



HAL
open science

The HiPEAC Vision 2019

Marc Duranton, Koen de Bosschere, Bart Coppens, Christian Gamrat,
Madeleine Gray, Harm Munk, Emre Ozer, Tullio Vardanega, Olivier Zendra

► **To cite this version:**

Marc Duranton, Koen de Bosschere, Bart Coppens, Christian Gamrat, Madeleine Gray, et al. (Dir.). The HiPEAC Vision 2019. Duranton, Marc; De Bosschere, Koen; Coppens, Bart; Gamrat, Christian; Gray, Madeleine; Munk, Harm; Ozer, Emre; Vardanega, Tullio; Zendra, Olivier. HiPEAC CSA, pp.178, 2019, 978-90-90-31364-1. hal-02314184

HAL Id: hal-02314184

<https://inria.hal.science/hal-02314184>

Submitted on 11 Oct 2019

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.



HiPEAC Vision 2019

HIGH PERFORMANCE AND EMBEDDED ARCHITECTURE AND COMPILATION



Editorial board:

**Marc Duranton, Koen De Bosschere, Bart Coppens,
Christian Gamrat, Madeleine Gray, Harm Munk, Emre Ozer,
Tullio Vardanega, Olivier Zendra**

This document was produced as a deliverable of the H2020 HiPEAC CSA under grant agreement 779656.

The editorial board is indebted to Dr Max Lemke and to Sandro D'Elia of the Technology & Systems for Digitising Industry unit of the Directorate-General for Communication Networks, Content and Technology of the European Commission for their active support to this work.

Design: www.magelaan.be

Source cover picture: © Alphaspirt | Dreamstime.com

© January 2019 HiPEAC

ISBN 978-90-90-31364-1

FOREWORD

This is the 7th edition of the HiPEAC Vision. The first one was issued in 2008. Over this decade, the performance of computing devices has increased dramatically, despite the increasing limitations of silicon technology. Computing technology has also had a profound impact on our way of life: 10 years ago, smartphones were virtually non-existent; today, they are such an important part of our existence that people feel uncomfortable without them and, in some cases, even find they have become addicted to them. Social networks have – for better or worse – changed the way we interact and share our lives with one another, with privacy seeming less and less important.

The big vertical companies (Google, Apple, Facebook, Amazon, Microsoft = GAFAM - and Baidu, Alibaba, Tencent, Xiaomi = BATX) are now the most profitable in the world, having overtaken energy companies. They are now even developing their own integrated circuits and cover the complete value chain from hardware to services. Thanks to ICT, new business models have sprung up in different domains: transportation (Uber), hotels (Airbnb), goods (Amazon, Alibaba), media (Netflix, Spotify, Deezer) etc.

The PC is now a commodity and its market is eroding due to the omnipresence of the more and more powerful smartphones and tablets. Complexity and cost are higher and higher for making high-end chips; large markets, like smartphones, are driving the industry for now, but this market is starting to saturate and new markets are sought for further growth. Programming has

changed from writing a complete application in C to gluing together libraries of various functionalities with interpreted languages such as Python.

Last but not least, recent progress in artificial intelligence, especially deep learning, is allowing the computer to move out of cyberspace and interact with the real world. Deep learning allows computers to see, hear and understand, enabling them to morph from their original form of grey boxes with keyboards and screens to new forms like cars, assistants in loudspeakers, and other devices integrated in the fabric of our life. If the expectations about AI become a reality, it will have a drastic impact on our civilization, including HiPEAC domains, because “intelligent” computers could help us make better hardware, software, operating systems and applications. However, new challenges are also opening up, such as how to convince people that they can trust these new machines, or how to guarantee that they will do what they are supposed to do, respecting safety, security and energy constraints? ICT systems are now heterogeneous, distributed and so complex that it is difficult for the human brain to comprehend them.

Each time the editorial board start a new HiPEAC Vision, we think it will be a “small” increment over the previous one. But we always discover that our field is evolving so fast and new facets of ICT are emerging so rapidly that we end up dealing with more and more topics. This release is no exception and we hope you will have as much pleasure reading it as we had producing it.

Figure 1: The HiPEAC vision documents.



CONTENTS

FOREWORD	1	2. PART 2: RATIONALE	15
INTRODUCTION	4	2.1 INTRODUCTION:	15
1. PART 1: KEY MESSAGES	9	2.1.1 STRUCTURE OF THE DOCUMENT	15
1.1 EFFICIENCY	9	2.2 BUSINESS DIMENSIONS	16
1.1.1 DEVELOP DOMAIN-SPECIFIC ACCELERATORS AND DESIGN KNOWLEDGE	10	2.2.1 BUSINESS TRENDS	16
1.1.2 DESIGNING HARDWARE PLATFORMS IS ONLY ECONOMICALLY VIABLE IF IT CAN BE AUTOMATED	10	2.2.1.1 THE AI BANDWAGON	16
1.1.3 INTEGRATION OF ACCELERATORS AND OTHER ELEMENTS IN AN EFFICIENT SYSTEM SHOULD BE FACILITATED.	10	2.2.1.2 HUMAN IN THE LOOP	24
1.1.4 SOFTWARE SHOULD BE WRITTEN BY SOFTWARE, NOT BY PROGRAMMERS	10	2.2.1.3 THE CONTINUUM: CLOUD, FOG AND EDGE COMPUTING	27
1.1.5 ELEMENTS IN A SYSTEM, OR IN A SYSTEM OF SYSTEMS, SHOULD BE ABLE TO ADAPT TO THEIR ENVIRONMENT DYNAMICALLY.	10	2.2.1.4 POST-EXASCALE HPC	28
1.2 CREDIBILITY, SECURITY, SAFETY AND ACCEPTABILITY	11	2.2.1.5 CYBER-PHYSICAL SYSTEMS AND THE IOT	29
1.2.1 EUROPE SHOULD INVEST IN TOOLS AND TECHNOLOGIES THAT ALLOW US TO CREATE SECURE AND SAFE SOLUTIONS	11	2.2.1.6 VIRTUAL, AUGMENTED AND MIXED REALITY	31
1.2.2 EUROPE SHOULD DEVELOP SYSTEMS THAT CAN BE UNDERSTOOD ENOUGH	11	2.2.2 BUSINESS MODELS	31
1.3 THE POSITION OF EUROPE	11	2.2.2.1 RENTING INSTEAD OF BUYING	32
1.3.1 EUROPE SHOULD BE A LEADER IN "INTELLIGENCE AT THE EDGE" SOLUTIONS AND COGNITIVE CYBER-PHYSICAL SYSTEMS	11	2.2.2.2 VERTICALIZATION AND DOMINANCE OF GLOBAL PLATFORMS (GAFAM + BATX)	34
1.3.2 EUROPE SHOULD CONSIDER ICT DOMAINS AS A CONTINUUM, AND NOT SILOS	12	2.2.2.3 OPEN SOURCE	35
1.3.3 EUROPE SHOULD LEAD ON THE USE OF COLLECTIVE DATA	12	2.2.2.4 CREATING ECOSYSTEMS	38
1.3.4 EUROPE SHOULD BE A LEADER IN ENERGY EFFICIENT, SUSTAINABLE AND LONG LIFETIME ICT	12	2.2.3 BUSINESS DOMAINS AND OPPORTUNITIES	39
1.3.5 EUROPE SHOULD DEVELOP SOLUTIONS USING MATURE TECHNOLOGY NODES	12	2.2.3.1 AUTOMOTIVE: THE NEXT FRONTIER?	39
1.3.6 CONTINUE RESEARCH ON POST-CMOS TECHNOLOGIES WHILE MAINTAINING A LINK WITH EXISTING ICT TECHNOLOGIES.	13	2.2.3.2 MEDICAL AND WELLBEING	40
1.3.7 DEVELOPMENT OF INNOVATIVE ALTERNATIVE ARCHITECTURES	13	2.2.3.3 GAMING: TESTBED FOR CONSUMER ADVANCED TECHNOLOGIES	41
1.4 SOCIETY	13	2.2.3.4 DIGITAL TWINS: MASTERING REALITY	43
1.4.1 DIGITAL ETHICS SHOULD GUIDE US TO THE FUTURE	13	2.3 REQUIREMENTS FOR ACCEPTABILITY	44
1.4.2 EMPLOYMENT	13	2.3.1 THE TRUSTABLE COMPUTER	44
1.4.3 DIGITAL SKILLS ARE THE FUEL OF INNOVATION	13	2.3.1.1 THE SAFE COMPUTER	44
1.4.4 SUSTAINABILITY	14	2.3.1.2 THE SECURE COMPUTER	45
		2.3.1.3 THE EXPLAINABLE COMPUTER	50
		2.3.2 THE ENERGY CHALLENGE	52
		2.3.2.1 FOR DATA CENTRES	52
		2.3.2.2 FOR CONNECTIVITY	53
		2.3.2.3 FOR SYSTEMS	54
		2.4 TECHNOLOGY DIRECTIONS	57
		2.4.1 TECHNOLOGY	57
		2.4.1.1 LIMITATIONS OF THE CURRENT CMOS TECHNOLOGY AND SILICON ROADMAP	57
		2.4.1.2 3D STACKING: AN ANSWER TO CMOS SCALABILITY CHALLENGES	61
		2.4.1.3 CRYOGENIC COMPUTING	63
		2.4.1.4 PHOTONICS FOR COMPUTING	64
		2.4.1.5 QUANTUM COMPUTING	66
		2.4.2 EMERGING TECHNOLOGIES: BEYOND SILICON	71
		2.4.2.1 FLEXIBLE ELECTRONICS	72
		2.4.2.2 SYNTHETIC BIOLOGY	73
		2.4.2.3 OTHER NEW MATERIALS	76
		2.4.3 ARCHITECTURE: HETEROGENEITY, ACCELERATORS AND IN-MEMORY COMPUTING	78
		2.4.3.1 MORE SPECIALIZATION THROUGH ACCELERATORS	78

2.4.3.2	NEAR/IN MEMORY COMPUTING	79
2.4.3.3	HW/SW CODESIGN	80
2.4.4	COMMUNICATION AND NETWORKING TRENDS	80
2.4.4.1	WIRED: FROM BETWEEN DIES TO BETWEEN RACKS	80
2.4.4.2	WIRELESS	82
2.4.4.3	SATELLITE COMMUNICATIONS	83
2.4.5	STORAGE TRENDS	84
2.4.5.1	VOLATILE MEMORIES	84
2.4.5.2	NON-VOLATILE MEMORIES	84
2.4.5.3	FUTURISTIC STORAGE	87
2.4.6	COMPUTATIONAL MODELS	87
2.4.6.1	NEUROMORPHIC COMPUTING	87
2.4.6.2	RESERVOIR COMPUTING	89
2.4.6.3	AI BEYOND DEEP LEARNING	90
2.5	SYSTEM-LEVEL DIRECTIONS	92
2.5.1	THE CONTINUUM OF COMPUTING	92
2.5.1.1	OPEN SOFTWARE ARCHITECTURE	94
2.5.1.2	SOFTWARE COMPOSITION	95
2.5.2	SOFTWARE IMPLEMENTATION: THE LIMITATIONS OF TRADITIONAL PROGRAMMING	96
2.5.2.1	CONTEXT AND INTRODUCTION	96
2.5.2.2	TECHNOLOGY TRENDS	96
2.5.2.3	THE OVERARCHING CHALLENGE: MASTERING COMPLEXITY	97
2.5.2.4	SOUGHT ENHANCEMENTS: ASSERTING CORRECTNESS	99
2.5.2.5	SOUGHT ENHANCEMENTS: ACCOMMODATING LEGACY, REUSABILITY AND EVOLUTION	99
2.5.2.6	SOUGHT ENHANCEMENTS: SECURITY, RESILIENCE, TRUST	100
2.5.2.7	SOUGHT ENHANCEMENTS: PREDICTABILITY, SAFETY, AND CONFORMANCE WITH SPECIFICATIONS	101
2.5.2.8	SOUGHT ENHANCEMENTS: BALANCING EFFICIENCY AND PERFORMANCE WITH PORTABILITY	101
2.5.2.9	SOUGHT ENHANCEMENTS: INCREASING PRODUCTIVITY FOR FASTER, CHEAPER AND BETTER PRODUCTS	102
2.5.3	SOFTWARE IMPLEMENTATION: TIME TO REINVENT PROGRAMMING	102
2.5.3.1	INTRODUCTION	102
2.5.3.2	NON-FUNCTIONAL PROPERTIES AS FIRST CLASS CITIZENS	102
2.5.3.3	BETTER ABSTRACTION AT BOUNDARIES	103
2.5.3.4	THE NEW PROGRAMMING LANGUAGES	103
2.5.3.5	NEW DOMAIN SPECIFIC LANGUAGES	104
2.5.3.6	ATTENUATING THE HUMAN FACTOR: COMPUTER PROGRAMS GENERATING PROGRAMS	105
2.5.4	SMART DESIGN TOOLS	107
2.5.5	THE OPPORTUNITIES AHEAD: THE SOFTWARE ROADMAP	108

2.6	THE SOCIETAL DIMENSION	110
2.6.1	IMPACT OF COMPUTING TECHNOLOGY ON SOCIETY	110
2.6.1.1	PRIVACY EROSION	111
2.6.1.2	FAKE INFORMATION	112
2.6.1.3	DIVIDE AND CONQUER	114
2.6.2	IMPACT OF COMPUTING TECHNOLOGY ON PEOPLE	115
2.6.3	COMPUTING TECHNOLOGY AND THE FUTURE JOB MARKET	117
2.6.4	COMPUTING TECHNOLOGY AND FUTURE OF EDUCATION	120
2.6.5	COMPUTING TECHNOLOGY AND THE FUTURE OF EUROPE	121
2.6.5.1	HIGH-PERFORMANCE COMPUTING	121
2.6.5.2	SECURITY	123
2.6.6	COMPUTING TECHNOLOGY AND PLANET EARTH	124
2.6.7	THE NEED FOR DIGITAL ETHICS	131
2.7	THE POSITION OF EUROPE IN THE WORLD	135
2.7.1	EUROPEAN POSITION (SWOT)	135
2.7.1.1	STRENGTHS	136
2.7.1.2	WEAKNESSES	139
2.7.1.3	OPPORTUNITIES	142
2.7.1.4	THREATS	143
2.7.1.5	CONCLUSION	145
3.	GLOSSARY	147
4.	REFERENCES	154
5.	PROCESS	171
6.	ACKNOWLEDGEMENTS	172
7.	HIGHLIGHTS OF THE HIPEAC VISION 2019	173

INTRODUCTION

In the introduction of the letter sent to investors for the year 2017 [257], the founders of Alphabet quoted A Tale of Two Cities by C. Dickens:

*“It was the best of times,
it was the worst of times,
it was the age of wisdom,
it was the age of foolishness,
it was the epoch of belief,
it was the epoch of incredulity,
it was the season of Light,
it was the season of Darkness,
it was the spring of hope,
it was the winter of despair ...”*

This quote is indeed appropriate to the current evolution of information and communication technology (ICT) and its impact on society. ICT offers an ever-increasing range of opportunities, but it might also be a threat to our current way of life. The semiconductor technology CMOS, which fuelled the digital era, is **nearly out of steam**, and will require gigantic investments to allow further improvement to its performance, while no other technology is on the horizon as a possible replacement, at least for the foreseeable future.

Artificial intelligence (AI) is now at the **top of the hype curve**, and everybody is claiming it for their domain: from high-performance computing (HPC) to the internet of things (IoT) (renamed “intelligence of things” in this instance), everyone is riding the AI wave. The term is an unfortunate choice, because it sounds threatening to some people, and the actual systems are far from being “intelligent” in the human sense. In fact, artificial general intelligence (AGI) is what really scares people. Another term (“cognitive?”) might be more appropriate: this covers many different technologies, both novel ones and also the more traditional approaches (operational research, Bayesian, etc...).

Cognition opens doors to a lot of things, such as awareness of the context, of the environment, considering the content and responding appropriately to the environment, including humans. So-called AI is also **changing the way we interact with computers**, incorporating voice, gestures, images and so on, and it could improve the efficiency of many processes, including hardware and **software development**, which are becoming so complex that optimal solutions are eluding the capacity of human brains.

We are finally seeing the emergence of what the HiPEAC roadmap

proposed in 2009: **“keep it simple for humans, and let the computer do the hard work.”** (hipec.net/v10). While already commercially available for mechanical engineering, asking machines to explore gigantic spaces of parameters to find a good solution to a given problem is a current research topic for hardware design (“auto-design?”), for software programming – sometimes referred to as software 2.0, and even for finding the parameters of the AI techniques themselves (cf. automatic machine learning, or “*auto-ML*”, where reinforcement learning, among other techniques, is used to design deep-learning networks).

Auto-ML potentially saves development time, improves results and might enable autonomous systems to use ML as a component. It will also have an impact on making AI and data science available to everyone, with the major drawback that people may use it without understanding it!

AlphaZero shocked the Go and chess communities, not only because of *“the ease with which #AlphaZero crushed human players, but the ease with which it crushed human AI researchers, who’d spent decades hand-crafting ever better chess software”* [163]. In addition to this “*epoch of belief*”, AI and related techniques open a “*spring of hope*” that we will eventually devise solutions able to **help us to improve efficiency** both for machines (managing their complexity, improving their efficiency in operations per watt) and humans (productivity in all domains, including software), and that this will contribute to **solving** a vast range of **societal challenges**.

The current success of deep learning relies of course on the algorithmic side, but also on the crucial availability of large amounts of labelled data and extremely powerful computing infrastructures. We observe that this is leading to a shift in innovation from academia to private organizations that possess both of those resources and that may hire the best experts in the field. For example, the “start-up” SenseTime in China has 8000 graphics processing units (GPUs) for its algorithmic training [402] and access to the wealth of data provided by its direct and indirect customers.

Hopefully, ICT will not be the catalyst of a “*season of Darkness*” as forecast in many sci-fi movies and books. If AI-related techniques even partially fulfil expectations, this might lead to a revolution that some have compared to the industrial revolution; but this will happen very rapidly, perhaps faster than our society can adapt to. Jobs will be destroyed in some sectors and created in

other fields, with all of the social consequences that changes of this magnitude bring about. Even our basic rights like privacy and liberty of choice may be at risk, as shown by the “Cambridge Analytica” scandal.

Techniques to manipulate people can nowadays be developed without people even noticing that they are being manipulated. **Fake news is indistinguishable from real facts.** Voice assistants may make our lives easier, but they do a lot of things in the background that we do not directly see. Similarly, social networks and tools that analyse user’s behaviour, “**lock us into our own bubbles**” and perpetuate our habits. People increasingly receive information through a single information channel, the internet: radios are shutting down and moving over to internet streaming, and the broadcast radio spectrum is being repurposed to provide internet connectivity by 3G, 4G and soon 5G. Internet providers have the potential to check (and control/modify) what anyone is accessing; as one example, e-books can be remotely erased from one’s e-reader.

Nonetheless, artificial intelligence can be genuinely helpful, releasing humans from the need to perform dull, menial and annoying tasks. The presentation of Google Duplex reserving a table at a restaurant is spectacular, and listening to the exchange, it is hard to tell who is the human and which is the machine. However, in the ICT domain, there is a quest for detailed explainability, and here a balance needs to be struck, ensuring acceptability without blocking progress. The key point here is that the system should be **explainable up to the level** that its intended users **need for their understanding** and to **allow trust** and confidence to be built on it. Evidently, this level varies depending on the person. In practice, trust is gained by experience (“I have used it several times, and it has worked every time”) or when a trusted party shows us that we can trust the system (which entails the problem of certification, and validation).

Trust is key for the social acceptance of the innovations created by ICT. This is a **major challenge that encompasses security, privacy and safety**, as current ICT systems can already directly control physical and potentially lethal systems such as cars, planes, etc. As humans are not always trustworthy, security (which is caused by malevolent humans) is a major challenge for ICT systems, as more is now at stake due to the omnipresence of modern-day ICT in our lives. Computing has become such a powerful commodity that it is now time to invest in digital ethics as a discipline, and to make sure that all professionals in computing receive basic training in it.

In recent times, there has also been a tendency for more and more countries to become more inward-looking, with potential consequences for global trade: if key hardware components are no longer freely available on the market, an increasing number of countries may be (and some already are) compelled to build their own ICT solutions, e.g. processors. Following this trend, the

European Union (EU) should **regain sovereignty** and self-sufficiency in ICT. Initiatives like the European Processor Initiative [284] are along these lines. The EU should build capacity to create **smart systems**, especially **safe and secure** ones (where security and safety should be addressed upfront, instead of being an afterthought), drawing upon its own knowhow in embedded systems and its automotive and aerospace markets. Open source allows accessibility by everyone and the development of innovative solutions, both hardware and software. In the HiPEAC domain, Linux, GCC and RISC-V are good examples of the interest of the Open Source community.

As shown in the 2017 HiPEAC Vision (hipeac.net/v17), ICT is expanding from cyberspace to interact with us directly, for example in self-driving cars, driverless underground and overground trains, factories and even cities. We are now in the **era of cyber-physical systems (CPS)**, and these systems will be increasingly enhanced with artificial intelligence, so that we could call them **cognitive cyber-physical systems**, or C²PS. This evolution will further increase the constraints of trust and autonomy. It will swing the pendulum more towards having “intelligence at the edge” rather than just in the cloud, as in the “big data” era.

