university laboratories. The availability of large numbers of graduates with research degrees has also led to a professionalization of industry and government agencies.

More recently, universities have been encouraged to broaden the I-shaped profiles of their graduates into T-shaped profiles [6].

This change means that students should have a broad base of general supporting knowledge and skills, supplemented with deep knowledge and skills in one or more areas. In the broad base, the student must learn complex problem solving, critical thinking, creativity, people management, coordination with others, emotional intelligence, judgement and decision-making, service orientation, negotiation and cognitive flexibility [7]. These are the competences that set humans apart from computers and robots.

The deep knowledge and skills element must encourage the student to learn how to push forward the state of the art in a subject, and to create new knowledge and to innovate. The harder students are pushed to stretch themselves in the deep elements,

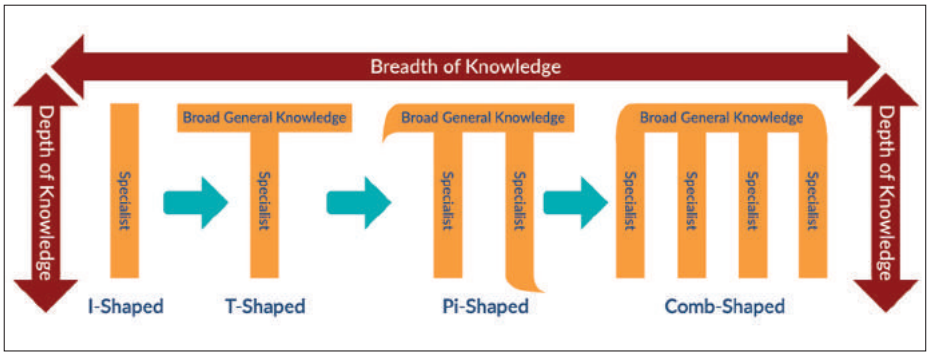


Figure 2: Broadening I-shaped to T-shaped profiles [6].

the more they will learn, and better placed they will be for the challenges of the 21st century. One thing is certain: there is little value in receiving specialist training only to end up doing routine tasks. Such jobs are disappearing because they are the easiest and fastest to automate. T-shaped education offers better guarantees for living well, self-fulfilment and wellbeing.

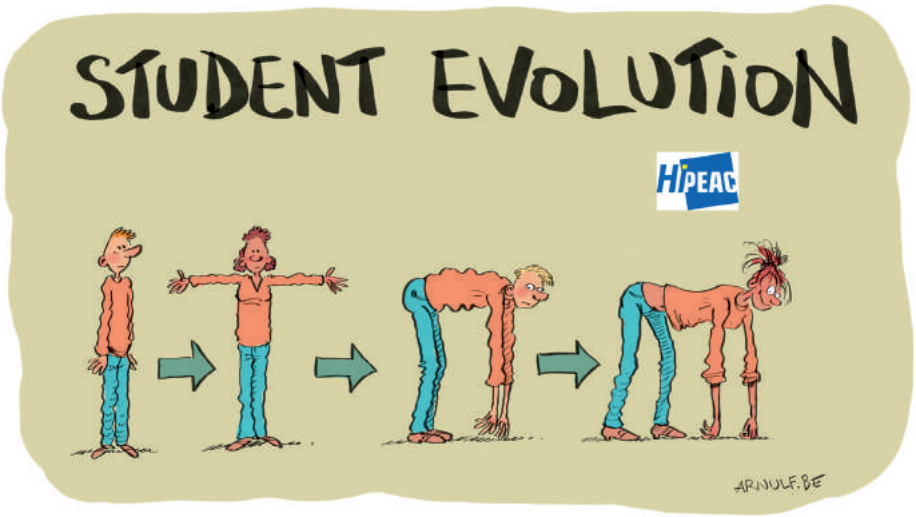
T-shaped profiles can also help to reduce the shortage of ICT workers in Europe (which is estimated at 1 million). It is not only computing specialists that stand to broaden and update their competences; other disciplines could also broaden their graduates' skill sets by providing them with a basic understanding of computing, big data analytics and artificial intelligence, on a par with the prescribed basics of sciences, history and foreign languages. Such graduates could easily contribute to the ICT sector in roles that are technically less demanding.

There will be a need for a broad range of graduate profiles because society increasingly depends on ICT (for digitizing industry, securing ICT systems, designing smart grids for the transport of renewable energy, development of precision agriculture to reduce the use of pesticides and irrigation, etc.). This measure would also improve the gender balance of the ICT sector in general.

Another recent evolution is that universities are encouraged to actively monetize their research via intellectual property (IP) portfolios and the creation of spin-offs. Most universities now have a technology transfer office tasked with helping researchers to protect their intellectual property and to exploit it, either via an agreement with existing companies, or via the creation of a spin-off company.

In some places, this evolution has led to two additional changes:

- The active promotion of an entrepreneurial mindset in students, via curricular activities that stimulate them to leave their comfort zone, and try to think more like an entrepreneur.
- Becoming an active partner in the regional entrepreneurial ecosystems, supporting and developing them by creating spin-offs and by training the next generation of entrepreneurs through study programmes. This activity also creates attractive jobs for highly specialized graduates (often with PhDs). This job market can slow down the brain drain to other parts of the world, and guarantee that Europe maintains a critical mass of expertise in important sectors.



This is a transition through which traditional research universities become entrepreneurial universities that produce not only graduates and research results, but also innovations and economic prosperity. This is the response of higher education to the needs of the post-industrial economy.

The need for digital ethics

In Luciano Floridi's formulation [10], humans have lived through prehistory (where communication was verbal and the clan was the information agent) to history (where all that someone has been able and allowed to write is written down: this is where the state is the information agent), to "hyperhistory" (where machines automatically produce the majority of data and have become the primary information agent).

By acting as an information agent in hyperhistory, ICT does more than actors did in history: it "wraps" the world in a digital envelope designed to facilitate computerized operation within it. We humans are being progressively absorbed into that envelope of which we are less and less in control. A few examples illustrate this situation.

1. The combined effect of exponential increase in processing power and equally exponential decrease in processing cost is the generation of an exponential amount of digital data. Exponential phenomena are problematic: when you begin to notice them, it is because they have become sufficiently large to catch the eye. Yet it then only takes a few more steps for them to ramp to gigantic proportions, entirely beyond control.
   The quantity of data being produced today already vastly exceeds the available storage capacity, and this will only worsen in the future. This means that massive amounts of data should be deleted daily or never even recorded. Deciding what to delete or what to skip is not a merely technical decision if that data is subsequently used to train AI systems, which, increasingly, is tantamount to "writing history", because deletion or skipping may cause harmful bias in the learned inferences.
2. Digitalization has broken the link between place and law. Geographical space and cyberspace do not coincide: the old legislative foundations no longer apply and the spreading of digitized information in cyberspace is very hard to control. Not surprisingly, therefore, some legislators have reacted to this situation by pushing forward the notion of "data sovereignty", which enforces the rights of the state (place) over the data originating within its boundaries. Examples of such efforts are the GDPR and the recently invalidated EU-US Privacy Shield [11]. This move, however, goes counter to the

original fabric of the internet, the web and the logic of their pipelines. Adopting sovereignty-driven norms calls for a rethink of the architecture of the internet and of the web above it.

3. Current ethical frameworks often assume that there are humans, or at least living biological beings, involved in an action. When two computers attack each other with malware, there are no living beings involved in the action, but malicious people are at the origin of the problem. This is why digitalization requires us to reconsider existing ethical frameworks and to adapt them for the various world envelopes that ICT keeps creating. There is a need to develop digital ethics as a discipline (similar to e.g. bio-ethics or medical ethics). The fact that some universities are setting up chairs and institutes in digital ethics is a promising evolution.

We are at a point at which computing engineers who develop hardware and software for decision-making systems should be more acutely aware of how ethics "flows" in their systems, e.g. the fact that a seemingly insignificant decision to store only relevant events in a database and leave others out might eventually lead to a bias in systems that use the database. Unfortunately, however, current computer science (CS) and computer engineering (CE) education scarcely extends to understanding the workings of ethics. We therefore believe it is necessary to include the basics of digital ethics in all higher education study programs, and certainly in computing curricula.

As part of ethics education, it is essential to include quality courses on sustainable computing. Rather than assuming there are no limits other than physical laws, computing education should start with the premise that any solutions should be delivered within the capacity of Planet Earth, from sustainable materials to energy-efficient software. Indeed, Europe could be a leader in the field of sustainable and ethical computing studies.

Incorporating ethics and sustainability into computing education will not be achieved by adding siloed ethics, philosophy or law optional courses within CS and CE curricula. Such courses would risk being isolated and not "germinating" in the students' minds. It is reported [12] that top-quality schools have begun recruiting philosophers in their CS departments and had them contribute to the design the syllabi of curricular courses. This is a promising approach that Europe should explore further.

Conclusion

Education has always been important. Excellent education is key to solving the global challenges of the 21st century. Without a well-trained workforce, Europe will not be able to compete with the rest of the world. Human minds are the most important natural resource that we have in Europe.

Europe was the birthplace of modern education, and it has one of the best education systems in the world: free, often excellent, education for all children up to the age of 18, and affordable higher education. Thanks to its education system, Europe has one of the best-trained and most professional workforces in the world.

An effective education system is also an enabler of social mobility and a prerequisite for sovereignty. It is an asset that we should cherish, protect and make future proof. This requires adapting it to a changing reality and evolving needs, and making it capable of driving future progress instead of undergoing it. It is the only way to stay on top in a knowledge-based global economy and to safeguard European sovereignty.

It is time to develop digital ethics as a separate discipline and to integrate it into computing curricula.

References

[1] "Martin Luther University of Halle-Wittenberg", https://en.wikipedia.org/wiki/Martin_Luther_University_of_Halle-Wittenberg

[2] Peter Watson, "The German Genius", https://www.amazon.com/German-Genius-Renaissance-Scientific-Revolution/dp/0060760230

[3] Samuel Arbesman, "Half-life facts", https://www.wired.com/2013/03/half-life-of-facts/

[4] "EU Key Competences for Lifelong Learning 2018", http://keepcalmandteachenglish.blogspot.com/2018/03/eu-key-competences-for-lifelong.html

[5] World Economic Forum, "The Future of Jobs", http://www3.weforum.org/docs/WEF_Future_of_Jobs.pdf

[6] "From I-Shaped to T-Shaped – Why IT Professionals Need to be Multi-Skilled", https://certibanks.com/KnowledgeArea.aspx?articleid=11

[7] World Economic Forum, "The 10 skills you need to thrive in the Fourth Industrial Revolution", https://www.weforum.org/agenda/2016/01/the-10-skills-you-need-to-thrive-in-the-fourth-industrial-revolution/

[8] Neil Selwyn, "The Internet and Education", https://www.bbvaopenmind.com/en/articles/the-internet-and-education/

[9] D. Schugurensky, "The forms of informal learning: towards a conceptualization of the field", https://pdfs.semanticscholar.org/6315/0f9e5376503715b1c0175f2a5354dd78fbcf.pdf

[10] L. Floridi, "Ethics & Social Science: Ethics in the Age of Information", https://www.youtube.com/watch?v=lLH70qkROWQ

[11] V. Manancourt, "EU court ruling strikes hammer blow to transatlantic data flows", https://www.politico.eu/article/eu-court-ruling-strikes-hammer-blow-to-transatlantic-data-flows/

[12] M. Wagner, "An AI and Computer Science Dilemma: Could I? Should I?", https://informatics.tuwien.ac.at/news/1896?utm_campaign=feed&utm_term=news

[13] "Volatility, uncertainty, complexity and ambiguity", https://en.wikipedia.org/wiki/Volatility,_uncertainty,_complexity_and_ambiguity

[14] "First school war", https://en.wikipedia.org/wiki/First_School_War

Koen De Bosschere is a professor in the electronics department of Ghent University, Ghent, Belgium.

Tullio Vardanega is an associate professor in the Department of Mathematics at the University of Padua, Italy.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2023.

Cite as: K. De Bosschere et al. Rethinking education. In M. Duranton et al., editors, HiPEAC Vision 2023, pages 188-192, Jan 2023.

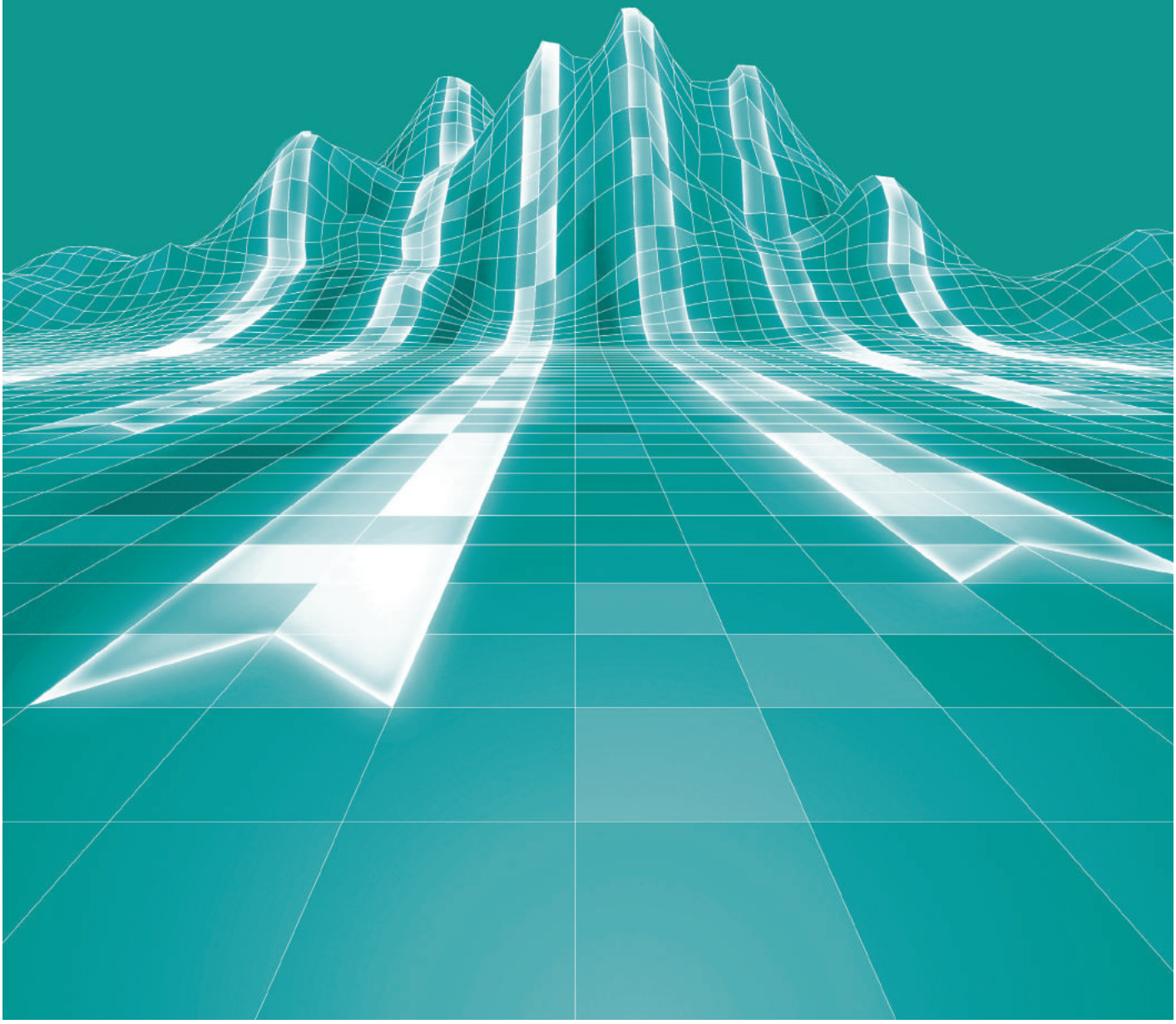
DOI: 10.5281/zenodo.7461976

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

© HiPEAC 2023

HiPEAC Vision 2023

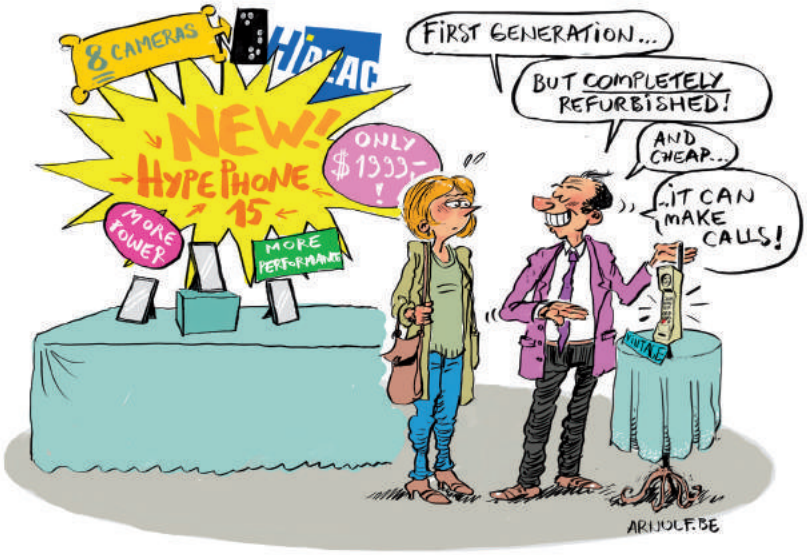
THE RACE FOR SUSTAINABILITY



Sustainability is the biggest challenge of the 21st century; without it, human civilization will be unable to continue. This chapter explores the relationship of computing technologies to sustainability.

The race for sustainability

By KOEN DE BOSSCHERE



Climate change is now felt all over the planet, leading to changing natural ecosystems, and eventually to loss of biodiversity. Climate change is not only about melting ice caps, and rising sea levels, but also about fresh drinking-water supplies, food security, and water needed for manufacturing, agriculture, transportation and energy production. Climate change affects many aspects of our daily lives.

However, sustainability is wider than climate change. It is about organizing the world within the limits of planet Earth both today and for the generations to come. One aspect of it is the use of raw materials, which are finite on a finite planet. Modern information and communication technology (ICT) devices require more than 60 minerals from all over the world, some of which are hard to mine, are scarce, or are mined in socially unacceptable conditions. Once used in a device, they are difficult to recycle; but without efficient reuse and recycling, there cannot be true sustainability.

Modern ICT devices are very power efficient, leading to a situation where more energy is needed to produce them than they will use during their entire lifetime ("embodied" versus "operational" energy use). Hence, integrating multiple functions into one device is more sustainable than creating separate devices per function. Contrary to common belief, continuing to use an old device is often more sustainable than buying a new, more power-efficient one. This shows that a full lifecycle assessment can lead to counterintuitive conclusions, and that such an assessment is essential in order to come up with sustainability solutions with real impact.

Sustainability is perhaps the most important grand challenge of the 21st century and it affects everything we do. This chapter looks into ways to make computing sustainable, and suggests how computing can contribute to sustainability goals.

The chapter contains five contributions.

- "*Climate change is here to stay, so we'd better prepare for it*"
  This article states that climate change is real, and we will have to slow down emissions while at the same time adapting to inevitable changes. Both actions will require digital technologies that are based on the internet of things (IoT), edge and cloud computing. Europe should invest in these technologies to stay at the forefront of technologies for climate change adaptation.
- "*Sustainability: A major challenge in Europe for the future*"
  This article discusses the definition of sustainability. It states that in order to be really sustainable, we should look (i) for solutions that can last indefinitely, (ii) for solutions that are globally sustainable (and not for solutions that export the environmental impact to a different part of the world), and (iii) for sustainable solutions that are also scalable so that everybody will get access to it (and not only the countries/people who can afford it).
- "*Towards circular ICT: From materials to components*"
  This article focusses on the environmental impact of ICT devices and advocates a full lifecycle assessment with maximal reduction of raw materials needed, repairability, maximal reuse of decommissioned hardware, and urban mining of e-waste. Urban mining is still in its infancy. Only a limited number of minerals can be extracted from e-waste although their concentration is often higher than in natural ore.

- "*Towards sustainable computer architecture: A holistic approach*"
  This article argues that sustainable development requires a holistic approach and involves multi-perspective thinking. Applied to computing, sustainable development means that we need to consider the entire lifecycle of a product. Analysing current trends reveals that the embodied energy footprint is, or will soon be, more significant compared to the operational energy footprint. The article summarizes what computer architects and engineers can and should do to better understand the environmental impact of computing, and to design sustainable computer systems.
- "*Everything as a service*"
  In the last 30 years, we have witnessed inreasing dematerialization in many economic sectors: streaming has replaced physical media, e-books have replaced paper books, newspapers have been replaced by apps, owning a car or a bike is gradually being replaced by renting or sharing one. Dematerialized goods require less space to store, and require less energy and materials to produce. Hence, they are reducing the environmental footprint of the world economy. Dematerialization depends on digital technologies.

Key insights

- Climate change affects all regions of the world, including the industrialized world. Everybody will have to adapt to the new normal.
- The environmental footprint of digital devices consists of the embodied energy and materials to produce the device, and the energy needed to use it (operational energy). Both are important when assessing the environmental impact of a technology.
- The more efficient a digital technology is, the greater the impact of its embodied emissions relative to its operational emissions.
- Even if all the energy used during production and use were 100% renewable, the environmental impact of computing would still be significant and growing due to the growing use of minerals.
- Digital devices are notoriously difficult to recycle and many end up in landfills. Better recycling will reduce the dependency of Europe on the import of minerals. It will, however, take another 20 years for recycling to seriously reduce Europe's dependency on the import of minerals (as the installed base is currently growing fast and the recycling technology for e-waste is still immature).
- We need a holistic approach to sustainability: a technology which shifts the emissions in time or in space is not sustainable. To have impact, sustainability solutions should be globally scalable.
- Digital technologies can help to dematerialize goods. Dematerialized goods require less space and energy over their lifetime.
- Reducing resource consumption can lead to a rebound effect (Jevons' paradox), which leads to higher consumption of goods, therefore undoing the savings.

Key recommendations

- We should use a full lifecycle assessment when designing computing systems (including easy repair and recycling). Ideally, we should be able to recycle all the material resources used to build computing systems infinitely. Therefore, they will have to be designed in such a way that they become easier to recycle. Europe could create rules to oblige companies that sell devices in Europe to prove that e.g. 50% of all the material resources of a device can be recycled (similar to emission regulations for cars, or rules on energy consumption for household appliances, …).
- Europe should stop using critical materials that are mined in negative social, geopolitical or environmental circumstances. One solution could be to require that all devices that are sold in Europe have to have a sustainability label (e.g. TCO Certified).
- We should extend the lifetime of devices as long as we can in order to reduce the environmental impact of the embodied energy. Planned obsolescence of hardware/software is not sustainable. Instead, we should make hardware products repairable, and make sure that software updates keep running on older hardware platforms.
- To reduce the carbon emissions of computing, computer architects should first and foremost focus on embodied energy because reducing operational energy consumption is of secondary importance. There are several ways to reduce the embodied energy: by using an older technology node where possible and by designing the smallest possible chip for a given function (e.g. by using a smaller processor, by avoiding dark silicon, by limiting the number of cores to the number actually needed, by avoiding accelerators that are only used sporadically, …).
- The European Union should consider introducing caps for the maximum energy consumption of ICT goods and services sold within the bloc.
- We should continue searching for environmentally friendly alternatives to goods and services with a substantial environmental footprint.
- We should develop business models and economic models that also take into account the lifecycle environmental costs in order to make the sustainable solutions more competitive with the non-sustainable ones.

Koen De Bosschere is a professor in the electronics department of Ghent University, Ghent, Belgium.

This document is part of the HiPEAC Vision available at hipeac.net/vision.
This is release v.1, January 2023.
Cite as: K. De Bosschere and T. Vardanega. The race for sustainability. In M. Duranton et al., editors, HiPEAC Vision 2023, pages 188-192, Jan 2023. DOI: 10.5281/zenodo.7461987
The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174. © HiPEAC 2023

Climate change is moving faster that we are. If we want to turn the tide, we will have to speed up.

Climate change is here to stay, so we'd better prepare for it

By KOEN DE BOSSCHERE

Over the past decade, we have begun to understand what climate change really means for the world. We have discovered that it is not limited to melting ice caps in the mountains and on the poles, or rising sea levels in coastal areas, but that it also impacts our backyard. There are more days with extreme temperatures; in some areas, there are huge wildfires, or severe flooding. Due to lower crop yields, food prices are rising, pushing up the cost of living.