Of course, clever **collaboration between devices**, sharing knowledge (and resources) locally (so-called “*fog computing*”) will not preclude collaboration with cloud resources; rather, we will see a **holistic continuum of systems**, going further away from the edge only when the system is not able to find or process information locally (leading to *connected cognitive cyber-physical systems*). **Intelligence at the edge** is a clear requirement, for example, for self-driving cars: communication with the cloud should not be a requirement for the vehicle to decide to brake to avoid hitting a pedestrian. Similarly, privacy will be a drive to keep and process private information locally.

Efficiency is also a driver: even if new communication standards like 5G offer larger bandwidth, lower latency and better quality of service, sending tens of thousands of 4K video streams to a server (for example, for video surveillance in a city) will still be insanely costly in term of bandwidth (and energy), leading to the need for local processing to decrease the bandwidth with the server. Other reasons include fast response time and latency, for example in industrial processes; the laws of physics, such as the latency due to the speed of light, will in some cases require local processing.

Yet some domains, such as gaming, could go the opposite way: we may be witnessing the last generation of game console hardware. Most of our digital world has already moved to the cloud, and uses a rental-based business model: along with music and movie streaming, most games are already downloaded, while the CD/DVD market has almost disappeared. This is also changing the mode of consumption from owning to renting (the “as a service” approach).

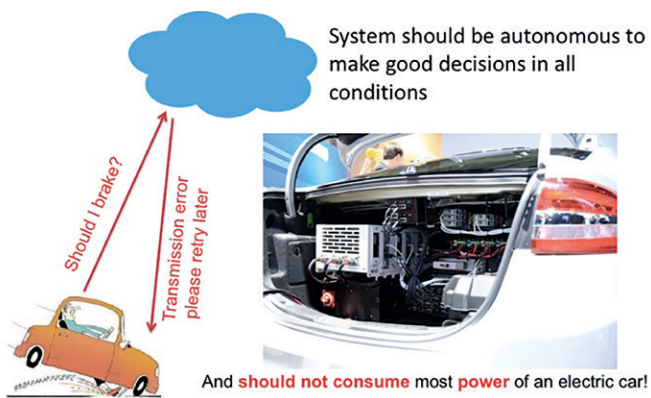


Figure 2: Systems should be autonomous to make good decisions in all conditions

Hopefully, we are also entering the “*age of wisdom*”: with recent studies revealing that technology is also having a profound negative impact on humans and on society, makers of smartphones operating systems (OS) are now starting to propose ways to limit their usage.

Awareness of the impact of ICT on the planet – consumption of scarce resources, energy consumption, recycling – is still not sufficiently widespread: most modern devices cannot be easily repaired and are thrown away or end up discarded in the environment. “Programmed obsolescence” is a problem that should be taken into consideration, and reparability or upgradability of hardware is becoming increasingly important.

The **energy consumption of ICT systems is growing rapidly**: some projections suggest that by 2030 ICT will consume the equivalent of half of the global electricity production of 2014; post-exascale systems might consume more than 80 MW: if not by wisdom, the cost of ownership – such as electricity bills – is becoming a major challenge. **Computing systems need to be orders of magnitude more energy efficient** if they want to succeed in their expectations. If not, the capital expenditure and operating costs of post-exascale systems will be higher than most countries are willing to spend, and a post-exascale computer may end up like large scientific instruments such as large space telescopes or particle accelerators, in the sense that only one exemplar is shared by scientists across the world, and that it can only be used to advance basic science.

A large proportion of the increase in ICT energy consumption is linked to data centres and associated resources (communication). Even if the end user doesn’t see it, sending a mail or requesting a service from a distant server consumes energy. If all our light switches at home were replaced by “intelligent” switches connected to a server 10,000 km away, we would be facing a major energy challenge (and an increase in the risk of being hacked, etc.). Here also, local processing would offer the same service. The total power consumption of all cloud data centres is already higher than the power consumption of small countries,

and it is expected to surpass that of a large country within the next five years. This evolution is not aligned with the international goal to reduce the climate footprint of modern society, nor is it in line with the need to decrease the dependence of the EU on imported energy. It can only be sustainable if the increased power consumption of ICT leads to even bigger savings in other domains like transportation, heating, manufacturing and so on.

One way to be more energy efficient is through **more efficient hardware**: an obvious way to do this is to **specialize the device** and to **reduce data communication**. **Heterogeneous computing and computing close to memory** (or even *in memory*) will be required, but at the cost of more complex programming. GPUs are the first step of specialization compared to general-purpose processors: they are designed for throughput and have less flexibility in terms of control units (which saves hardware). They are now key in computing: computers at the top of the TOP500 high-performance computing list are mostly powered by GPUs. Used in artificial intelligence for the learning phase, where “**adequate computing**” resources are already applied in switching to floating point coded in only 16 bits, they have also found applications in blockchain and the associated mining. As a result, GPUs are in huge demand and their cost has increased.

Another way to be more energy efficient lies at the software level, by ensuring that **program development** considers **energy as a first-class property**, and by having applications aware of energy consumption at runtime. At present, most program development phases still focus exclusively on functional properties (what the program does), without considering non-functional properties, including energy (and time constraints), to a sufficient extent. This omission leads to tremendous inefficiencies.

We should also see the **emergence of novel computing paradigms** and technologies, which could – in cooperation with the available hardware – result in better performance for certain domains. **Neuromorphic computing and quantum computing** are currently at the top of the list, but their integration in the current computing ecosystem is not contemplated yet as it would raise numerous and challenging software needs. Other technologies can also be helpful in particular use cases, such as photonics, plastic or organic electronics, etc.

Another important focus is the impact of the rapid evolution of ICT and the consequent **rapid obsolescence of entire systems**, besides and beyond the intended “planned obsolescence” of some products. So many of our **current systems are so complex** in terms of internal structure, installation and dependencies, that they are **phenomenally hard to repair and debug**. Accommodating legacy, both hardware and software (most frequently the latter), is a significant challenge. Obsolescence is an evident challenge for the new smart devices that replace non-smart devices; ironically, the latter (e.g. light switches, cars, electricity meters and so on) traditionally have a long to very long lifetime, which

their replacements do not match at all. Under these circumstances, which car manufacturer will be willing to continue servicing the software that controls a self-driving car (new features, addition of new traffic signs, security updates and so on) over time? They might have to move to different business models like renting or leasing cars in order to be able to get them off the road when they become too expensive to update.

An enormous amount of data is currently being generated each second: what will its real lifetime be? **A large proportion of data created** – so-called “dark data” – will be never read. Data stored using current means will not be readable in a few decades from now, owing to OS and hardware evolution. This is evidently causing a serious problem in sectors with systems built to last for many years, such as aeronautics. CERN – as an example of an organization with a long lifetime– has started using containers to store not only the data, but the full software stack necessary to read and write that data. INRIA, in cooperation with the United Nations (UN), has started a software heritage initiative [404] to store software in source code form in order to preserve it for posterity. If we cannot overcome this problem, the achievements of this phase of **civilization will be short lived**, as we will **not leave readable traces to our successors**.

Software development has gradually moved from producing single, monolithic, self-contained programs to **constructing programs comprised of a large number of independent parts glued together**, by compilers, builders and even by interpreted languages, to form a single, coherent system. Programming languages have followed this trend, shifting attention from serving the needs of monolithic, homogeneous, self-contained programs to being targeted at integrating, and possibly orchestrating, diverse software parts that are frequently heterogeneous and multi-vendor.

This trend reflects a genuine effort to avoid reinventing the wheel over and over again. However, it also **massively increases the complexity** of the resulting systems, and therefore raises the challenge of mastering and understanding that complexity. What do individual parts provide, under which conditions, what are their functional limits? These issues naturally lead back to the notion of **design by contract** and encapsulation, and to the necessity to contain both code and contracts at the library (component) level, addressing inter-operability issues at the boundaries of such components, thanks to contracts established on them and explicitly supported novel programming languages.

Unfortunately, the programming languages that are currently popular for such levels of system integration have the traits of the scripting languages from which they evolved, and miss out on specification capabilities, hence providing little to no support to contract specification and enforcement or containerization, which will undoubtedly be key assets in the quest to manage the complexity of ever-growing and increasingly complex software.

In addition, currently most of the **developments are using unsafe programming languages**, increasing the risks of memory leaks, pointers errors which are the entry points of most hackers.

Very much like energy, security and safety are non-functional properties that continue to be considered as an afterthought, except in a very few specific, demanding application domains such as aeronautics and, more recently, automotive. Giving designers and program developers the ability to manipulate these properties explicitly, in common with other attributes of value, would make it much easier to implement security policies, consider security and safety, and make informed trade-offs that would be explicitly stipulated (hence documented) in the programs.

Physics, chemistry, mathematics, all these disciplines have developed into different specialties: nuclear physics, organic chemistry, physical chemistry, calculus, group theory. Until now, computer science has been divided into different subject areas, but to some extent these have been closely related: system architecture, formal languages, operating systems and so forth. But now, as the relationship between computer science and other areas, such as mechanical engineering, biology and even psychology becomes closer, it may be time to acknowledge this diversity by developing **computer science specialties that are interdisciplinary**. With computing expanding out of the isolation of cyberspace, the integration between computer science and other sciences must be strengthened. It may be time to reinvent computer science.

PART 1: KEY MESSAGES

- **Efficiency** is essential and needs to be improved in all its aspects despite the drastic increase in complexity: energy efficiency for the system, and productivity (of humans) in developing new systems and software.
- **Trust and acceptance** are still a major challenge for ensuring the success of ICT systems.
- **Europe has to keep its place** in ICT, especially in the domain of CPS, edge or embedded systems where it is already in a good position.
- **The societal impact of ICT** should be considered.

The key messages are therefore clustered into each of these four themes.

1.1 EFFICIENCY

We have come a long way from the situation in which a computing system consisted of one computer core programmed in one or very few programming languages. The need for more computing power on energy-constrained computer platforms (from smartphone to datacentre) has forced computer vendors for a decade now to introduce multicores, and they have recently been compelled to leave homogeneous multicores for heterogeneous multicores consisting of different kinds of accelerators that are more efficient, but more difficult to program and efficiently use.

Modern computing systems consist of multiple heterogeneous cores and memories programmed in a multitude of programming languages, and they interact with the physical world. The end of Dennard scaling points towards more heterogeneity at both the hardware and the software level, further increasing the complexity of applications and more business-critical and safety-critical applications. This leads to more stringent non-functional requirements: performance, energy consumption, security, safety, privacy and so on.

Therefore, the design and implementation of modern computing systems has become **so complex** that it exceeds the cognitive capacity of even the best computer scientists. The development phase will either take too much time to bring the system to the market, or the resulting system will contain too many errors, some of which might lead to incur safety hazards. The current approach to managing complexity through adding layers of abstraction has reached its limit, due to the inefficiency introduced by each additional layer and the lack of global optimization. There is little hope that systems and the associated applications will become less complex in the future: they won't. Every complexity reduction earned by local optimizations will be seized upon to build even more complex systems. Hence, there is only one way forward: we have to find practical and efficient solutions to deal with the increasing complexity.

All of the above leads to the following set of high-level recommendations:

1.1.1 DEVELOP DOMAIN-SPECIFIC ACCELERATORS AND DESIGN KNOWLEDGE

As long as there is no breakthrough that will continue the (exponential) scaling at the technology level (continuing the so-called “free lunch”), the only way to continue performance scaling is to specialize hardware for important application domains. The use of accelerators incurs a cost though: for the design itself, to gain market share, and to develop relevant tools and software. If accelerators are the future of computing, Europe should invest heavily in accelerators, their design knowledge and their ecosystems. It is the only way to stay relevant in the global computing business.

1.1.2 DESIGNING HARDWARE PLATFORMS IS ONLY ECONOMICALLY VIABLE IF IT CAN BE AUTOMATED

Designing an accelerator is a complex task, which could take years if started from scratch. Given the economic importance of accelerators, we need to dramatically lower the cost of their design. Candidate solutions for lowering the cost include:

- i investing in sophisticated design environments for accelerators, and
- ii making use of (open-source) designs that are easily amenable to adaptation.

The combination of both solutions will bring accelerator design within the reach of medium-sized companies. To support the European computing industry, Europe should invest in an ecosystem of tools to design accelerators. The tools could use advanced AI-related techniques to facilitate their use by designers, and could explore the space of solutions under the control of the designers. The United States (US) Electronics Resurgence Initiative (ERI) from DARPA is a step in that direction.

The second important element is avoiding starting from scratch each time and being able to leverage similar designs in order to create more optimized solutions: systems are so complex that all components cannot be designed from scratch for each new system. Open source designs will also promote the appearance of innovative and new solutions in Europe. Open source solutions have the additional benefit that code (even code generating hardware) can be inspected for bugs by a large community. This builds trust and also, providing access, democratizes new solutions.

1.1.3 INTEGRATION OF ACCELERATORS AND OTHER ELEMENTS IN AN EFFICIENT SYSTEM SHOULD BE FACILITATED

Once all the building blocks of the system are available, their integration in a coherent system should be facilitated, both at the hardware and at the software level. Solutions should be defined and disseminated allowing the reuse, integration and

orchestration of white, black and grey boxes in a coherent way and with enough guarantees (security, reliability, bug-free, ...). Tools, API, meta-information, interface contracts, etc. are potential solutions to be investigated. At the hardware level, a library of silicon blocks (chipselets) with a shared interfacing method could be developed, transposing the approach of PCB (Printed Circuit Board) and components to the micro-scale using interposers (the “new” PCB) and chiplets.

1.1.4 SOFTWARE SHOULD BE WRITTEN BY SOFTWARE, NOT BY PROGRAMMERS

Writing quality code for modern general-purpose processors is already very challenging for qualified humans. It is beyond the capacity of humans to develop correct, efficient, and secure code for new-generation heterogeneous computer platforms, particularly in a viable way for lead time and cost. The only long-term solution is to develop production environments capable of automatically generating and optimizing code, out of a wide range of high-level specifications either written in a domain-specific language or codified in a large and comprehensive labelled data set for machine learning. To protect the future of its software industry, Europe should invest in the development of powerful integrated design environments capable of generating powerful, efficient, secure, safe and traceable code for heterogeneous computing systems.

1.1.5 ELEMENTS IN A SYSTEM, OR IN A SYSTEM OF SYSTEMS, SHOULD BE ABLE TO ADAPT TO THEIR ENVIRONMENT DYNAMICALLY

ICT solutions are increasingly a continuum ranging from deep-edge (microcontrollers linked to sensors or actuators), to edge, concentrators, micro-servers, servers and cloud or HPC. A system itself is now a component of a larger system. In this continuum of distributed computing devices, we could imagine the http protocol as a kind of assembly language for these new kinds of systems. Due to the complexity, size, and heterogeneity of the systems and their providers, interoperability is key. Of course, standardization could play an important role, but *de facto* approaches are likely to be winners due to their rapid introduction and acceptance.

In addition to static approaches, creating devices which are dynamic and “intelligent” – and thereby able to communicate with their peers, exchange capabilities and interface formats – will enable easy-to-build systems. However, this still entails challenges of defining and ensuring the quality of service (QoS) in various configurations and situations. “Self-X” might allow the reconfiguration of systems in a broader system and ensure minimum mode of operation even in degraded situations.

1.2 CREDIBILITY, SECURITY, SAFETY AND ACCEPTABILITY

As the number of ICT solutions keeps growing, it is important that people can trust these systems. For this to happen, systems need to be both secure and safe. This means that devices will not harm us when they interact with their environment, and that they cannot be influenced by outsiders, and should not leak any information in an unwanted manner. These issues are of immediate concern for all cyber-physical systems and network-connected devices.

1.2.1 EUROPE SHOULD INVEST IN TOOLS AND TECHNOLOGIES THAT ALLOW US TO CREATE SECURE AND SAFE SOLUTIONS

When our software is created by tools, rather than by humans, these tools should have security and safety as explicit requirements. However, there will still be (limited) amounts of code written by humans, which will also be security-critical code. (Memory) safe programming languages should be used, replacing the unsafe C/C++, for example. In addition to new code, we will also bear the burden of legacy code for quite some time to come. This software should also be protected, in such a way that we can keep the legacy code itself as unchanged as possible.

1.2.2 EUROPE SHOULD DEVELOP SYSTEMS THAT CAN BE UNDERSTOOD ENOUGH

Furthermore, with the rise of machine learning, ICT systems will make more and more decisions based on results generated by machine learning. In order to engender public trust in these systems and the decisions coming forth from them, these algorithms and their decisions need to be explainable, at least to a sufficient level to build trust. If we cannot explain why certain decisions are being made, at least part of the general public will rightfully distrust these decisions. But the role of humans, such as for building example databases, specifying the system, should also be clearly explained.

1.3 THE POSITION OF EUROPE

Computing contributes less to the European GDP than it does to the GDP of other industrialized countries. Although from an economic perspective, it might seem attractive to buy computing goods and services instead of developing them, this also creates risks. First of all, the employment associated with the production of the goods and services is exported outside Europe. Secondly, relying on non-European products and services makes Europe vulnerable to foreign espionage, theft of data, export limitations, etc. Finally, Europe might become less competitive in some areas because there is no build-up of local expertise (such as that needed to design accelerators, to use deep learning, to secure computing systems and so on).

Europe is potentially a large market but its lack of federation makes it difficult to develop big business-to-consumer (B2C) ICT companies like in the USA or China. But there are large business-to-business (B2B) ICT companies, including SAP, Atos, Thales and others. Europe cannot compete in term of very advanced technology (for example, STM and Global Foundries will not go below 7nm [352]) so it should focus on its strengths. Currently, Europe is no longer sovereign and autonomous in hardware for ICT. The “end of life” of the current silicon approach might be a chance to regain leadership in emerging solutions: non-silicon-based technologies, neuromorphic computing, deep-edge and cognitive cyber-physical systems (C2PS).

1.3.1 EUROPE SHOULD BE A LEADER IN “INTELLIGENCE AT THE EDGE” SOLUTIONS AND COGNITIVE CYBER-PHYSICAL SYSTEMS

Europe has a large but fragmented market, has a good education system, is strong in transportation (cars, planes, and trains), and in microcontrollers/micro-electrical-mechanical systems (MEMS)/sensors, embedded systems and embedded software, and has a high level of diversity. European culture is also very sensitive to privacy, security and safety requirements. Therefore, it should build on its strengths and become a leader in intelligence at the edge, allowing better control both of the devices and of privacy.

The cost and diversity of edge devices, together with energy efficiency, are challenges that Europe can tackle with its innovation and diversity, if the path from idea to market is improved. In those fields, the most advanced CMOS technology is not always the best choice, because of its cost and lack of long-term reliability (qualification for safety applications). Inter-disciplinarity and exchanges between and within (scientific and industrial) communities should also be further developed to enhance creative solutions adapted to new or emerging markets. The culture of analysis and understanding of Europe could lead to the development of the right level of “explainable AI” solutions, but this should not be slowing down its development compared to more pragmatic countries that don’t care so much about that.

1.3.2 EUROPE SHOULD CONSIDER ICT DOMAINS AS A CONTINUUM, AND NOT SILOS

Future ICT systems will be interconnected and on a continuum from deep-edge devices, edge devices such as cars or household appliances, local servers and cloud or HPC systems. With the emergence of fog computing, where small ICT systems collaborate with their neighbouring devices to enhance their performance, computation or storage), the distribution of computing and storage might be more distributed and not only centralized in the cloud. All systems and solutions should collaborate in order to give the best service to users.

It is therefore clear that Europe should encourage collaboration between ICT at the edge and cloud/ICT initiatives. For example, it is foreseen that future HPC loads will not only be simulations with large floating points applications, but will encompass more and more big data and artificial intelligence, from data that will be provided (in real time?) and pre-processed by edge devices. This convergence of simulation, machine learning and knowledge could allow the emergence of a 5th paradigm in science and technology, where machines could be directly used for the emergence of scientific discoveries.

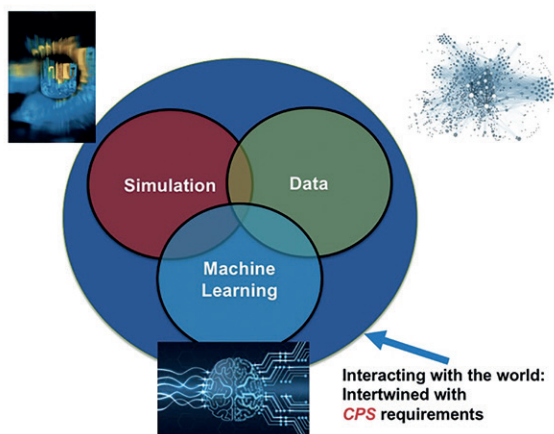


Figure 3: Convergence of simulation, machine learning and big data will unlock a broad class of problems

In term of communities, there is still a culture in silos that can also be observed in large companies. The people developing software should talk to people developing hardware, the cloud or HPC should know about the constraints of edge devices, etc.

1.3.3 EUROPE SHOULD LEAD ON THE USE OF COLLECTIVE DATA

Europe has lost the battle of private data to the big B2C companies (GAFAM and BATX), but it should not lose the battle on state-owned, collective or domain data. Developing the ethical use of data collected by cities, states, medical institutions and so on might allow Europe to develop its capabilities in artificial intelligence solutions based on large amounts of data, without requiring large B2C companies like Google, Facebook etc. Solutions which ensure the privacy and security of data should be developed and enforced.

1.3.4 EUROPE SHOULD BE A LEADER IN ENERGY EFFICIENT, SUSTAINABLE AND LONG LIFETIME ICT

ICT is having an increasing impact on the environment, not only due to the energy it consumes, but also because of the scarce minerals it needs and the waste it creates. Europe could lead in the design of sustainable electronics, recycling of computing devices and modularity allowing to increase the life-time of ICT systems (mainly consumer devices).

Key European industries also require long-lasting electronic devices: cars, planes and trains have a very long lifetime compared to consumer electronics devices like smartphones, and cost of re-qualifying or re-certifying new electronics is prohibitive. Innovative approaches should be developed in order to increase the lifespan of electronic systems (for example, certification and virtualization, modularity, specific supervision, etc.), not only at the hardware level, but also at the software level (economic impact of providing a software upgrade for an outdated system).

Europe should invest in urban mining of electronic waste to become less dependent on the import of minerals and metals required to produce ICT hardware and to limit the environmental impact of mining.

1.3.5 EUROPE SHOULD DEVELOP SOLUTIONS USING MATURE TECHNOLOGY NODES

Without exception, semiconductor manufacturers in Europe have announced that they will not go into sub 10nm technology. Advanced CMOS technology, even if density is increased, only results in slight performance improvements, requires huge investments and the design of chips will be expensive (due to complexity, and the cost of masks and technology). Therefore, it is not clear if the cost per transistor will still decrease (the original Moore's law).

More mature technology (above 10nm) will be less expensive while still having the right density and performance for certain applications. For example, further low-power consumption can be achieved by controlling the bias in fully depleted silicon on insulator (FDSOI) technology and is therefore suited for IoT devices which will require ultra-low power in standby/active listening mode, but high performance when fully activated.

The use of interposer and chiplet technology will allow a reduction in design costs and the mixing of different technologies, such as analogue, power converters and digital. It might become the sweet spot for edge devices, having the best performance/cost ratio. Europe should continue investing in those technology nodes in order to expand their efficiency and usefulness for devices, and not only at the edge. It should also encourage architectural development based on less aggressive technologies like these. For example, the computing chips of the Chinese

supercomputer Sunway TaiHuLight – ranked first in the TOP500 list of the world’s most powerful supercomputers in 2017 – were created using over 10nm technology.

1.3.6 CONTINUE RESEARCH ON POST-CMOS TECHNOLOGIES WHILE MAINTAINING A LINK WITH EXISTING ICT TECHNOLOGIES.

Europe should continue investing in research on post-CMOS technologies to lay the foundations for the ICT technology of the future. Research and innovation should be supported now, as there is no clear idea of which new technology will be used in practice in the future. Post-CMOS technologies are not intended to substitute advanced CMOS, but to complement it, allowing growth in performance and efficiency to be sustained.

A fast transition from research to industrialization should be also encouraged, so that Europe will not only be the origin of a new technology, but also be able to benefit from it when the market emerges. Integration with existing hardware and software technologies should also be taken into consideration early in the development.

These post-CMOS technologies might also be a good source of the innovative sensor/actuator/interface technologies that will play a crucial role in the future CPS and wireless sensor networks.

1.3.7 DEVELOPMENT OF INNOVATIVE ALTERNATIVE ARCHITECTURES

Due to the slowdown in silicon technology improvements, and to the challenges of energy and efficiency, it is time to develop innovative alternative architectures (non-von Neumann systems). For example, processing should be near data (computing near or in memory) and the communication bandwidth should be increased for a number of challenging applications.

Innovative architectures of the past should be re-evaluated in view of the new challenges and new progress in manufacturing and technology. For example, architectures for neural networks that were booming in the 1990 are now regaining interest due to deep learning-based applications. New computing models can be efficiently applied to specific applications and they could lead to benefits because of the slowdown of performance increase of general purpose programmable processors. One example of this is the forthcoming Configurable Spatial Accelerator from Intel, which is being described as a dataflow engine rather than an x86 or classical von Neumann system [86].

The use of accelerators will enable us to continue performance scaling without technology scaling. At some point, accelerators will also run out of steam, at which point, new concepts or new technology scaling should take over again.

1.4 SOCIETY

Digital technologies will continue to transform society. An increasing number of citizens and scientists are worried that this transformation will be so profound that it might disrupt society itself. Major areas of concern include the use of artificial intelligence to build weapon systems, the impact of computer-based automation on employment, the impact of access to or ownership of computing capacity on inequality and the impact of computing on sustainability.

1.4.1 DIGITAL ETHICS SHOULD GUIDE US TO THE FUTURE

Computing has become such a powerful commodity that we should start thinking whether everything that can be done should actually be done. Decades ago, similar questions led to the establishment of disciplines like medical ethics, bio-ethics, business ethics, military ethics and so forth. It is now time to invest in digital ethics as a discipline, and to make sure that all professionals in computing receive basic training in it. The creation of cyber armies in many countries might also call for some form of regulation. Europe should invest in the development of digital ethics and digital ethics should support policy makers to make decisions.

1.4.2 EMPLOYMENT WILL EVOLVE

The impact of computing, in particular artificial intelligence and robotics, on employment cannot be underestimated. Many routine manual and cognitive jobs will (continue to) disappear, new jobs will be created, and existing jobs will change due to automation. Whether this transformation will eventually lead to the net destruction of jobs, or to the net creation of jobs in the next decade is difficult to predict. What is clear is that the disappearing routine jobs are often medium-skilled jobs, and they will be replaced by a combination of low-skilled and high-skilled jobs. This evolution will exacerbate income inequality, and might lead to social unrest.

Europe should keep investing in training programmes to retrain workers that are at risk of losing their jobs, and to try to reintegrate them in the job market at the highest possible level. Given the longer-term demographic evolution of Europe, and the lack of consensus on the need of immigration, automation and the resulting productivity increase might help safeguard economic growth with a shrinking workforce.

1.4.3 DIGITAL SKILLS ARE THE FUEL OF INNOVATION

Without a sufficiently large workforce with the right digital skills, innovation will slow down. In the future, Europe will face fierce competition from US and, increasingly, Asian companies and universities. In order to stay relevant, Europe should invest heavily in the digital skills of its own population and in some strategic profiles in particular: security experts, machine learning experts,

blockchain experts, computer architects, system designers, and tool builders. Policies should also be created to help ensure that they stay in Europe. Demographically, Europe will not be able to beat Asia, but it can make sure that it maintains enough innovation potential.

1.4.4 SUSTAINABILITY

Computing plays a crucial role in the implementation of the 17 Sustainable Development Goals of the United Nations. In order to reach sustainability, we have to eradicate extreme poverty to reverse population growth, to completely stop the use of fossil fuels to halt climate change, and develop a circular economy to reduce the use of natural resources, at the least. In the best-case scenario, the world's population will continue to grow up to 10 billion people by 2075, and they will all have a middle-class lifestyle (meaning that they will have access to healthcare, education, fresh water, electricity, internet, mobility, and so on). Without advanced computing, this will not be possible within the capacity and means of one Planet Earth. Europe should invest in solutions and technologies that will bring the ecological footprint of Europe within the biocapacity of the continent (i.e. a 50% reduction).

PART 2: RATIONALE

2.1 INTRODUCTION

2.1.1 STRUCTURE OF THE DOCUMENT

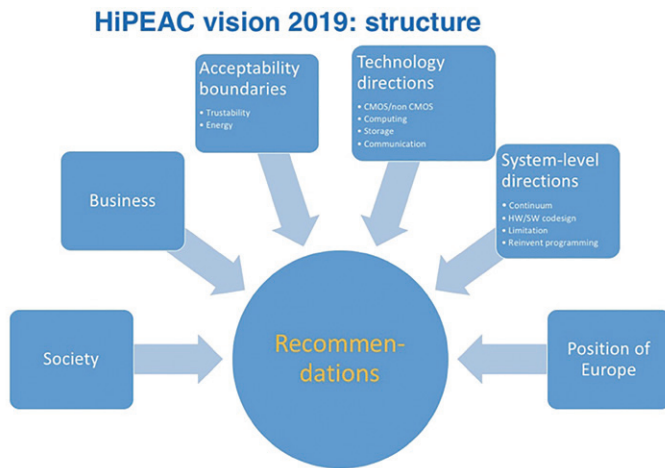


Figure 4: Structure of the HiPEAC Vision 2019

The rationale part of the document details the various elements that contributed to forging our recommendations.

As even the best scientific and technical solution will not emerge if there is no business backing it up, in section 2.2 we start by exploring the current situation in various aspects of business, and the implications of this.

However, even if there is good business potential, a new ICT solution will prevail only if it is accepted. Section 2.3 illustrates a few aspects which are required for an acceptable solution: it should be secure, safe, understandable and efficient.

Next, in section 2.4, we explore current silicon-based CMOS mainstream technology, its limitations and potential alternative approaches to keep improving our systems. As enhancements of CMOS technology will be more and more difficult, alternative solutions at the technological, architectural and implementation points need to be explored and further developed, but this will take time.

In addition to technology, improvements are required as well at system level and in software. Section 2.5 sets out current approaches and their limitations, before explaining where new solutions are required.

Bearing in mind the increasing prevalence of computing in every area of our lives, section 2.6 explores the impact of ICT technology on society and vice versa. Research and innovation in the HiPEAC domain should be done in this context.

Finally, in section 2.7, the document provides a SWOT analysis of the strengths, weaknesses, opportunities and treats of European ICT.

2.2 BUSINESS DIMENSIONS

2.2.1 BUSINESS TRENDS

Business likes to move from buzzword to buzzword, with a lot of hype in solutions that will “solve all your problems”. In previous years, it was cloud and big data; now artificial intelligence (AI) and deep learning are the topics of the moment. This is exemplified by the annual “hype cycle” published every year in the summer by Gartner [359]. Deep learning and digital twin are on the top of the hype curve for 2018, while quantum computing and deep neural networks application-specific integrated circuits (ASICs) are on the rise. Internet of Things (IoT) platforms, virtual assistants, blockchain, autonomous driving level 4, and augmented reality along with mixed reality are entering the era of disillusionment.

In comparison to the 2017 HiPEAC Vision, we see that in fact the IoT is not exploding in the consumer market: smart watches and connected lights have found a market, but other consumer IoT devices are struggling. Virtual reality and augmented reality helmets are on the market, but are still limited to a relatively small number of game players. The first accidents showed that self-driving cars (and their interactions with their drivers) still require quite some progress, while concerns rose about the power consumption of bitcoin that rippled down to blockchain.

Contradicting forecasts from a few years ago, for example in the 2017 HiPEAC Vision, smart robots are still not widely present in homes, with the exception of virtual assistants and smart speakers that are becoming increasingly popular. The home robot Kuri was cancelled and Jibo announced that it would be downsizing significantly [324, 335]. Pepper from Softbank is more for businesses and shops than for the home, and Buddy from Blue Frog has still not been released.



Figure 5: the new Aibo dog from Sony

Source: Sony

On the other hand, Sony reintroduced its Aibo dog, which has more connectivity and intelligence in the cloud. It sold 20,000 items in Japan and has now been introduced in the USA at a price of US\$2,900. According to Sony, “the biggest difference from previous models is a new cloud-based AI engine, which relies on a powerful on-board computer and advanced image sensors to make Aibo smarter and more lifelike. The new Aibo can recognize its owner’s face, detect smiles and words of praise, and learn new tricks over time” [460].

But there are various reasons that could explain why home robots are struggling to find their sweet spot: their price (still high for relatively few functionalities); no “killer app”; unrealistic customer expectations (who believe they are like sci-fi robots); and technical difficulties [325]. An additional factor is that some of their functionalities (verbal and pseudo social interaction, interface with the web and home automation) are in fact covered by smart assistants integrated in loudspeakers (such as Amazon Echo Dot, Google Home and Apple HomePod).

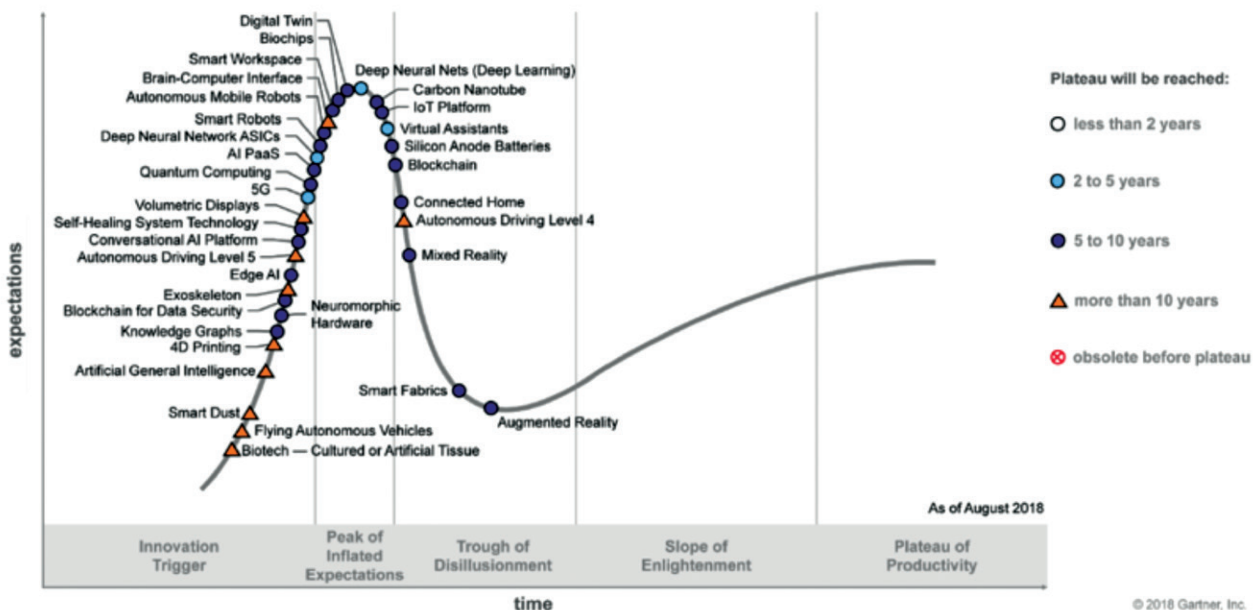


Figure 6: The Gartner Hype Cycle for Emerging Technologies, 2018 – Source: Gartner

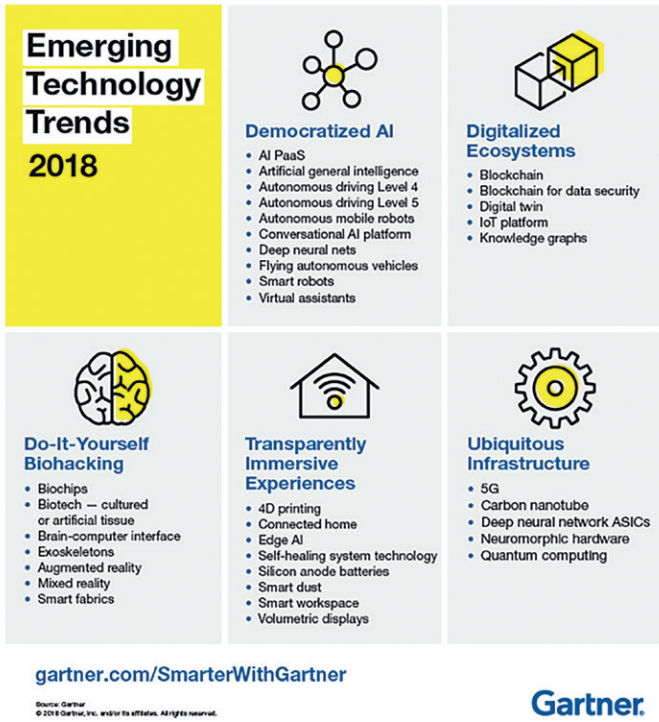


Figure 7: The Gartner Emerging Technology Trends 2018

Besides AI related technologies, we observe that ICT technologies are at the core of most technologies on the Gartner hype curve. Further detail on the business drivers of AI, virtual assistants and “as-a-service” themes may be found in section 2.2.

2.2.1.1 THE AI BANDWAGON

2.2.1.1.1 From cloud to deep learning



Figure 8: Cover Popular Electronics magazine

Over the last 10 years, we’ve seen an evolution of the most-hyped business buzzwords. The first driver was the consolidation of computing and storage resources on the “cloud” after having everything decentralized in personal computers (PCs); see 2.2.1.3 “Cloud, fog and edge computing”. In a way, this represented a return to the centralized computing centres of the 1970s, where computing and storage was so expensive that it was reserved for a select few cases and shared between users through “dumb” terminals (remember the “VT100” range?). Personal computers were in fact started by hobbyists and promoted through *Popular Electronics* magazine, whose January 1975 issue featured the Altair microcomputer.

As many readers will be aware, the introduction of the IBM PC in 1981 started the democratization of computing, with “minicomputers” becoming business tools for companies, thanks to the brand name of IBM. The network (which became mainstream as the internet) helped by interconnecting PCs. That was the first swing of the pendulum oscillating between centralized and decentralized computing.

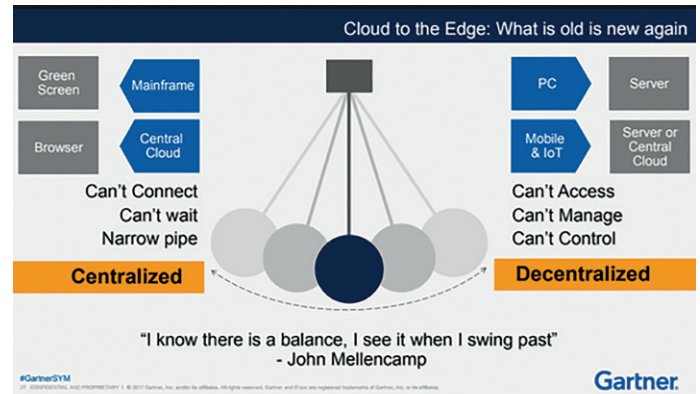


Figure 9: Gartner Cloud to Edge, the pendulum between centralized and decentralized computing will swing equally in IoT deployments

But a PC is not resource efficient: it is idle most of the time and a single user seldom creates a 100% central processing unit (CPU) load 24/7. Consolidating processors and storage in large numbers in data centres that can be shared by multiple users, on demand, reduces the cost of information technology (IT) infrastructure for companies. This prompted the emergence of “cloud computing”, where resources are delocalized and may even be in different locations for redundancy and safety. The cloud is also a location where the data and usage of thousands or even millions of users meet, and new services can be exploited thanks to this convergence. A large flow of emails, video, pictures, or other data began flowing from users to data centres, and between data centres, ushering in the so-called “big data” era.

Moving a step beyond connecting PCs and mobile devices, the Internet of Things (IoT) was also seen as a way to create more data, mostly issuing directly from the physical world, such as data from sensors and the transformation of physical measurements into ICT compatible data.

Therefore, IT managers were pushed to invest in cloud and big data, but they were not really clear on how to practically exploit all those data and computing resources. In a sense, artificial intelligence (AI) provided a response to this challenge, requiring minimal investment in human resources to exploit a large amount of data, extracting the relevant information from it.

Artificial intelligence is marketed as an easy way to exploit big data and large computer infrastructure to solve business processes, with the promise of finding optimizations to open up even unknown market potential. AI, and more specifically deep

learning, was first a necessity for the major technology companies in the USA – Google, Amazon, Facebook and Apple, or (GAFA) – and in China – Baidu, Alibaba, Tencent and Xiaomi (BATX): for example, to check if the millions of pictures uploaded everyday are “correct” (a typical Facebook deep learning use case). They have all the necessary resources: large and powerful computing infrastructure for learning and managing large databases, large sets of data and ways to attract the best scientists.

As explained in the insert “What is AI? A brief history of deep learning” on page 23, the renaissance of AI was triggered by the superior performance of the deep learning approach for image classification, initiated by the work of Hinton *et al.* in 2012. As deep learning provides relatively good results (good enough) when applied to various application domains, with relatively low human effort, it has really taken off since then; now it is on 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”.

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.

2.2.1.1.2 Personal assistants

In addition to image classification, progress in AI, particularly deep learning, is 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.

Improvements in the accuracy of recognition have also triggered the development of specific accelerator hardware (in the case of Google):

“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!” [274].

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.

First introduced by Amazon with the Echo in November 2014, the assistant Alexa can be considered a success, being integrated into over 20,000 devices as of September 2018. Amazon has sold more than 50 million Alexa-enabled devices. According to [343] “*there are now 50,000 Alexa skills - what Amazon calls its voice apps - 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.

Google followed Amazon in 2016 with its Google home speaker, which is rapidly growing in the market (Google sold 5.4 million smart speakers versus 4.1 million for Amazon in the first half of 2018 [427]). Apple was next, introducing its HomePod embedding Siri in 2018.



Figure 10: From left to Right: Amazon Echo, Apple HomePod and Google home. Size of devices not respected

Besides providing basic functions of web interaction – such as giving weather forecasts, traffic reports or the time, making to-do lists, setting alarms, and fetching information from Wikipedia – these home assistants have found their market in streaming music, as they provide an ideal interface for paid subscriptions for streamed music. They are also slowly 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). Ecosystems are being created around them, with more and more devices becoming interoperable and therefore able to be controlled by voice.