As temperatures continue to rise in the coming decades, the climate will continue to change, and we will have to adapt to this new normal. In order to do so, we will need technology to monitor and control strategic processes and services that are at the foundation of our modern society (power grids, water supply, food production, …).

Key insights

- Climate change affects all regions of the world, including the industrialized world, in ways that most people did not anticipate.
- To safeguard our modern society, we will have to adapt and to become more resilient by preparing for wildfires, megadroughts, flooding, …
- To slow down climate change, we will have to stop using fossil fuels. Defossilizing the global economy will require major investment in digital technologies. The internet of things (IoT), edge and cloud computing can play a major role in adapting to climate change.

Key recommendations

- We should invest in digital solutions that will help defossilize the economy: smart grids, smart mobility, smart cities, further dematerialization, …
- Europe should invest in the IoT, edge and cloud computing as key digital technologies for the European Green Deal.

According to [1], the term "sustainability" was introduced in the 18th century in the context of forestry. At that time, it referred to managing forests to ensure that future generations had access to enough wood for fuel and for timber. In the 20th century it was broadened beyond forests. The first report of the club of Rome, "The Limits to Growth" [2], made very clear that exponential growth of resource consumption could not be sustainable on a finite planet, and that one day natural resources would be depleted. The same report already warned about the exponential increase of CO_2 in the atmosphere in 1972. Later, the definition of sustainability was broadened with social and economic aspects into the formulation of the "triple bottom line" consisting of people, planet and profit [3]. Today's best-known definition of sustainability are the 17 sustainable development goals defined by the United Nations in 2015 [4].

Climate change

Over the past decade, it has become apparent that climate change is increasingly impacting the planet, society and the economy, and that slowing down and eventually stopping or even reversing climate change is probably the biggest challenge of the next 50 years. The UN Secretary-General António Guterres already warned the world in 2018 that "Climate change is moving faster than we are"; since then, carbon emissions have only increased.

As of today, in many countries, evolution rather than revolution is considered the appropriate approach for fighting climate change, but evolution takes time. In addition, many countries require solutions to be affordable in order to be politically realistic. The fact that doing too little or acting too late could in the end be more expensive is not the first concern. In 2022, the United Nations climate change conference COP27 agreed on a loss and damage fund for vulnerable countries hit hard by climate disasters but excluded any mention of winding down the use of fossil fuels [5]. This is a clear sign that countries find it easier to pay for the consequences of climate change than to tackle the root cause.

While there are still people unsure about the causes of global warming (manmade or part of natural fluctuations in the climate), there is a growing awareness that climate change is real, that we will definitely have to adapt, and we will have to do it fast because the climate is changing faster that we could have imagined a decade ago.

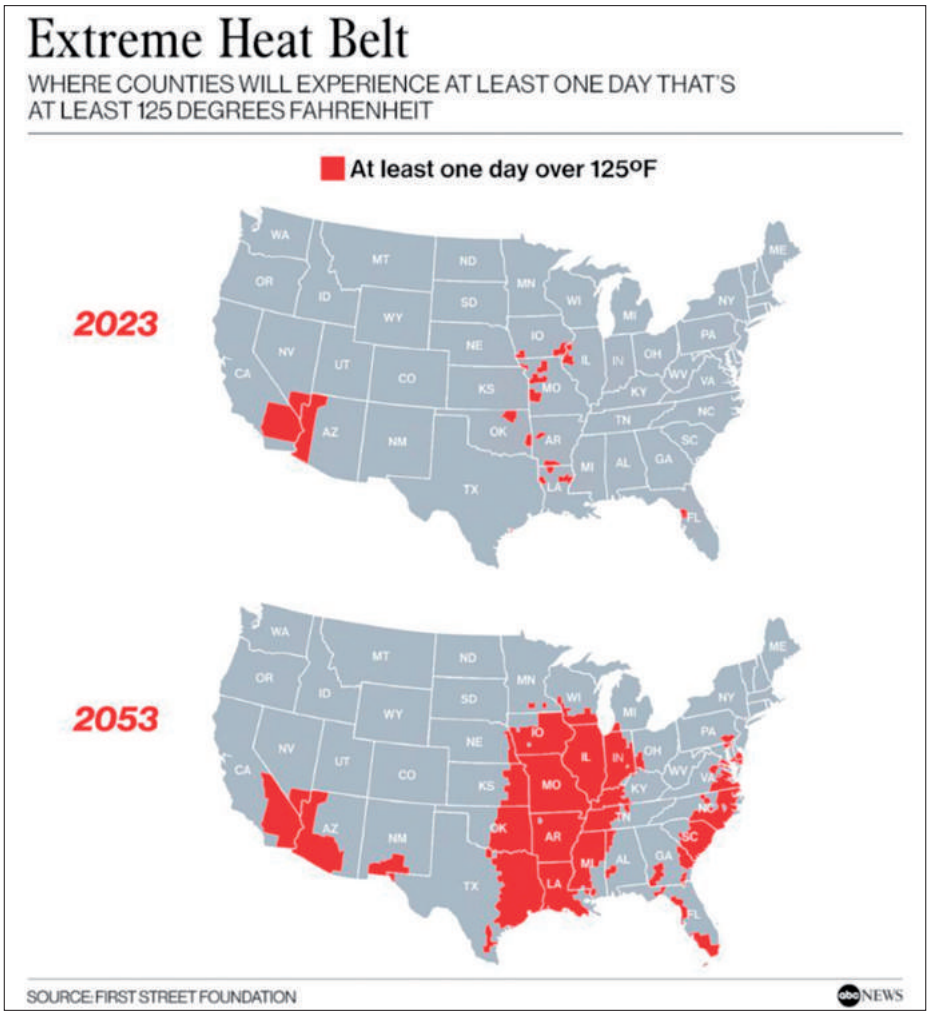


Figure 1: US counties with at least one day of 51°C per year, from [7]

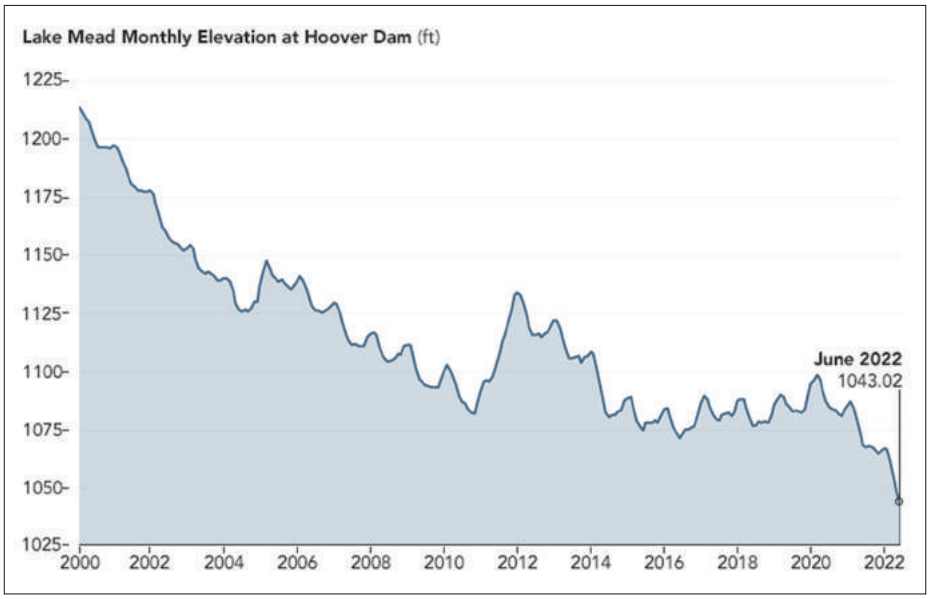


Figure 2: Lake Mead water levels [8]. Power generation halts at 950 ft. The dam stops functioning at 895 ft.

Another recent insight is that climate change impacts all parts of the world, not just small islands in the Pacific or poor countries, but also the industrialized world. Hence, while some will be much worse affected than others, everybody on the planet is going to pay the price, be it directly (fires, flood, heat, drought, …) or indirectly (higher cost of living, broken value chains, higher insurance premiums, contributions to the loss and damage fund, …).

According to a study released by the First Street Foundation [6], in 2053, 107 million Americans will experience at least one day of >51°C (125°F) per year, compared to 8 million today; see Figure 1. All over the US, the number of extremely hot days will increase from one week to one month. Some US citizens might decide to leave the hot areas and move to areas with more moderate temperatures where living is more affordable (lower heating and cooling costs). This will have an impact on the local economy of the states involved.

Economic consequences

The examples are plentiful, and they all illustrate how a mild increase of the average temperature can have serious economic consequences.

Lake Mead, the largest freshwater reservoir of the US, providing water and electric power to 40 million people living in the desert, was in the summer of 2022 at 27% of its full capacity (this is the lowest level since 1937 when the lake was filling up for the first time). It took seven years to fill up the lake, and the last year that it was completely full was in the summer of 1999; see Figure 1 [8].

The water level in the lake is nearing the level where the Hoover Dam will stop producing electricity (950 ft) [9]. In 2020, electricity production was about a third of the maximum capability with a full lake.

Moreover, in 2023, Arizona will receive 25% less water from the Colorado river. This means the end of hundreds of desert farms that depend on the irrigation water from the Colorado river. Some will survive on pumped up ground water, but since it

rains very little in the desert (and most rain runs off or evaporates), this water resource will eventually get depleted too. California is a major producer (and exporter) of agricultural products. Without irrigation, the productivity of the whole area will shrink and farmers might have to switch to crops that depend less on water. Food prices will increase, and given the US' role as an exporter, this will have consequences beyond national borders.

The Colorado river is not the only case. The Mississippi river was at its lowest level in a century in October 2022. In some places, the river could be crossed on foot [10], while seawater entered its estuary, threatening the collection of drinking water in Louisiana [11]. The Mississippi is one of the major transportation arteries of the US used to transport crops from the central US. If the megadrought of 2022 becomes the "new normal", this will have serious consequences for the states that depend on the river for transportation, drinking water and irrigation.

The Yangtze River, the largest and most important waterway in China, was 40% lower than normal in August 2022. 84 million people depend on the hydropower generated by the Three Gorges dam on the Yangtze River. Major companies in the area like Toyota and Foxconn and a major vehicle battery maker had to suspend their operations in August 2022 to save energy. This will impact China's GDP in 2022. The lack of precipitation is also threatening crops all over the area [12].

In Europe, the River Po carried 10% of its usual flow in the summer of 2022, and the Po Valley, the food basket of Italy (with 30-40% of the country's total agricultural production), ran out of irrigation water, leading to a substantially lower yield (e.g. 60% less risotto rice).

Extreme drought is not only a problem for food production, but also for livestock farming. As an example, some farms in the US have had to sell cattle because of increasing feed prices, low-quality or insufficient forage [13]. This led to a reduction in the supply of dairy products. Vineyards are also suffering from drought, extreme heat and fires [14].

As mentioned above, major transportation arteries in Europe and the US (the Rhine and Danube rivers and the Mississippi) reached such low levels in 2022 that barges could only be loaded up to 25%, leading to delays and fivefold transportation costs. This has a severe impact on industry (and will cost an estimated 0.2 points of economic growth in Germany).

Rivers also provide cooling water for fossil and nuclear power plants to condense the used steam back to water. When the flow of the river is too low (or when the water gets too hot), the plants have to reduce the power production, or to shut down. In the worst case, this leads to suspended operations in energy-intensive factories too. This is perfect storm: a megadrought leads to less precipitation and hydropower, and problems with the operation of large thermoelectric power plants, all while during a time of peak water and electricity consumption.

Ecological consequences

Global warming does not only impact people and livestock; severe drought obviously also impacts natural ecosystems. Animals cannot survive in the wild without water and food. In places where water disappears, wild animals die. In some parts of Africa, more large animals were killed by the drought and wildfires than by poaching in 2022. This poses a major challenge for nature conservation.

Due to the changing climate, natural ecosystems are changing across the world, and some species will have to adapt in order to survive. In any case, climate change is going to impact biodiversity because the climate is changing faster than the biological capacity of species to adapt to the new situation [15]. Since humans are part of and depend on the natural ecosystem too, some global areas might become too harsh (or too expensive) for us to thrive too.

Climate change and biodiversity are intimately connected: climate change will impact biodiversity, while protecting biodiversity might also slow down climate change. An example is the protection of (tropical) forests and wetlands. Since they

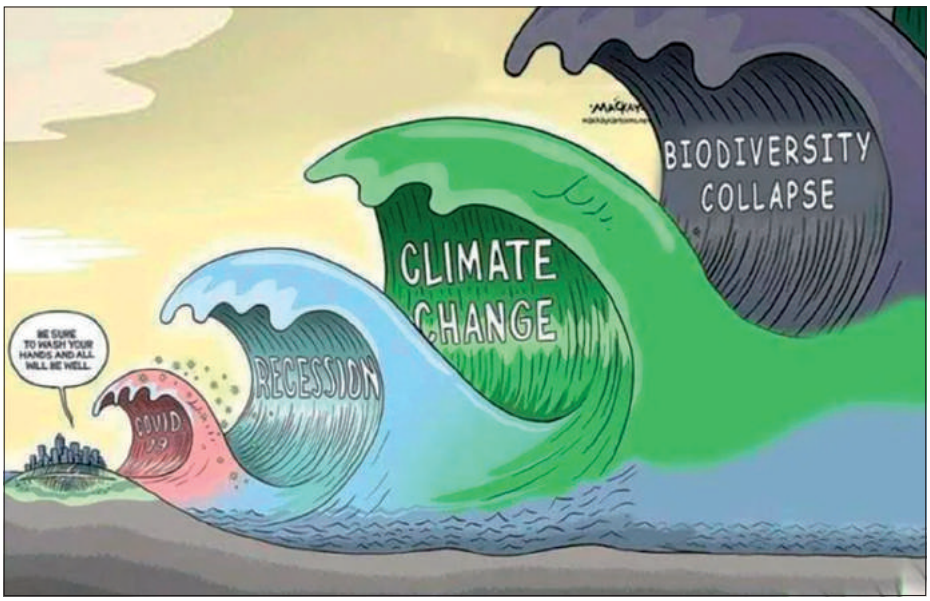


Figure 3: Climate change and loss of biodiversty are connected in a positive feedback loop

	Asia	Europe	N-America	World	Asia/Europe/N-America
Potatoes	178,52	107,69	24,09	359,07	86%
Corn	373,81	365,31	123,94	1160	74%
Rice	676,61	10,32	4,07	756,74	91%
Wheat	347,89	255,02	84,87	760,93	90%

Table 1: Staple food production around the world (2020) [17]

Illegally abandoned tyres in the River Po visible due to drought. Carmagnola, Italy, July 2022. Source: Mike Dot on Adobe Stock

soak up CO_2, they slow down climate change. Should they disappear, huge amounts of CO_2 will be released, further accelerating climate change. Such positive feedback loops are dangerous mechanisms, and once activated, they are extremely hard to stop [16]; see Figure 3. Hence protecting biodiversity is a very important priority in efforts to control climate change.

Given the recent weather evolutions, weather in moderate climate zones seems to be becoming less moderate: longer periods of drought, heavier rainfall and floods, more violent storms, higher temperatures (both on average and at the extremes). The northern hemisphere contains 67% of the land of the planet, and almost 90% of the population. It also produces the staple foods for a large part of the world as shown in Table 1.

If food production decreases by, for example, 20% by the changed climate in the coming years (without taking into account the effects of wars), this will lead to increased food prices. Since the food market is a global market, the increased prices will be felt all over world, including in areas where climate change has a milder impact. In poor areas of the world, this will lead to famine.

This effect will be aggravated by the growing world population which will require 50% more food by 2050 (a considerable part to feed farm animals). Hence, before anything else, climate change is creating a food security challenge for the world. Hunger inevitably leads to social unrest, and political instability. This might destabilize some countries.

Technical solutions for the food security challenge include new crops that are better adapted to the new climate, vertical farms in urban areas, lab meat, and reducing food waste in the food supply chain; lifestyle changes include consuming food with a smaller ecological footprint, like vegetarian alternatives to meat.

The way forward

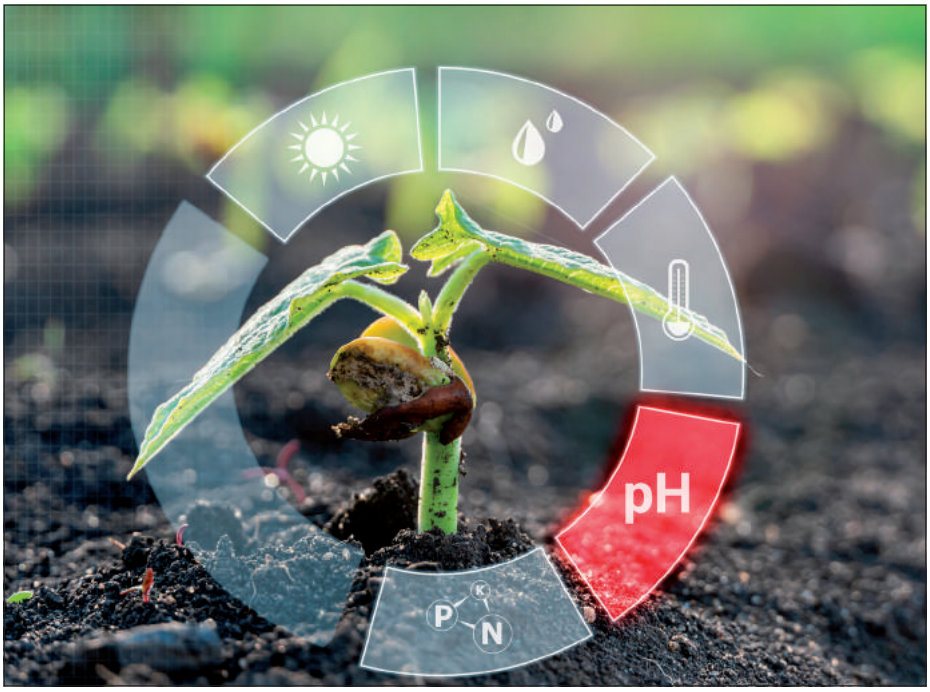
Since we cannot stop climate change in the short term, we will have to adapt. In fact, we will need two types of adaptation: (i) adaptations to decarbonize the world to slow down climate change, and (ii) adaptations to deal with the existing consequences of climate change.

Given the size of the problem, there are no easy fixes to deal with climate change, and business as usual will not suffice. Adaptations to decarbonize the world will require investment, and lifestyle changes. Some recent studies indicate that 2050 for net zero will be too late to limit climate change to 1.5°C, and that the world should try to reach net zero sooner. Many major corporations have announced plans to work towards net zero in 2040.

Digital technologies will play a crucial role in facilitating lifestyle changes and adapting to climate change.

- Decarbonizing the world will require advanced digital technologies to design a new power grid, merge and balance renewable sources of energy, storage systems, distributed power generation, …
- Lifestyle changes will be facilitated by digital technologies too. Video conferences are in some cases a good alternative for in-person business meetings, immersive virtual experiences might in some cases be a low-carbon alternative to travel.
- Finally, there are adaptations to climate change. Advanced sensing technologies can help monitor extreme weather conditions, precision agriculture can help increase the yield of crops, genetically modified crops could be made more resistant to extreme weather conditions, a digital twin of planet earth can help predict the effects of climate change.

These digital systems will be geographically distributed, connected, cognitive, and the size of whole countries. The key digital technology used to design and build them



will be the IoT, edge and cloud computing. Europe should invest in these technologies in order to secure itself a place in the global market of climate change adaptation technologies.

Conclusion

Climate change is real, and it will not go away, so we will have to adapt. Digital technologies are one of the key technologies that will help humanity to adapt to the new climate normal. The biggest challenge is to decarbonize the world as fast as we can. This will require huge investment in power savings, and new power-generation and power-transportation systems. All these solutions will require lots of for digital technologies based on the IoT, edge and cloud computing. Europe should invest in these technologies to order to stay at the forefront of climate change adaptation technologies.

References

[1] U. Grober, "Deep roots – A conceptual history of 'sustainable development' (Nachhaltigkeit)," 2007. [Online]. Available: https://bibliothek.wzb.eu/pdf/2007/p07-002.pdf. [Accessed 27 November 2022].

[2] D.H. Meadows et al. , "The Limits to Growth," Club of Rome, 2 March 1972. [Online]. Available: https://www.clubofrome.org/publication/the-limits-to-growth/.

[3] "Triple bottom line," The Economist, 17 November 2008. [Online]. Available: https://www.economist.com/news/2009/11/17/triple-bottom-line. [Accessed 26 November 2022].

[4] "The 17 Goals," United Nations, [Online]. Available: https://sdgs.un.org/goals . [Accessed 27 November 2022].

[5] "COP27 ends with announcement of historic loss and damage fund," United Nations, 22 November 2022. [Online]. Available: https://www.unep.org/news-and-stories/story/cop27-ends-announcement-historic-loss-and-damage-fund. [Accessed 27 November 2022].

[6] T. Grant, "Millions of people in Midwest to experience 'extreme heat belt' by 2053: Report," ABC News, 16 August 2022. [Online]. Available: https://abcnews.go.com/US/millions-people-midwest-experience-extreme-heat-belt-2053/story?id=88410076. [Accessed 27 November 2022].

[7] "The 6th National Climate Risk Assessment: Hazardous Heat," First Street Foundation, 15 August 2022. [Online]. Available: https://report.firststreet.org/heat. [Accessed 27 November 2022].

[8] "Lake Mead Keeps Dropping," NASA , [Online]. Available: https://earthobservatory.nasa.gov/images/150111/lake-mead-keeps-dropping. [Accessed 27 November 2022].

[9] R. Ramirez, "The West's historic drought is threatening hydropower at Hoover Dam," CNN, 16 August 2022. [Online]. Available: https://edition.cnn.com/2022/08/16/us/hoover-dam-hydropower-drought-climate/index.html. [Accessed 27 November 2022].

[10] A. Chinchar, B. Miller and D. Alsup, "The mighty Mississippi is so low, people are walking to a unique rock formation rarely accessible by foot," CNN, 16 October 2022. [Online]. Available: https://edition.cnn.com/2022/10/15/weather/mississippi-river-low-water-tower-rock-climate/index.html. [Accessed 27 November 2022].

[11] R. Rojas, "As Drought Drops Water Level in the Mississippi, Shipwrecks Surface and Worries Rise," The New York Times, 3 November 2022. [Online]. Available: https://www.nytimes.com/2022/11/03/us/mississippi-river-drought.html. [Accessed 27 November 2022].

[12] H. Davidson, "China drought causes Yangtze to dry up, sparking shortage of hydropower," The Guardian, 22 August 2022. [Online]. Available: https://www.theguardian.com/world/2022/aug/22/china-drought-causes-yangtze-river-to-dry-up-sparking-shortage-of-hydropower. [Accessed 27 November 2022].

[13] "New AFBF Survey Shows Drought's Increasing Toll on Farmers and Ranchers," 14 August 2022. [Online]. Available: https://www.fb.org/market-intel/new-afbf-survey-shows-droughts-increasing-toll-on-farmers-and-ranchers. [Accessed 27 November 2022].

[14] M. Scott, "Hard-hit by climate change, winemakers turn to sustainability to ride the storms," Reuters, 14 September 2022. [Online]. Available: https://www.reuters.com/business/sustainable-business/hard-hit-by-climate-change-winemakers-turn-sustainability-ride-storms-2022-09-14/. [Accessed 27 November 2022].

[15] I. Quintero et al. , "Rates of projected climate change dramatically exceed past rates of climatic niche evolution among vertebrate species," Ecology Letters, vol. 16, no. 8, pp. 1095-1103, https://doi.org/10.1111/ele.12144, 2013.

[16] I. B. Ruiz, "Five worst climate feedback loops," Deutsche Welle, 5 April 2018. [Online]. Available: https://www.dw.com/en/when-nature-harms-itself-five-scary-climate-feedback-loops/a-43649814. [Accessed 27 November 2022].

[17] Govind Bhutada, "Mapped: Food Production Around the World," 12 2021. [Online]. Available: https://www.visualcapitalist.com/cp/mapped-food-production-around-the-world/.