However, those devices have not proved very appealing for making online purchases: “*According to a report published by The Information, only 2% of people who own Amazon Alexa-enabled devices like the Echo have used them to make an online purchase in 2018. Of those 2% who bought something, 90% of them didn’t make any additional purchases through Alexa*” [413]. This can probably be attributed to the fact that it is more convenient to

view the articles you'd like to buy rather than listen to a long list of items.

Alexa is still the most popular, closely followed by Google Assistant, which delivers better overall performance thanks to the large databases that Google can access to train it.

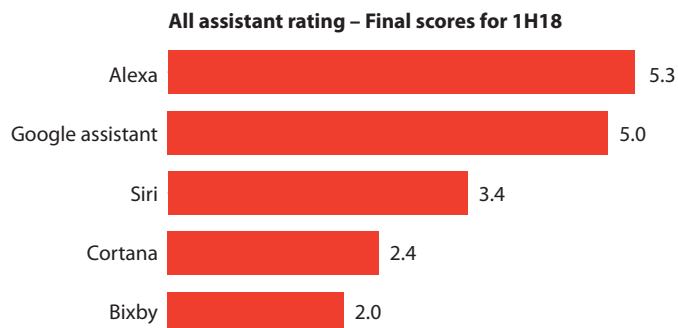


Figure 11: AI assistant rating – Final scores for 1H18

MARKET ESTIMATE FOR VOICE-ACTIVATED SPEAKERS

“According to the Future of Tech report from the global information company, The NPD Group, sales of voice-activated speakers will add an incremental \$1.6 billion dollars to the US technology industry through 2019, as these devices are leveraged both as an interface to smart home services and as digital assistants. Alongside voice-activated speakers, sales of home automation devices will add an additional \$1.7 billion dollars to the technology industry through 2019, with 19 percent of consumers planning to purchase a device in the next 12 months.”

“As consumers are increasingly interested in leveraging voice-activated speakers to control smart home products, voice-activated speaker sales are expected to experience 50 percent US dollar growth from 2016-2017 to 2018-2019. According to the report, demand for voice control in streaming speakers will grow the segment to nearly \$2.7 billion by 2019.”

From <https://www.npd.com/wps/portal/npd/us/news/press-releases/2018/us-voice-activated-speaker-sales-to-see-50-percent-growth-by-2019-according-to-npd-forecast/>

VOICE ASSISTANTS AVAILABLE EVERYWHERE...

Amazon “is already adding Alexa into cars, office spaces and hotels, building on its vision of making Alexa available everywhere you are. That work is already introducing the new world of voice computing to millions more people. Going forward, it could bring about the futuristic notion of having an intelligent, digital assistant with you at all times to help you get through your day and even chat with you if you want”.

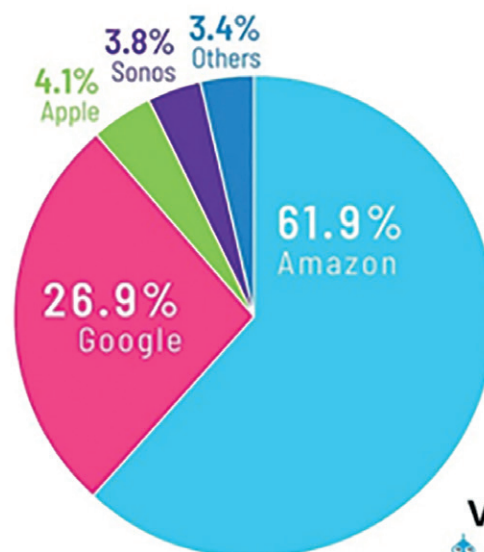
From

<https://www.cnet.com/news/amazon-alexa-assistant-is-now-in-20k-devices/>

We can predict that the next generation of voice assistants will be more customizable, with customers being able to choose the sex and regional accent etc. of the voice, and that they will evolve to understand and operate in a larger context. They will help users by automating routine tasks and helping organize their lives. As an example, in 2018, Google demonstrated a voice assistant reserving a table at a restaurant and booking a haircut without human intervention. This demonstration also gave rise to concern 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; see 2.3.1.1, “The explainable computer”, for more on this topic.

With access to the internet and company data, voice assistants could also participate in meetings as required, contributing relevant information during the discussion [422]. They show their full usefulness when connected to the IoT or smart devices such as lights and appliances, controlling them remotely.

In addition, voice assistants provide a way for people unfamiliar with computers to access resources without having to know how



Source: Voicebot Voice Shopping US Consumer Adoption and Attitudes 2018 Report



Figure 12: U.S. Market Speaker Share May 2018

to use a computer or a particular user interface. Current voice assistants will probably evolve towards being more personal and offering multiple modalities in addition to voice recognition, such as gesture recognition and mood analysis.

One reason for their growing acceptance is that voice is a natural way for humans to interact. In addition, many people like having a constantly available “butler” who obeys orders. While smartphones and social networks fulfil the basic need of communication, assistants will try to be personal “butlers” for everybody.

Nonetheless, voice assistants, especially in Europe, have also attracted criticism for being “spies”, recording everything that is said in a house. Technically, this is not accurate and the behaviour is not the same in all assistants. Normally, the processing of the triggering word is done locally, on the device, and the streaming of sound recorded to the outside servers is done when the “magic” 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. However, the 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 [321] are promoting local processing to 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.

Even big players in the domain are aware of this; for example, Amazon recently announced “local voice control” which will allow its new Echo Plus assistant to still be able to recognize a set of commands for controlling lights or other local devices even if the

internet connection is lost [419]. It is also a smart home hub, so it can directly control smart devices without accessing the internet for sending commands; everything is done on a local network.

With regard to voice recognition, it should be remembered that basic voice commands on some mobile phones can still be recognized even if the connection is lost.

In this, we see the “intelligence at the edge” idea starting to become reality, even for the big players. See 2.2.1.2, “Cloud, fog and edge computing” for more on this topic.

2.2.1.1.3 More details on deep learning

Thanks to the fact that a deep network is formed by learning and is not explicitly programmed, it is applied in many applications where it is difficult to define explicitly an algorithm, such as image recognition (essential for autonomous vehicles), speech comprehension (all personal assistants, from Siri to Alexa or Google Now, use deep networks, often recursive), lip reading and participation in various games. A large “labelled” (indexed) database is all that is needed; these are often available from major internet players (Google, Baidu, Facebook, Microsoft, Apple, etc.), explaining why they conduct in-depth research of learning. For example, more than two billion photos pass every day through two types of deep networks at Facebook, Instagram, Messenger, WhatsApp for image/index recognition and face recognition (although not enabled in Europe).

Networks and techniques are becoming more complex, combining several approaches – such as in the case of the AlphaGo program developed by Google DeepMind that beat Lee Sedol (a 9-dan professional in the Go game) in March 2016, generating a lot of publicity for deep learning and AI techniques.

The image shows a screenshot of the SNIPS website. At the top, there is a navigation bar with the SNIPS logo on the left and links for Snips AIR, Developers, Enterprise, Technology, and Token Sale on the right, along with a green 'Sign Up' button. Below the navigation bar is a large heading 'Voice assistants are broken'. Underneath this heading are four circular icons, each representing a different issue: an eye (privacy), a padlock (security), a person with a mask (developers), and a person in a cage (users). Below each icon is a bolded title and a short paragraph explaining the issue.

Icon	Issue	Description
Eye	They offer no privacy	Sending conversations to the cloud means anyone could access your private life and that of your family.
Padlock	They offer no security	Centralizing a large amount of user data increases the risk of massive data breaches and mass surveillance.
Masked person	They exploit developers	Developers have no access to their users' data, and are at the mercy of app stores that can kill their apps.
Person in cage	They exploit users	Companies building assistants use and monetize their users' data without giving them back.

Figure 13: SNIPS is promoting local processing to keep privacy intact

In general, there are two phases in the use of deep networks: the *learning phase*, in which the network parameters (connection weights) are determined by the learning rule, and the *inference phase* in which the network is used to classify the data.

The learning phase is the most demanding, with millions or billions of example presentations and changes in network settings. It is now generally done on 16-bit floating point graphics processing units (GPUs) or on specialized circuits such as Google's Tensor Processing Units (TPUs). The inference phase is less demanding and can be performed with less precision (integer, even reduced to eight bits). It is usually this phase that is implemented in embedded devices for image recognition, etc. Synaptic weights are downloaded after learning and can be updated after a new learning, extending the number of recognized objects.

For example, Supervision, the network developed by the University of Toronto's Geoffrey Hinton and colleagues is composed of 650,000 artificial neurons connected by 630,000,000 shared connections (synapses). On today's networks, the learning stage could require a few exaflops (more than a billion billion operations).

There are a large number of approaches for the learning phase, but they can be classified into three main classes:

- i supervised learning (presentation of inputs AND desired outcomes corresponding to the particular class of input presented);
- ii unsupervised learning (the network determines its output from different inputs which then do not need to be labelled and tries to automatically discriminate entries into different classes);
- iii reinforcement learning, which focuses on maximizing a reward.

The third class, reinforcement learning, was used to train the AlphaGo program and its successors, like Alpha Zero, which, in a few hours, and without knowledge of the field except the rules, beats all its predecessors both at the game of Go and also at chess.

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

We are even beginning to see research using deep learning approaches to create other, more optimized deep learning networks. This is called Auto-ML (see section 2.4.4, Design tools).

The major players in the field provide their deep network development tools as free software. Examples include TensorFlow (Google), CNTK (Microsoft), DSSTNE (Amazon), Theano, Caffe (Berkeley) and Caffe2, Torch (Facebook with open-source) and

PyTorch (Python interface), N2D2 (CEA), Torchnet learning modules, OpenAI Gym (Open AI), MXNet, etc. In fact, software is a non-critical element in creating an effective system of in-depth learning. A large database and neural network topology are the main ingredients: the value lies in the neural network topology and its weights, determined after learning on a particular database.

2.2.1.1.4 Artificial intelligence: strategic for companies and for countries

Artificial intelligence investments are expected to reach nearly US\$232 billion by 2025 [406]. Many start-ups working in the field of AI have recently been acquired by large companies. For example, in 2014, Google bought DeepMind in the UK (the company that created AlphaGo and AlphaZero), while in 2016, Intel bought Movidius in Ireland and the USA (specializing in low-power vision systems, used for example in unmanned aerial vehicles (UAVs), familiarly known as drones) and Nervana.

In total, Google, IBM, Yahoo, Intel, Apple and Salesforce have acquired more than 30 companies working on AI over the past five years.

"The AI (chipsets) market is expected to grow from USD 7.06 Billion in 2018 to USD 59.26 Billion by 2025, at a CAGR of 35.5% from 2018 to 2025." (from [339])

Well-known scientists and large corporations are investing heavily in AI and deep learning. Countries like the United States of America, China and Japan are launching major AI projects, confident that new breakthroughs will occur and will certainly have a profound impact on our society in the years to come. President Obama said "my successor will govern a country transformed by AI", showing the impact that AI could have in the future.

There is currently an international battle for who will be the leader in artificial intelligence: Russia's President Putin has said the nation that leads in AI "will be the ruler of the world" [416]. China is making huge investments in AI [415] and is feared by the USA, which is also investing in AI through DARPA, for example [41]. As Eric Schmidt has explained, "It's pretty simple. By 2020, they (China) will have caught up. By 2025, they will be better than us. By 2030, they will dominate the industries of AI." (From [320, 340])

China has several assets which could enable it to become the global leader in AI:

- China has a lot of data: it has developed its own internet ecosystem, and applications like WeChat can do a lot of things, and it can then collect information on their use and users. Digital identifiers are registered by the government, and cameras are omnipresent.

- China is developing its own computing infrastructure for AI: thanks to the US ban on exporting strategic processors, accelerators and interconnects in the HPC domain to China, in a short amount of time China developed the capability to produce computing systems in the TOP500 list of the most powerful supercomputers in the world (in fact, that topped the list until mid-2018). This allowed “pipe cleaning” of foundries, chip design and system realization. Now this knowledge can be applied to develop dedicated hardware for AI. China also has a lot of start-ups developing deep learning or AI accelerators, with or without the active support of the government.
- The Chinese government has shown how keen it is to develop AI, issuing a plan and providing large amounts of funding. There is even competition between local governments to be more attractive for AI entrepreneurs and start-ups on AI. In his book *AI Superpowers: China, Silicon Valley, and the New World Order* [260], Kai-Fu Lee traces this belief in AI back to the success of AlphaGo (and Alpha Zero) winning against human champions at Go, an ancient Chinese game that is deeply rooted in the country’s culture. This acted as an electroshock, proving that AI can really lead to great results, triggering numerous actions on AI development in China.
- China has a lot of entrepreneurs. With deep learning, good ideas and the willingness to test them in reality (providing they have enough data), this is enough to create great AI products, even without the top scientists in the field. That said, the quality of AI researchers in China is rising, and now China is top ranked in scientific publications and patents about AI [334]. AI is an enabler to better performance in a lot of application domains.
- AI could create a “winner takes all” phenomenon: the first results from AI will result in economic benefits, then a quasi-monopolistic status because the AI-designed approach will allow more margins, so that the winner could reduce prices and kill the competition until it achieves a monopoly.

In contrast, in the USA most activity around AI is undertaken by the major technology companies (GAFAM), which are also draining universities in the rest of the world of AI experts. See 2.7.1.4.2 “Brain drain” for more on this topic.

If one day, artificial general intelligence becomes a reality, and if that artificial general intelligence is more powerful than human intelligence, Europe will only be able to compete with the rest of the world by building ever smarter computing systems. Instead of the war for talent (fought by companies, universities and countries), in order to improve competitiveness, Europe will have to invest in intelligent systems that will help create better products and do better research. There is a belief that “the future information society will not be built on human brains but on artificial brains”. The societal values of Europe should be built into systems, in order to ensure its future existence.

A list of AI initiatives in different countries can be found in [297], for example. More details on national AI strategies can be found in [298] or in [311].

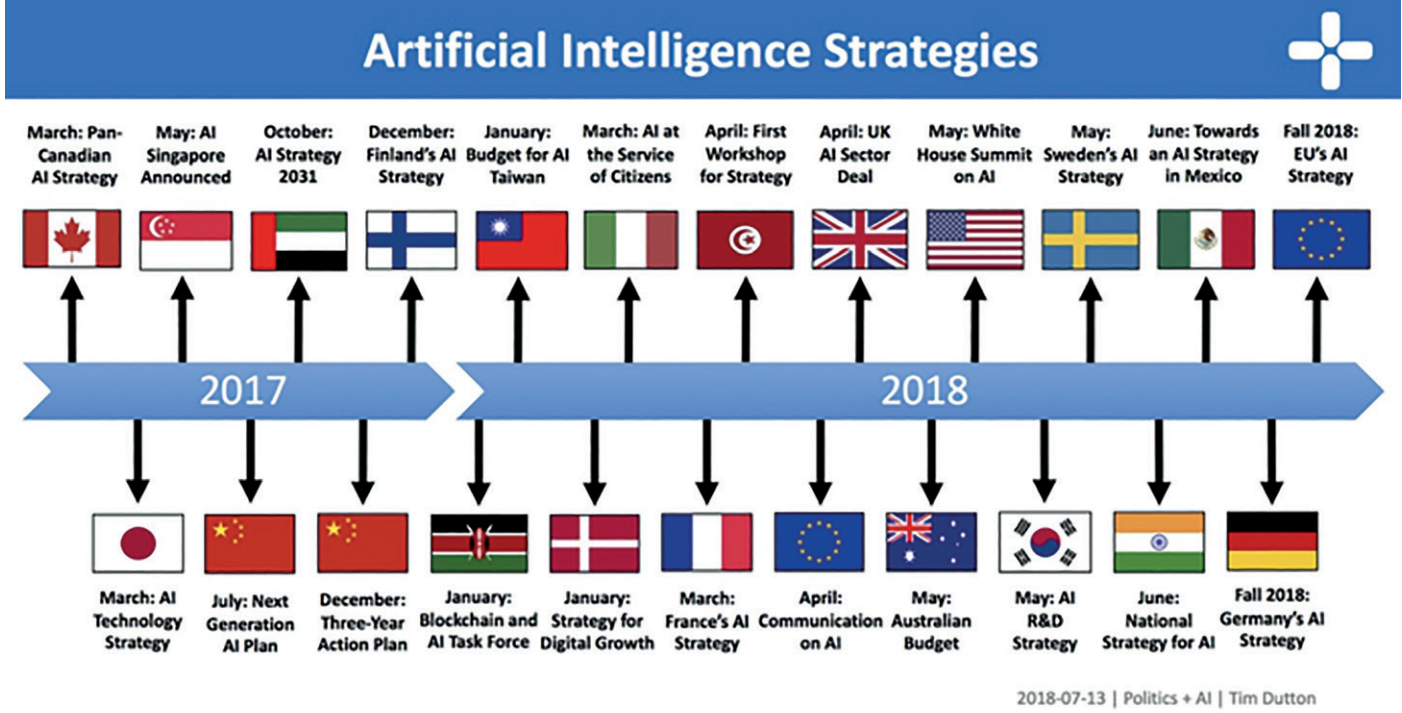


Figure 14: International strategies on AI

WHAT IS AI? A BRIEF HISTORY OF DEEP LEARNING

Throughout history, people have sought to make machines that amplify their physical, then mental abilities. The brain was not always considered the centre of intelligence: Aristotle believed that it was only used to cool the heart. However, the approach advocated by Plato, Hippocrates and Democritus, for whom the brain was the centre of awareness of sensations and the guardian of intelligence, finally prevailed and many generations of researchers have sought, and still seek to analyse its functioning. The idea of imitating it to make “intelligent” systems is not new, but it was the discoveries of the 20th century that triggered the first results.

Drawing on the knowledge of biologists of their time, in 1943 Warren Sturgis McCulloch, an American neurologist, and Walter Pitts, a mathematician and psychologist, proposed a mathematical model of the simplified functioning of biological neurons, cells which form one of the components of the brain. Their paper, “A Logical Calculus of Ideas Immanent in Nervous Activity”, was published in 1943 in the “Bulletin of Mathematical Biophysics” (5:115-133) and remains the basis of formal neural networks. Their model is simple: a neuron performs a binary function that compares the weighted sum of its inputs (connected to the other neurons) to a threshold.

They have shown that a sufficiently complex network can “calculate” any function. John von Neumann, whose “First Draft of a Report on the EDVAC” is considered to be the first description of a modern computer (von Neumann’s machine) cites only this McCulloch and Pitts paper in this 1945 report and infers from McCulloch and Pitts’ article that “*everything that can be described exhaustively and unambiguously... can be conceived as an appropriate neural network*”. It confirms that a neural network can represent a universal Turing machine, and therefore a universal calculator. Unfortunately, the limitations of the technology of the time did not allow him to develop the highly parallel approach of neural networks, and thus it resulted in an architecture with memory, a control unit, an arithmetic unit and input and output units, which are found in today’s computers.

In 1957, psychologist Frank Rosenblatt invented an algorithm called a “perceptron”. For this classifier, the weighting between neurons is inspired by the Hebb rule, which considers that when two neurons are excited together, their link is strengthened. The perceptron rule takes into account the observed error when propagating an input whose output function is calculated by the perceptron. The first winter of neural networks was caused by Marvin Minsky and Seymour Papert’s book *Perceptrons: an introduction to computational geometry*, which shows limitations of perceptrons. The 1986 book *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* by David Everett Rumelhart and James McClelland relaunched the

field with a testable approach of multilayer networks (essentially with an intermediate layer, called the “hidden layer”) called multi-layer-perceptrons (MLPs).

A learning rule (called gradient backpropagation) for calculating the weights of intermediate layers was published in his thesis in 1985 by Yann LeCun (now at Facebook), then widely distributed by David Rumelhart, Geoffrey Hinton (now at Google Brain) and Ronald Williams in 1986. This led to an initial explosion in the use of neural networks in the 1990s. They were used for the recognition of handwritten characters (to recognize postcodes), for image analysis etc. A first era of specialized circuit development followed, but the techniques of the time allowed only limited parallelism, and the rapid advance of general-purpose processors limited their expansion.

The uptake of support vector machine (SVM) then signalled the beginning of a new winter of neural networks by offering better performance than MLPs for image classification. The principles were explored between 1963 and 1970 by Vladimir Vapnik, but it was only in 1992 that an article by Boser, Guyon and Vapnik synthesized the results and allowed broad development of SVMs for classification.

Meanwhile, neural networks became deeper (with more layers), thanks to methods allowing the use of back-propagation approaches to gradient networks with more than one hidden layer. The networks became more complex, specializing the layers as in the visual cortex. The results of neuroscientists David Marr, David Hubel and Torsten Wiesel (the latter two were awarded the Nobel Prize in Physiology or Medicine in 1981 for their discoveries concerning information processing in the visual system) inspired researchers to make networks more suitable for object recognition. Their predecessor is the “neocognitron” invented in the 1980s by Kunihiko Fukushima. Deep convolutional networks

Algorithm name	Date	Error on test set
Supervision	2012	15.3%
Clarifai	2013	11.7%
GoogLeNet	2014	6.66%
Human level (Andrej Karpathy, now Director of AI at Tesla)		5%
Microsoft	05/02/2015	4.94%
Google	02/03/2015	4.82%
Baidu / Deep Image	10/05/2015	4.58%
Shenzen Institutes of Advanced Technology	10/12/2015 (the CNN has 152 layers)	3.57%
Google Inception-v3 (Arxiv)	2015	3.5%
WMW (Momenta)	2017	2.2%

as currently used are more than 20 years old, but thanks to the dramatic increase of data availability and computer power, more complex networks are now possible, which unlock a complete new range of performance.

The most recent renaissance was brought about in 2012 by Alex Krizhevsky, Ilya Sutskever and Geoffrey Hinton, who used deep convolutional neural networks for the ImageNet challenge, which consists in classifying images in the ImageNet image database. The Hinton Supervision Network beats the other approaches with an error rate of 15.3% compared to 26.1% for the second. From 2013, the top eight approaches in the challenge are based on deep neural networks. Indeed, deep networks are better than a human on this challenge, with less than 3.5% errors. The following table shows the very rapid improvement of deep learning algorithms, until being better than humans.

THE 5TH RESEARCH PARADIGM?

The convergence of simulation, machine learning and knowledge allows the emergence of a 5th paradigm in science and technology: as explained in the previous HiPEAC Vision: “The first three paradigms were experimental (empirical description of phenomena), theoretical (discovery of laws, models, etc. able to predict results) and, more recently, computational science (computer simulations). The fourth paradigm of scientific discovery is the analysis of massive data sets, enabled, e.g. by data capture, curation, mining and analytics techniques and thus permitting new scientific discoveries.

In the fourth paradigm, computers are used to extract information from raw data, but it is still humans who perform the analysis of the information and make the scientific discovery. We believe that within the next decade there will be a fifth paradigm, in which computers will not only extract information from data, but will also formulate a hypothesis, design new experiments and simulations or make a formal proof and finally make scientific discoveries without human intervention. We already have examples of this with formal provers, data analytics, and approaches like IBM’s Watson. Potentially, the Ultra-Intelligent machine could solve problems that are beyond the reach of human intelligence”

HiPEAC Vision 2017 pp.59

2.2.1.2 HUMAN IN THE LOOP

The human aspect needs to be increasingly taken into consideration in the development of ICT systems. Not only from the acceptability of ICT point of view (credibility, ethics), but also

because devices and systems will more and more interact directly with humans, and not only through keyboards, touchscreens and displays. Humans will be an active part of the systems and will be part of the equation that new ICT systems will have to solve. For example, self-driving cars will co-exist with cars driven by humans, and they should be prepared for human reactions. Cobots are “aware” of the presence of humans and adapt accordingly. Systems should adapt to their users; for example, user interfaces should learn their owners’ habits or most common actions. Voice assistants should recognize the voice of their “master” and should adapt to their habits. We are moving towards human-aware systems. Artificial intelligence and its techniques will have a lot to do in this process.

2.2.1.2.1 Explainability and/or transparency

One of the main complaints about machine learning, particularly deep learning, is that their models are opaque, non-intuitive, and difficult for people to understand and that the machines are unable to “explain” their decisions, leading to a lack of confidence and trust in the system. Their results can also be totally different when altering just a small part of their input, such as a few pixels in an image [344]. This is an important problem, leading to new kinds of piracy tuned for this kind of processing. To address this, there are two developing fields of research concerning deep neural networks:

- Explainable AI (especially deep neural networks)
- Creating robust solutions which are impervious to deliberately introduced fake data.

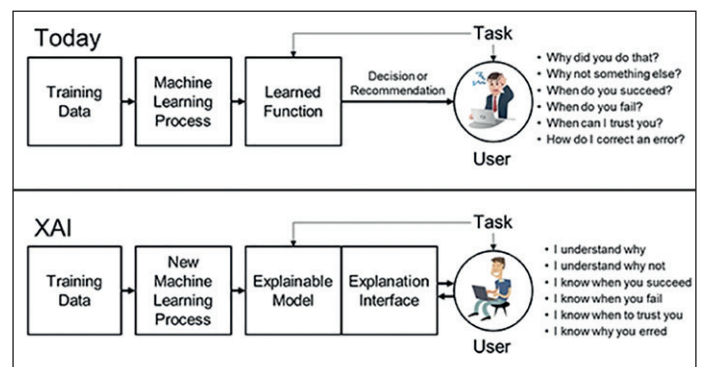


Figure 15: XAI concept

Source: Defense Advanced Research Projects Agency, U.S. Department of Defense

The explainability of the results of deep learning is an important topic for the acceptability of solutions, but this should not be taken too far; by way of comparison, there are a number of industrial processes that are not fully understood but this does not prevent them being used in everyday life. What is more important is to ensure that a deep learning neural network effectively learns what it is supposed to learn, and not something else. A good example of work on explaining the prediction of classifiers may be found at [262]. The setting was the following: a deep learning network was trained to classify images of dogs

and wolves. When a husky dog image was presented, it was misclassified as a wolf. In fact, the deep learning network had been trained to recognize snow; pictures with wolves have snow, and the husky image also had a snowy background.



(a) Husky classified as wolf

(b) Explanation

Figure 16: Raw data and explanation of a bad model's prediction in the 'Husky vs. Wolf' task

Source: [129]

It is therefore very important to develop approaches and tools to check that the learning databases are free from bias, whether introduced deliberately or accidentally. This is perhaps easier to achieve than full "explainability" of deep learning decisions, which is important but difficult to achieve without clear breakthroughs.

The most important points are to ensure that the specifications that led to the learning databases are as complete and exhaustive as possible, with minimum bias, and checking after learning that the system has effectively learnt what it was supposed to learn, rather than other artefacts present in the learning database. It is also important to expose the system to counter examples, that is, things it should not do. Most of the time, designers focus on what the system should do (recognition rate) and not what it should not do (false alarm). Sometimes, modern databases are "too good", with only clear images, making the system more sensitive to noise or other artificially introduced artefacts. With this "classical approach" of supervised learning, humans are still in the loop and ultimately responsible for the design of the learning database, therefore for the resulting deep network and what it will do in the inference phase.

Another idea to consider is not using deep learning alone for a task, but combining several approaches (including other deep learning solutions or symbolic ones) and adding a kind of supervisory process that checks whether the results are coherent.

The rise of artificial intelligence also entails an important psychological impact. Our civilization has always tried to augment humans through science and technology. Now, we accept that machines can be stronger and faster than humans. However, people consider that "intelligence" is the last part unique to humans and AI is starting, at a very low level, to challenge this.

Computers in their current form are less frightening, first because most people today have always been aware of them, and second because they are "dumb", simply executing lists of instructions provided by humans. Even when people complain about a machine if it is doesn't do what it is supposed to do, they know deep down that it is ultimately the fault of human programmers.

With machine learning, the responsibility of the human "programmer" – in this case the human that sets up the learning database – is not so clear. Systems like AlphaZero, which didn't have to learn from a large database of game examples, could be even more frightening. The overreaction in asking for conditions, explanations and so on for AI systems that would not been required for human-programmed or other systems might arise from these situations. Additional explainability and transparency is indeed very important, but it should not block progress if existing requirements can be applied.

Legal liability in the event of an AI system failing is important, but, in the case of deep learning, it applies more to the initial specifications and definition of the learning database (done by humans – at least for now) than on how the deep learning system works by itself.

Breaking it down, we can identify a number of historical steps towards making systems "intelligent":

- 1 Algorithms and classical programming: here the "intelligence" is given by the programmer who has to define the steps to solve the problem.
- 2 Symbolic AI or expert systems: here there is a split between the "engine", which is generic, and the database of rules, which are specific to the problem. This was meant to decouple the technical computer science problem into two aspects: making the engine and knowledge engineering, the latter allowing people who are not programmers to input their "knowledge" into the machine. The responsibility for "bugs" in the engine lies with the computer scientist, while the responsibility for the correct rule set lies with the expert.
- 3 Machine learning (with approaches like deep learning): here also there is a decoupling between the engine (for example,

how the deep learning software works) and the application knowledge, which, contrary to the previous approach, is not a set of logic transitions and rules (the “how” to solve the problem), but a set of examples (the human defines by the choice of the examples “what” needs to be done). The responsibility for “bad” results is ultimately linked to humans, either because they provided incomplete specifications to design the learning database, or because they misused some properties of the system, or because they didn’t check that the system had correctly “learnt” the right function. It is clear that a methodology still needs to be developed to avoid system “misbehaviour”, but it is not the fact that the system is artificially intelligent that means that humans are no longer in the loop and responsible for how the system was trained.

Taking this to the extreme, we can see that in all of the above four cases, humans should be held ultimately responsible in the event of errors by artificial intelligence. The main problem will be to identify potential errors, and to be able to correct them.

Needless to say, as shown in the picture below, humans are also prone to error (such as optical illusions) and sometimes have difficulties checking whether a system is unrealistic.

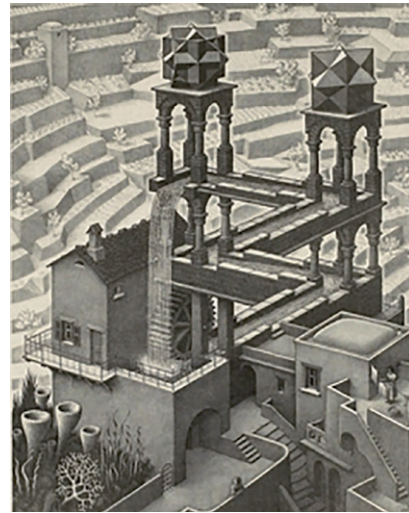


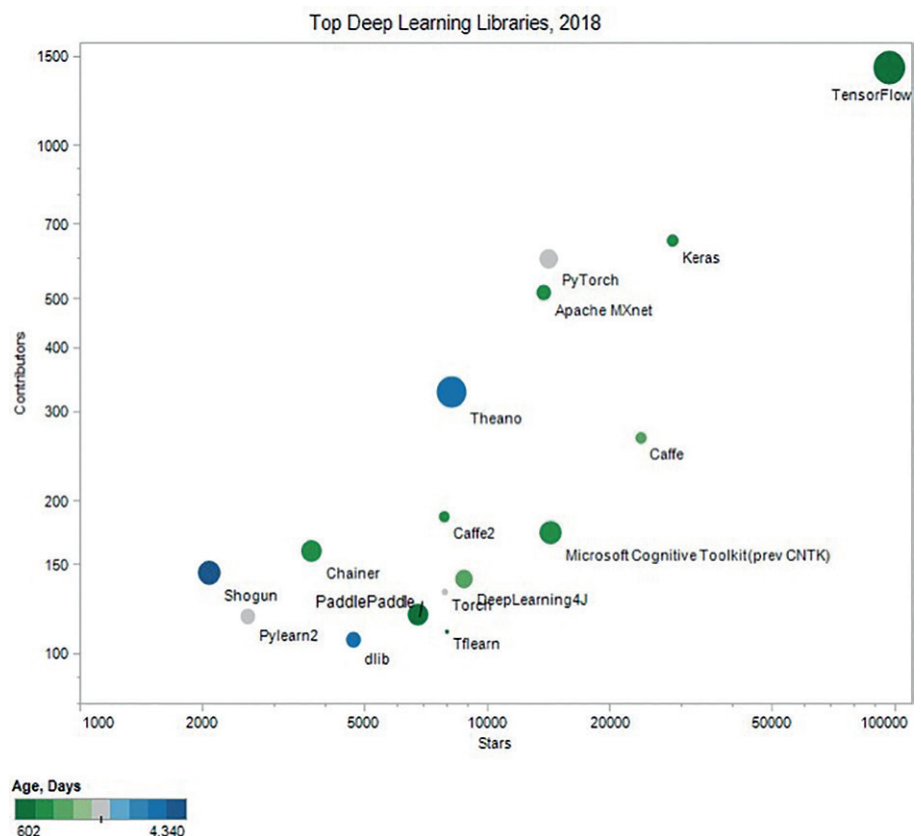
Figure 17: M.C. Escher, Cascade, 1961

- 4 Machine learning (with approaches like AlphaZero): here the engine still comes from computer scientists, but there is no choice of examples for the training: only the specifications – the rules of the game – are provided by humans, which only requires good specifications of the problem to be solved. The system (for example by reinforcement learning) internally generates examples of states to solve the problem. After a large space exploration, it can deliver answers, but the complexity and the number of steps will make it very difficult to grasp for humans. One verification can be to execute the solution found by the system and check that it “works”. It is still not yet artificial general intelligence, therefore humans are still in the loop at the beginning to give the correct constraints and rules to the system. If the system gives a wrong answer, it is likely that the initial specifications were not correct.

2.2.1.2.2 Making AI familiar to humans: the hobbyist way?

As explained in the previous part, there is a need to make AI solutions, with their limitations and capabilities, more understandable and accepted by people. A promising approach is to allow interested people to make their own AI system, play with it and learn from it. In addition to classical training approaches,

Figure 18: Top 16 open source deep learning libraries by Github stars and contributors, using log scale for both axes. The colour of the circle shows the age in days (greener - younger, bluer - older), computed from Start date given on github under Insights / Contributors. Source: Dan Clark, KDnuggets



one that could be efficient is enabling AI solutions for hobbyists, so that they can build their “own” AI system. There is clear momentum in this direction, with a multiplicity of open source and cheap hardware available on the market. Amazon has opened up its application programming interface (API) so that hobbyists can embed the Alexa system in their Raspberry Pi or other devices. Google has launched a line of DIY kits [261] and its new edge TPU chip will also be available as a USB stick or a small board, same for the neural computer stick 2 from Intel [322]. SNIPS [https://snips.ai/], Mycroft [314], Gladys [299], Jarvis [238], etc. are proposing software which allows users to develop personal assistants themselves, and most software development environments for deep learning are also open source.

2.2.1.3 THE CONTINUUM: CLOUD, FOG AND EDGE COMPUTING

As noted above, general public computing has switched from stand-alone desktop computers to mobile devices connected to the cloud. This addresses the new usages and needs created by smartphones and similar devices that allow people to be connected at all times and get access to a huge amount of information – potentially, the whole of internet – and keep in touch on social networks.

In **clouds**, data is mainly stored and processed on remote servers and can be accessed by numerous terminals of various types. Current computing and storage clouds, both for private and business users, are mainly hosted by large third-party providers like Google, Amazon, Microsoft and DropBox. When the cloud computing model first surfaced, it was hailed as offering huge resource savings for customers as compared to in-house servers. Today, cloud computing providers can tune their hardware and software stacks to customer and periodic usage patterns, and offer very attractive conditions and expertise to their customers.

One of these attractive features is that cloud computing offers essentially elastic resources: you only pay for the resources you use, and these can grow or decrease to meet demand. As discussed in 2.2.1.1.1 “From cloud to deep learning”, if you own your own computing resources, they are fixed and provisioned to meet the worst case (maximum occupation), despite not being used at a rate of 100% the majority of the time, leading to extra costs. You also need to pay for the maintenance and management of the system.

Elastic resources are so attractive that even some banks, while very reluctant to share the private data of their customers with outside companies, are moving from having their own data centres to renting resources in public clouds. As an example, in 2017 UBS, the world’s largest wealth management company, moved its risk-management platform to Microsoft’s Azure cloud [269].

Cloud computing has also entered the high-performance computing (HPC) market, with services such as Amazon Web Services (AWS) offering on-demand, scalable resources for HPC workloads via their elastic HPC clusters [266]. Although unlikely to replace in-house HPC facilities, elastic resources such as these mean that a larger number of people are able to access significant computational resources without huge capital investment or the need to undergo a peer-review process to use existing HPC facilities.

The cloud has also been used to offer access to (pseudo-)quantum machines: D-Wave, IBM and Rigetti already offer quantum clouds, and more are expected over the next few years [385].

However, there are a number of issues with the cloud computing model. Perhaps the most obvious is the enormous amount of energy required to power the world’s data centres, which used 416.2 terawatt hours of electricity in 2015 [198], a figure set to increase as greater numbers of devices become connected. Cutting the energy consumption of computation and, crucially, communication will continue to be an important area for data centres over the next ten years. See 2.3.2 “The energy challenge” for a detailed analysis of this topic.

Part of the issue here is that cloud resources could be managed more efficiently: users tend to overestimate their resource needs, providers often keep resources back to meet peak traffic, and the architecture of cloud systems creates resource fragmentation, with resources scattered around compute nodes and data centres [74].

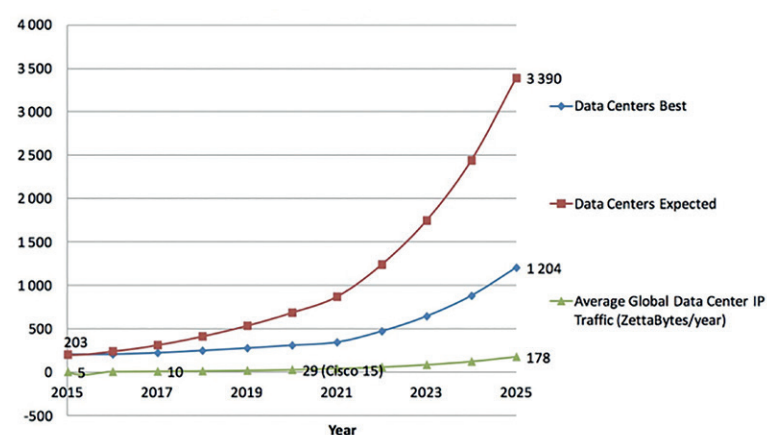


Figure 19: Electricity usage of Data Centres 2015-2025
Source: Anders S.G. Andrae, Total Consumer Power Consumption Forecast, October 2017

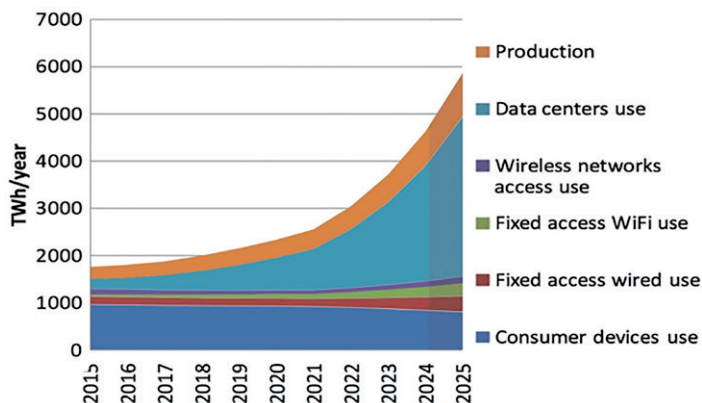


Figure 20: Expected case ratio
 Source: Anders S.G. Andrae, Total Consumer Power Consumption Forecast, October 2017

Another issue of growing concern is confidentiality. Local information is rightly considered to offer greater privacy protection than information in a cloud. After the Snowden case, there is growing awareness that non-local data can easily be, and in fact often is, abused by spy agencies, malevolent hackers or even private companies that store or handle data for customers. The EU General Data Protection Regulation, or GDPR, has also ushered in a new awareness of privacy rights and prompted businesses to make changes to the way they handle users' data.

As a result, many users now prefer to keep their data in private personal NAS/SAN devices, or in locally distributed environments (IoT) at home. They may also opt for private but mutualized data stores, shared among a limited number of trusted users, in which case they could be called federated, or distributed, clouds, that bear similarities with current home media centres for the family.

Thanks to disruptive new storage technologies, the data storage and distribution landscape is also changing. For example, with affordable, large storage capacity in a small form factor, some users and small companies prefer to store their data on a local device that they own and of which they know the location, preferring the fog computing [218] approach rather than clouds. This could also extend to performing most computation locally, which is known as edge computing [127].

Fog and edge computing are gaining traction for applications which are particularly latency sensitive, such as real-time data processing in smart city applications, and in cases where privacy is of particular concern, such as intelligent toys used in therapy sessions with children [59], or where security is paramount, as in the case of intelligent vehicles (see 2.2.3.1 "Automotive: the next frontier?").

Such forms of local storage and processing can be offered on a continuum with cloud computing, with only the most intensive computing, or that which requires access to a lot of non-local (meta)information, taking place on remote servers. Processing

along the continuum opens the door to analysing huge amounts of data in real time, as in the EU-funded CLASS project [272].

It should be pointed out that cloud computing models could still provide enhanced privacy for users. One area of potential business opportunities is to send encrypted data (i.e. by homomorphic encryption) to the remote application that then performs its operations without ever decrypting. As a result, the application can never know the actual data nor the meaning of the results it computes. This would be the ultimate solution for keeping data private, but it runs against the current business model of companies such as Facebook and Google that are built on gathering and reselling as much information about their users as possible ('If the product is for free, you are the product'). As an alternative, the business model could go back to selling computing capabilities to users.

2.2.1.4 POST-EXASCALE HPC

The term "high-performance computing" (HPC) needs to be redefined: In the past, it was synonymous with "technical computing using supercomputers" to model and simulate complex scientific phenomena. In the future, HPC will become the convergence of traditional HPC (simulation) with processing and storage of big data and processing of artificial intelligence (AI) applications in the same data centre, along with ways of orchestrating computing resources for the different workloads. This will also concern the interfaces of this structure with external devices (distributed and edge devices).

This "converged HPC" system will satisfy the requirements for simulations (high-precision floating point, for example), big data (fast interconnect between nodes, large repository of data) and AI (memory per node, lower precision arithmetic, fast interconnect).