Koen De Bosschere is a professor in the electronics department of Ghent University, Ghent, Belgium.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2023.

Cite as: K. De Bosschere. Climate change is here to stay, so we'd better prepare for it. In M. Duranton et al., editors, HiPEAC Vision 2023, pages 196-200, Jan 2023.

DOI: 10.5281/zenodo.7461991

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

© HiPEAC 2023

Sustainability is a multi-faceted problem with ecological, economical, and societal aspects. We need a holistic approach to tackle it.

Why sustainability is crucial for Europe's future

By PATRICK BLOUET

Sustainability is a major challenge for society, and the need for it has been made clear by recent events linked to climate change, geopolitical tensions and war. However, sustainability is more than the capacity to maintain our living standards. Any approach to sustainability should consider three important aspects: (i) a systemic approach, (ii) the scalability of the solutions, and (iii) the need to reconcile society, the environment and the economy.

A systemic approach means that any solutions considered must result in improvements for the world as a whole. It is not enough to work on something that delivers local improvements while at the same time making the situation worse at the global level. For instance, developing energy transition solutions that reduce the CO_2 emissions in one place while at the same time increasing emissions due to intense mining for critical materials elsewhere is not a globally sustainable solution.

Linked to this systemic approach is the scalability of the solution; it is important to consider whether a solution can be scaled up. It is good to have local solutions, but ideally these solutions should be able to scale up to national, European or global level. Solutions that, for example, increase the demand for particular minerals far beyond what the planet is able to provide are not globally scalable.

It is also important that all potential solutions are viable in a given economic model. Clearly, our current economic model is not designed to support robust environmental responsibility compatible with planetary limits. New economic models will have to be deployed to align the needs of the human population, environmental limits and a viable economy.

Finally, solutions should also be palatable to the population; for example, lifestyle changes are very hard to sell. Without addressing all these aspects, it is difficult to make a clear assessment of the potential of any given solution to the sustainability issue.

Key insights

- Sustainability solutions should be taken in a global context. As an example, defossilizing European industry should not lead to increased emissions due to mining essential minerals elsewhere.
- Sustainability solutions should be scalable to a global level to have impact. Electric cars are beneficial for CO_2 emissions, but they need such large quantities of rare minerals that we won't be able to transition to 100% electric mobility over the next decade.
- Sustainability solutions should also consider societal and economic constraints. Solutions that require lifestyle changes are hard to sell, as are solutions that lead to a shrinking economy.
- Sustainability solutions must clearly express how long they will last for. In parallel, more durable products and recycling materials should be developed, thereby creating a virtuous circular economy cycle.

Key recommendations

- Always look at solutions from a multi-dimensional point of view, including features, performance and environmental aspects (resource usage, the impact on the climate, the impact on biodiversity, etc.)
- Consider end-of-life management at the very beginning of the design cycle for information and communication technology (ICT) products.
- Extending the lifetime of hardware is very dependent on software. Ensure that software is flexible enough to support a 'second life' for hardware (for example through auto diagnosis, auto testing, and a modular structure).
- Develop design frameworks that integrate full product lifecycle management constraints.
- Develop technologies that reduce the use of raw materials and a recycling process that allows true sustainability with very low environmental impact.
- Adapt the manufacturing processes of semiconductor and electronic systems to allow massive usage of recycled materials and strengthening European sovereignty and enabling a real circular economy.
- Revisit business models such that environmental constraints are no longer seen as a drawback but as a huge opportunity for innovation. Create economic value from extending the lifetime of electronics.
- Set-up a strong European training program to develop skills supporting the repair, reuse and the other topics of the 6R framework supporting electronic lifetime extension.

Defining sustainability

Sustainability is a word that is now used everywhere and is a concern for almost everybody. This is perfectly understandable at a time when the continuity of our society is under threat from environmental pressures, geopolitical tensions, and even a war in Europe.

To understand sustainability, it is important to start from its definition. An internet search returns numerous definitions of sustainability, as illustrated by a few examples below:

> *The quality of being able to continue over a period of time*
> https://dictionary.cambridge.org/dictionary/english/sustainability
>
> *Sustainability consists of fulfilling the needs of current generations without compromising the needs of future generations, while ensuring a balance between economic growth, environmental care and social well-being*
> https://www.becas-santander.com/en/blog/what-is-sustainability.html
>
> *Sustainability is [the] ability to maintain or support a process over time. Sustainability is often broken into three core concepts: economic, environmental, and social.*
> https://www.investopedia.com/terms/s/sustainability.asp
>
> *Development must meet the needs of the present without compromising the ability of future generations to meet their own needs.*
> https://ec.europa.eu/environment/sustainable-development/

What all these definitions have in common is the idea of a process which is capable of lasting over a certain period. However, it is important to note that the timescale is not precisely defined; in fact, sustainability is never associated with a duration (such as 10 years, 100 years, or even eternity). There is sometimes a target date to achieve 'sustainability', like 2030 or 2050, but never the idea that the solutions could last forever, or that they could create a long-lasting situation for future generations.

The lack of the time dimension in sustainability creates a situation where we don't know what we are talking about. Actions taken by industry or government give a false sense of security, as we don't know how long these solutions will last. This has a perverse impact, as decisions are often taken for short-term beneficial effect with potentially very detrimental consequences in the medium or long term.

It is therefore mandatory to integrate the time dimension in sustainability, clearly indicating the period during which the proposed solution is sustainable. It would help greatly to identify the areas where the greatest efforts are needed.

Overconsumption and biodiversity

To seriously address sustainability, there are two major issues which need to be considered very carefully:

1. Overconsumption of natural resources

This refers either to the overconsumption of finite, non-renewable resources or to renewable resources that are consumed faster than they can regenerate. In both cases, sustainability is limited to the consumption of the initial reserve of resources. This is the key structuring hypothesis in a finite world.

Many different materials are used in our most popular electronic devices. Moreover, the more advanced the silicon technology used to power these systems, the more new materials are needed to build chips and other elements. Despite the increase in the use of raw materials, the level of recycling is still very low, as shown in Figure 1 from a European Union (EU) survey about critical materials in Europe [1].

In 2021, the amount of waste electronic and electrical equipment (WEEE) worldwide totalled an estimated 57.4 Mt – greater than the weight of the Great Wall of China, Earth's heaviest artificial object. Of this waste, only a small proportion (less than 20%) is properly documented and recycled. Figure 2 shows that there is a huge potential for urban mining.

2. Massive extermination of biodiversity

Regarding biodiversity, electronics has a negative impact in two main aspects:

- *Extraction of materials:* This is mainly due to activities related to mining and refinement of raw materials. In some cases, the concentration of material in ore is very low (for instance 18 to 24 ppm for rare earth materials which means 1000 tonnes of ore need to be extracted to generate between 18 to 24 kg of noble material). This has consequences of impacting a large ground area to extract all these materials. To make matters worse, very often the refinement processes are very water and energy consuming, and polluting. The concentration of these materials per tonne of electronic waste is much higher, and urban mining takes less

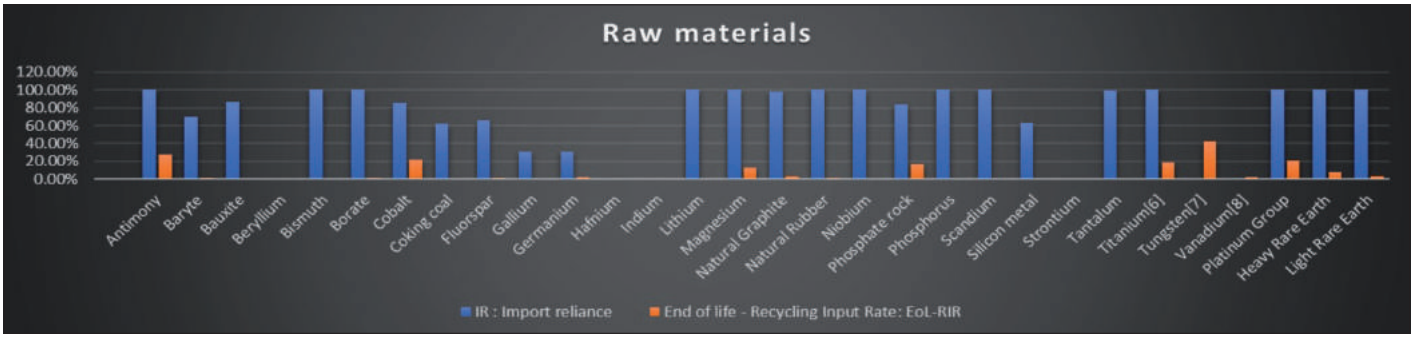


Figure 1: Critical raw materials in Europe (IR = (Import – Export) / (Domestic production + Import – Export) EoL-RIR is the percentage of overall demand that can be satisfied through secondary raw materials)

ground area, and can be done in a more controlled environment with less pollution in the environment.

- *Electricity generation to power the electronic devices:* The transition to green energy has a major impact on the ground occupation of electricity generation systems. 1 GW of nuclear power requires around 0.1 km^2 while 1 GW of solar power requires around 10 km^2 and 1 GW of wind power requires between 150 km^2 to 200 km^2 [2] depending on the size of each turbine. This shows that the impact on the ground surface, and therefore on biodiversity, is much more important than just that linked to mining activity for mineral extraction.

Figure 2: Key materials in use in different electronic appliances

Rethinking electronics

To improve the sustainability of the electronic and semiconductor industry, two main dimensions must be investigated:

- Increase product lifetimes.
- Improve the level of recycling of elementary materials through a true circular industry.

The **6R** Framework (**R**educe, **R**ecycle, **R**eliable, Repair, **R**euse, **R**efurbish, (see Table 1) provides a scheme to address both dimensions. All aspects linked to product durability and end-of-life management must be taken into account in the design phase. Trade-offs must be made not only between performance, features, and cost but also – and maybe primarily – with environmental constraints. We are at the beginning of a paradigm shift where the algorithm has to be changed from: "What does technology allow me to do?" followed by: "Try to reduce environmental impact" to: "What are the environmental constraints?" followed by: "What can be done within these constraints?"

The 6R framework mainly covers the design and end-of-life of the product, as detailed in Table 1.

A systemic approach

Problems linked to the environment are systemic problems. This means that they encompass various domains, each with very specific dynamics and constraints. Very often, climate change is presented as the main problem to fix. While it is fundamental, plenty of others – which are tightly linked to climate change – also need to be fixed, such as pollution, water use, materials overconsumption, overpopulation, food and agriculture, biodiversity and many others.

Implementing a solution in one domain must take into account the impact in other domains to ensure that the local optimization proposed in one place doesn't create new problems elsewhere. The overall situation should not be worse than the local problem we want to solve.

In the example of Figure 3, it is clear that focusing only on CO_2 emissions could give a very positive view of the situation. But looking at energy consumption in detail, the diagram shows that the situation is deteriorating, with all the consequent negative impacts this has on the environment, beyond net CO_2 emission. The same can be said from carbon extraction systems. They extract CO_2 from the air, but in the process, they also consume significant amounts of energy.

Solving systemic problems requires aggregated indicators which can give an objective view of the evolution of the situation to ensure that adaptations are always going in the right direction. The most difficult part of this approach is creating models which are representative enough of the various domains which need to be addressed, and then feeding these models with relevant data.

Scalability of solutions

Even though one solution seldom fits all, it is important to consider the extent

Framework component	Description	Examples
Reliability	This is the first step toward product lifetime extension. The time of programmed obsolescence is over. During the design process reliability must be given the same importance as performance, features, cost, …	•Avoid operating points of systems that are too close to the operational limits of the technologies (power, thermal, ...). •Select a technology with better ageing performance.
Reduce	This is a different approach which consists of doing the same as before with fewer materials, with materials which will be easier to manage in the end-of-life of the product, or with new manufacturing paradigms.	•Make thinner PCB with less epoxy resin and thinner copper tracks. •Use alternative materials to PCB (biodegradable plastics, paper) with bio-sourced conductive inks. •Move from a subtractive to an additive manufacturing process.
Recycle	Given the increasing use of rare materials, recycling is becoming mandatory to improve sustainability, although it should be remembered that this will be only for a limited period as the recycling rate is never 100%. Once again, recycling must be addressed during the design phase in making decisions which will try to reduce as much as possible the number of materials and their quantity, while selecting materials which have the highest potential for reuse.	•Improve the collection rate of electronic devices. •Improve the recycling rate of metals. •Increase the reuse of recycled material in industrial processes in the electronics and semiconductor industry. •Reduce the environmental impact of recycling processes.
Repair	This is a very important step in product lifetime extension. Very often a product is thrown away just because a simple broken element cannot be replaced to make the system work again. The ability to repair must be addressed during the complete design process of a product, from the semiconductor up to the external product packaging. It also needs to address the supply chain to ensure the availability of spare parts, as well as the creation of a repair network with the right skills.	•Design systems that are easy to put together and take apart. •Make spare parts available over a long period. •Allow easy failure diagnostics. •Allow easy retesting/recharacterization. •Develop an adequate network of repairers with the right skills. •Develop a repairability index, as already implemented in France, at European level.
Reuse	This is the next step after repair, meaning that if it is no longer possible to repair a product, at least part of the product could still be repurposed as spare parts for other systems. In these conditions it is important to be able to evaluate the remaining lifetime and health of the recovered parts. As with repairing, a specific reverse supply chain is needed to be able to collect product and provide spare parts in the repair network.	•In addition to the properties required for repairing, reuse also needs additional properties linked to modularity.
Refurbish	Refurbishing is the step which allows a second life of a product. This means that a component or subsystem is not good enough to be used in its initial application but can perfectly fit in a new, less demanding one in term of features or performance.	•All the characteristics already defined for repair and reuse apply for refurbishing. •Make sure that unused functions do not create problems for the remaining ones and are securely isolated. •Allow partial diagnosis, retesting and recharacterization of a system.

Table 1: 6R Framework

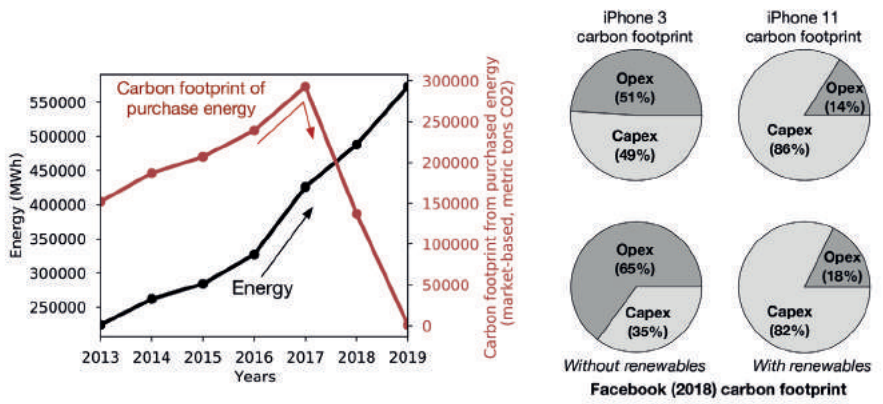


Fig. 2. Carbon footprint depends on more than just energy consumption (left). Although the energy consumption of Facebook's Prineville data center increased between 2013 and 2019, its net operational carbon output decreased because of renewable-energy purchases. The carbon-emission breakdown has shifted from primarily opex-related activities to overwhelmingly capex-related activities (right). The top two pie charts show the breakdown for the iPhone 3 (2008) versus the iPhone 11 (2019); the bottom two show the breakdown for Facebook's data centers with and without renewable energy.

Figure 3: Carbon footprint and electricity consumption [3]

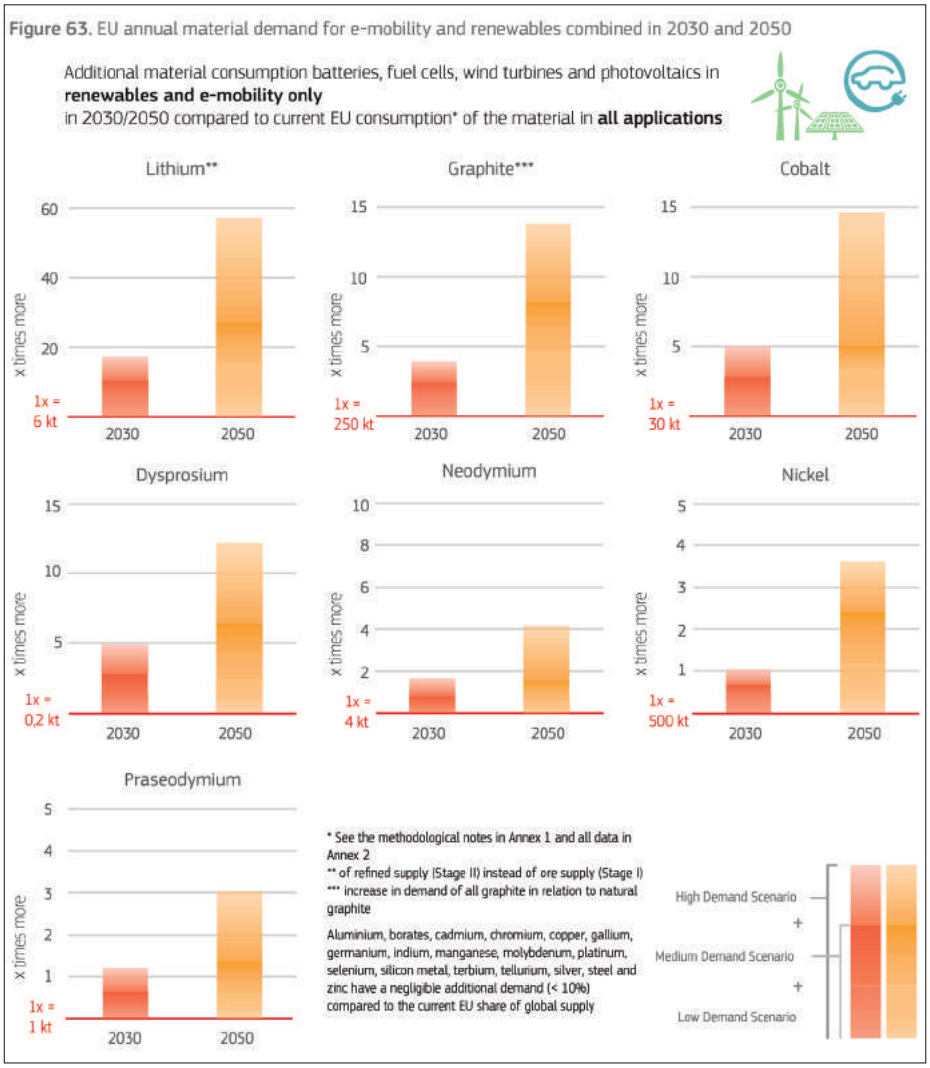


Figure 4: Additional materials needed to support electric mobility and the energy transition in Europe [4]

to which a solution can scale up without resulting in more disadvantages than benefits. This aspect is very tightly linked to the systemic approach, as the impact of a problem and its solutions must be examined together. Too often, a proposed fix is seized upon as the ultimate solution, while considering its constraints in depth shows this is far to be the case.

To give an example, today, electric mobility is seen as the solution to CO_2 reduction in transportation. This may be true if you only consider CO_2 emissions; overall, electric vehicles generate lower emissions, mainly during the lifetime of the vehicle.

Unfortunately, however, when looking at the need in terms of raw materials to support the massive deployment of electric cars, coupled with the need to support the energy transition and electric mobility at the European level, scale-up is not possible; see Figure 4. In order to satisfy European demand, the pressure on critical materials would be so great that there is a high probability it will not happen; at the very least, it would provoke fierce competition for resources with other parts of the world.

This makes it very clear that European sovereignty is strongly dependent on Europe's capacity for recycling materials and subsystems at a very high rate, while at the same time increasing the use of recycled materials in our industrial processes. It is also important to notice that recycling processes must be designed with a low environmental impact in mind and without creating extra dependencies on other materials or components.

This strategy must be applied to all industrial domains where Europe's dependency on raw materials is very important. It can only be achieved if product end-of-life and recycling processes are defined simultaneously during the design phase of a new product.

Reconcile society, the environment and the economy

Sustainability is always linked to society and the economy. It is extremely difficult to decorrelate the environment from

the economy, which makes the concept of green growth, green economy and, more specifically for Europe, the European Green Deal very challenging.

Analysing the situation using Kaya's Equation or directly from raw data about gross domestic product (GDP) (in 2016 constant dollars) and energy consumption (in million tonnes of oil equivalent, or Mtoe) to evaluate the correlation between GDP and energy consumption, the conclusion is almost always the same: technology and process improvements are always undone by increases in population size and living standards, while worldwide GDP is fully correlated to the amount of energy we consume every year, see Figure 5.

Unfortunately, this pattern has not changed over the past 60 years, which creates a real challenge for the European Green Deal to really decorrelate economic growth from environmental impact while continuing to offer an overall high standard of living to the European population.

To account for the complexity of the situation, some new economic models have been proposed. One of them is "Doughnut economics" invented and promoted by Kate Raworth; see Figure 6. In this model, the three dimensions – society, the environment and the economy – are integrated and trade-offs must be done while simultaneously managing all three. The safe area is the doughnut in green, where everything reaches an equilibrium. Any overshoot beyond the doughnut indicates pressure which is not sustainable for the environment. Any red sector inside the doughnut indicates the potential risk of social instability. The figure shows the 2017 estimation for different parameters [5], and it clearly shows we are already beyond the limits for the environment but also regarding the social stability of our modern society.

Moving from the situation today to this kind of model is not at all easy, as the changes needed are enormous. While the model may be interesting, it is far from clear that the population will accept such a new model.

Perhaps a first step towards such a new way of thinking in the technology realm could be to define first our environmental limitations and then define what could be done within these limits in terms of technology. It is too often the case that a particular technology is developed because engineers can do it, without considering the overall consequences. Changing the algorithm in our development, taking the limitations of our world as a starting point, could result in developing the things we really need and producing an environmentally acceptable solution to a real problem.

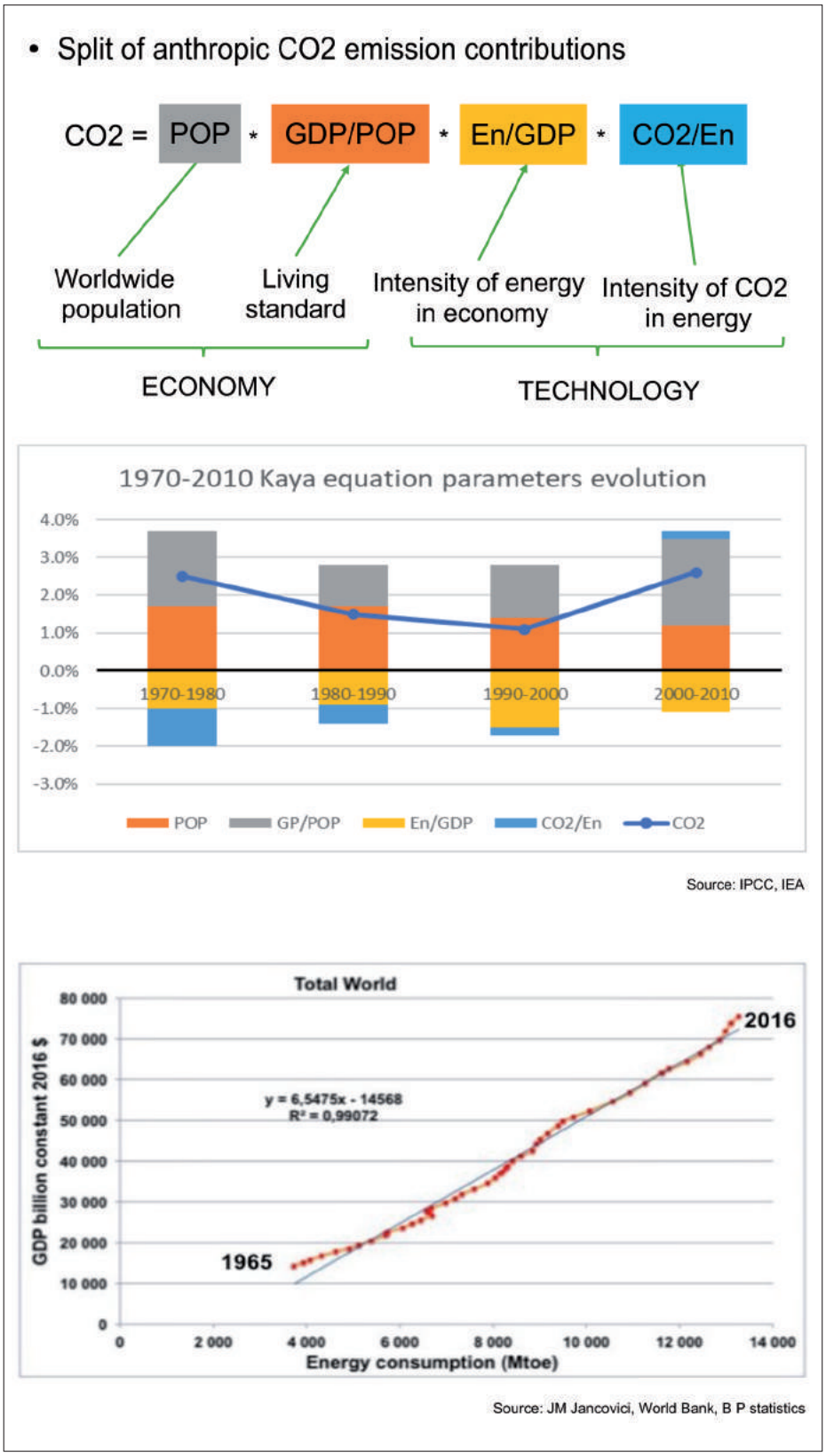


Figure 5: split of CO_2 contribution based on Kaya's equation. Link of energy use and worldwide GDP.

Conclusion

Sustainability is a major issue for Europe and, more generally, at the worldwide level. Better transparency on the period for which a solution is sustainable is needed. Scale-up and a systemic approach must be addressed to allow solutions which could be deployed extensively and deliver global environmental benefits. Nevertheless, there are many opportunities for innovation to extend product lifetime and make devices more compatible with a circular economy and lower environmental impact.

References

[1] "Critical Raw Materials Resilience: Charting a Path towards greater Security and Sustainability", EU 2020 COM 474 2020, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020DC0474

[2] P. Blouet, J.-M. Jancovici, C. Pitrou, "Peut-on améliorer les éoliennes?", 2020. https://lederniercarbone.org/efficacite-eoliennes/

[3] Udit Gupta, Young Geun Kim, Sylvia Lee, Jordan Tse, Hsien-Hsin S. Lee, Gu-Yeon Wei, David Brooks, Carole-Jean Wu, "Chasing Carbon: The Elusive Environmental Footprint of Computing", IEEE Micro, vol.42, no.4, pp.37-47, 2022. DOI: 10.1109/MM.2022.3163226, https://ieeexplore.ieee.org/document/9744492.

[4] "Critical Raw Materials for Strategic Technologies and Sectors in the EU: A Foresight Study", European Union, 2020. DOI: 10.2873/58081. https://rmis.jrc.ec.europa.eu/uploads/CRMs_for_Strategic_Technologies_and_Sectors_in_the_EU_2020.pdf

[5] Raworth, Kate. "Doughnut Economics: Seven Ways to Think Like a 21st-century Economist." London: Random House, 2017. https://www.kateraworth.com/doughnut/

Patrick Blouet is a research and development program manager at STMicroelectronics with responsibility for European projects. For more than 25 years, he has been interested in environmental issues as a personal concern.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2023.

Cite as: P. Blouet. Why sustainability is crucial for Europe's future. In M. Duranton et al., editors, HiPEAC Vision 2023, pages 201-208, Jan 2023.

DOI: 10.5281/zenodo.7461997

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

© HiPEAC 2023

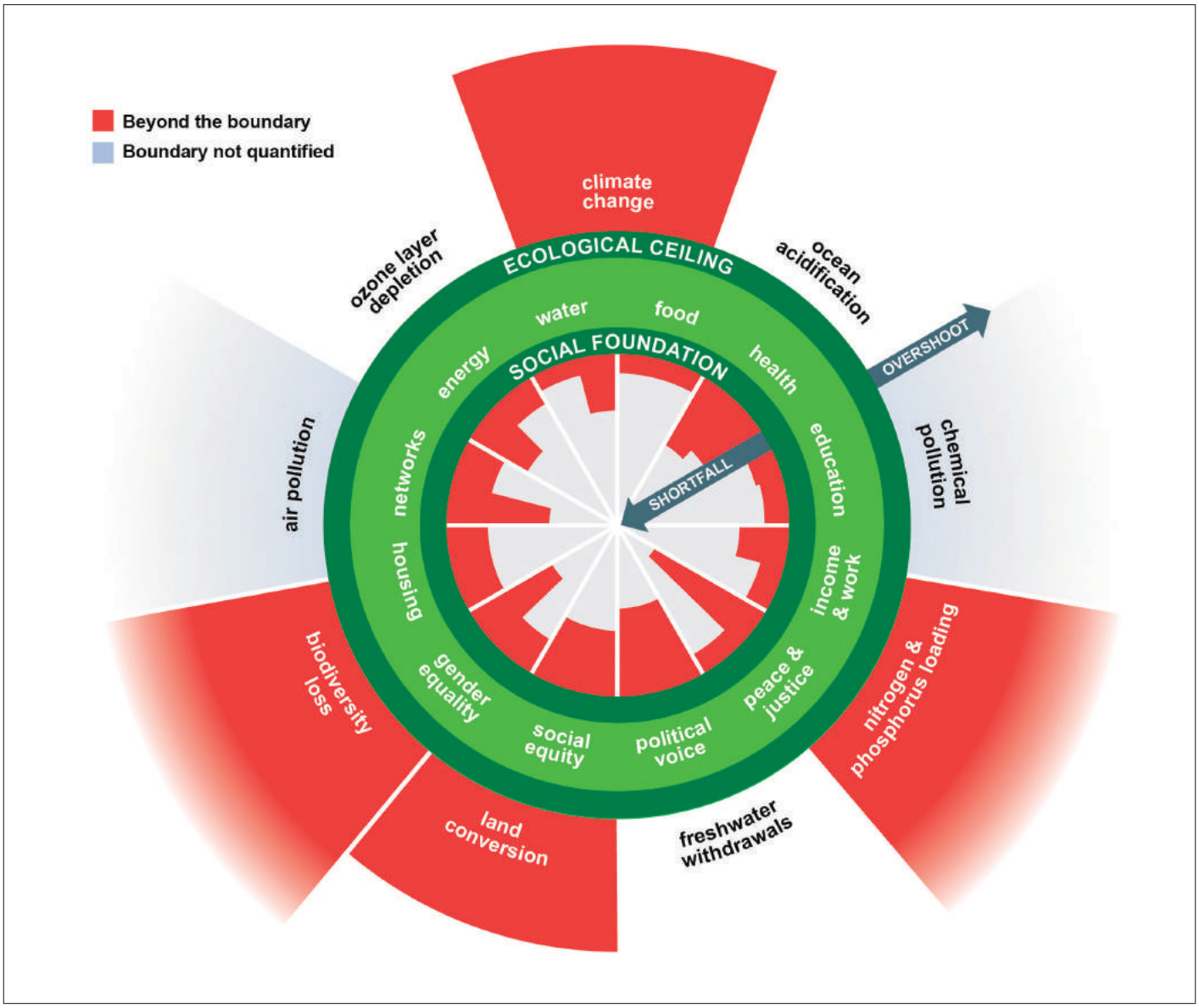


Figure 6: Doughnut economic values for 2017

Information and communication technologies (ICT) are the fuel of the digital society. This article reviews the big challenges to make ICT (more) circular.

Towards circular ICT: From materials to components

By THOMAS ERNST and JEAN-PIERRE RASKIN

Europe is facing one of the biggest challenges of its history: that of preserving a viable environment for the decades to come. On human consciousness of its environmental impact, Albert Allen Bartlett [1] said: "The greatest shortcoming of the human race is our inability to understand the exponential function". ICT was, and still is, driven by exponential functions (Moore's law, ops/Watt, data exchanges volume worldwide,…). However, each exponential has its sustainability limits which are often reached sooner than expected.

In ICT history to date, when those limits are reached, they are overcome by a technology breakthrough. By way of example, there was the CMOS (complementary metal-oxide-semiconductor) technology, a product of laboratory curiosity of the 1980s when bipolar transistor power consumption was too high. New integrated system (multi-core processor) and device (SOI, FinFET,…) architectures entered the market when energy consumption limits were reached by central processing units (CPUs). Today, CMOS is still the cheapest and most reliably power-efficient technology for the development and growth of the internet of things (IoT).

Awareness of the rapid proliferation of wirelessly connected objects around us gives rise to questions and fears among the public. There is a need for more transparency on the environmental footprint of ICT and a need to establish an ambitious roadmap, at research and industrial levels, for moving into a virtuous cycle of eco-innovation.

In this article, we propose pathways to maintain or even reduce global (exponential) use of energy and materials for ICT.

Key insights

- ICT is energy demanding and contributes to approximately 4% of greenhouse gas emissions. Approximately half of the emissions associated with ICT are generated at the production stage.
- ICT is based on a large number of critical materials and Europe imports nearly all of them.
- Recycling of ICT is technically difficult, consumes energy, generates pollution, and is not currently profitable because of the low price of primary raw materials. Only 10-15% of electronics waste is recycled.
- A product-centric economy favours the "produce, consume and waste" vision and thus leads to planned obsolescence of digital objects and infrastructure. A service-centred model would be more in line with the principles of the circular economy.
- The "rebound effect" cancels out the efficiency gains of any ICT system.
- It is time to design differently, to design within limits.

Key recommendations

- Expand the lifetime of devices through better design, both by enhancing the intrinsic durability of components and by adopting a modular approach in which the replacement of faulty or obsolete components is made easy.
- Rethink software and applications to make them less resource consuming (data, energy, hardware) and compatible with still usable hardware. Make software reconfigurable and evolutive.
- Integrate circular economy concepts, eco-design and full life cycle assessment (including fabrication, use and end-of-first-life phases) at the early stage of research and development of new ICT technologies.
- Avoid toxic and substitute critical materials to protect environmental and human health, as well as facilitating the recycling of materials at the end of the device or component life. Use secondary (recycled) materials and use more renewable energy, and develop bio-based and greener chemistry in the ICT industry. This requires enhanced basic and applied research into materials and fabrication processes.
- Draft guidelines and legal policies to enforce supply-chain transparency for all imported and non-imported materials and components.
- Develop new economic models which lead to the decoupling of economic profitability from the depletion of natural resources, for instance the economy of the functionality applied to ICT.

An increasing impact of ICT growth on natural resources

Behind our screens, we have the impression of living in a dematerialized world, where everything is fast, clean, and reconfigurable. The reality is different. The digital society we live in has never been so energy- and material-intensive and this is leading to increasing pressure on natural resources, ecosystems and the climate. Today, ICT is a growing economic activity, similarly to transport, energy production, manufacturing and agriculture. Information and communication technologies consume around 5% of the world's electricity production and are responsible for around 4% of greenhouse gases, a level equivalent to air transport. Our smartphone contains electronic circuits that require more than sixty different materials. We are talking about virtually every element on Mendeleev's periodic table except radioactive materials. At the end of their life, the recycling rate of electronic equipment is very low (less than 15%).

It is extremely difficult to separate the sixty materials that make up electronic circuits.

Umicore, one of the most advanced companies in the field of electronic materials recycling, manages to extract 17 elements out of the 60. Material recovery from obsolete equipment is not profitable given the low cost of raw materials imported from countries south of the equator. We do not pay for the environmental and social costs. This leads to a double penalty for the Southern countries: they suffer from environmental pollution (loss of biodiversity, pollution of the air and groundwater, etc.) during the extraction of raw materials, as well as receiving 75% of our electronic waste.

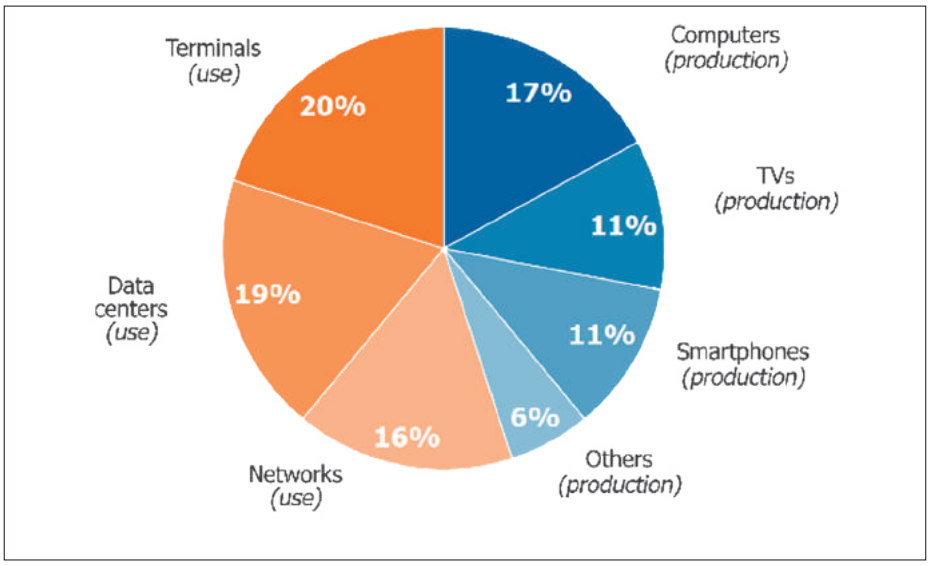


*Figure 1: Estimation of distribution of the energy consumption of digital technologies for **production (45%)** and **use (55%)** Source: The Shift Project, 2018*

Towards low-waste production

Efforts are being made to reduce power consumption and industrial waste during the manufacturing of electronic devices and their components. For example, low temperature processes, reduction of heat dissipation in and from the oven, and reduction and recycling of chemicals and water cleaning, are evaluated for each step of the manufacturing of nanoelectronics devices. However, beyond resource efficiency and recycling waste, the transition towards a cost-effective circular economy needs to be implemented and the general design methodologies of materials for sustainable development proposed by Ashby [19] adapted to specific ICT domains such nanoelectronics.

Some companies are beginning to adopt the lifecycle approach including, for example, in the production and recycling of the ultra-pure water needed [2] for the microelectronics industry. There are active research programmes, both on the part of equipment makers and in research labs, seeking to reuse exhausted gas or fluids within the fab [11], and also to develop much more efficient material deposition techniques.

There is still room for improvement and high potential for reduction in use of raw materials. While progress is continually being made towards greener chemistry, this approach is still far from being environmentally friendly and etches away a significant fraction of materials, as well requiring very high consumption of water for rinsing. The reduction of toxic chemicals such as solvents in the microelectronics foundries must be a priority. Tomorrow, bio-based [12] chemicals and materials may be used to reduce significantly the use of solvents and chemicals in lithography, as shown in preliminary results [13]; more research efforts are required. Bio-based materials are also being investigated for use in packaging in the ICT domain.

More circularity between companies must be facilitated by the coherent installation of industrial activities in the same area, thus developing eco-systems in which the by-products or waste from one industry could be the supply material for another. For instance, the hafnium required in CMOS production is a by-product of ultrapure zirconium used by the nuclear industry and produced mainly in France and the United States [14].

To further reduce the waste of energy and materials, the ICT industry must adopt a holistic approach to developing sustainable products. Several initiatives already exist in the private sector. As an example, we can point out the significant and long-term efforts by STMicroelectronics to evaluate (by lifecycle assessment) the carbon footprint of their microcontrollers [15] and to establish a clear material declaration available online.

To envision a more sustainable future, Europe must:

- Take clear actions to make the ICT supply chain more transparent.
- Make lifecycle assessment and declarations of materials systematic (including for imported products) with shared methodology worldwide.
- Implement a clear and ambitious plan to maximize product lifetime and anticipate its end of life.

This will encourage both research and industry sectors to innovate for the good of everyone.

Looking at minerals

The electronic industry needs a wide variety of minerals. For example, since the 1970s the silicon-based complementary metal-oxide-semiconductor (CMOS) field-effect transistor (FET) has been the mainstream technology for most transistor applications, thus making today's digital economy possible. Over the years, the number of elements exploited in their manufacture has increased greatly (Figure 2), especially since 2000 with the implementation of high-k dielectrics and metal gate stacks which are essential to minimize short-channel effects and gate-leakage current of short transistors (today gate length shorter than 20 nm).

A growing awareness of the limited nature of the supplies of some elements that have specialized and important uses is reflected in the proliferation of terms to describe them and the ores from which they are derived, including 'gateway minerals' and 'critical' and 'endangered' elements. Some countries have adopted policies recognizing the high strategic importance of some of these for their physical and economic security.

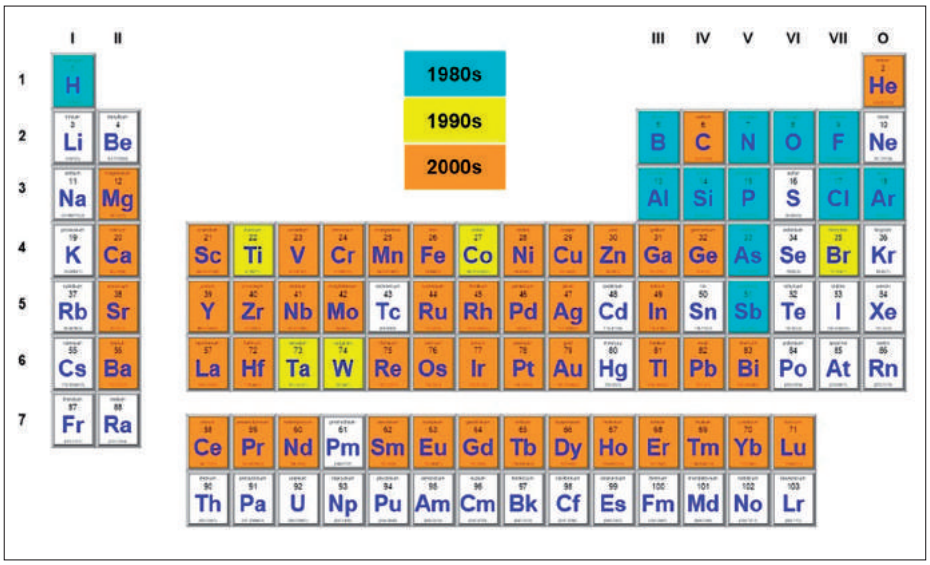


Figure 2: Introduction of elements in the manufacture of CMOS transistors: complementary metal-oxide-semiconductor (CMOS) transistors mainly involved silicon, oxygen, boron, phosphorus and integrated circuit interconnections were made of aluminium in the 1980s. There were relatively few changes in the 1990s, but a large diversity of elements was introduced in early 2000 and many integrated circuit interconnections were switched to copper.

In 2010, 14 elements were considered as critical by the European Commission (EC) according to both their strategic importance for future technology and their scarcity, while in 2017 the number rose to 17, including metals (such as tungsten and some rare earth elements), and also other raw materials such as phosphates, natural graphite and magnesite[3].

Modern devices and systems rely heavily on a high degree of control of material properties and a mastery of manufacturing techniques and, to date, the ICT industry has been remarkably successful in fulfilling these needs.

The manufacturing process is a "top-down" or "subtractive" one based on UV photolithography, etching and many sequential, highly organized and efficient steps of chemical and physical treatment of the chip, layer after layer.

Although silicon is abundant on Earth, the sand used to produce high-purity silicon is already being viewed as a scarce commodity due to its multiple additional uses in concrete, asphalt and glass.

Work is being undertaken on substituting or decreasing the use of toxic, hazardous and critical raw materials. For currently crucial elements such as indium, ruthenium, platinum, gallium, arsenic and gold, new technologies and materials are being investigated with a view to replacing them or drastically limiting their use in some critical devices (e.g. in sensors, memories, optoelectronics and spintronics).

Other examples of the move towards sustainable electronics include the avoidance of lead in micro-components like actuators included in cell phones, use of 2D mono-atomic or ultra-thin atomic-deposition layers to reduce the use of some active materials by a factor of up to 106, and use of silicon-based substrates such as silicon-on-insulator (SOI), instead of materials made from combinations of group III and group V elements, for radiofrequency (RF) technologies.

The scaling down of dimensions of high-tech devices in recent decades and the multiplication of materials in the components – some of them in extremely small quantities of a few micrograms – are leading to new challenges in recycling. The need for large amounts of power and the use of aggressive acids and solvents can make recycling of such electronics impractical.

Approaches to increasing the sustainability of microelectronic devices must include extending their lifetimes through better design, by both enhancing the intrinsic durability of components and adopting a modular approach in which replacement of faulty or obsolete components is made easy.

These approaches can draw on the experience of, for example, some European microelectronics manufacturers and R&D laboratories (e.g. On-Semi, X-FAB, Infineon, ST-Microelectronics, NXP) that are designing or fabricating highly reliable components for automotive, energy management and security applications.

Challenging the 'top-down' or subtractive approach, 'bottom-up', additive fabrication processes may offer a new paradigm for ultra-small circuit design. This could involve 3D printing and consist in layer-on-layer depositing of 2D materials such as graphene, molybdenum sulphide and hexagonal boron nitride [4]. Nanomaterials offer potential alternatives to the widely used indium tin oxide transparent conducting layers. Proposed alternatives include a transparent plastic sheet on which nanocarbons have been deposited, or which contains a loose conducting nanomesh [5]. It is expected that technology based on nanostructures will require far fewer raw materials that any traditional approach. However, recycling nanomaterials at the end of the life of the device will bring new, challenging problems [6].

Steps towards the ambitious goal of achieving the sustainability of the material basis of the digital society will require concerted action covering a range of interlocked approaches. These will address the entire lifecycle of not only the digital devices themselves but also the services that support them, paying attention to their energy and environmental footprints as well as economy and efficiency in the utilization of resources.

Research into substituting critical materials with more abundant ones or which leads to breakthroughs to reduce significantly their use, for instance by localizing active atoms where needed, will open up the path to more responsible electronics.

Waste hierarchies and multiple Rs

The concept of waste hierarchies has become a central feature of the 3R (reduce, reuse, recycle) and circular economy approaches, in which materials recovered across the lifecycle are re-applied at the highest possible level of utility and with minimization of waste or environmental damage. The potential of the approach is evidenced by the virtual disappearance of waste landfilling in, for example, the Netherlands [7].

However, the need to take a comprehensive view of the entire global system of recycling is illustrated by the case of microelectronic devices. Around 50 million tonnes of electronic waste, or e-waste, is thrown away each year. A large amount of e-waste (40% of that produced by Canada, the EU and the United States) is shipped in containers to countries in Africa and Asia, where people manually dismantle the appliances and burn them in the open, producing dangerous levels of hazardous substances which severely impact their health.

Over the last few decades, the number of Rs has grown and at least eleven have been cited in relation to the operation of circular chemistry. Their relevance to the challenge of the sustainability of the material basis of the digital society is outlined in Table 1.