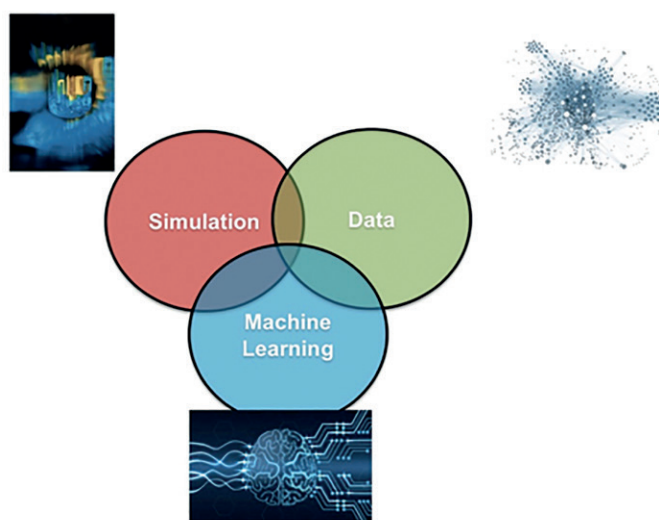


Figure 21: The 3 future pillars of HPC applications

The converged systems will also allow the acceleration of simulations with AI techniques:

- forecasting results with lower compute requirements;
- interpolation and extrapolation;
- setting parameters in long simulations – similar to what is done in auto-machine learning (auto-ML) techniques;
- reducing the parameter space;
- and validation to check that the results are correct.

AI can also replace simulation in certain cases, where exact models are not available. Conversely, AI can also be accelerated by HPC technology:

- pre-processing of large datasets;
- data cleansing;
- massive and fast training of deep neural networks;
- more than real-time inference phase (there is even contest on the shortest learning time for ImageNet, which is now in the range of few minutes see [312]).

team	hardware	software	batch size	time	accuracy
Facebook[4]	Tesla P100*256	Caffe2	8192	1hr	76.3%
IBM[8]	Tesla P100*256	Torch	8192	50mins	75.01%
UC Berkeley[3]	KNL 7250*2048	Intel Caffe	32768	20mins	75.3%
Perferred Network[9]	Tesla P100*1024	Chainer	32768	15mins	74.9%
腾讯机智	Tesla P40*1024	TensorFlow	65536	8.6mins	76.18%
腾讯机智	Tesla P40*2048	TensorFlow	65536	6.6mins	75.76%

Training settings and performance of ResNet-50 on ImageNet by different teams

Figure 22: contest on the shortest learning time for ImageNet

Source: [142]

For more details on the next generation HPC systems, the reader may refer to the companion to the HiPEAC Vision, a joint document produced by the European Technology Platform for High Performance Computing (ETP4HPC), the Big Data Value Association (BDVA) and HiPEAC. This overview document outlines the overall key research challenges for the 2021-2027 timeframe in the area of HPC and high-performance data analytics, with strong links to the internet of things (IoT), cyber-physical systems (CPS) and AI. It will be available in early 2019.

2.2.1.5 CYBER-PHYSICAL SYSTEMS AND THE IOT

As noted in the previous HiPEAC Vision, we have entered an era where the traditional computing system, recognizable by the keyboard and screen as interfaces, is being complemented and to some extent supplanted by mobile computing models characterized by machine-to-machine communication, and comprising a vast array of sensors and actuators. These are collectively known as the *internet of things* (IoT) and *cyber-physical systems* (CPS).

In the model of the IoT, smart sensors in the environment communicate via gateways, or specialized computing devices, with remote servers in the cloud. They generate an enormous amount of data, which is analysed to extract information to provide new and better services. An IoT system is a distributed system composed of a number of physically separated, communicating devices which do not usually involve a human in the loop. An example would be a voice-activated assistant (as discussed in 2.2.1.1.2 “Personal assistants”) or a wearable fitness device (more on this in 2.2.3.2 “Medical and wellbeing”).

Cyber-physical systems take the integration with the physical world a step further by directly interacting with the physical world based on the results of data analytics. Examples include steering or braking a self-driving car, moving a factory robot arm or simply switching on a light.

THE INTERNET OF THINGS VS. CYBER-PHYSICAL SYSTEMS

In our definition, a cyber-physical system (CPS) is characterized as having an actuator that directly affects the physical world (a screen is not considered an actuator in this definition), while an IoT system is distributed and composed of physical objects that communicate, typically via the internet.

Within this definition, CPS and IoT are not mutually exclusive. For example, a self-driving car that is not connected and makes all its decisions locally is a CPS device, but not an IoT device. It would become an IoT device as well if it is connected, for example to get maps from a server. A smart sensor transmitting the local temperature to a smartphone is an IoT device, but not part of a CPS. If it is connected to a thermostat that controls heating, the combination – that is, the system composed of the sensor, the various servers, and the thermostat – becomes a CPS, while the sensor remains an IoT device.

These new computing paradigms throw up new challenges. In the case of the IoT, **security** – or protecting the system from malevolent attacks – and **privacy** – where the data generated may be used for purposes not authorized by the subject, or unauthorized data accesses take place – are major challenges. Cyber-physical systems have the additional challenge of ensuring **safety**, that is, that the system will not harm the environment.

Moreover, these systems are constrained by properties of the physical world such as **time**. In a CPS and some IoT systems, if the system isn’t sufficiently fast or if it is busy, it will lose data and cannot ask the environment to re-send the data. If the computer doesn’t respond in time, this could lead to accidents, for example a self-driving car failing to brake in time.

Energy is another constraining factor: communication is inherent to these systems, and the energy cost of communication is usually higher than that of computation. This can be overcome to some extent by processing as much data as possible locally, rather than transmitting raw data, as in the **edge computing** model (as discussed in 2.2.1.3 “Cloud, fog and edge computing”), which has the additional benefit of enhancing privacy by keeping data on the device. We expect processing at the edge, providing artificial intelligence functions, to become increasingly prevalent. It should be noted, however, that edge computing requires increased local storage and processing power, which can push up the financial cost of a device.

Zero-power computing for the IoT could provide a solution to the energy challenge and is an area with a tremendous potential market, for example with smart tags for asset management. The challenges here are ultra-low power, cost and security. A detailed discussion of this topic may be found in section 2.3.2.3.2 IoT.

Interoperability throws up further issues for the IoT and CPS, particularly in the consumer IoT. Unless a customer buys all his devices from the same company, they need to download a special app to control a new device. In addition, the IoT ecosystem is still largely divided into different domains, while users generally seek cross-domain applications – the bonus of your connected car is that it can find a parking space and charging stations while linking to weather services and your calendar, and so on.

The IoT and CPS markets have grown significantly and we expect that trend to continue. The **automotive industry** is a major driver of the IoT and CPS, with some estimates predicting that the automotive IoT will reach \$100.93 billion by 2023 [398]. **Smart cities** are another key growth area, with applications ranging from smart street lighting to traffic management.

However, given the issues outlined above and since no “killer app” has yet emerged from these domains, we expect to see steady growth rather than the explosion which was forecast in the early days of the IoT. This is particularly relevant for the **consumer IoT**: although voice-activated assistants have become increasingly common, devices such as wearables still have not taken off as expected, perhaps due to the short life of the battery, privacy concerns, and, above all, a failure to see the benefit of such devices.

Performance issues have also damaged consumer confidence, with major implications for the roll-out of self-driving cars, for example [369]. The complex task of providing security updates for many different devices, all with their own proprietary code, coupled with pressure to get new products on the market means that security issues may not be sufficiently dealt with, and high-profile hacks continue to plague IoT products, deterring customers [423].

With regard to the **industrial IoT and CPS**, where the data from smart sensors can be used to drive global process improvements, the benefits in terms of cost savings and efficiency improvements are often clearer. Indeed, cyber-physical systems have become so prevalent in industry that they are now the norm rather than the exception.

Driving the so-called “**Industry 4.0**”, these systems range from turning off a machine which is overheating to more sophisticated artificial intelligence applications. One of the challenges here is managing the complexity of computing architectures capable of delivering the necessary processing power while complying with energy and time restrictions [337].

Results from a number of European initiatives aimed at promoting the uptake of cognitive cyber-physical systems have demonstrated that these have helped improve efficiency in a range of European industries, from glass production to aerospace systems [363].

There has also been increasing take-up of the IoT and CPS in **agriculture**, ranging from plant irrigation and disease detection to reducing pesticide use.

However, the stakes are even higher in the industrial IoT and CPS in terms of **security and safety**. While a consumer breach may result in inconvenience for an individual, breaches in the industrial IoT could have results as serious as turning off the power for an entire country, malicious large-scale damage to industrial equipment or turning off healthcare systems in hospitals [351].

The industrial IoT is particularly vulnerable to such attacks, due to factors such as a larger attack surface from the increase in connected sensors and devices, ageing operational equipment and control systems from a pre-connected era which were designed without security in mind, an extremely complicated landscape of operating systems and poor cybersecurity practices such as software updates [350]. As an example, the WannaCry ransomware attack was estimated to have affected more than 200,000 computers across 150 countries, with a cost estimated between hundreds of millions to billions [294]. See 2.2.3.2 *The secure computer*, below, for further detail on this topic.

In summary, for the IoT and CPS to really fulfil their considerable potential, a holistic approach needs to be taken ensuring that the user is able to take advantage of enhanced functionalities while having a seamless experience which masks the complexity of the system. Providing artificial intelligence using edge computing overcomes some of the energy and privacy issues associated with the IoT while enabling a range of new application areas.

There is a growing need for standards and regulations in the CPS and IoT domain. Air, rail, shipping, manufacturing, energy, medical, health, i.e. almost any industrial sector could benefit from IoT and CPS technology [77].

2.2.1.6 VIRTUAL, AUGMENTED AND MIXED REALITY

Virtual reality (VR) refers to the creation of a completely artificial visual world. *Augmented reality* (AR) refers to the addition of some artificial graphical elements to the (picture of) reality. The experiences that overlay graphics on video streams of the physical world are *augmented reality*, and the experiences that occlude your view to present a digital experience are *virtual reality*.

The experiences enabled between these two extremes is *mixed reality* (MR), a term originally introduced in a 1994 paper by Pal Milgram and Fumio Kishino [234]. Mixed reality [282] thus blends real-world and virtual content into hybrid environments where physical and digital objects coexist and interact.

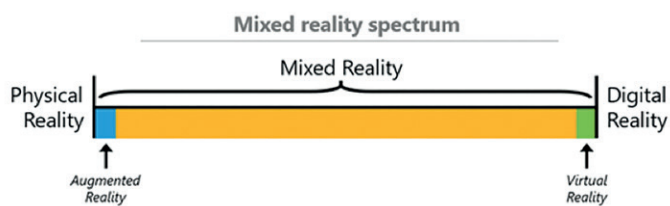


Figure 23: Mixed Reality Spectrum
Source: Microsoft

In recent years, VR, AR and now MR devices have become much more numerous, yet they are still not widespread. Currently, VR is mainly used in games, simulators, and movies; while augmented reality is found in some games and many industrial applications, including simulators. These devices include head-up displays (HUDs), smart glasses, VR/AR headsets, etc. provided by various vendors: Google Glasses, Microsoft HoloLens, Sony SmartEyeglass, HTC Vive, Oculus Rift, Google Cardboard, etc.



Figure 24: Google Cardboard is a handheld VR headset designed to be used with smartphones – Source: Google

Although VR is of course common in games, Pokémon Go was the first widespread AR game. Released in July 2016, it still has a significant user base (147 million monthly active users as of May 2018).

The rise of VR/AR/MR devices and applications has taken place steadily over the past few years, although no “killer app” has yet been found. However, in the last few years, a number of tools like libraries, development kits, etc. have been released by major companies to help developers produce VR/AR/MR applications much more easily.

- **Apple ARKit** (API) [278], released with iOS 11 in September 2017, combines device motion tracking, camera scene capture, advanced scene processing, and display conveniences to simplify the task of building an AR experience on iOS.
- **Google ARCore** [279], released in March 2018, aims at building new augmented reality experiences that seamlessly blend the digital and physical worlds on Android platforms. It relies on three key technologies to integrate virtual content with the real world as seen through one’s phone camera: motion tracking to allow the phone to understand and track its position relative to the world; environmental understanding to allow the phone to detect the size and location of flat horizontal surfaces like the ground or a coffee table, and light estimation to allow the phone to estimate the environment’s current lighting conditions.
- **Google VR** [280] helps create immersive VR experiences. It is multiplatform (Android, Unity, Unreal, iOS., and even web browser), thus covering a very large spectrum on small and large devices and computers. It provides native APIs for key VR features like user input, controller support, and rendering.

In November 2017, Google also released Poly [319], a website for users to browse, distribute, and download 3D objects. It features a free library containing thousands of 3D objects for use in virtual reality and augmented reality applications.

These sample releases show that a complete, accessible, development ecosystem is now becoming available to build VR/AR/MR applications. This is likely to result in a much larger number of such applications in the coming years, significantly increasing the take-up rate of these technologies, and creating numerous business opportunities.

2.2.2 BUSINESS MODELS

Business approaches are also evolving rapidly. Access to different kinds of media is now obtained via the internet and physical media have largely disappeared. Vertical companies control a significant part of the market, while open source has become a credible alternative. Ecosystems are becoming stronger and stronger and serve to cluster the market.

2.2.2.1 RENTING INSTEAD OF BUYING

The digitization of media (audio, video, books, programs, video games...) has profoundly changed civilization; indeed, 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, at an extremely low cost. Thanks

to the worldwide internet, they can potentially be accessed from anywhere in the world.

This was made possible with the performance increase in processing, storage and digital communications, fuelled by Moore's law. For music and video, the process took place in several steps.

DEMATERIALIZED SOUND AND IMAGE: STEPS TOWARDS DIGITIZATION

The first step towards digitization was the replacement of analogue vinyl discs and audio tapes with their digital counterpart, the **compact disc (CD)**. Intermediate steps were taken with the digital cassette (DCC) from the European company Philips and with the Minidisc from the Japanese Sony, but both joined to share patents and initial technology to establish the CD as a standard.

Next, the 650 MB *storage* provided by the CD and the cheap digital processing hardware that followed drove new media applications: storing computer data (**CD-ROM**) and movies (**video-CD**). Specialized processors were able to decompress such a large amount of data in real time that a movie could be also stored on the medium. *Optical technology* improved, allowing more data (4.7 GB) to be stored on the same physical **digital video disc (DVD)**, which increased to 25 GB with **Blu-ray** (in fact, it went up to 100 GB on four layers for 3D movies).

Note that up to this point, the business model was similar to the one for analogue media: people had to buy the physical disc to get the content, with all the consequences this entailed, such as going to the shop and having space to store the discs.

As *digital processing* became cheaper, it was possible to further compress the media (**MP3 format**, developed by the German institute Fraunhofer-Gesellschaft, more or less in the context of the EUREKA project EU147, Digital Audio Broadcasting (DAB) by 1987). The availability of *cheap local storage* then meant that data could be stored locally; a small, low-power hard disk drive (HDD) triggered the first iPod from Apple, then Moore's Law with the increased capacity of solid-state storage with flash memories provided further improvements.

Through its **iTunes music store**, Apple made it easy to get music remotely on your device, avoiding the need to physically move to a shop to buy your music. This was also made possible with the improvement of *servers and communication* with the higher throughput of the internet. But the bandwidth and latency, fine for downloading, were not good enough to ensure high quality streaming.

Digitization also allowed a shift to renting, rather than buying, since listening or viewing quality doesn't degrade with the number of uses, as opposed to long-playing (LP) vinyl records, video and tape cassettes, and other analogue media. Netflix started its mail rental service so that customers wouldn't have to go back to the rental shop or pay late return fees. The notion of owning a physical medium storing a movie or audio track began to gradually fade away, because of the vast choice and the convenience of getting the content remotely.

As *bandwidth* increased, and the servers and all the transmission chain were able to deliver streams in real-time, the streaming model started, avoiding the need for local storage and the associated costs. Theoretically this allowed people to access video and music instantaneously from anywhere. Streaming also enabled users to share their own movies, for example on YouTube, creating a new generation of "prosumers".

The business shifted from hardware manufacturers to content providers, with large data centres being set up. Companies like Amazon first built large data centres for their own use, but later started renting computing resources to external users in order to have them constantly loaded at maximum. Customers no longer had to maintain the (hardware) infrastructure, and could adapt the computing resource to their need (elastic computing), as discussed in 2.2.1.3 "Cloud, fog and edge computing". The consolidation of services on large data centres for rental is a common business practice now.

Big media distributors companies are now pushing the subscription model and streaming of media, because it locks-in users and allow companies to gather more information from their users and their preferences. This is the new gold for so-called "surveillance capitalism" [223]. Knowing the preferences of their users allowed Netflix to begin producing their own television series, "knowing" they will be watched from the analysis of their customers' practices. Digitization has also had a huge impact on the music industry [147]



Figure 25: Vinyl



Figure 26: Cassette



Figure 27: Compact disc

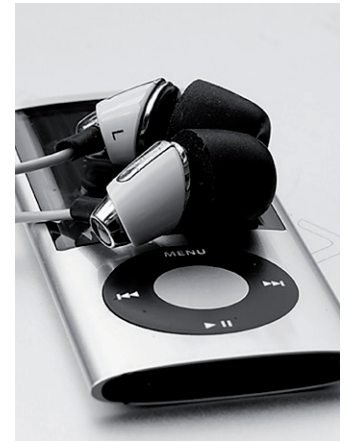


Figure 28: MP4 Player

These new trends in business have had a number of consequences:

- First, and most obviously, representatives of the “old” business models, such as shops selling CDs or DVDs, DVD rental shops and even book shops, are vanishing. This is particularly visible in the USA, where the big media chains of shops of the 2000s have now closed down.
- Even the gaming sector, with its high processing power and bandwidth requirements, is increasingly moving towards streamed games, as discussed in 2.2.3.3 “Gaming: testbed for consumer advanced technologies”.
- The consumer hardware market now focuses almost exclusively on video and audio interfaces for human senses, such as TVs, headphones, loudspeakers (and mobile phones), with more and more homogeneity in the devices and less diversity. A box from your internet provider, a media hub, a good TV, (intelligent) speakers is all you will need at home, along with a tablet to read books and access social media, while a mobile phone with good headphones will be sufficient when you are on the move.

If you are a gamer, you may still need a game console, but there are signs that these are coming to the end of their natural life, as explored in 2.2.3.3 “Gaming: testbed for consumer advanced technologies”. Virtual reality devices could also emerge, but CD players, MP3 players, DVD players and so on are essentially dead.

LAST STORE STANDING

At the time of writing, there was one Blockbuster store still open in the USA, located in Bend, Oregon. A combination of loyal customers and nostalgic tourists seem to make up the bulk of the customers [338].

This technology is also having an impact on how we view the world. With things increasingly becoming non-physical, younger generations have a different view to their parents on the notion of owning.

Three of the five “essential characteristics” of cloud computing as stipulated by the US National Institute of Science and Technology (NIST) [276] – i.e: on-demand self-service, network-accessed, resource pooling, rapid elasticity, measured service – have prompted a major change in the general notion of possession, much beyond their scope as initially conceived. In fact, there is less and less point in seeking permanent possession of a physical good that has a digital equivalent (a video, a book, etc.) or whose availability can be summoned instantly (transport, look-ups in a dictionary or knowledge base, etc.). The physical good occupies physical space, which is a scarce resource for many, implies direct costs in terms of money and of time for maintenance and care, and tends to rapidly become obsolete, rarely acquiring value in that process. The digital equivalent has none of those limitations and has one single, vital, prerequisite, which can hardly be renounced in most part of our active life (i.e., connectivity); it is therefore considerably more attractive.

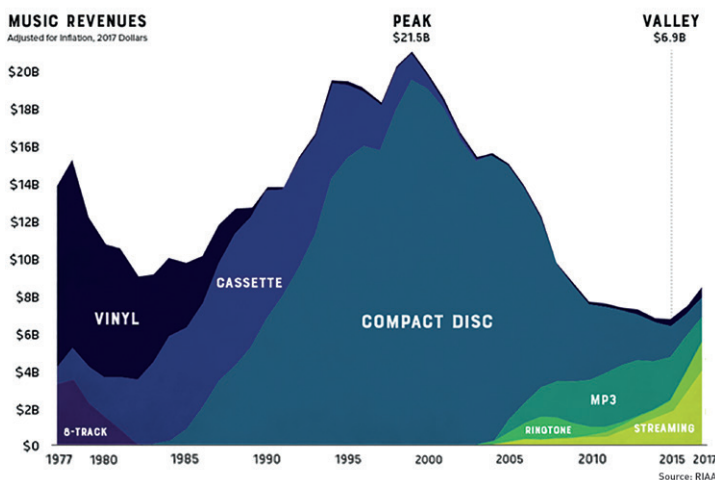


Figure 29: Music revenues 2015-2017

This observation is at the heart of the *as-a-service economy* [186], which is sure to expand far beyond the cloud as we know it and enter 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 device, form a gateway to a gigantic and ever-growing wealth of potentially cooperating services.

A number of social and technical challenges stem from this notion:

- The increasingly critical dependence on connectivity, which has both psychological and functional traits; a full discussion of the psychological effects may be found in 2.6.2 “Impact of computing technology on people”. If not suitably addressed at both levels, this vulnerability may become hard to sustain. Users have learned ways to mitigate the loss of connectivity: for example, they download instead of streaming when they fear break-ups of connectivity. Service providers have also provided fixes: they cache local copies close to the user, without promising that they are up to date, and do the update as soon as connectivity is restored. Evidently, these mitigations can be improved.
- The efficiency of the architecture of the application and its service infrastructure in the continuum from the user device to the cloud. Much finer-grained criteria than in use today, many of which are non-functional (e.g., privacy, energy, predictability), should be employed to determine where to deploy the individual parts of the application system.
- The interoperability, complementarity and contract-based orchestration of the apps, without which there is bound to be a tremendous amount of unnecessary duplication.