Table 1: Levels of resource hierarchy and relevance to sustainability of the digital society

Level	Level definition	Explanation and application to sustainability of the digital society
1	Reject	Rejecting use of a material or process, whether to conserve scarce resources, reduce energy demand or avoid serious pollution, is considered the highest hierarchy level. A broad approach to sustainability will require questions such as: "do we really want this component, product or process at all?" and "can this substance be replaced by one that is more available, less toxic, or more readily recyclable?" • EU directives restrict the use of harmful substances, such as lead in solders used in microelectronics circuitry. • It has been shown to be possible to replace rare-earth metals used in electronic devices with combinations of more common elements [8]. • Harmful organic solvents can be substituted with more eco-friendly ones [9].
2	Reduce	Overall reduction in the quantities of materials and energy consumed can be achieved with a combination of strategies including product simplification (e.g. fewer materials), miniaturization of the constitutive transistors, interconnects, electrical components, improved battery performance and lower energy consumption, increased product life span and reduced distance for transport. • 3D printing and other additive processes are being introduced in microelectronics to lower the consumption of raw materials and chemicals for etching. • Graphene is being explored [16] as a replacement for transparent conductive films such as indium tin oxide widely used in touchscreen applications. • Efforts have been made to replace or at least to reduce the amounts of 'III-V' materials such as InP, GaAs, GaN with high carrier mobility, which are used in high frequency devices. Thanks to the silicon-on-insulator technology introduced to the market in 2012, more than 95% of RF switches used in mobile devices are now silicon- and not III-V-based. Furthermore, III-V localization by local wafer bonding (for instance CEA-Leti [17]) or local epitaxial growth (for instance imec [18]), only where this material is needed, drastically reduces the amount used.
3	Reuse	Reuse enables an object to remain in service for a long period of time. The library is a very widely known example. • Second-hand shops are a familiar way of enabling reuse of items for their original purpose, including laptops, tablets, smartphones, and many other electronics appliances. • Leasing of ICT equipment (e.g. smartphones provided on a use/return basis as part of a mobile network contract) can also be a channel for reuse, depending on the approach of the leasing company. For example, nearly all the internet provider companies do not sell their WiFi box to clients but rent them. The company stays responsible for the box. They take care of its maintenance and upgrade. Thanks to this move to the economic model of functionality instead of ownership, the lifetime of each electronics component is greatly extended.
4	Redistribute	Redistribution is also concerned with reusing resources, but involves transport over longer distances – for example, to bring a second-hand product to a new market. • Rates of mobile phone usage have risen very steeply in Africa in recent years, with Nigeria providing the biggest African market for second-hand smartphones.
5	Repair	Repairing and maintenance are generally the least resource-intensive solutions to extend the life of devices that are damaged or cease to perform effectively. Unfortunately, however, the way that most ICT devices are manufactured does not facilitate their repair. Moreover, access to spare parts is often impossible or only guaranteed for a short period of time. Once a device needs repair, critical factors are a good diagnosis of the problem and easily removable elements. There is growing interest from both consumers and some producers in repair of ICT devices. • Citizens' initiatives such as 'repair cafés' are flourishing in many countries, providing locations where specialist tools are available and expert advice may be offered by volunteers. Thanks to the internet, there are also several websites which bring people together to form a community for helping each other repair things, for example iFixit. • Consumer pressure for the redesign of electronic products to promote repair rather than disposal of broken items has been growing and has been reflected in the EU's circular economy approach. • Fairphone, the world's first ethical modular smartphone on the market, has been designed for a high level of reparability. Most of the spare parts can be ordered online and the product webpage has many tutorials to assist consumers to repair or upgrade their device.

6	Refurbish	For electronic materials, refurbishment means combining repaired and redistributed products. It may also involve updating a product to current standards. Availability of components and accessibility for repair and replacement are critical. • Reconditioned cellular phones are sometimes returned to the market by companies that initially produced them and recovered them after a period of use. In 2019, Fairphone launched its programme "Refurbished phones give valuable resources a new life". These phones have the same high-quality standards as a brand new Fairphone and the same two-year warranty – but a lower price (half the price of a brand new one).
7	Repurpose	Repurposing is the updating or adaptation of a product such that it can be used to serve a new function or within another context. • There are initiatives such as Puzzlephone from which, at the end of a smartphone life, parts can be used in a new context or application such as in a computer. The repurposing of an object will be favoured if retrieval of its functional parts is possible.
8	Remanufacture	Remanufacture is generally a more thorough process of disassembling a product, replacing worn and broken parts with new ones and reassembling it. • An assessment of remanufacturing of end-of-life computers identified the potential to enhance resource conservation and prevent natural resource degradation, and that remanufactured computers could be technically, environmentally, economically and socially feasible if there is an adequate supply of quality cores, involvement of highly skilled workers, incorporation of a standardization process, and the use of advanced machinery.
9	Recycle	As well as the simple reuse of materials, recycling also encompasses the recovery of component materials to act as a feedstock for new material or device production. • Electronic scrap comprises a highly complex and heterogeneous group of materials. Only a few precious elements present in integrated circuits are recovered by current recycling techniques because they require a large amount of energy and the use of highly polluting chemicals, while freshly mined minerals are relatively cheap. However, Umicore refines e-scrap containing precious metals such as silver, gold, palladium and copper. • The problem of flame-retardant additives needs to be solved in order to recover plastic components of e-waste. • Better technology is needed to improve the purity of recycled metals such as noble metals used in digital equipment. • More sustainable chemical and biological processes may be applicable to both synthesis and recycling of organic components such as polymers.
10	Recover	Recovery involves retrieving the lowest forms of energy or feedstock for energy production from a material. The actual material is broken down and cannot be recycled further. • There is a huge number of waste-to-energy plants in the world. Burning waste and particularly e-waste produces, in addition to CO2, toxic smoke and fine particles which are difficult to prevent from being released into the atmosphere. The efficiency of these waste-to-energy plants and their impacts on human health and the environment remain challenging. • Production of hydrogen as a fuel from printed circuit boards by steam gasification has been examined [10].
11	Return	Most hierarchies call this landfill. However, this stage is more than that – it encompasses the return (sometimes in chemically modified form) not only of solids, but also liquids and gases, back to the environment after use. • More than 70% of e-waste makes its way into landfill, leading to potentially dangerous contamination of groundwater.

It is emphasized that attention to the Rs alone will not solve the problems of waste and resource depletion in the next couple of decades, but will help to extend the timeframe of a transition period during which optimal solutions to sustainable sourcing, production, use and recycling can be developed. In terms of the mining of the currently required minerals, diversified approaches must be developed during this transition period. These include at least three complementary avenues:

1. Develop further mining of current mine waste in order to use the maximum amount possible of the extracted rocks, minerals and chemical elements. This will require thorough knowledge of each deposit, information/teaching to mining companies and, where possible, new regulations.
2. Encourage refining of by-products in the mining of the main deposits of ores (i.e. low-abundance chemical elements associated with a much more abundant

one). For example, potentially valuable levels of gallium and scandium are associated with ores of aluminium. Similarly, useful quantities of copper, antimony, silver, arsenic and tellurium may be recovered from gold deposits. However, at the same time, there is a need to develop independent supplies of technologically important minerals, beyond their 'companionality' as a by-product of one or more host metals from geologic ores, to ensure reliability of supplies.

3. Promote tolerance of new, responsible mining of metals in countries or regions which have become averse to mining. This will require education, with frank and balanced information provision and discussion that engages civil society, scientists and engineers, government and the media. The discussions must include assessment of the true price of raw materials, including the ethical and environmental prices (which is not the case at the moment), and would lead to increased costs of raw material in the coming years, with concomitant increased costs of the associated products. Taking account of the "true" price would, in turn, help engender a redefining and reordering of priorities for production and consumption.

Acknowledgements

The present article summarizes the main discussion outcomes from a panel of experts in the fields of geology, materials science, micro and nanoelectronics fabrication process, electronics circuits design, electronics packaging, supply chain management, and ICT systems: Mathilde Billaud, (Fraunhofer), David Bol, (UCLouvain), Thierry Baron (CNRS), Patrice Christmann (BRGM), Marie Garcia-Bardon (imec), Tapani Jokinen (Fraunhofer), Francois Martin (CEA-Leti), Bertrand Parvais, (imec), Karine Samuel (UGA), Lutz Stobbe (Fraunhofer) and Olivier Vergeynst (Green IT).

References

[1] Al Bartlett, https://www.albartlett.org/

[2] W. Den, C.-H. Chen, Y.-C. Luo. "Revisiting the water-use efficiency performance for microelectronics manufacturing facilities: Using Taiwan's Science Parks as a case study". Water-Energy Nexus 1 (2018) 116-133. https://doi.org/10.1016/j.wen.2018.12.002.

[3] "List of Critical Raw Materials for the EU". Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions on the 2017 European Commission, Brussels, COM/2017/0490 final, 2017. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52017DC0490 (accessed 29 August 2020).

[4] X. Ling, Y. Lin, Q. Ma, Z. Wang, Y. Song, L. Yu, S. Huang, W. Fang, X. Zhang, A.L. Hsu, Y. Bie, Y. Lee, Y. Zhu, L. Wu, J. Li, P. Jarillo-Herrero, M. Dresselhaus, T. Palacios, J. Kong. "Parallel stitching of 2D materials". Adv. Mater. 28 (2016) 2322-2329. https://doi.org/10.1002/adma.201505070.

[5] A. Khan, S. Lee, T. Jang, Z. Xiong, C. Zhang, J. Tang, L.J. Guo, W.D. Li. "High-performance flexible transparent electrode with an embedded metal mesh fabricated by cost-effective solution process". Small 12(22) (2016) 3021-3030. https://doi.org/10.1002/smll.201600309.

[6] S.A. Younis, E.M. El-Fawal, "P. Serp. Nano-wastes and the environment: Potential challenges and opportunities of nano-waste management paradigm for greener nanotechnologies", in: C. Hussain (Ed.), Handbook of Environmental Materials Management. Springer, Cham, 2018. https://doi.org/10.1007/978-3-319-58538-3_53-1.

[7] A. Lansink. "Challenging changes – Connecting waste hierarchy and circular economy". LEA, Nijmegen ISBN/EAN 978-90-821783-5-7, 2017. https://www.challengingchanges.org/the-book/ (accessed 29 August 2020).

[8] M. Irving. "Common element combos could replace rare-Earth metals in electronics". New Atlas 5 July 2019. https://newatlas.com/common-elements-replace-rare-earth-metals-electronics/60447/ (accessed 29 August 2020).

[9] F. Pena-Pereira, A. Kloskowski, J. Namieśnik. "Perspectives on the replacement of harmful organic solvents in analytical methodologies: a framework toward the implementation of a generation of eco-friendly alternatives". Green Chemistry 17 (2015) 3687-3705. https://doi.org/10.1039/C5GC00611B.

[10] J.A. Salbidegoitia, E.G. Fuentes-Ordóñez, M.P. González-Marcos, J.R. González-Velasco, T.K. Bhaskar. "Steam gasification of printed circuit board from e-waste: Effect of coexisting nickel to hydrogen production". Fuel Processing Technol. 133 (2015) 69-74. https://doi.org/10.1016/j.fuproc.2015.01.006.

[11] "CVD-semens monosilane reactor process with complete utilization of feed gases and total recycle", patent US 8,657,958 B2, 2014

[12] M. Caillau et al, "Sub-micron lines patterning into silica using water developable chitosan bioresist films for eco-friendly positive tone e-beam and UV lithography", SPIE Advanced Lithography, 2018, San Jose, California

[13] Mathieu Caillau, Pierre Crémillieu, Emmanuelle Laurenceau, Yann Chevolot and Jean-Louis Leclercq, "Fifty nanometer lines patterned into silica using water developable chitosan bioresist and electron beam lithography", https://doi.org/10.1116/1.4996870

[14] Framatome, "Fuel Business Unit Jarrie", https://www.framatome.com/EN/businessnews-142/framatome-fuel-business-unit—jarrie.html

[15] ST, "Footprint of a Microcontroller", https://www.st.com/content/st_com/en/about/st_approach_to_sustainability/sustainability-priorities/sustainable-tec hnology/eco-design/footprint-of-a-microcontroller.html

[16] Dexter Johnston, "The Market for Nanomaterial Solutions for ITO Replacement Gets Crowded", https://spectrum.ieee.org/nanoclast/semiconductors/nanotechnology/nanomaterial-solutions-for-ito-replacement-gets-crowded

[17] Bertrand Szelag, Karim Hassan, Laetitia Adelmini, Elodie Ghegin, Philippe Rodriguez, Fabrice Nemouchi, Pierre Brianceau, Elisa Vermande, Antoine Schembri, David Carrara, Pierrick Cavalié, Florent Franchin, Christophe Jany, Segolene Olivier, "Hybrid III–V/Silicon Technology for Laser Integration on a 200-mm Fully CMOS-Compatible Silicon Photonics Platform", https://doi.org/10.1109/JSTQE.2019.2904445

[18] Alan Y. Liu, John Bowers, "Photonic Integration With Epitaxial III–V on Silicon" https://doi.org/10.1109/JSTQE.2018.2854542

[19] Ashby, "M. Materials and Sustainable Development". Butterworth-Heinemann (2015). https://www.elsevier.com/books/materials-and-sustainable-development/ashby/978-0-08-100176-9 (accessed 29 August 2020).

Thomas Ernst is scientific director at Leti, CEA tech, France

Jean-Pierre Raskin is a professor at the Université catholique de Louvain, Belgium.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2023.

Cite as: T. Ernst and JP-Pierre Raskin. Towards circular ICT: From materials to components. In M. Duranton et al., editors, HiPEAC Vision 2023, pages 209-215, Jan 2023.

DOI: 10.5281/zenodo.7462003

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

© HiPEAC 2023

Ensuring that computing systems are sustainable is a highly complex area that requires a holistic approach. Given the importance of sustainability, though, the computing systems community must rise to this challenge.

Towards sustainable computer architecture: A holistic approach

By LIEVEN EECKHOUT

Sustainability and climate change represent a major challenge for our generation. This article argues that sustainable development requires a holistic approach and involves multi-perspective thinking.

Applied to computing, sustainable development means that we need to consider the entire environmental impact of computing, including raw-material extraction, component manufacturing, product assembly, transportation, use, repair/maintenance, and end-of-life processing (disassembly and recycling/reuse). Analysing current trends reveals that the embodied footprint is, or will soon be, more significant compared to the operational footprint.

The article summarizes what computer architects and engineers can and should do to better understand the sustainability impact of computing, and design sustainable computer systems.

Key insights

- Improving computing-system sustainability is **more involved than minimizing carbon emissions during production and usage**. Material use (including rare-earth elements and/or minerals from politically unstable regions in the world) and ultra-pure water consumption are significant sustainability concerns related to chip production. **Even if all the energy consumed during production and use were green, the environmental impact of computing would still be significant, and growing**.
- Sustainable development requires multi-perspective thinking along at least six **dimensions: materials, energy, environment, regulation, society, and economics**.
- The environmental footprint of computing continues to grow under current scaling trends. **When focusing on carbon emissions, embodied emissions are, or will soon become, the biggest contributor compared to operational emissions** across the broad range of computing devices.
- **Embodied emissions are** growing at a fast pace because of increasing demand for chips and **increasing energy intensity of semiconductor manufacturing**. Perhaps contradictory to popular belief, improving the energy efficiency of computing systems does not necessarily make them more sustainable.
- Improving the energy and power efficiency of computing systems may lead to a rebound effect (Jevons' paradox) which may be counterproductive to the environmental impact if the resulting **increase in demand outweighs the efficiency improvement**.
- Improving computing-system sustainability requires a **holistic approach to computer architecture design and development**, requiring **multi-dimensional optimization including chip area, power, energy, performance, reliability, and fault tolerance**.

Key recommendations

- Computer architects should take a **holistic approach when designing sustainable computer systems, and not solely focus on carbon emissions**.
- Computer architects and engineers should primarily focus on **reducing the embodied footprint of computer systems**. Reducing the operational footprint is of secondary importance, although still significant.
- Reducing the **embodied footprint of computing** can be achieved through a variety of options:
  - producing **fewer chips** (e.g. by consolidating functionality)
  - **extending the lifetime of chips** (e.g. by deploying fault tolerance and/or reconfigurability)
  - designing **smaller chips** (i.e. using additional transistors in accordance with Moore's Law, as long as it holds, in a frugal way)
  - manufacturing chips in **older technology nodes**
- **Decarbonizing the manufacturing process is not a panacea** as it does not affect other sustainability concerns related to material use and extraction, chemicals and gases emitted, and ultra-pure water consumed during production.

- Computer scientists and engineers should be wary of **Jevons' paradox. Efficiency improvements most often lead to a significant rebound effect.** Collaborating with entrepreneurs may yield new, **more sustainable business models** for computing.
- Computer architects should collaborate with various partners along the supply chain, user groups, and end-of-life recyclers to obtain **high-quality data to assess the environmental impact of raw material extraction, manufacturing, production, assembly, transportation, product use, maintenance, recycling**, etc.
- **Sustainability modelling tools** (both detailed models and high-abstraction analytical models) need to be developed, finetuned and validated to be able to **holistically balance the embodied and operational footprint** of computing devices.
- **Existing and emerging architecture paradigms** (multicore processing, hardware specialization, chiplet-based integration, etc.) need to be **assessed and re-evaluated from a sustainability perspective**.

Sustainability versus climate change

Sustainability is one of the grand challenges of our generation. Climate change is happening. A recent United Nations Climate Change report [1] in preparation for COP 27, the Sharm el-Sheikh Climate Change Conference, in November 2022 alerts that, while countries are making progress to trend down global greenhouse gas (GHG) emissions, these efforts are insufficient to limit global temperature rise to 1.5 °C by the end of the century. Much more effort is needed to keep this threshold within reach.

Virtually all economic sectors contribute to global emissions. The five economic sectors that contribute most to GHG emissions are industry, electricity, agriculture, transportation and buildings, accounting for nearly 90% of emissions, according to the Organization for Economic Co-operation and Development (OECD) [2]. Freitag et al. [3] recently reported that information and communication technology (ICT) is estimated to contribute 2.1% to 3.9% of worldwide GHG emissions, and this contribution is rising. As computer scientists and engineers, it is our responsibility to limit ICT's contribution to global warming, and, if possible, even decrease it.

While climate change is receiving increasingly wide attention – rightfully so! – it is important that we keep the broader picture in mind when reasoning about potential solutions. The broader picture relates to sustainability. To give just one concrete example: the transition towards green energy sources relies heavily on battery technology, which should be produced with the lowest possible environmental impact, using materials obtained in a safe, responsible, social, and ecological way. Moreover, at the end of their life, batteries should be repurposed, remanufactured, or recycled. In other words, and put more bluntly, solving the climate problem should not create an environmental problem. For similar points, see also Patrick Blouet's article on sustainability in this HiPEAC Vision.

Sustainability is about the efficient use of materials and energy, environmental impact, emissions, (socio-)economics, impact on human health, human rights, politics, policy, legislation, etc. Global warming is an important aspect of sustainable development, but it should not be the only focus. The extraction and use of raw materials and energy sources is another important aspect of sustainability. What to do when devices reach their end of life and how to repurpose e-waste is equally critical. Sustainability also affects and may require new legislation and business models to reduce pressure on raw-material extraction and to incentivize a circular economy with a reduced environmental footprint.

The Brundtland report of the World Council on Economic Development from 1987 provides a broad, yet useful, definition of sustainable development. It states the following: *"Sustainable development is development that meets the needs of the present without compromising the ability of future generations to meet their own needs."* This definition is an extremely powerful and unambiguous statement. It is a call to action for our generation: when generating economic activity and developing new devices and services, we should be wary of the impact this may have on future generations.

Sustainable development

Sustainable development is often narrowed down to reducing energy consumption and/or transitioning towards green energy sources. However, sustainable development is much more involved than providing and using green energy. If we are solely aiming for carbon-neutral computing, this may not necessarily bring us to a more sustainable future.

The fundamental reason is that sustainable development is extremely complicated and multi-faceted. It requires multi-perspective thinking and reasoning, involving many stakeholders with (often) conflicting interests. Moreover, the problem statement is often poorly defined. Overall, in many cases, there is no "right" answer to questions of sustainable development. The question then is how to assess sustainable development. We hence need a framework for critical thinking that recognizes the complexity and the interdependences between the various goals, interests, and constraints.

There are at least six dimensions to consider when reasoning about and assessing sustainable development [4]:

1. **Materials**: What materials do we need? How many materials do we need? How efficiently are we using these materials? Is there a supply chain? Is the supply chain reliable and secure? Important

concerns to account for include the growing demand for materials (as an example, the European Union (EU) would need 60 times more lithium by 2050 to be climate neutral [5]), the availability of critical materials on earth (e.g. rare-earth materials), price volatility, monopoly of supply, supply-chain risks, geopolitics, import and export regulations, etc.

2. **Energy**: How much energy is needed for material extraction, transportation, production, use, repair, and end-of-life processing? What is the energy source? Is the energy source reliable and secure? Some materials require substantial amounts of energy to extract, and if material extraction is done mostly using brown energy sources, there is a non-negligible carbon footprint associated with material extraction. For example, the extraction of 1 kg gold requires around 250 billion joules of energy and leads to around 15 tonnes of CO_2 emissions [6].

3. Environment: What is the carbon footprint of a device throughout the entire lifetime of a device from production to use to end-of-life processing? Are the carbon emissions during production offset by reducing carbon emissions during a device's lifetime? What is the environmental impact to the air, water, and land? How many land resources are needed? What is the impact on biodiversity?

4. **Regulation**: What are the national and international regulations regarding material use? Are there export and import rules for materials and/or components? What do legislation and directives stipulate regarding the collection and recycling of devices at the end of their lifetimes? Legislation and subsidies to stimulate the green energy transition may have significant impact on how the economy invests in its decarbonization.

5. **Society**: Will the development create jobs and welfare? Will the affected communities along the entire chain (from material supply to end-of-life recycling) benefit from the development? Are there potential concerns regarding health during the production process, usage, and end-of-life processing?

6. **Economics**: Is the development economically viable? What is the cost-benefit balance? Is the upfront investment going to generate revenue and financial benefits?

To illustrate the inherent complexity and multi-dimensionality of sustainable development, let us provide a concrete real-life example: Ireland has decided to limit data-centre construction until 2028 [7]. The reason is that allowing more data centres to be deployed would compromise the country's commitment that 80 percent of the nation's electricity grid should come from renewables by 2030, i.e. Ireland is unable to build renewable capacity fast enough to meet all demands and at the same decarbonize the grid. This example illustrates how sustainable development is a multi-objective optimization problem affecting all six dimensions:

- **materials** (to build renewable energy capacity)
- **energy provision** (to decarbonize the grid)
- **environment** (to reduce air pollution),
- **regulation** (the moratorium on data-centre construction)
- **society** and **economics** (the decision affects employment and the nation's welfare)

The Ireland example is not an isolated case; in fact, several countries are pushing for European legislation for tighter control over the instalment of data centres that consume vast amounts of electricity [8]. Furthermore, following newly adopted legislation, data-centre operators (like any other large company in Europe) will be required to report how their business activities affect sustainability [9].

Another example illustrates that carbon-free operation does not necessarily imply the most carbon-efficient solution. Acun et al. [10] point out that a data centre that operates solely on renewable energy does not minimize the total carbon footprint because of the large number of solar panels, wind farms, and batteries needed to enable carbon-free operation. The reason is that the embodied carbon emissions to produce and manufacture the renewable-energy devices (solar panels, wind farms and batteries) outweigh the operational carbon emissions saved during the lifetime of the data centre. This implies that, to minimize the total carbon footprint of a data centre, a more holistic approach is needed that accounts for both the embodied and operational emissions, rather than just focusing on the operational side.

Understanding trends in environmental impact

To understand the overall environmental impact of humankind, it is enlightening to go back to a simple yet informative formula developed by the biologist Paul Ehrlich and environmental scientist John Holdren in 1971: $I=P{\cdot}A{\cdot}T$. This formula quantifies the impact I of human activity on the environment as a function of the population P, the affluence per person A, and the impact technology T has on the environment per unit of affluence.

The world's population is growing and so is the average affluence per person. If the growth rate of the world's population and the average per-person affluence exceeds the reduction by technology, the environmental impact increases. This is happening today: the earth overshoot day – the date when the world's population has used all the biological resources that the Earth regenerates during the entire year – moved from a day towards the end of December in 1971 to end of July in 2022 [11].