2.2.2.2 VERTICALIZATION AND DOMINANCE OF GLOBAL PLATFORMS (GAFAM + BATX)

The new giants of the economy (Google, Apple, Facebook, Amazon, Microsoft = GAFAM - and Baidu, Alibaba, Tencent, Xiaomi = BATX) are companies that encompass more and more domains, have high added value and try to cover a large part of the value chain. Their business model is based on digital and processing, and they rely on large computing infrastructure or computing devices. For example, while they did not start out in the processor business, they are increasingly tending to design their own processors and/or accelerators.

Apple designs its own processors for its iPhone line of products, with its own very efficient implementation of the ARM instruction set [357]. Famously, this resulted in their previous supplier, Imagination Technologies, being sold off to a Chinese-owned private equity firm.

Google has designed accelerators for deep learning: the TPU line of chips (TPU and cloud TPU for servers, and edge TPU for edge devices). While TPUs and cloud TPUs are for internal use, the edge TPUs will be available to external customers.

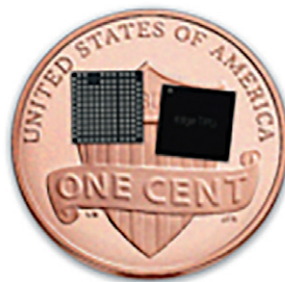


Figure 30: Edge TPU from Google

Amazon and Facebook have also reportedly gone into hardware development. Facebook has been very active in the Open Compute Project (OCP) [292], whose mission is to design and enable the delivery of the most efficient server, storage and data centre hardware designs for scalable computing. Facebook said it has saved about US\$2 billion in three years thanks to this project.

Alibaba has formed a chip subsidiary, Pingtougou [246], and Huawei has its own chip subsidiary (HiSilicon).

By having control of the complete ICT chain including hardware, they can obtain more efficient solutions for their needs and also save money. It is the same rationale that drove Tesla to develop its own chip for its self-driving cars, removing the need to rely on NVIDIA's chips [425].

The following graph shows the revenues of technology companies in the world whose revenue is at least US\$10 billion. The graph shows 23 technology companies, and their aggregate revenue is \$1.6 trillion, which, if it were a country's gross domestic product (GDP), would put them right after Canada on the global scale. Note that this revenue is generated by only 3.4 million employees. The figure shows that the top seven highest revenue companies are vertical.

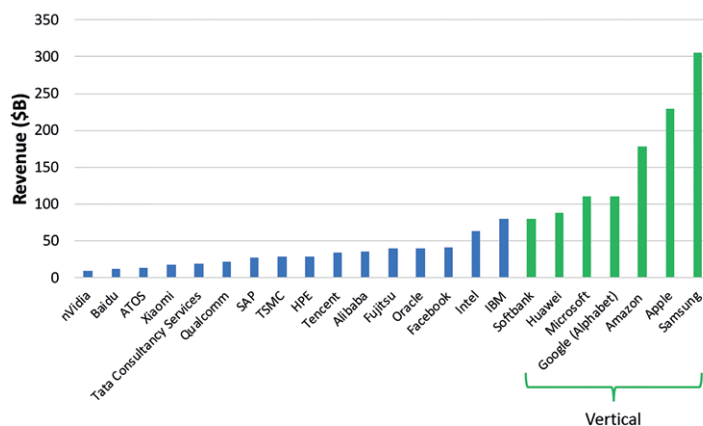


Figure 31: Revenues of technology companies in the world

The following graph shows the productivity in terms of revenue per employee, which is calculated by the total revenue divided by the number of employees. With the exception of Facebook, the most productive five companies are vertical.

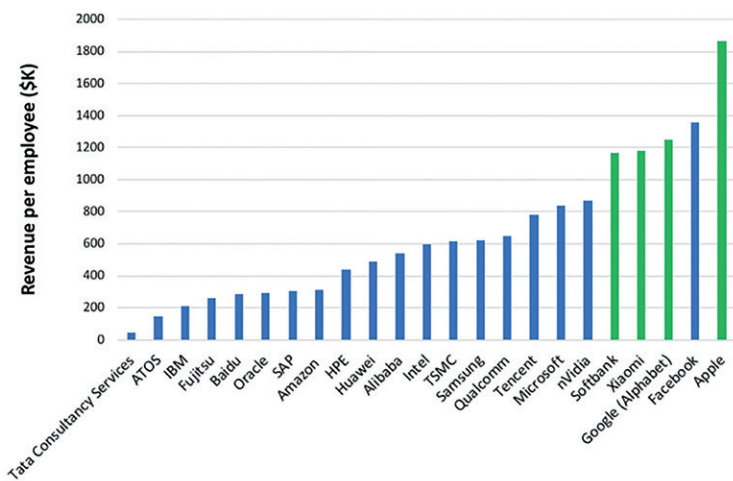


Figure 32: Revenue per employee of technology companies in the world

The following graph shows the total revenues per country on which these companies are based. The USA leads the nearest country by a factor of three. The position of the EU is not very optimistic. In fact, the EU does not have a single vertical computing company.

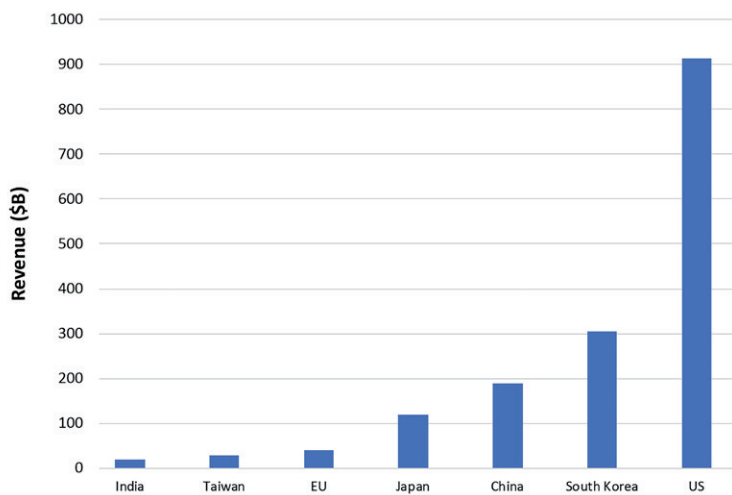


Figure 33: Total revenues per country in which technology companies are based

There are fundamental reasons why it is more difficult to create a global company in Europe. European scale-up companies have to expand in all major European countries, which means hiring local staff, localizing the product or service, and setting up a local sales team. There are no such scale-up barriers in large countries like the USA or China, where the first product can immediately reach hundreds of millions of customers.

In 2015, 99% of all European businesses were small or medium enterprises (SMEs), while 94% were independent (not controlled by another company/not controlling another company), and spread over the whole European territory [73]. SMEs form the backbone of the European economy; they create 85% of new jobs, and represent two thirds of private sector employment in the EU, and 57% of GDP [440]. Instead of focusing on the creation of a few globally leading companies, Europe should focus on the creation of many SMEs, and help them to expand. See section 2.7.1.2.3 for further discussion on this topic.

2.2.2.3 OPEN SOURCE

The term open source is applicable to the sharing of technical information, and predates the information technology revolution that started in the 1970s. More than a century ago, the automobile industry formed the Automobile Manufacturers Association, in which each manufacturer could develop new technology and file patents, but the technology and patents were shared openly between all members of the association without the exchange of money or the filing of lawsuits.

IBM shared the source code of its operating systems in the 1950s and 1960s, an early example of the sharing of software. With the advent of the microcomputer in the mid 1970s, many computer hobbyists shared their own developed software.

2.2.2.3.1 Free Software, Open Source Software, and Open Source Hardware

In the mid-1980s, in an attempt to recreate the spirit of the early days of microcomputer software and hardware development, Richard Stallman started the free software movement. Software is considered free software if its license allows anyone all of the following four rights:

- 0 Run the software
- 1 Study and change the software
- 2 (Re-)distribute the software
- 3 Improve the software.

The term “free” is to be considered to have the same meaning as in “free speech”, as in “to have the liberty to”. This is not the same as “free of charge”, which may well limit the use to running a binary copy of the program.

Note that the third right: (re-)distribute the software, is not restricted to distribution free of charge. An organization may very well sell a piece of free software for a charge that is larger than the cost of distribution. It is up to the buyer if they want to pay for something that could be obtained for less.

Many of the views held by champions of free software are also held by proponents of open source software. The key difference between these two groups, free software proponents and open

source proponents, is the attitude towards proprietary software such as Windows. Whereas proponents of free software regard software as freedom of speech and view proprietary software thus as unethical, proponents of open software have no objection to the existence of proprietary software.

The reader should be aware of the fact that, although companies producing proprietary software may seem in opposition to free and open software, these companies also will allow for the existence of free and open software as it prevents other companies from gaining a monopoly and thereby threatening their existence. While open source software proponents are happy to coexist on this basis with the makers of proprietary software, proponents of free software are not.

2.2.2.3.2 Open source software economics

At first glance, it seems strange that an individual or a company would contribute to open source or free software development, as there are no apparent economic gains. Some people engage in open source development because they enjoy developing software, the incentive being personal satisfaction. Others cooperate in open software development out of idealism to create a better community or world. Being able to show a person's ability to develop quality software is also a means of building a reputation in the software community, opening the possibility to a (better) job. In addition, developing a piece of open software might be a means to acquire that software's functionality without having to pay for it through buying it as proprietary software.

In fact, companies as well as individuals contribute to open source software development. It might act as a means to attract talented developers, it may speed up a system's development by mobilizing more developers and it might also improve the quality. Companies can use open source software as a vehicle to sell services, a business model used by Red Hat, for example. The acquisition of Red Hat by IBM for US\$34 billion underlines the point that open source and big business need not be mutually exclusive.

For some software system categories, there seems to be a need, but hardly any market. Compilers are a good example of this, with the GNU Compiler Collection and LLVM as instances. In some cases, there might still exist a niche market for specialized versions of such software: Intel sells a C-compiler that produces highly optimized code for Intel processors. In this particular example, Intel uses its knowledge of the details of the hardware platform to reach otherwise unattainable levels of code efficiency.

There is a general trend visible: for a particular, often highly specialized application area, there is no market for tooling, but the tool application itself creates a market. There is no market for compilers, but there is a market for the usage of tools (programmers); there is hardly any market for system modelling tools, but

there is a market for model development, and in some cases even for models. There is hardly any market for modelling software for 3D-printers, but there is a market for 3D models of objects.

For Europe to capitalize in this area, it should maintain and enhance its tool building capability, thereby also increasing its tool application and modelling expertise.

2.2.2.3.3 Open source hardware

The hardware situation is significantly different to that of software. The development of software, from idea to executable binary, is nowadays 99% design and just 1% production. The design consists of what is colloquially called software development, and encompasses the definition of the functionality of the system, its translation into a system architecture, and finally to code. The production of the software is in fact "no more" than the compilation step. This last step requires computer infrastructure, which is very cheap compared to the human resources doing all the design steps.

Hardware development is different in its production steps: the largest investment here is in chip production equipment. The cost of an extreme ultraviolet lithography (EUV) wafer scanner is around €100M. There is a shorter, cheaper way to produce hardware by using field-programmable gate arrays (FPGAs), but the resulting devices are in many respects less efficient than a dedicated chip implementation.

Around 2013, in the wake of the Snowden disclosures, the realization that proprietary hardware might contain undesired or even unwanted functionality got a foothold. As a reaction, several governments outside the USA, notably Russia, initiated the development of processors to get full control of nationally deployed government hardware. See 2.6.5.1 "High-performance computing" for more on this topic.

Open source hardware takes the idea of having full control over hardware designs a step further. It is founded on the belief that the free exchange of hardware designs will in the end produce efficient and safe hardware.

One of the big advantages of open source hardware is the openness of the hardware to security auditing, a very important aspect in the quickly changing cyber security landscape. With the design of a hardware component freely accessible, it is also open for scrutiny by many designers, which reduces the risk of potential security holes going undetected.

CHINESE DEVELOPMENTS

The Loongson (formerly called Godson), is a MIPS architecture based processor family, developed since 2001 by the Chinese Institute of Computing Technology and the Chinese private company BLX Design. Loongson is used as an embedded and as a standalone processor. The latest versions of the Loongson have instructions that allow for efficient x86 architecture instruction emulation, allowing Loongson based systems to execute native x86 with 70% average performance. It is fabricated by STMicroelectronics. A few Loongson based systems have been developed, but have not achieved commercial success. The Loongson has been used in several supercomputer designs. The latest Loongson based supercomputer, the Dawning 6000, is marketed by the Lugon corporation.

The China based Jiāngnán Computing Lab started developing processors around 2005. The development was spurred by the USA threatening to ban the export of processors to China because of their possible use in weapons development. The Sunway, or Shengwei processors are RISC-based, but the architectural details are largely unknown, as these processors are for military purposes. The Sunway based Sunway TaiHulight is a supercomputer ranking at second place in the TOP500 list as of June 2018. It is based on the 64 bit 260-core manycore processor designed by the National High-Performance Integrated Circuit Design Center in Beijing.

In 2010, a project at the University of California, Berkeley, started the development of a RISC architecture that is completely open. This platform called RISC-V, is aimed at designing a versatile instruction set architecture, targeting embedded, personal computing, high-performance vector computing, and parallel computing applications. But is also designed for computer architecture education and as a platform for academic research.

Backed by several large companies including AMD, BAE, Google, Hewlett Packard Enterprises, IBM, NVIDIA, and Micron among others, this architecture is gaining traction, not in the least because of the availability of supporting software such as compilers and a version of Linux. In Europe, the low-power PULP platform [394] created by the University of Bologna and ETH Zürich is based on RISC-V, while the European Processor Initiative includes a RISC-V strand [403].

Three years after RISC-V, in 2013, IBM launched the OpenPOWER foundation, based on its POWER architecture. OpenPOWER is not open source hardware. The foundation is based on a partner model, in which a partner brings in its own Intellectual Property to gain access to the architecture.

The Open Compute project, or OCP, originated in Facebook's realization that serving its exponentially growing user community would require a close guard on computer infrastructure investments, both in terms of building and in running costs. After the development and building of a new data centre that is claimed to be about 25% less expensive to build and about 40% more energy efficient, Facebook, together with four partners, including Intel, founded the Open Compute Project Foundation in 2011, hoping to spark the same kind of creativity in the hardware world as open software had done for the software community.

In the following years, the number of partners grew to 14 members. The goal of the Open Compute project is to design scalable data centre products, such as servers and switches. The project advocates the free exchange of component designs with the goal of producing scalable and thereby efficient hardware for data centres.

RISC-V and the Open Compute platform are not the first open source hardware platforms to be launched. Their most notable predecessors are the DLX, developed at Berkeley as an academic platform; ARM cores up to and including v2.0, developed by ARM; and OpenRISC, developed in 2000. However, with the exception of ARM, these designs met with limited commercial success.

LinkedIn has recently founded the Open19 Foundation which aims at standardizing datacentre infrastructure designs, such as racks, power distribution, and networking links.

These initiatives show that large web based companies are driving data centre technology in many aspects, and are starting to make their way into other markets as well, as illustrated by the fact that OCP is also finding its way in the telecom provider market [348].

2.2.2.3.4 Open source machine learning

The rapid rise in interest in artificial intelligence (AI) in the last couple of years has also created an open source movement for this area. The AI world at this moment is mostly focused on machine learning (ML), a subfield of AI. A score of open source ML packages is available; see for example [290]. About half of the packages listed contain a set of ML algorithms, while some packages concentrate on one particular algorithm. The packages also differ in the implementation language used, such as Python, Java, and C/C++, which also indicates differences in intended audience. Some packages are meant for educational purposes, others for production code.

TENSORFLOW

A notable example of an open source machine learning package is TensorFlow from Google. TensorFlow is a dataflow programming package for a range of tasks. Besides symbolic mathematics, it can also be used for neural networks. The Google TensorFlow processing unit (TPU) is tailored for TensorFlow.

TensorFlow grew out of Google's proprietary DistBelief package, which development started in 2011. TensorFlow was released in 2017. TensorFlow is used inside Google both for experimentation and for production. There is also a stripped down version for Android, called TensorFlow Lite.

TensorFlow has a number of APIs, some with backwards compatibility guarantee (Python, C), some without (C++, Go, Java, JavaScript, Swift).

OPENAI

As noted in this HiPEAC Vision, AI is seen both as a blessing and a threat. The threat of AI, in particular, has sparked an open source movement which is unusual in its motivation. Some researchers believe that development of AI can result in human extinction or other unrecoverable global catastrophe. This belief is based on the observation that currently the human species dominates all other species through distinctive features of the human brain. If a technology surpasses the human brain in general intelligence, it may dominate all other organisms, including humans, making humans dependent on this superintelligence.

Developing AI technology in the open will expose its features, potential use, and potential threats to the whole world, preventing anyone or any group to take sole advantage from the technology.

A notable example of this movement is OpenAI [318], a non-profit research institute founded in late 2015 by Elon Musk, among others. The goal of the OpenAI institute is to focus on long-term positive impact of AI on the human species. Although Musk acknowledges the fact that developing AI poses a risk anyway, he believes in "empower(ing) as many people as possible to have AI. If everyone has AI powers, then there's not any one person or a small set of individuals who can have AI superpower." [399].

OpenAI has released an AI benchmark, a virtual meta-learning robot, a debate game machine learning application, a robotized team for a video game, and an application for training a robot hand.

Open source hardware is definitely making traction in defining standards for massive compute intensive infrastructure such as data centres, driving both construction and operating costs down. Although Europe lacks web giants comparable to Google and Amazon, the lowered cost enables Europe to develop new applications based on this equipment.

In the AI application area, a lot of emphasis is on machine learning, with accompanying open source initiatives. It is, however, likely that other AI application area will also grow in the coming years. The current emphasis on neural networks has shown these applications to be hardware and power hungry when implemented with digital technology. But research is starting to direct its focus to non-digital technology as well, such as neuromorphic computing. Although in itself not the target area of the HiPEAC community, interfacing and embedding these new technologies in cyber-physical systems (CPS) is an area the community should work on.

2.2.2.4 CREATING ECOSYSTEMS

In ICT, ecosystems are important and drive the industry. A company has to create an ecosystem because it cannot do everything. For example, the success of Apple is largely due to the App Store, where independent developers can create new applications and usages of the Apple hardware. The success of the PC was also due to the numerous software that could run on the platform. Android and iOS with their relatively open development environments and easy way to sell and distribute apps (thanks to app stores) are the major platforms for mobile phone. The two ecosystems coexist, meaning that developers often have more work to support their work on both platforms.

For the IoT, Apple is promoting its HomeKit with the aim of creating an ecosystem of devices interoperable and controlled by Apple products, taking security into account.

Amazon and Google are trying also to create ecosystems around their personal assistant (Alexa and Google) by adding "skills" (i.e. small interface programs running on the cloud) that allow the creation of *ad hoc* interfaces with various devices. The pioneer is IFTTT [304] that allows the API of various products to be linked together. At the time of writing, there was an important difference between the two companies' approaches:

- Google interfaces with devices through their cloud interface, i.e. from Google servers to the web interface of the service of the device (for example, to control Philips Hue lights, you need to open a Hue account so that the Google server can talk to the Hue server that will send a command to the lights).

- For certain devices, Amazon uses the same approach as Google, but their terminal can also directly talk to the IoT device on the local network without using a cloud-to-web interface. The Amazon Echo plus has a ZigBee interface allowing this direct link with ZigBee devices (for example, to control Philips Hue lights, not even a Hue bridge is required as the Echo “talks” directly to the light bulbs).

All these ecosystems are based on *de facto* standards and APIs that are provided by companies that are able to attract enough customers and developers to create the need for a published API and development tools, in order to have more and more solutions integrated in their ecosystems. They are very different to solutions provided by standardization committees, but there are more and more devices, and interoperability while ensuring security and privacy will become more and more important. We are still in the infancy of IoT devices, and it is still a kind of jungle where *ad hoc* interfaces are developed with security and privacy as an afterthought. We often see a large number of home surveillance cameras or devices being hacked because of basic security measures not being implemented.

Current IoT solutions are often not even scalable: you can buy smart bulbs that you can control with your smartphone, which in fact are controlled by a server on the other side of the world: your phone has to communicate to a server located several thousand kilometres away to switch on a light which is few metres from you! Besides latency, this is not energy efficient (due to the energy cost of transmitting information to the other side of earth) and has a large surface of attack for hackers, while another solution could be to directly connect the bulb to your smartphone, using a means of local communication (ZigBee, Bluetooth or WiFi). There are several billion lightbulbs on earth, and even if they do not often communicate, the accumulated requirements in bandwidth may be also a limitation to this model of using the cloud as a controller.

Europe should enforce security of the devices used in Europe, and work to ensure interoperability. This can be done through regulation, but also by helping the development of (*de facto*) standards, which can be done using open source software repositories (for developers) but also using an “app store” with apps meeting a certain number of criteria to be accepted. Digital Innovation Hubs could also help by creating synergies between innovative companies, developers and users.