The environmental impact in the IPAT equation can be measured along several dimensions, including materials used, GHG emissions, water pollution, biodiversity, etc. Yoichi Kaya, an energy economist, reformulates the IPAT equation to specifically focus on carbon dioxide (CO_2) emissions:

$$F=P\cdot G/P\cdot E/G\cdot F/E$$

where P represents the world's population, G/P the gross domestic product (GDP) per capita, E/G the energy intensity or the amount of energy consumed per unit

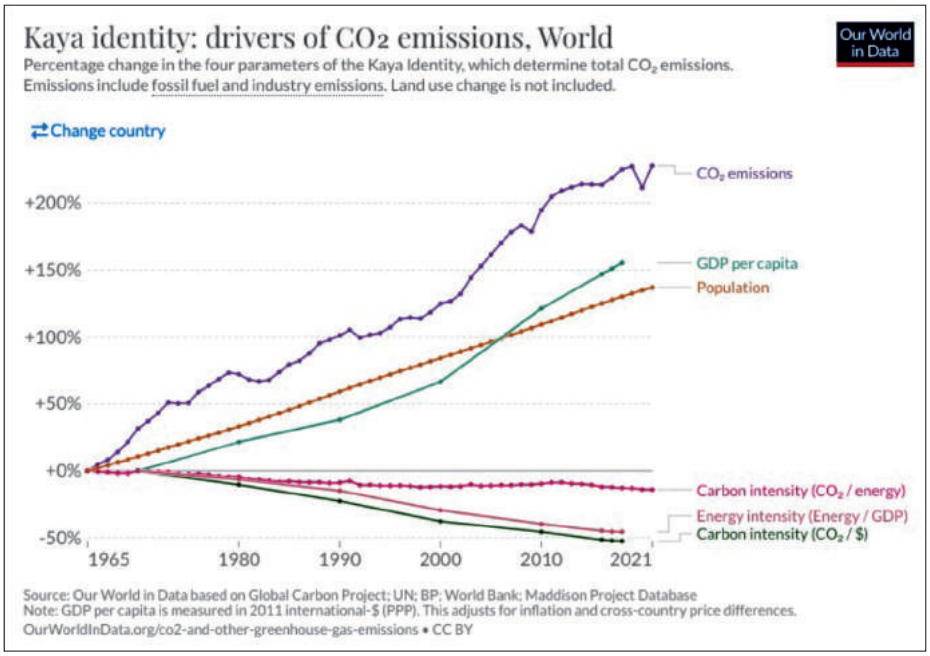


Figure 1: Kaya identity: population growth and affluence per capita growth outpace the decrease in energy and carbon intensity, leading to a net overall increase in carbon emissions. Taken from [12].

of GDP, and *F/E* the carbon intensity per unit of energy. The growth rate in in the global population and GPD per capita is currently outpacing the decrease in energy intensity and carbon intensity, leading to a net annual increase in carbon emissions, as illustrated in Figure 1.

While the IPAT and Kaya equations are insightful and widely used, they should be interpreted with care. In particular, the equations suggest that the different variables are independent of each other. However, they are not. For example, reducing the energy intensity of a device or service typically leads to a price reduction, which in turn may stimulate consumption. If the increased consumption outweighs the energy intensity reduction, we end up with a net increase in environmental impact – exactly opposite of what we had envisioned!

This is the well-known Jevons' Paradox, named after Williams Stanley Jevons, who was the first to observe the rebound effect of the steam engine's improved coal efficiency leading to an overall increase in coal consumption [13]. This rebound effect is also (part of) the reason why improving energy or power efficiency of a computing device does not necessarily lead to a net reduction in environmental impact. Most often, an energy- or power-efficiency gain leads to increased usage and deployment, effectively increasing the environmental impact of computing – as we will discuss further.

ICT's environmental impact

We will now leverage the above equations to analyse the environmental impact of computing. To do so, it is important to make a distinction between embodied versus operational emissions [14]. For the discussion that follows, we will mostly focus on GHG emissions, but several aspects also pertain to other environmental concerns, as we will point out.

Embodied emissions relate to raw-material extraction, manufacturing, assembly, transportation, repair, maintenance, and end-of-life processing. Operational emissions relate to product use during a device's lifetime.

Embodied emissions can be further categorized in scope-1, scope-2, and scope-3. Scope-1 refers to the chemicals and gases used during manufacturing – this includes fluorinated greenhouse gases with orders of magnitude higher global warming potential than CO_2. Scope-2 refers to the energy consumption during chip manufacturing – this includes empowering the extensive production facilities with hundreds of manufacturing tools and requiring climate and humidity control. Scope-3 pertains to the energy consumption for the extraction and production of materials used for integrated circuit manufacturing. For the purposes of this discussion, we will focus on scope 1 and scope-2, because scope-3 follows a similar pattern. See [15] for more details.

The embodied scope-2 emissions of computing can be modelled as follows:

$$F_{(scope\text{-}2)} = C \cdot W/C \cdot E/W \cdot F/E$$

where C represents the number of chips produced, *W/C* the number of wafers needed per chip, *E/W* the amount of energy needed per wafer, and *F/E* the carbon intensity of chip manufacturing. The number of wafers needed per chip *W/C* tends to stagnate as we approach the reticle limit of chip manufacturing, i.e. the maximum size that lithography machines can process with a single mask. In contrast, the number of chips *C* tends to increase at a growth rate of 9% per year [16], and the amount of energy needed for manufacturing *E/W* increases at a rate of 11.9% per year as we transition to new technology nodes, according to recent data provided by imec [17]. The carbon intensity of chip manufacturing is not improving fast enough to compensate for the increase in chip demand and energy intensity of manufacturing, which as a result leads to an overall annual increase in scope-2 emissions due to chip manufacturing.

The embodied scope-1 emissions can be modelled similarly:

$$F_{(scope\text{-}1)} = C \cdot W/C \cdot F/W$$

where *F/W* represents the carbon dioxide equivalents due to fluorinated compounds per wafer. Imec data reports that this factor is increasing by slightly more than 9.3% per year [17]. With the number of chips increasing by 9% per year and chip die size being constant, embodied scope-1 emissions are hence trending up.

Operational emissions can be modelled as follows:

$$F_{operational} = C \cdot E/C \cdot F/E$$

where E/C represents the total electricity usage of a chip over its entire lifetime and F/E the carbon intensity during device use. When there is work to be done, a chip's operational emissions are proportional to its energy consumption. When the chip is idle, operational emissions are proportional to the chip's idle power. A variety of optimizations across the system stack improve the energy and power efficiency of individual devices: from transistor tuning to dynamic voltage and frequency scaling, clock gating, power gating, etc. The question is whether the per-device energy and power efficiency improvements are outweighed, or even worse, outpaced, by the increase in the number of chips deployed.

What do the trends look like?

Gupta et al. [14] performed a survey of consumer devices from vendors including Apple, Google, Huawei and Microsoft. They conclude that embodied emissions dominate for battery-operated devices such as wearables, smartphones, tablets, and laptops, while operational emissions dominate for always-connected devices such as speakers, desktop computers and gaming consoles. For data centres, most emissions are related to construction, infrastructure, and hardware manufacturing: interestingly, while total energy usage is trending up – presumably because of increased server count and/or higher degree of consolidation (cf. Jevons' paradox) – total operational emissions are decreasing for Facebook and Google, thanks to their policy of contracting and securing green energy sources to power their hyperscale data centres.

Making decisive conclusions about the environmental impact of specific computing devices is inherently difficult because of the variety of use cases in which computer systems are manufactured, deployed, and used. For example, the use of green energy sources during chip manufacturing may shift the contribution from embodied to operational emissions. Nevertheless, the overall conclusion that can be reached based on the above equations [15] indicates that embodied emissions are continuing to grow under current scaling trends, and that embodied emissions already are, or will soon be, the biggest contributor; see also Figure 2. The fundamental reason is the increasing demand for chips (because of economic dynamics based on selling products) and the growing energy intensity of semiconductor manufacturing (because of advancements in chip technology), which do not seem to be counterbalanced by the transition to green energy sources and improvements in per-device energy and power efficiency.

Looking forward

There are several important conclusions to be taken from the above analysis.

First and foremost, to reduce both the embodied and operational emissions of computing, we could reduce the number of chips that we produce and sell. This could possibly be achieved by integrating more functionality within individual chips. Modern-day heterogeneous system-on-chip (SoC) designs integrate a couple of dozen accelerators in addition to central processing unit (CPU) and graphics processing unit (GPU) cores, yet this has not led to a reduction in carbon footprint, on the contrary (yet another example of Jevons' paradox).

Current business models are based on selling devices and hence stand in the way of reducing the number of chips that we produce. The number of connected devices is rapidly increasing: Cisco estimates that the internet of things (IoT) was born between 2008 and 2009 when there started to be more connected devices than people; today there are more than seven connected devices per person – this number is even higher in the Western world (up to 12.9 and 8.9 devices per person in North America and Western Europe, respectively) [18].

Service-model based business models, such as leasing a smartphone as recently offered by Fairphone [19], may incentiv-

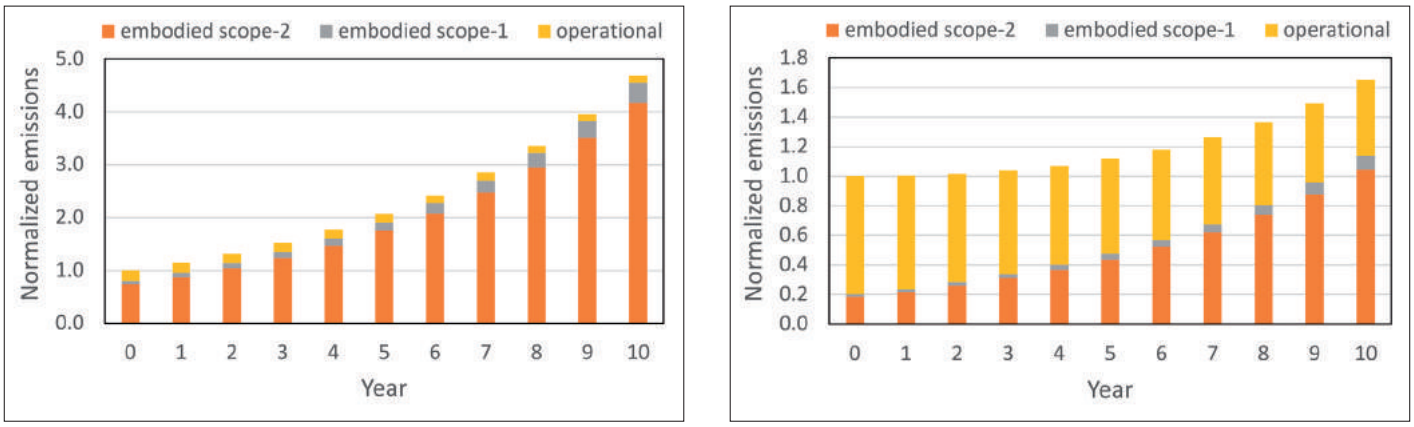


Figure 2: Projection for total emissions over the next decade given current scaling trends for two scenarios: (left) embodied emissions dominate initially (80% of total emissions in year zero), and (right) operational emissions dominate initially (80% of total emissions in year zero). Total emissions are increasing dramatically, and embodied emissions are, or will start, dominating. This analysis assumes the following annual growth rates: +9% number of chips C, +11.9% wafer energy intensity (E/W), +9.3% wafer chemical/gas intensity (F/W), -2.5% carbon intensity (F/E), -10% operational energy intensity (E/C), and 0% die size (W/C) – note the assumed negative growth rates for carbon and operational energy intensity, and the zero growth rate for chip die size. Taken from [15].

ize manufacturers to design systems with a longer lifetime that can be repaired, remanufactured, reassembled, reused, recycled, etc. (See the article "Everything as a service" in this HiPEAC Vision for more examples of service-based offerings and their contribution to sustainability.)

Another (complementary) approach to reduce the demand for more chips could be to prolong the lifetime of a device by deploying fault-tolerance techniques to fix errors or to enable graceful degradation (e.g. disabling a faulty core in multicore CPUs or GPUs). Reprogrammable hardware, e.g. field-programmable gate arrays (FPGAs), and hardware reconfigurability could also be viable enablers to provide hardware acceleration in a more sustainable way.

Indeed, it is a fact that electronic devices have a short lifespan and e-waste is a major problem. In particular, the average lifetime of a smartphone is around four years, but young people tend to replace their smartphones more quickly: 40% of 18- to 24-year-olds keep their phones for less than a year [20]. As a result, e-waste is a major issue, with more than 7 kg of e-waste per person on average and more than 16 kg of e-waste per person in Europe [21].

We could design smaller chips to reduce the embodied footprint per chip. There is some leeway in fact because of Moore's Law. Moving to a new chip technology node offers twice the number of transistors for the same chip area, but because of the increased energy intensity of new technology nodes, the total embodied footprint per chip in fact increases. There is a middle ground, though in which computer architects use only a fraction of the additional transistors (for as long as Moore's Law continues to hold) to add functionality such that the total embodied footprint does not increase or, even better, decreases.

In fact, this implies that we need to design smaller chips, albeit with a (slightly) higher transistor count than the previous generation. As projected in [15], reducing die size by 25% each year still allows 6% more transistors each year to add new functionality, and yet embodied scope-2 emissions would decrease by 12% each year. (Note that smaller chips not only help reduce the embodied footprint; they also improve manufacturing yield.) This path, unfortunately, is not what industry is currently pursuing, in part presumably because of market dynamics and competition.

Manufacturing chips in older, less energy-demanding chip technologies could also help reduce embodied emissions. That potentially comes at the cost of higher operational emissions because of a less energy-efficient chip technology. Whether the trade-off balances towards more versus less sustainable system design remains to be seen. In any case, if this pathway turns out to be promising, now could be the right time given the recent emergence of chiplet-based integration in which different chiplets may be manufactured in different chip technologies, thereby reducing the overall embodied footprint of chiplet-integrated devices.

Rapidly transitioning to green energy sources for manufacturing will drastically reduce carbon emissions. In fact, TSMC has engaged itself to supply 25% of its fabrication plant power supply from renewable energy and be carbon-neutral by 2050 [22]. While this greatly affects embodied scope-2 emissions, it does not affect scope-1 or scope-3 emissions. In addition, it does not reduce other sustainability concerns such as those around raw materials and ultra-pure water supply needed for manufacturing, e-waste, etc.

Moreover, if the green energy supply is appropriated from the global market through green energy contracts, this does not fundamentally reduce the carbon footprint at the global societal scale because other users are hence deprived from green energy. Finally, green energy sources are not carbon free either – solar panels, wind turbines, etc. also incur an embodied carbon footprint for manufacturing, maintenance, transportation, and end-of-life handling.

Reducing a chip's operational emissions, while less important than reducing its embodied emissions, is still an important optimization criterion. Lowering energy consumption when there is work to be done reduces a chip's operational emissions. Lowering idle power consumption when there is no work to be done, also reduces a chip's operational emissions. Note though that efficiency improvements are possibly (in practice, frequently) subject to Jevons' paradox. For example, energy optimization due to higher performance enables more jobs per unit of time to be executed, thereby increasing overall energy consumption. Likewise, a power-saving optimization may enable more concurrent jobs within the available power envelope, which may lead to an overall increase in energy consumption.

Sustainable design and inherent data uncertainty

It is clear from the above discussion that improving the sustainability of computing systems is complicated and requires a holistic design approach that touches upon a variety of design criteria including design complexity and chip area, performance, energy and power efficiency, reliability, and fault tolerance.

Computer architects are well versed in optimizing along a single design criterion while taking other design criteria into consideration, for example, optimizing performance with limited impact on power consumption and design complexity, or improving reliability with limited impact on performance and energy consumption. Optimizing for sustainability, on the other hand, requires a more holistic approach, considering all design criteria and stakeholders at the same time while optimizing the impact on the overall embodied and operational footprint and being subject to significant degrees of uncertainty.

For example, while a fault-tolerance technique that fixes hard errors or enables graceful degradation may prolong the lifetime of a device, thereby damping the quest for more chips, it comes at the cost of increased embodied footprint (to provide the fault-tolerance hardware circuitry) and operational footprint (to dynamically monitor the operation during the device's lifetime). Whether a fault-tolerance technique leads to net overall reduction in environmental footprint depends on the

relative importance of the embodied versus operational footprint, the likelihood of an error, and the typical use case of the device under design. Optimizing sustainable computer systems is without any doubt a challenging design problem.

This design challenge is further complicated by the large degree of uncertainty in a variety of dimensions. While companies' sustainability reports and product lifecycle-assessment (LCA) reports provide a wealth of data, there remain many unknowns and data limitations, in part because of industry secretiveness, or simply because of lack of reliable data. For example, a recent study by imec [17], which attempts to quantify the environmental footprint of modern-day chip manufacturing, makes assumptions regarding the energy consumption of a fab's facility equipment (i.e. it is "*assumed to contribute to 40% of the total energy*"); furthermore, the degree of abatement of fluorinated GHGs (scope-1) is unknown, as well as the use of materials and the energy needed for material extraction (scope-3). As another example, the Apple iPhone12 LCA report [23] uses industry averages when parameters are unknown for the production process, i.e. a company may not know the sustainability impact of its suppliers.

The operational footprint and its importance relative to the embodied footprint is even harder to assess, as it depends on typical user behaviour, product lifetime, and the geographic location of the user (which determines the carbon intensity of the user's power grid mix). Historical data could be insightful, but it only provides a hint. Note further that product use may be subject to the infamous rebound effect, which may significantly shift the relative importance of the operational versus embodied footprint.

Overall, it is safe to conclude that there is inherent data uncertainty. Gupta et al. [24] recently proposed the ACT model to analyse a computer system's sustainability at design time. This model relies on detailed numbers from production processes in industry. This is an important step for our community at large (both in industry and academia). Nevertheless, the authors note that there is "*lack of up-to-date carbon emission data for the latest compute, memory, and storage technologies*". Furthermore, they hope to "*encourage industry to publish more detailed carbon characterizations to standardize carbon footprint accounting*". Imec's sustainable semiconductor technology and systems (SSTS) program aims at addressing exactly this issue by collaborating with major industry players to quantify the environmental impact of integrated circuit manufacturing [25]. While significant progress is being made regarding the embodied footprint of computing, more is needed. Moreover, an equally substantial effort needs to be made to quantify the operational footprint of computing, which may turn out to be even more challenging.

Sustainable design based on first principles

And yet, despite the large degrees of uncertainty and the multi-faceted design problem, computer architects need to make design decisions to make computer systems more sustainable. One option may be to revert to first principles and guide sustainable design decisions using a first-order model. First-order modelling should not be viewed as a replacement for, but rather as a useful complement to, detailed models like ACT and others. In fact, a detailed sustainability accounting method can provide initial data for a first-order model, and

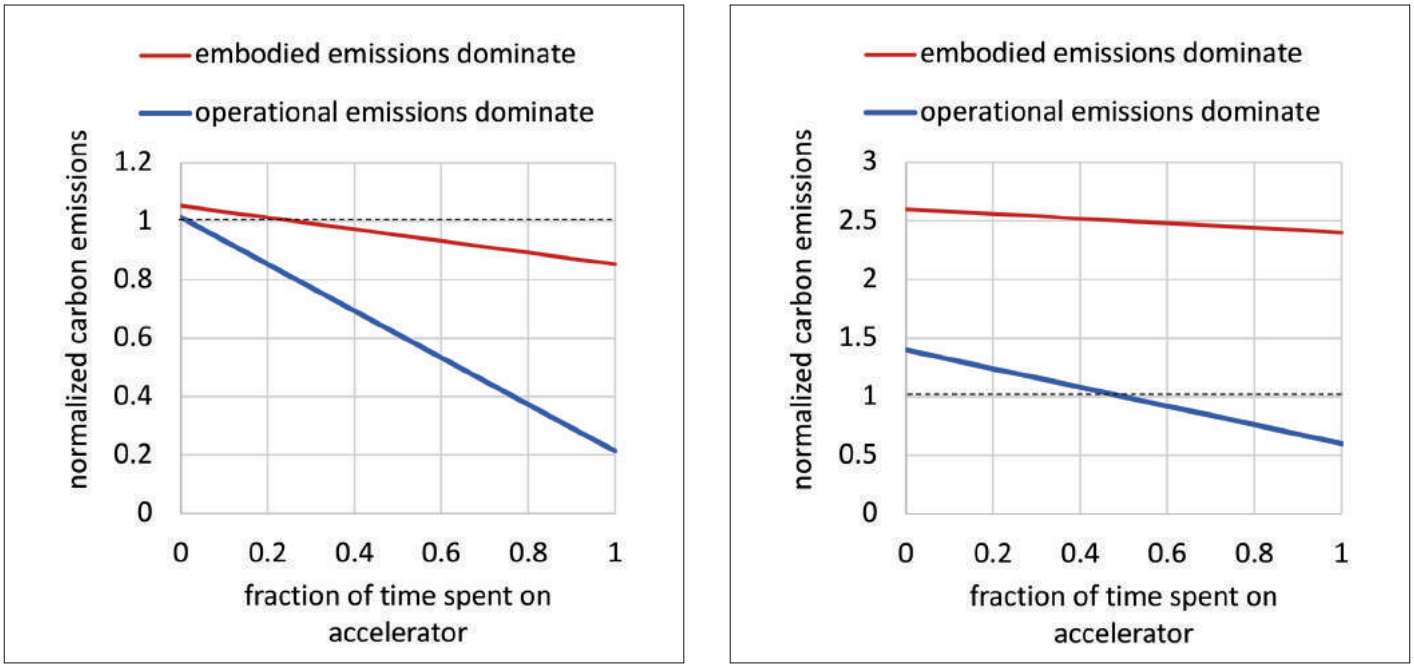


Figure 3: Total carbon footprint of a general-purpose CPU plus accelerator as a function of its degree of use, assuming that the accelerator takes up 6.5% extra chip area (left) versus 2x extra chip area (right), normalized to a general-purpose CPU without an accelerator. The accelerator is assumed to consume 500x less energy than the general-purpose CPU for performing the same work. Two scenarios are considered: embodied emissions account for 80% of total emissions versus 20% of total emissions. The larger the chip area of the accelerator, the more frequently it needs to be used and the higher the relative weight of the operational emissions need to be for the accelerator to be sustainable. Taken from [10].

vice versa, a first-order model can provide directions where the detailed model should be further refined.

A first-order model uses proxies for the embodied and operational footprint that computer architects have control over, see for example [10] for more details about a first-order model for computer chips. A useful, first-order proxy for the embodied footprint of a chip is its die size, i.e. the larger the chip, the higher the embodied footprint for a given chip technology in terms of the energy and materials needed and the chemicals and gases emitted during production. A useful proxy for the operational footprint of a chip is energy consumption assuming a fixed-work scenario (i.e. a device performs a fixed amount of work during its entire lifetime) and power consumption assuming a fixed-time scenario (i.e. a device is used for the same amount of time, and hence performs more work). The relative importance of embodied versus operational emissions can be captured via a parameter the architect can vary to explore different use case scenarios.

Although (deliberately) simple, a first-order sustainability model can reveal a variety of interesting insights which computer architects can take forward to design more sustainable computer systems. There is a fruitful avenue of future work to explore how computer architectures can be made more sustainable. Analysing to what extent archetypical CPU and GPU design paradigms and solutions (e.g. caching, speculation, microarchitecture, acceleration, etc.) affect computer system sustainability would be extremely valuable.

For example, as reported in [10], the first-order model can be used to assess whether hardware specialization is sustainable. Integrating a hardware accelerator next to a general-purpose processor incurs a cost in terms of embodied footprint (because of a larger chip) which may be compensated for by the reduced operational footprint (because of lower energy consumption when using the special-purpose accelerator rather than a general-purpose CPU). In other words, the reduced operational footprint amortizes the increased embodied footprint.

The question is where the tipping point is. The larger the accelerator, the more frequently the accelerator needs to be used and the higher the relative weight of the operational emissions needs to be for the accelerator design to be sustainable, as illustrated in Figure 3; if the accelerator is taking up significant chip area, and the embodied emissions dominate, the reduction in operational emissions does not compensate for the increased embodied emissions.

This suggests that the current trend towards large system-on-chip (SoC) designs with dozens of accelerators that occupy a significant fraction of the chip and that are not powered on all the time due to dark-silicon constraints, may not be a sustainable design paradigm. A more fruitful, sustainable design paradigm might be to consolidate accelerator designs to a common-denominator accelerator that can serve multiple critical applications while incurring less chip area, thereby reducing the embodied footprint at the expense of an increased operational footprint, with a net improvement in sustainability. Investigating these (and other) architecture trade-offs in more detail is a promising research avenue for computer architects in industry and academia.

Conclusion

Improving computing-system sustainability is a challenging and multi-faceted problem. The embodied footprint is, or will soon be, a more important contributor than the operational footprint, primarily due to an increasing demand for chips and increased energy intensity of integrated circuit manufacturing. Decarbonizing the production process and use phase of compute devices is not a panacea, though, because it does not address other sustainability concerns including raw material extraction, chemicals and gases emitted, and ultra-pure water used during production.

What makes sustainable computer system design unique compared to traditional optimization criteria is that it requires a holistic approach considering chip area, power, energy, performance, lifetime, reliability, etc. The field of computer architecture specifically, and computer science and engineering in general, has only recently embarked on this endeavour.

Computer architects should continue to (1) collect high-quality data to assess the sustainability impact across the entire lifetime of a computing device, from raw-material extraction, transportation, manufacturing, assembly, use, repair, end-of-life processing, etc., (2) develop detailed and high-abstraction models to help designers evaluate the impact on sustainability at design time, and (3) analyse and revisit architecture design paradigms considering their sustainability impact. Overall, sustainable system design is an extremely timely and societally important topic where substantial innovation is to be achieved and expected in the following years.

References

[1] "Climate Plans Remain Insufficient: More Ambitious Action Needed Now," United Nations, 26 October 2022. [Online]. Available: https://unfccc.int/news/climate-plans-remain-insufficient-more-ambitious-action-needed-now. [Accessed 28 November 2022].

[2] "Climate Action Explore policy solutions by key economic sector," OECD, [Online]. Available: https://www.oecd.org/stories/climate-action/key-sectors/. [Accessed 28 November 2022].

[3] C. Freitag, M. Berners-Lee, K. Widdicks, B. Knowles, G. S. Blair, A. Friday, "The Real Climate and Transformative Impact of ICT: A Critique of Estimates, Trends, and Regulations," Patterns, vol. 2, no. 9, pp. 100340, https://doi.org/10.1016/j.patter.2021.100340, 2021.

[4] M. Ashby, Materials and Sustainable Development, 1st edition, Elsevier, 2015.

[5] "Critical Raw Materials Resilience: Charting a Path towards greater Security and Sustainability," 2 September 2020. [Online]. Available: https://ec.europa.eu/docsroom/documents/42849. [Accessed 28 November 2022].

[6] L. Eeckhout, "A First-Order Model to Assess Computer Architecture Sustainability," IEEE Computer Architecture Letters, vol. 21, no. 2, pp. 137-40, July-Dec 2022.

[7] P. Judge, "EirGrid pulls plug on 30 Irish data center projects," Datacenter dynamics , 24 May 2022. [Online]. Available: https://www.datacenterdynamics.com/en/news/eirgrid-pulls-plug-on-30-irish-data-center-projects/. [Accessed 28 November 2022].

[8] A. Roach and E. Krukowska, "Big Tech Gets Caught Up in Europe's Energy Politics," Bloomberg, 23 June 2022. [Online]. Available: https://www.bloomberg.com/news/articles/2022-06-23/google-facebook-data-centers-face-europe-political-snags-over-in-energy-crisis?sref=FwE94DUF. [Accessed 28 November 2022].

[9] "Corporate Sustainability Reporting Directive (CSRD) in "A European Green Deal"," European Parliament, 20 October 2022. [Online]. Available: https://www.europarl.europa.eu/legislative-train/theme-a-european-green-deal/file-review-of-the-non-financial-reporting-directive. [Accessed 28 November 2022].

[10] B. Acun, B. Lee, K. Maeng, M. Chakkaravarthy, U. Gupta, D. Brooks, C.-J. W, "Carbon Explorer: A Holistic Approach for Designing Carbon-Aware Datacenters," in ACM International Conference on Architecture Support for Programming Languages and Operating Systems (ASPLOS), Vancouver, 2023.

[11] "Earth Overshoot Day," [Online]. Available: https://www.overshootday.org/. [Accessed 28 November 2022].

[12] H. Ritchie and M. Roser, "Emissions drivers," Our World in Data, [Online]. Available: https://ourworldindata.org/emissions-drivers. [Accessed 28 November 2022].

[13] "W. Stanley Jevons, "The Coal Question," 1865," Yale University, [Online]. Available: https://energyhistory.yale.edu/library-item/w-stanley-jevons-coal-question-1865. [Accessed 28 November 2022].

[14] U. Gupta, Y. G. Kim, S. Lee, J. Tse, H.-H. S. Lee, G.-Y. Wei, D. Brooks, C.-J. Wu, "Chasing Carbon: The Elusive Environmental Footprint of Computing," in IEEE International Symposium on High-Performance Computer Architecture (HPCA), Virtual, 2021.

[15] L. Eeckhout, "Kaya for Computer Architects: Towards Sustainable Computer Systems," IEEE Micro, pp. 1-8, https://ieeexplore.ieee.org/document/9932869, 2022.

[16] "The McClean Report 2022," IC Insights, January 2022. [Online]. Available: https://www.icinsights.com/services/mcclean-report/. [Accessed 28 November 2022].

[17] M. Garcia Bardon, P. Wuytens, L.-A. Ragnarsson, G. Mirabelli, D. Jang, G. Willens, A. Mallik, S. Spessot, J. Ryckaert, B. Parvais, "DTCO including Sustainability: Power-Performance-Area-Cost-Environmental score (PPACE) Analysis for Logic Technologies," in 2020 IEEE International Electron Devices Meeting (IEDM), Virtual, 2020.

[18] "Cisco Annual Internet Report (2018–2023) White Paper," Cisco, 9 March 2020. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html. [Accessed 28 November 2022].

[19] "Fairphone Easy: The sustainable smartphone solution," Fairphone, [Online]. Available: https://shop.fairphone.com/en/fairphone-easy. [Accessed 28 November 2022].

[20] Z. Muhammad, "People have started changing their smartphone less often," Digital Information World, 28 August 2019. [Online]. Available: https://www.digitalinformationworld.com/2019/08/how-often-should-you-upgrade-your-phone-infographic.html. [Accessed 28 November 2022].

[21] I. Tiseo, "Per capita electronic waste generation worldwide from 2010 to 2019," Statista, 4 October 2022. [Online]. Available: https://www.statista.com/statistics/499904/projection-ewaste-generation-per-capita-worldwide/ . [Accessed 28 November 2022].

[22] P. Lin and L. Sun, "TSMC Becomes the World's First Semiconductor Company to Join RE100, Committed to 100% Renewable Energy Usage," TSMC ESG, 27 July 2020. [Online]. Available: https://esg.tsmc.com/en/update/greenManufacturing/caseStudy/37/index.html. [Accessed 28 November 2022].

[23] "Product Environmental Report: iPhone 12," 13 October 2020. [Online]. Available: https://www.apple.com/environment/pdf/products/iphone/iPhone_12_PER_Oct2020.pdf.

[24] 8. U. Gupta, M. Elgamal, G. Hills, G.-Y. Wei, H.-H. S. Lee, D. Brooks, C.-J. Wu, "ACT: Designing Sustainable Computeer Systems with an Architectural Carbon Modeling Tool," in ISCA '22: Proceedings of the 49th Annual International Symposium on Computer Architecture, New York, 2022.

[25] L.-A. Ragnarsson, C. Rolin, S. Shamuilia, E. Parton, "The green transition of the IC industry," Imec, [Online]. Available: https://www.imec-int.com/en/expertise/cmos-advanced/sustainable-semiconductor-technologies-and-systems-ssts/stss-white-paper. [Accessed 28 November 2022].

Lieven Eeckhout is a senior full professor in the department of electronics and information systems at Ghent University, Belgium.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2023.

Cite as: Lieven Eeckhout. Towards sustainable computer architecture: A holistic approach. In M. Duranton et al., editors, HiPEAC Vision 2023, pages 216-224, Jan 2023.

DOI: 10.5281/zenodo.7461989

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

© HiPEAC 2023

Dematerialization and the "as-a-service" (XaaS) business model are a strong business trend. They are key to reducing the carbon footprint of the economy.

Everything as a service

By KOEN DE BOSSCHERE and MARC DURANTON

Over the last 30 years, almost everything that can be represented by bits and bytes has been digitized: music, movies, photos, books, news, … This digitization has led to full dematerialization of production, transmission and consumption. Thanks to increased compute power, large and affordable digital storage capacity, and fast networks, digital solutions have not only replaced their physical counterparts; in many cases, they are better.

This trend has led to disruption in several economic sectors, which have had to reinvent their business models so as to transition from selling physical goods to selling a service. The "as-a-service" business model also comes with a variety of service revenue models that generate recurring revenue. On top of that, the service platforms generate a wealth of real-time data that the service company can use to optimize the service offering.

Key insights

- Digitization has transformed the entire media industry. Consumers and distribution platforms have become content creators too.
- The "as-a-service" business model has changed users' worldviews. Ownership is being gradually replaced by "24/7 access" and renting. This is leading to the rapid dematerialization of the economy, and might have impact on the long-term existence of creations.
- The impact of dematerialized services like streaming, videoconferencing and cloud gaming on the environment is moderate, which makes them ecologically less damaging than travelling (by car) to a cinema, a meeting or a gaming event. The fact that the energy consumption in data centres is increasing is actually good news if it means that there are overall energy savings thanks to reductions in other sectors of the economy.

Key recommendations

- Keep investing in ultra-low-power computing technology (data centres, networks, devices) so as to reduce the carbon footprint of digital services and offset the environmental impact of their exponential growth.
- If the above recommendation is executed, keep investing in the dematerialization of additional services (based on e.g. renting shared machines like tools and cars instead of buying them).
- Create digital libraries and archives, in order to preserve digital-only creations.
- Ensure that European ethical standards are properly taken into account by content providers, especially when it comes to exploiting service usage data.

Very few people in 1982 realized that the introduction of the compact disc (CD) was the start of a new trend (the digitization of analogue information) that was going to disrupt whole industries. For the customers of 1982, it was just a convenient and higher quality carrier of music.

In 1988, Fujifilm introduced the first fully digital camera, able to store up to ten photographs on a memory card. This represented another major analogue-to-digital transition. The first camera phone was the Kyocera Visual Phone VP-210 in 1999. By 2010 all smartphones could record and play media, and resolution and storage capacity were no longer a serious constraint for most users. In the end, it took almost 30 years to evolve from the first digital music player to a powerful multimedia device in pocket format that people are prepared to spend a couple of hundred euros per year on, and that they carry with them at all times. Today, for millions of people it is the last device they see at night, and the first they see in the morning.

Digitization and the "as-a-service" business model leads to disruption

In the process of digitization, the business model of the content providers has also changed. Instead of selling physical content (like they did in the times of the CD), they started selling digital content that could be downloaded. When the networks became better, they moved to a subscription model for a streaming service where the user has full access to millions of songs for a flat monthly rate of less than €10, or less than the cost of one CD per month. Subscriptions to video streaming services are of the same ilk. Per streaming service, there are several plans, and access to premium content requires a higher monthly subscription rate. Some also have a free plan where content is regularly interrupted by adverts, which is called the "freemium" business model (a portmanteau of free and premium).

The advantage of this model for the provider to *lock in* subscribers and thereby provide more stable revenues. As, most of the time, the digital contents are subject to digital rights management (DRM) which can be revoked at any time, the owners don't own the content, and their subscription is only a rental licence. For the first time in history, intellectual content can be erased with one click (as in 2009, when Amazon remotely deleted some digital editions of the books of George Orwell – including 1984 – from the Kindle devices of readers who had bought them) [7].

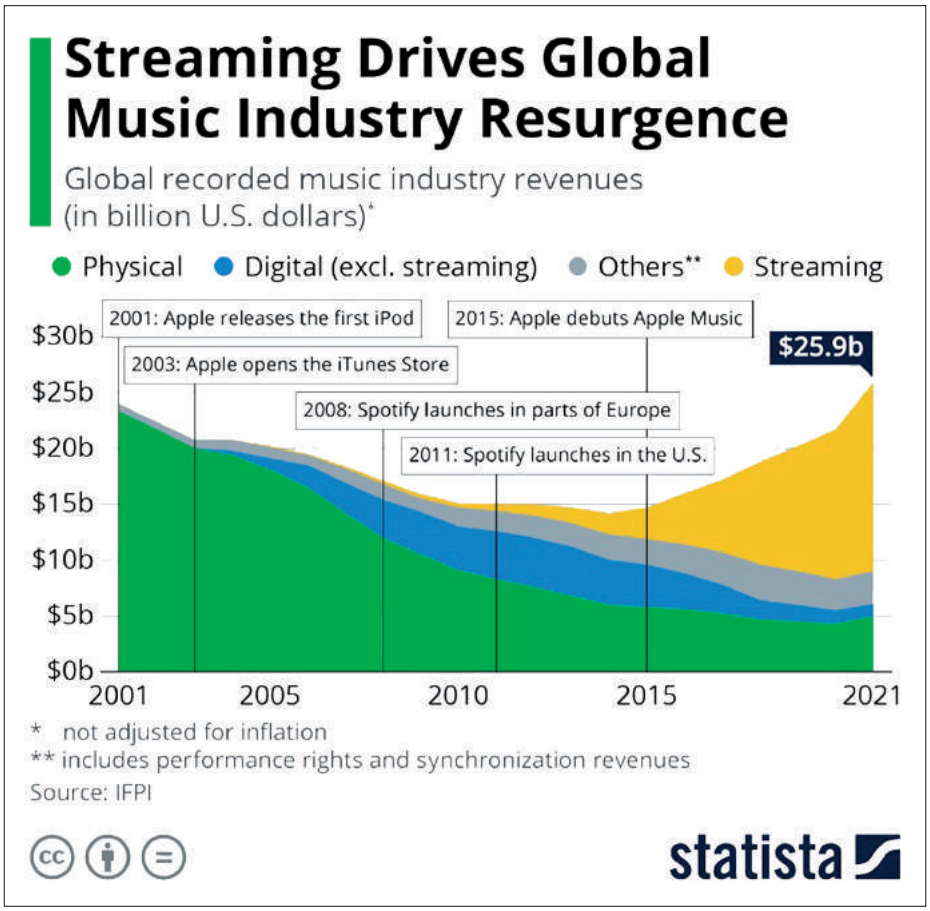


Figure 1: The fast growing music streaming industry brings music industry revenue back to levels seen in the 2000s [1]

Streaming services have put music shops and video rental services out of business in a very short time (Figure 1). Users now have access to all the content they can dream of without having to invest in a collection, and music is available on every platform at an affordable price. The choice of what is available now depends on the choices of the provider; for example, some classic movies can't be found at all on platforms that keep only what is fashionable and immediately profitable. Distributors no longer have to invest in the production and distribution of physical media and – very importantly – have full access to the behaviour of their customers.

Over the past decade, some distributors (especially in the video streaming sector) have started producing their own content and have become very successful. In 2022, Netflix spent $14 billion on content creation and is the fourth largest entertainment producer of 2022, see Figure 2. This content earned no fewer than 24 Oscar nominations in 2020.

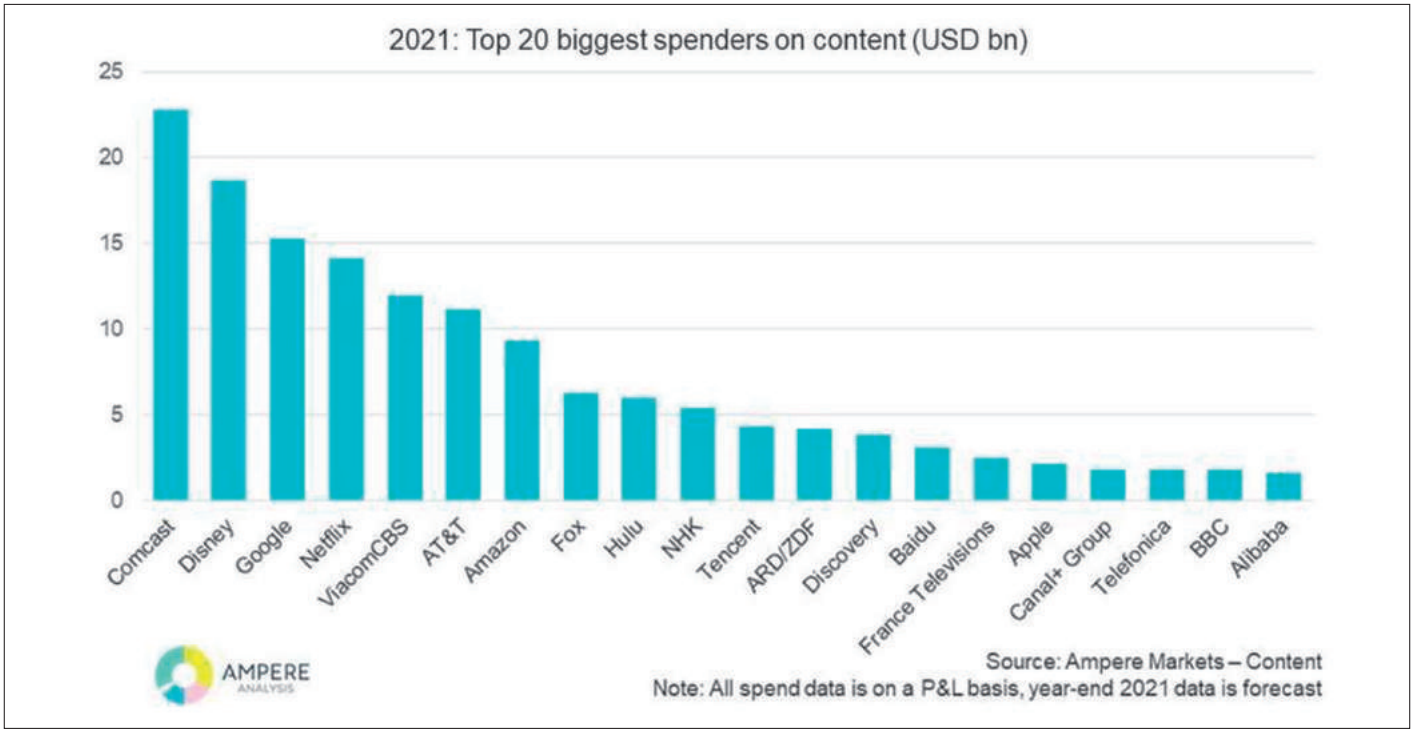


Figure 2: Content creation spending [2]

One of Netflix's key assets is that it has access to the behaviour of its viewers and can tailor its offering to their interests and preferences. It is a dominant market leader, making it more difficult for smaller players to survive. "Recommender systems" used by streaming services use artificial intelligence to best profile the users. The side effect is that users may be exposed to less diversity of content; such systems could even be, in malevolent hands, a way to manipulate people, based on their individual behaviour – i.e. customized manipulation – and victims may well be unaware that this is happening.

More generally, the digitization of media (audio, video, photos, newspapers, books, games…) has profoundly changed society. The impact of this technology may be compared to the introduction of the printing press in Europe by Johannes Gutenberg. Today everything from books to audio recordings and movies can be duplicated forever without loss of quality, in next to no time, and at an extremely low cost. Thanks to the internet and the cloud, content can be accessed from nearly anywhere in the world – if not blocked by regional licences and DRMs.

This was made possible thanks to the giant leaps made in performance in processing, storage and digital communications, fuelled by Moore's Law. Modern streaming media companies are made possible by reliable broadband access and huge data centres distributed across the globe.

But there is more. The printing press enabled more people to print and distribute their ideas in print format. Digital media enables everyone with a smartphone to produce and distribute audio, video, photos, text, games, … We are all now "prosumers" who produce and consume at the same time. YouTube is currently the market leader for video sharing.

However, there are significant issues associated with controlling the quality, sensitivity or veracity of the content. Facebook and YouTube use algorithms based on artificial intelligence to check the most obvious characteristics of the content and identify pornography or violence for example, but recent history has shown that this is not a perfect system and that it can also incur problems. These social media applications and their massive data sets are a key driver for the development of modern AI. In the meantime, human content moderators are employed to act as a filter, often in lower income countries, and often on contracts that offer low pay and with poor conditions. They are required to review hundreds of videos a day, many of which are psychologically disturbing, and are not always offered adequate psychological support [10, 11, 12].

The consumer market has reacted to this evolution by offering a wide range of devices to improve the media experience (headsets, headphones, loudspeakers, large screens, …). The computing part is almost invisible: a smartphone, tablet or smartphone suffices to carry out the necessary processing.

Traditional print media like newspapers and magazines have followed the trend. They cannot survive without a digital channel. In the beginning, the content was merely a digital copy of the printed version, but that no longer suffices. Today, media outlets start from the digital content that feeds into a news website, offering video, audio, blogs, vlogs, digital puzzles, etc. They process information as fast as television networks: in real time with breaking news and updates. The print version follows the next day.

The benefits for the consumer are clear: digital delivery is more convenient, faster, available 24/7 on the platform of choice, cheaper and takes up less space. Since it is dematerialized, it is available worldwide, and converts the use of physical resources (paper, ink, fuel for transportation) into the use of energy.

However, it is much harder to come up with a profitable business model: many users do not want to pay for premium content they can find for free elsewhere on the internet, and the income from advertising is also shrinking because many advertisers prefer larger platforms like Google or Facebook. Print newspapers are struggling to survive and are primarily bought by those over fifty (Figure 3). 73% of the young generation gets its news from digital media; their smartphone, rather than a television, radio or printed newspaper, is their window to the world.

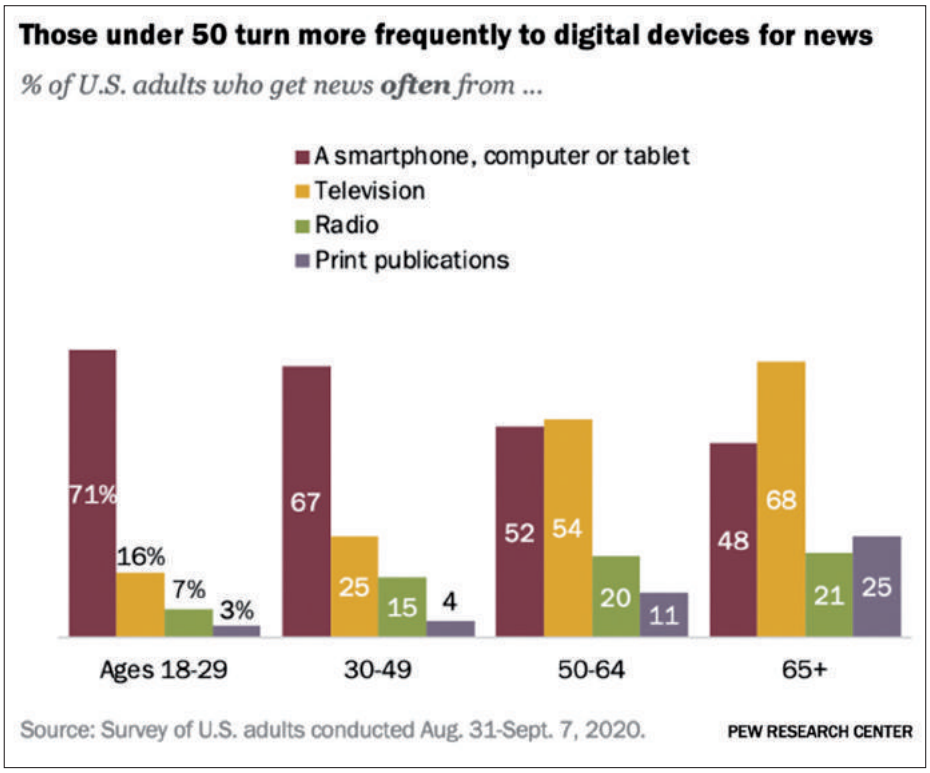


Figure 3: Platform usage by U.S. Adults [3]

The availability of large amounts of content also means that users spend less time on each item, switching from one to another, at the expense of paying close attention. This results in content that is short, superficial and very appealing, if not sensational, rather than offering in-depth analysis.

The as-a-service economy is changing people's worldview

This technology is also having an impact on how people view the world. With things increasingly becoming non-physical, younger generations have developed a different view of possession. Having 24/7 access to a service is perceived as being as good as possessing it, even for emotional assets like family pictures and movies. Many people are no longer concerned about keeping such treasures safe at home.

In fact, there is no point in seeking permanent possession of a physical good that has a digital equivalent (a video, a picture, a book, etc.) or that can be made instantly available (a shared car, look-up in a dictionary or knowledge base, etc.). The physical good occupies physical space, which is a scarce resource for many. An upfront investment is necessary to acquire it along with, in some cases, the cost of maintenance and care. Physically owned goods tend to become rapidly obsolete, rarely acquiring value in the process. The digital equivalent has none of those limitations but has one single, vital, prerequisite: connectivity for the users, and storage for the providers.

This even applies to the ownership of a car. An increasing number of (mostly urban) people no longer feel the need to own a car, but instead rent one from a car sharing company. It is hassle free. For a fixed subscription and a reasonable price per kilometre, one can use the car whenever they like without having to worry about insurance, maintenance, cleaning the car, renting a garage, … A shared car is used more intensively than a privately owned car, which means that there will be fewer cars moving around and parked in a city. Since the cars are replaced often, the user will always use a car with the lowest emission rates.

This observation is at the heart of the as-a-service economy [4], which is expanding far beyond the cloud as we know it and entering our everyday lives through the simple appendix of a connected device. The as-a-service economy materializes in apps that, once installed in the user's device, form a gateway to a gigantic and ever-growing wealth of potentially cooperating services.

Access to an almost unlimited amount of information also alters the value that is attributed to that information. When people had to go to the effort of visiting a shop or library, for example, to access a song, a movie, a newspaper article or a book, information had value. Today, with near-free access to almost limitless content and information, their perceived value is lower and people discard them more easily. When buying a newspaper, people will normally read at least part of it. With a subscription, more articles will go unread, while people are not necessarily listening to more music than they did in the past, because they find it tiring to discover it. Most websites therefore have recommender systems that suggest a small range of selected items to the user, hoping he or she will click on them, like them… and keep paying the subscription.

COVID-19 has acted as the digital transformation officer of the world

In 2020, COVID-19 spurred the transition to the as-a-service economy. At a record pace, yet more physical objects and activities were digitized.

- Money: most people have now learned how to pay cashless (and sometimes even contactless) in shops. It is easy, more hygienic and more convenient. There is little chance that cash payments will make a great comeback in the future.
- Physical meetings: experience during the pandemic has shown that in order to meet, there is no need to move a body to another location. A meeting can happen in dematerialized cyberspace instead of in a physical meeting room. Although most of us are well aware of the limitations, we also appreciate the advantages (no need to move to another location, switching between meetings takes less than one minute, it is very convenient for sharing documents, virtual meetings are easier to schedule, everybody is more equal – there are no reserved seats at the table, it is easier to do some other work if

an item on the agenda does not require one's attention, some people even manage to be in two meetings simultaneously,…). There are also clear disadvantages: most people find virtual meetings more exhausting than physical meetings, online interaction is less effective for meetings in which the participants do not know each other, and scheduling a meeting between distant timezones is a challenge, for example. One study even found that "videoconferencing hampers idea generation because it focuses communicators on a screen, which prompts a narrower cognitive focus" [13]. It is, however, clear that many meetings that used to take place in person will remain permanently substituted by virtual meetings in the future.

- Schools: although not ideal for compulsory education (from preschool to the end of high school), distance learning is suitable for knowledge transfer in higher education and lifelong learning. This is an example of the dematerialization of a classroom. It is, however, not ideal for teaching skills or attitudes and it is definitely not good on a social level, for making friends or building a network, which is also an important goal of higher education.

Users are also discovering the impact of this new economy: if they have an internet outage, or the provider of a service is down, they are stuck without any options until the service starts again. Some people are discovering that physical books are more reliable than services that can be disrupted.

Maybe surprisingly, dematerialization is not only applicable to goods and services that can be digitized: it also making material goods smaller and lighter, meaning that they require less (and lighter) physical material to manufacture [5], can be cheaper to produce, and have a smaller environmental footprint.

The impact of streaming on the environment

There have been concerns about the climate impact of the use of digital technologies, fuelled for example by a report by the *Shift Project* indicating that watching one hour of NetFlix generated 3.2 kg CO_2. Although clearly too high because this number was based on old assumptions and on an error in which bit rate was confused with byte rate, there is a lot of controversy about this kind of study because of the huge economic interests involved (big tech and entertainment companies). This study generated a worldwide discussion on the climate impact of streaming.

A follow-up study by George Kamiya from the International Energy Agency resulted in much lower numbers, but was also based on assumptions (only the operational costs were taken into account, not the investments in devices and data centres) and best estimates. The conclusions of this study are [6]:

- The energy efficiency of computing doubles every 2.7 years, and it doubles every two years for transmission networks. It is important to base an analysis on recent data. Using five-year old data will lead to an error of 400%.
- The energy consumption of hyperscale data centres has remained flat at 1% of global electricity use, while the traffic has tripled, and the workload has doubled. The increased demand has apparently been offset by efficiency improvements.
- The biggest energy cost of viewing streamed content is the viewing device itself. A smartphone is five times more efficient than a laptop, and 100 times more efficient than a 50-inch TV screen.
- According to George Kamiya, one hour of Netflix streaming consumes the following amount of energy (Table 1).

The total energy consumption is, in the first place, determined by the energy consumption of the device (smaller is better), and the resolution (smaller is better too). Depending on the mix of devices for Netflix, this leads to an average of 76.9 Wh, of which 70% is used in the device, 25% in the transmission, and 5% in the data centre.

- Expressed in terms of emissions, one hour of streaming is equivalent to 35.6 g CO_2 (based on the global average CO_2 intensity of 463 g/kWh in 2019). This is the equivalent of driving 250m (and eight times less on a smartphone). Only 30% is due to the streaming itself; the remaining 70% would still be used whatever the device was used for (like watching a broadcasted movie, or playing a game). These numbers are confirmed by other recent studies [8,9].
- The numbers also depend heavily on the origin of the electricity, and on the time of the day the streaming happens. In France, it would only be 4g CO_2 or driving 28m.
- These numbers will keep falling in the coming years due to efficiency gains in data centres, transmission networks and devices, and the decarbonization of electricity generation.
-

The conclusion is that it is very difficult to draw a definitive conclusion on such studies, in part because there are a range of commercial implications of their results. However, if we follow George Kamiya's study, the carbon footprint of streaming a movie is moderate – certainly much lower than taking a car to go to a cinema – and the gap between the two will continue to widen in the future. Given the similarity between streaming video content and videoconferencing systems, meeting virtually will always consume less energy than attending a physical meeting.

In 2021, HiPEAC decided to calculate the cost of our virtual conference. To calculate the footprint of the virtual conference, we measured the number of hours of streaming video and the number of participants per video stream. We then

Energy consumption (Wh)	TV	Laptop	Smartphone	Average
	(Wifi, 4K)	(Wifi, HD)	(4G, Auto)	-
Data centre	13.9	6	0.5	3.7
Transmission	18.8	18.3	8.5	17.7
Device	120	22	1.2	55.5
Total	**152.7**	**46.3**	**10.2**	**76.9**

Table 1: Energy consumption of 1h streaming on different platforms

estimated the energy required in the Zoom data centres and the network, and finally we made assumptions about the devices used based on information gathered from the post-conference survey. The energy consumption was calculated and then converted in to CO_2 (based on the average European carbon footprint per kWh).

The total conference generated 50 kg CO_2 or 56 g per delegate. A comparable calculation for an in-person conference (based on carbon calculators for trips), resulted in an average of 400 kg CO_2 to attend the conference, that is more than 7000 x the emissions of virtual attendance. Even if the estimates were incorrect by a factor of five, there would still be a huge reduction in emissions, by a factor of 300. For this study, we only compared operational emissions, not embedded ones (required to produce the infrastructure, be it either servers and end user equipment, or planes, trains and cars).

Hence, it is obvious that dematerialized meetings (when possible, and when effective) offer a huge decarbonization benefit. The benefits are lower in the case of streaming media. When the gain is small, we should be aware of Jevon's paradox: this states that more efficient use of a resource can lead to lower cost and therefore increased demand, undoing the effects of the efficiency gains. This seems to be the case with the streaming market, which is apparently growing faster than the efficiency gains in computing.

Conclusion

Dematerialization has been taking place for the last 40 years, and there is no reason why it would or should stop now. Dematerialized services consume fewer resources than physical ones and are hence are more sustainable. They are also cheaper and are available 24/7. However, there are many concrete challenges to work on. The environmental footprint (including energy use) should be further reduced and existing solutions should be improved; for example, the tools for virtual meetings, lectures and conferences that do not yet offer the immersive experience that physical events offer.

In addition, there are still huge dematerialization opportunities: for large events, for tourism, and for museum visits. How cool would it be to "visit" a city or a museum with an interactive video guide who shows you all the interesting places or objects, and where you can determine how the tour will evolve? This could happen at home, but perhaps also in a virtual tourism facility with a fully immersive experience including the sounds, the smells, the burning sun, and maybe a meal with local food afterwards. After a while, the brain might no longer be able to make the distinction between memories created by the realistic virtual visit, and a real visit.

References

[1] Felix Richter, "Streaming Drives Global Music Industry Resurgence", 2020, https://www.statista.com/chart/4713/global-recorded-music-industry-revenues/

[2] Adrian Pennington, "Looks like 230 billion will be spent on content in 2022", 2022, https://amplify.nabshow.com/articles/230-billion-will-be-spent-on-content-in-2022/

[3] Elisa Shearer, "More than eight-in-ten Americans get news from digital devices", 2021, https://www.pewresearch.org/fact-tank/2021/01/12/more-than-eight-in-ten-americans-get-news-from-digital-devices/

[4] "Servitization in Industry", Gunter Lay (editor), 2014, Springer, ISBN 978-3-319-06934-0, doi: 10.1007/978-3-319-06935-7.

[5] "Dematerialization", https://en.wikipedia.org/wiki/Dematerialization_(economics)#

[6] George Kamiya, "Factcheck: What is the carbon footprint of streaming video on Netflix?", 2020, https://www.carbonbrief.org/factcheck-what-is-the-carbon-footprint-of-streaming-video-on-netflix

[7] Brad Stone, "Amazon Erases Orwell Books From Kindle", 2009, https://www.nytimes.com/2009/07/18/technology/companies/18amazon.html

[8] "The carbon footprint of (nearly) everything", https://www.viessmann.co.uk/company/blog/the-carbon-footprint-of-nearly-everything

[9] Louis-Philippe P.-V.P. Clément, Quentin E.S. Jacquemotte, Lorenz M. Hilty, Sources of variation in life cycle assessments of smartphones and tablet computers, Environmental Impact Assessment Review, Volume 84, 2020.

[10] Niamh McIntyre , Rosie Bradbury , Billy Perrigo, "Behind TikTok's boom: A legion of traumatised, $10-a-day content moderators", 2022, https://www.thebureauinvestigates.com/stories/2022-10-20/behind-tiktoks-boom-a-legion-of-traumatised-10-a-day-content-moderators

[11] Casey Newton, "The Trauma Floor: The secret lives of Facebook moderators in America", 2019, https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona

[12] Miriah Steiger, Timir J. Bharucha, Sukrit Venkatagiri, Martin J. Riedl, Matthew Lease, "The Psychological Well-Being of Content Moderators: The Emotional Labor of Commercial Moderation and Avenues for Improving Support", CHI '21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, May 2021,https://dl.acm.org/doi/10.1145/3411764.3445092

[13] Melanie S Brucks, Jonathan Levav, "Virtual communication curbs creative idea generation", Nature, June 2022, https://doi.org/10.1038/s41586-022-04852-5

Koen De Bosschere is a professor in the electronics department of Ghent University, Ghent, Belgium.

Marc Duranton is a researcher in the research and technology department at CEA (the French Atomic Energy Commission) and the coordinator of the HiPEAC Vision 2023.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2023.

Cite as: K. De Bosschere and M. Duranton. Everything as a service. In M. Duranton et al., editors, HiPEAC Vision 2023, pages 225-230, Jan 2023.

DOI: 10.5281/zenodo.7462005

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

© HiPEAC 2023

Glossary

ACO	Ant colony optimization
AES	Advanced Encryption Standard
AGI	Artificial general intelligence
AI	Artificial intelligence
ALU	Arithmetic logic unit
Ambient intelligence	http://echo.iat.sfu.ca/library/philips_02_ambientintelligence.pdf
ANPR	Automatic number-plate recognition
API	Application programming interface
ASIC	Application-specific integrated circuit
ASLR	Address space layout randomization
B2C	Business-to-consumer
Big Data	https://en.wikipedia.org/wiki/Big_data
CAGR	Compound Annual Growth Rate
CD	Compact Disc
CE	Computer engineering
CGI	Computer-generated imaging
Cloud computing	https://en.wikipedia.org/wiki/Cloud_computing
CMOS	Complementary metal-oxide semiconductor
COVID-19	https://en.wikipedia.org/wiki/COVID-19
CPS	Cyber-physical systems
CPU	Central processing unit
CRUD	Create, Read, Update, Delete
CS	Computer science
CSA	Coordination and support action
CXL	Compute Express Link
Cyberspace	https://en.wikipedia.org/wiki/Cyberspace
DALL·E	https://openai.com/dall-e-2/
DDoS	Distributed denial-of-service
Deep fake	https://en.wikipedia.org/wiki/Deepfake
Deep learning	https://en.wikipedia.org/wiki/Deep_learning
DevOps	https://en.wikipedia.org/wiki/DevOps
Digital Twins	https://en.wikipedia.org/wiki/Digital_twin
DIH	Digital Innovation Hub
DLT	Distributed Ledger Technology
DNN	Deep neural network
DPU	Data processing unit
DRM	Digital rights management
DSP	Digital signal processor
EC	European Commission
ECA	European Chips Act
EDA	Electronic design and automation
EDRi	European Digital Rights
EU	European Union
EUV	Extreme ultraviolet
FET	Field-effect transistor
FHE	Fully homomorphic encryption
FLOPS	Floating point operations per second
FMCG	Fast-moving consumer goods
FOSS	Free and open-source software
FPGA	Field-programmable gate array
GAN	Generative Adversarial Network
GDP	Gross domestic product
GDPR	General Data Protection Regulation (EU)
GHG	Greenhouse gas
GPGPU	General-Purpose GPU
GPT-3	Generative Pre-trained Transformer 3
GPU	Graphics processing unit
HCPS	Human-CPS
HDL	Hardware description language
HPC	High-performance computing
HSML	Hyperspace Modeling Language, see https://spatialwebfoundation.org/protocols/contracts/
HSTP	Hyperspace Transaction Protocol, see https://spatialwebfoundation.org/protocols/standards/
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
IaC	Infrastructure as code
IC	Integrated circuit
ICT	Information and communication technology
IDE	Integrated Development Environment
IGZO	Indium gallium zinc oxide
IMC	In (or near) memory computing
IoT	Internet of things
IP	Intellectual property
IPU	Image processing unit
ISA	Instruction set architecture
IT	Information technology
JU	Joint undertaking
KDT	Key Digital Technologies
KET	Key enabling technology
LCA	(Product) Life cycle assessment
LCD	Liquid-crystal display
LED	Light-emitting diode
LLM	Large language model
M2M	Machine-to-machine communication

MAS	Multi-agent systems
Metaverse	For industrial metaverse, see https://www.technologyreview.com/2022/12/05/1063828/the-industrial-metaverse-a-game-changer-for-operational-technology/
MIMD	Multiple instruction, multiple data
ML	Machine learning
Moore's law	https://en.wikipedia.org/wiki/Moore%27s_law
MRAM	Magnetic random access memory
NFP	Natively flexible processor
NISQ	Noisy intermediate-scale quantum
NLP	Natural language processing
NPU	Neural processing unit
NSA	US National Security Agency
OADR	Old age dependency ratio
OCP	Open Compute Project
OECD	Organisation for Economic Co-operation and Development
OLED	Organic light-emitting diode
OPD	Organic photodiode
OS	Orchestration Systems
OSI	Open Source Initiative
OSS	Open-source software
PC	Personal computer
Photonics	https://en.wikipedia.org/wiki/Photonics
PII	Personally identifiable information
PISQ	Perfect intermediate-scale quantum computing
PQC	Post-quantum cryptography
PSO	Particle swarm optimization
QC	Quantum computing
QT	Quantum technologies
Qubit	https://en.wikipedia.org/wiki/Qubit
R&D	Research and development
RaaS	Ransomware as a Service
RF	Radio frequency
ROI	Return-on-investment
ROM	Read-Only Memory
SaaS	Software as a service
SDK	Software development kit
SiC	Silicon carbide
SIMD	Single instruction, multiple data
SME	Small and medium-sized enterprises
SoC	System on a chip
SOI	Silicon on insulator
SoS	Systems of Systems
Spatial Web	https://spatialwebfoundation.org/
Spintronics	https://en.wikipedia.org/wiki/Spintronics
SRAM	Static random access memory
STEM	Science, technology, engineering and mathematics
SWOT	Strengths, weaknesses, opportunities and threats
TCB	Trusted computing base
TFT	Thin-film transistor
TinyML	https://www.tinyml.org/
TLS	Transport Layer Security
TOF	Time of Flight
TPU	Tensor Processing Unit
TransContinuum	https://www.etp4hpc.eu/transcontinuum-initiative.html
TRL	Technology readiness levels
UCIe	Universal Chiplet Interconnect Express
UN	United Nations
US	United States
VC	Venture capital
VCU	Video compression unit
VFX	Visual effects software
VLSI	Very Large Scale Integration
VPU	Visual processing units
VRAM	Video Random Access Memory
VUCA	Volatile, uncertain, complex and ambiguous
WEEE	Waste electronic and electrical equipment
WWvW	World Wide Voice Web
WWW	World Wide Web
XaaS	Everything as a Service
xPU	Special purpose processing unit
XR	Extended reality

Process

The HiPEAC Vision presents the trends that have an impact on the high-performance and embedded architecture, compilation and related communities. Recently, it evolved into a more agile document consisting of standalone, independent articles that underpin its recommendations. The articles and recommendations are based on information collected through a number of channels.

- Meetings with teachers and industrial partners at the ACACES 2022 Summer School;
- A survey circulated to all HiPEAC members, which received 36 responses;
- Several consultation meetings:
  - 4 May 2022: Consultation meeting on future safety-critical collaborative systems,
  - 13 May 2022: Consultation meeting on "the convergence of AI, IoT, Edge computing, CPS, continuum of computing, x-verse: the next computing paradigms",
  - 16 May 2022: Consultation meeting on edge / federation of resources / fog / swarm + "Guardian Angels",
  - 31 May 2022: Consultation meeting on emerging paradigms.
  - 14 November 2022 (online): EU-US workshop on the continuum of computing (technical / scientific focus).
  - 16 November 2022 (online): EU-US workshop on the continuum of computing (programmatic focus)
- Six in-person editorial board meetings and 13 by Webex;
- Exchanges with other organizations, such as ECS/KDT, AIOT and ETP4HPC;
- Feedback from presentation of the HiPEAC Vision 2021 at several conferences and workshops (SC21 for example) and from exchanges with the DG CNECT;
- Feedback from a presentation of an early version of the HiPEAC Vision 2023 at CONNECT University (streamed to 125 people) on 7 December 2022.

Note: this document is called a "Vision" because it is the result of the interpretation of the trends and directions as seen by the HiPEAC community. As HiPEAC has no direct power to enforce the recommendations, the timeline associated with the potential implementation of the recommendations is uncertain; this is why the document is not a roadmap per se.

Acknowledgements

This document is based on the valuable input of HiPEAC members. The editorial board, composed of Marc Duranton (CEA), Koen De Bosschere (Ghent University), Bart Coppens (Ghent University), Christian Gamrat (CEA), Madeleine Gray, Harm Munk (TNO), Thomas Hoberg (ATOS), Charles Robinson (Thales), Tullio Vardanega (University of Padua) and Olivier Zendra (INRIA), would like to thank all the authors of the individual articles: Philippe Sayegh and Capm Petersen (VERSES.io), Francesca Palumbo, Raquel Lazcano and Daniel Madroñal (Università degli Studi di Sassari), Jason O. Hallstrom (National Science Foundation), Koen Bertels (QBee / Ghent University), Gaël Blondelle and Philippe Krief (Eclipse Foundation), Patrick Blouet (STMicroelectronics), Thomas Ernst (CEA), Jean-Pierre Raskin (Université catholique de Louvain) and Lieven Eeckhout (Ghent University).

The editorial board would also like to thank Vicky Wandels (Ghent University), Eneko Illarramendi (Ghent University) and Catherine Roderick, for their useful comments and their support.

In addition, the editorial board would like to thank the following people for their valued contributions to specific articles: Rajendra Akerkar (Western Norway Research Institute), Djamila Aouada (Université du Luxembourg), Alessandra Bagnato (Softeam), Miklós Györffi (Eötvös Loránd Research Network), Michael Henshaw (Loughborough University), Peter Gorm Larsen (Aarhus University), Carles Hernandez Luz (Universitat Politècnica de València), Hugo Daniel Macedo (Aarhus University), Claudio Pastrone (LINKS Foundation), Peter Popov (City University London), Claudio Sassanelli (Politecnico di Bari), Marcus Völp (Université du Luxembourg), Thorsten Weyer (Technical Hochschule Mittelhessen), Sebastian Feld (Delft University of Technology), Laurent Larger (FEMTO-ST Institute), Menno Lindwer (GrAI Matter Labs), Ioan Lucian Prejbeanu (SPINTEC), Carlo Reita (CEA-LETI, Sanaz Mahmoodi Takaghaj (Pennsylvania State University), Emre Ozer (PragmatIC), Yulia Sandamirskaya (Intel Labs Munich), Mathilde Billaud, (Fraunhofer), David Bol, (UCLouvain), Thierry Baron (CNRS), Patrice Christmann (BRGM), Marie Garcia-Bardon (imec), Tapani Jokinen (Fraunhofer), Francois Martin (CEA-Leti), Bertrand Parvais, (imec), Karine Samuel (UGA), Lutz Stobbe (Fraunhofer) and Olivier Vergeynst (Green IT).

Special thanks to the Belgian comic artist Arnulf who provided a light-hearted, often satirical take on the content of this Vision document.

HiPEAC Vision 2023

With technology evolving faster than humans can adapt, while geopolitical and environmental pressures intensify, it has never been more important for Europe to identify clear priorities for computing. The HiPEAC Vision 2023 explores the urgency of these themes in terms of six races for leadership, both with other world regions and against time.

Global recommendations

- **Break silos** to gain a holistic view, which is necessary for global optimizations. Promote **collaboration between teams**, launch joined **project calls to create synergies between domains**, establish **cross-disciplinary European competence centres** and promote **open source (hardware and software)**.
- **Develop tooling for cross-disciplinary and multi-dimensional challenges**. A global view is necessary for enhanced optimization, but it is complex. **AI should be used as a "helper"** to propose solutions.
- **Develop trustable runtime orchestrators** able to **manage complex systems**. In some cases, systems will have to select a large number of options "in real-time". This will require the development of trustable orchestrators which are loyal to their users.

Technology races

The race for the "next web"

Increasing amounts of data are being generated by machines that are interconnected and linked with the physical world to provide new optimizations and services. The "next web" will have to take this into account and smoothly integrate the "web of machines" with the "web of humans", executing computations where they make most sense and taking into account new non-functional requirements. Europe should develop, standardize, test and validate such related technologies running on the continuum of computing, including trustable orchestrators to manage this complexity and serve personal users as well as industrial users by enabling an European industrial metaverse.

The race for artificial intelligence (AI)

Power and data-hungry large language models are delivering incredible results. At the other end of the spectrum, it is increasingly important to have efficient AI systems at the edge. Europe should focus on AI solutions that enhance its strengths in the embedded sector, while also keeping up with developments in large models, and prepare for their squeezing into edge devices. AI should be used judiciously as a "helper", although AI-generated results should always be validated before use.

The race for innovative and new hardware

Emerging paradigms, such as quantum computing, neuromorphic computing, spintronics and photonic devices, appear to offer efficient solutions for particular problem areas, at a much lower energy consumption than current devices. Research in this area should be coupled to practical applications, environmental impacts and should tackle heterogeneous integration with current technologies (in hardware and software).

The race for cybersecurity

Computing systems will only be practically usable if they are secure. Critical infrastructure and supply chains should have both their hardware and software components hardened against cyberattacks. The EU should build on its strengths in cybersecurity, including for post-quantum systems, and should broaden mandatory security and privacy, EU-based, audit and certification of IT systems.

The race for sovereignty

Europe should strive for digital sovereignty, both in hardware and in software and should promote open source. It is also vital to keep investing in talent, research and innovation, and in a more entrepreneurial ecosystem.

The race for sustainability

Designers should use a full lifecycle assessment when designing new computing systems, and computer architects should focus on embodied energy. Europe should develop new economic models that take lifecycle environmental costs into account and should continue searching for ways to develop environmentally friendlier (digital) goods and services e.g. by dematerialization.

hipeac.net/vision

9 789078 427032