2.2.3 BUSINESS DOMAINS AND OPPORTUNITIES

2.2.3.1 AUTOMOTIVE: THE NEXT FRONTIER?

Few inventions have had a more profound impact on society than the car. Originally a replacement for horses and carriages, it soon became a status symbol. It led to new mass production techniques, the creation of millions of jobs, the emergence of the oil industry, the creation of a highway network, etc. It also allowed people to travel longer distances to go to work or visit family and friends, and it was the start of mass tourism as we know it. Today it is the cause of traffic jams, air pollution and 1.25 million traffic deaths per year. Hence, the economic and social impact cannot be underestimated.

Cars are also an expression of the owner’s personality and status; sports car owners, for example, tend to have a different profile to minivan drivers. Many people spend quite a lot of time carefully picking a car that reflects their personality. In that search, they are not looking for the cheapest car, but they are willing to spend more money to buy the right car.

Car manufacturers have created a strategy to survive in this market. They have specialized in particular market segments: luxury cars, sports cars, cheap cars and so on. They innovate continuously to gain market share within their segment. Their current focus is on minimizing the negative effects of using a car, such as by adding satellite navigation systems in order to reduce the stress of driving in an unknown area, or to avoid traffic jams; by selling electric cars to people who are concerned about the environment; or by integrating driver assistance system to make driving more relaxed and safer for the driver, the passengers and the people in the street.

Technically speaking, a car is an engineering marvel. It is an example of the compact integration of many disciplines of mechanical, chemical (fuel), materials and electrical engineering in a small and affordable consumer product. Cars have always been at the forefront of modern engineering. It comes to no surprise that car manufacturers are currently heavily investing in reducing emissions, electrical cars, self-driving cars and so forth, because they believe that this will give them the competitive advantage they need in order to grow.

A modern electric self-driving car is a further integration of computer engineering and software engineering into a traditional car. In the future, a car might become a computer on wheels; its value might be more in the software than in the mechanical hardware. All the key enabling technologies to develop self-driving cars are available: accurate positioning and sensing, high-performance embedded computing, high-bandwidth wireless connectivity for traffic management, car-to-car communication and infotainment, artificial intelligence to interpret scenes.

Behind-the-Scenes Software Will Capture the Largest Slice of the Autonomous Car Opportunity

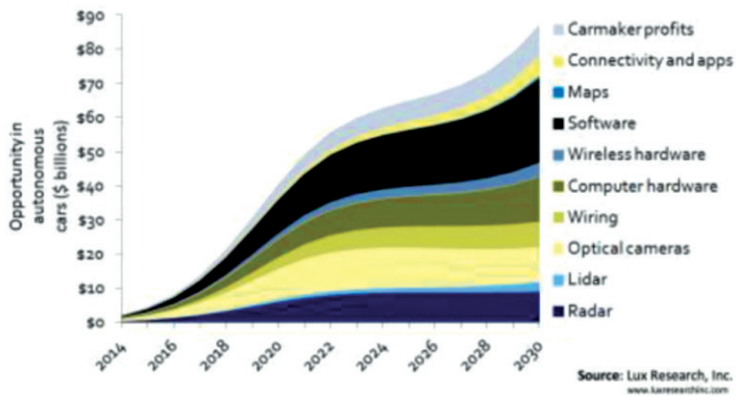


Figure 34: Software will capture the largest slice of the Autonomous Car Opportunity

Given the position of Europe in the car industry, this also represents an opportunity for the European computer and semiconductor industry. It is a matter of time before autonomous vehicles become mainstream. The biggest market for autonomous cars in 2035 will be Asia Pacific (estimated 45%), followed by the USA (20%) and Europe (20%), with smaller markets in the Middle East and Africa. Nevertheless, car manufacturing will remain a very important industry sector in Europe and a large provider of jobs all over Europe.

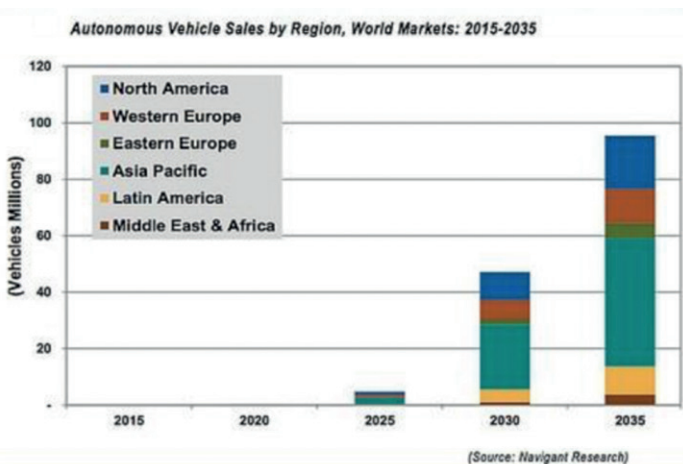


Figure 35: Autonomous Vehicle Sales by Region

This transition – which is certain to happen – will lead to profound changes in the way we organize transportation system and cities. It might also be the start for the transition from product-based to service-based car mobility.

2.2.3.2 MEDICAL AND WELLBEING

For a long time now, there has been a distinct link between healthcare systems and technological evolutions. These include advances in all kinds of medical imaging (magnetic resonance imaging (MRI), computed tomography (CT), etc.) that have allowed much more precise, safe, and fast diagnostics for a plethora of diseases. Similarly, advances in the technology to analyse the human genome allow people to be warned of certain hereditary diseases, and help advance the research of diseases

themselves. These medical technology innovations have already had an enormous impact.

This trend will not stop. First of all, the ageing population will not necessarily want to be confined to a hospital bed or care facilities; this is furthermore not scalable to the aging population. Advances in medical technology will allow older people to stay in their homes for longer.

However, care for older people is not the only booming business in the healthcare sector. Improvements in energy efficiency, battery technology, miniaturization, sensor technologies, etc. enable the development of low-power sensor nodes that track (and possibly influence) different health-related aspects of our bodies around the clock. These personal health technologies can be targeted towards all ages. The most obvious example are pace makers.

Another example that is targeted towards a broader audience are smart watches and sports watches that track the wearer's heartrate. This allows the user to tune their sport workouts to their individual bodies. Recently, Apple announced two new features for their Apple watches, claiming to be able to make an electrocardiogram of the heart and to detect an irregular heart rhythm. While these features have a relatively low bar for regulatory clearance [216], they nevertheless show a very interesting direction in which consumer-targeted personal health applications can evolve. Similarly, a Slovenian start-up is looking to produce a different wearable to detect heart arrhythmias [2].



Figure 36: Apple Watch Series 4, which has a sensor to measure an ECG.

Source: Apple

LAB-ON-A-CHIP

A lab-on-a-chip or microfluidic chip [68] is a device consisting of channels, chambers and valves to analyse fluids. Channels are sub-millimetre in diameter, and fluids can be directed, mixed and separated using the channels, chambers, mixers and valves.

The applications of microfluidics are very diverse; they include DNA extraction, on-chip polymerase chain reaction (PCR), cell analysis (e.g. sorting), single cell imaging, disease diagnosis, drug delivery, pathogen detection, microbial fuel cell and more. Microfluidics are extremely low cost, high throughput and fast measurement and analysis.

Today, microfluidic devices are isolated devices, and are not connected. Analysis results are normally observed using a fluorescent microscope. In the future, they will be used in highly integrated biomedical cyber-physical systems in which a microfluidic engine will be a part of a cyber-physical system that is tightly integrated to the compute engine, memory, image and other sensors and communications.

The microfluidic engine in an integrated system can be used to analyse fluid samples and make diagnoses using the integrated sensors and compute engine, and the results are communicated through a radio communication component in the integrated device or it can be used as an energy harvester [53] to generate energy to power the entire device. These devices will have a high impact on the healthcare market, in particular, in the area of low-cost point-of-care diagnosis and implantable monitoring devices [16].

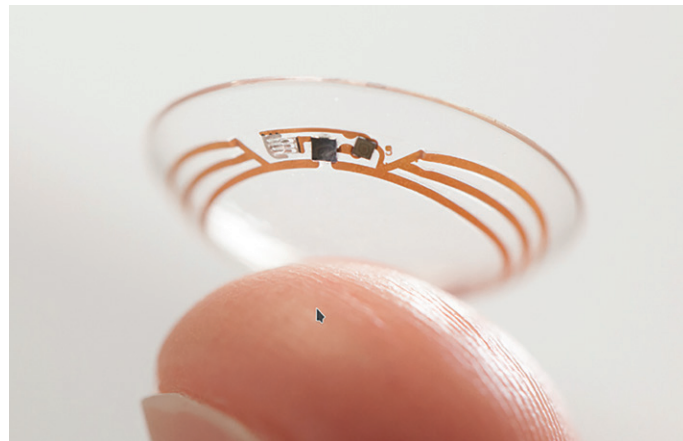


Figure 37: smart contact lens.

Source: Google

These examples show that healthcare is, and will remain, a key area for businesses to invest in. The adoption of such technologies will not be immediate: in addition to unavoidable safety and regulatory aspects when bringing such devices to the market, not everyone will immediately adopt these technologies. The early adopters will be those for whom such technologies directly affect their quality of life, and those who have an immediate interest in monitoring their health. These will ensure the momentum for the general population to adopt such technologies.

2.2.3.3 GAMING: TESTBED FOR CONSUMER ADVANCED TECHNOLOGIES

Gaming is a major driver of the IT industry, worth about US\$138 billion globally, with roughly 50% of the market on mobile, 25% on console and 25% on PCs.

As shown in in figure 38, it is also believed to grow significantly in the future.

The gaming industry has for years relied on the use of cutting-edge, powerful computing systems to provide for the needs of fast gaming and rich virtual world environments. “Gamer PC” has become synonymous with a high-end, boosted machine, compared to those used by most professional people daily. Gaming is thus a relatively low-risk, yet high market-value, testbed for advanced consumer technologies.

One clear example is how gaming acts as a driver for innovation in the semiconductor industry: graphics processing units (GPUs) were initially developed for rendering 3D games, while recently GPU company NVIDIA announced a new breakthrough with accelerators for ray tracing, allowing photorealistic games in real time, as detailed in 2.2.3.3.2 “Real-time ray tracing”.

As discussed in 2.2.1.6 “Virtual, augmented and mixed reality”, virtual reality or augmented reality is another area where gaming is at the forefront of innovation. While not as widespread as expected a few years ago, the augmented/virtual reality ecosystem will certainly grow in the future, mainly by reducing the price of the devices.

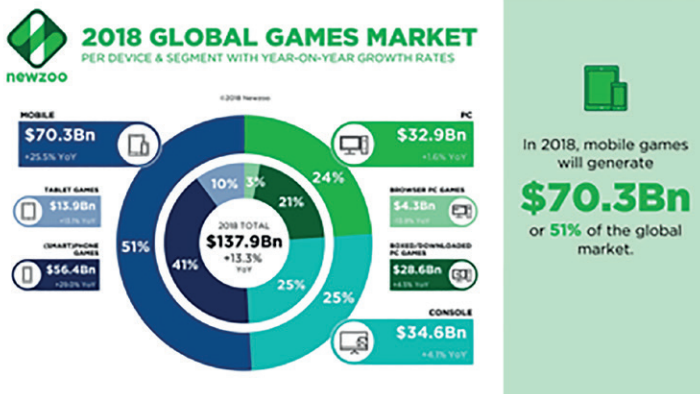
Such sensors can also improve the quality of life of patients with chronic diseases such as diabetes. As an example, Senseonics has developed the Eversense XL continuous glucose monitoring device [20]. This is an implantable device that allows wireless monitoring of a patient’s glucose level during a period of 180 days without the need for blood samples. A less invasive technique is the Freestyle Libre continuous glucose monitoring device developed by Abbott Diabetes Care [64]. This device can be worn continuously on the patient’s arm for a period of up to ten days and nights without needing to be replaced (and they can sleep, swim, etc. while wearing it), and allows patients to monitor their glucose levels discreetly. This is a significant improvement of such patients’ quality of life.

Perhaps even more futuristic is Google’s patent application for a smart contact lens that senses the wearer’s glucose level [442]. Unfortunately, this project was recently cancelled [461], but research on smart contact lenses continues.



these two sides). Nowadays, the trend is cloud gaming [289], where more is executed on the server and less on the local machine. Microsoft has announced that the next Xbox would be cloud-based, while NVIDIA also offers its own cloud gaming GeForce Now, like Sony's Playstation Now service.

The next evolution of this is to have the whole game executed on the server, with the local machine serving only as a display machine, very similar to streaming services for music (Spotify-like services) or for watching movies and series (Netflix-like services). This is also very much a reminder of past client-server architectures.



This way, there would be no more need to have very powerful high-end machines to play the games, since all computations would be performed on remote cloud servers. Portability issues would significantly decrease, as developers would only have to stream their images to an appropriate, high-speed, connection, with a simple display on the end. This would imply a subscription-based business model, popular with editors as it is anti-piracy, since the end-user no longer owns, hence can no longer copy, the game.

However, having games fully executed on server and streamed to the end-user machine poses a number of challenges, because of the graphically-rich content that requires near-instant interaction between the game controller and the graphics on the gamer's screen. When streaming TV or movies, consumers are comfortable with a few seconds of buffering at the start, but streaming high-quality games requires latency measured in milliseconds, with no graphic degradation.

To this end, Google, together with Ubisoft, have recently unveiled Project Stream (former code name Yeti) [267], a technical test to solve some of the biggest challenges of streaming, by pushing the limits with one of the most demanding applications for streaming: a blockbuster video game. This project aims at initially making Assassin's Creed Odyssey playable in Chrome through their game streaming service, then to have this available for all Chromecast devices. All that would be needed from consumers would be a browser, an internet connection, and not much else.

2.2.3.3.2 Real-time ray tracing

NVIDIA recently announced their Turing GPU architecture [364] that it claims is able to for the first time to deliver real-time ray tracing.

RT-C2, OR THE PERILS OF ACRONYMS

Note that NVIDIA often uses RT as an acronym for "ray tracing", while RT is often used in the computing community for "real time".

Figure 38: Global Market share of gaming, Source: [205]

Even with its high requirement on processing power and display bandwidth, gaming is no longer limited to PCs or game consoles with large storage: the content is becoming increasingly dematerialized, with games being downloaded from internet shops such as Steam; even the game engine can be run on the cloud (Cloud gaming [386]) thanks to the low latency of the modern internet.

In this section, we discuss two of these trends in gaming: the rise of *cloud gaming* (or *streaming gaming*) and the arrival of *real-time ray-tracing* on desktop machines.

2.2.3.3.1 Cloud gaming or streaming gaming

Mass consumer gaming has historically gone through a number of phases. First there were hardware-coded gaming machines, then consoles with hardware-coded game cartridges, then PCs with games on floppy disks; in each case, a physical medium was required to access the game. Then came the internet and downloaded games, with (so-called at the time) streaming services like Steam [330] that provided game subscriptions and download capabilities, through an online shop and game repository.

More recently, online gaming made it possible to connect millions of consoles and PCs to massive multiplayer online games, for which part of the game was locally executed and another part executed on a server (with all the synchronization issues between

GPU	Memory (GDDR6)	Memory w/ NVLink	Ray Tracing bandwidth	CUDA Cores	Tensor Cores	Estimated Price
Quadro RTX 8000	48GB	96GB	10 GigaRays/sec	4,608	576	\$10,000
Quadro RTX 6000	24GB	48GB	10 GigaRays/sec	4,608	576	\$6,300
Quadro RTX 5000	16GB	32GB	6 GigaRays/sec	3,072	384	\$2,300

Figure 39: Specification of the NVIDIA Turing-based products – Source: HPCwire

Turing features NVIDIA’s new “RT Cores” to accelerate ray tracing and new “Tensor Cores” for AI inferencing which make real-time ray tracing possible. These two engines, along with more powerful computing power for simulation and enhanced rasterization, allow a game-changing increase in performance, according to NVIDIA. More details on the whole architecture can be found at [392].

applications for the consumer market, both in virtual reality and augmented reality domains.

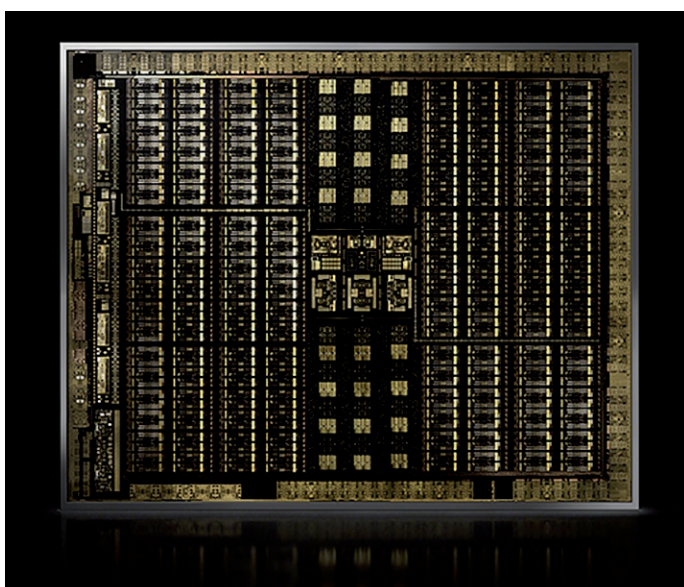


Figure 40: NVIDIA Turing die shot
Source: Nvidia

NVIDIA can thus provide a new generation of hybrid rendering to address complex simulations, such as particles or fluid dynamics for scientific visualization, virtual environments and special effects.

The US\$250 billion visual effects industry is a prime target. Indeed, hybrid rendering enables cinematic-quality interactive experiences, new effects and fluid interactivity on highly complex models. Initial Turing-based products — the NVIDIA® Quadro® RTX™ 8000, Quadro RTX 6000 and Quadro RTX 5000 GPUs — target 50 million designers and artists across multiple industries, and is also bound to interest the gaming industry.

Indeed, having full real-time ray tracing on desktop machines would make it possible to further increase the precision and realism of virtual scene rendering, offering a whole new range of

THE NES CLASSIC EDITION

It is not always the most realistic and powerful game machine that has the most success. In June 2018, the Nintendo NES Classic was the highest unit-selling hardware platform in the USA, beating the PlayStation 4, Nintendo’s Switch, and the Xbox One. “This is the first time a Nintendo Entertainment System console has led in monthly unit sales since NPD tracking began in 1995,” NPD analyst Mat Piscatella said. The NES classic is a modern and smaller version of the Nintendo Entertainment System (NES) of 1983. It is an emulation running on ARM processors with 30 built-in games of the licensed NES library and costs about US\$60.



Figure 41: Nintendo NES Classic
Source: Nintendo

2.2.3.4 DIGITAL TWINS: MASTERING REALITY

The advent of smart manufacturing, one of the flagships of the “Industry 4.0” initiative, entails the concept of “digital twin”, an identical copy of a manufacturing product or process which only exists in the digital space. Digital twins are especially useful for prototyping, testing, and diagnostics, where they can be used to direct actions, settings and modifications to be made to the physical counterpart in the real world, with less risk of unknowns and better efficacy. Software plays a central role in making digital twins usable in manufacturing industry, joining together a variety of competences that include augmented and virtual reality, 3D modelling and big-data analysis.

2.3 REQUIREMENTS FOR ACCEPTABILITY

Europeans are becoming increasingly demanding in terms of non-functional requirements, and security, privacy, safety together with more ecological awareness are more and more necessary to ensure the success of an ICT product. The following sections detail some of the aspects that are required for an ICT product to be accepted.

2.3.1 THE TRUSTABLE COMPUTER

With the ever-increasing interactions humans have with computers, it is important that computers can be trusted. The trustworthiness of a computer encompasses different aspects. We want these devices to be *safe*, which means that when these devices interact with us, they will not harm us. This is particularly important with the proliferation of cyber-physical systems (CPS). Furthermore, these devices should be *secure*, meaning that they cannot be influenced by outsiders and should not leak any information. With the increased connectivity of all devices, and the rise of the internet of things (IoT) in particular, making computers secure against attackers is also of paramount importance. When something unexpected arrives, people want to know why, so that being able to explain to them the why of the decision (of the machine) is important, also for being able to correct it (debug).

2.3.1.1 THE SAFE COMPUTER

That CPS need to be safe, immediately leads to the fact that they should not rely on an instantaneous connection with remote servers for safety-critical decisions; instead, they should be fully

autonomous. This means that both the availability of an internet connection (i.e., the computer can temporarily be cut off from the internet) and its latency cannot be depended upon. Thus, such processing needs to be performed by local processing based on locally available data. The sensor data can, if need be, still be sent to remote servers for further processing and improved services, but the safety of those in the environment of the CPS should not depend on this.

We give a few examples in which the need for such local processing is obvious:

Automotive

The computers in smart cars and self-driving cars need to take split-second decisions on when to brake and how to steer, depending on a large amount of sensor inputs. This information cannot be sent to the cloud to be processed; any such processing would delay the action for too long, if the car is even in an area with connectivity.

Avionics (automatic take-off / landing)

Similarly, but in a 3D environment, aeroplanes heavily rely on even more time-critical automatic systems. These include autopilots [288] routinely auto-landing commercial airliners at speeds above 240 km/h, as well as airborne collision avoidance systems (ACAS) [287]) that must typically manage two (or more) aeroplanes flying toward each other at 800 km/h each. Delays caused by communicating to an external server, not to mention failing communications, are unacceptable in such cases.

53% of consumers experienced cybercrime or know someone who has

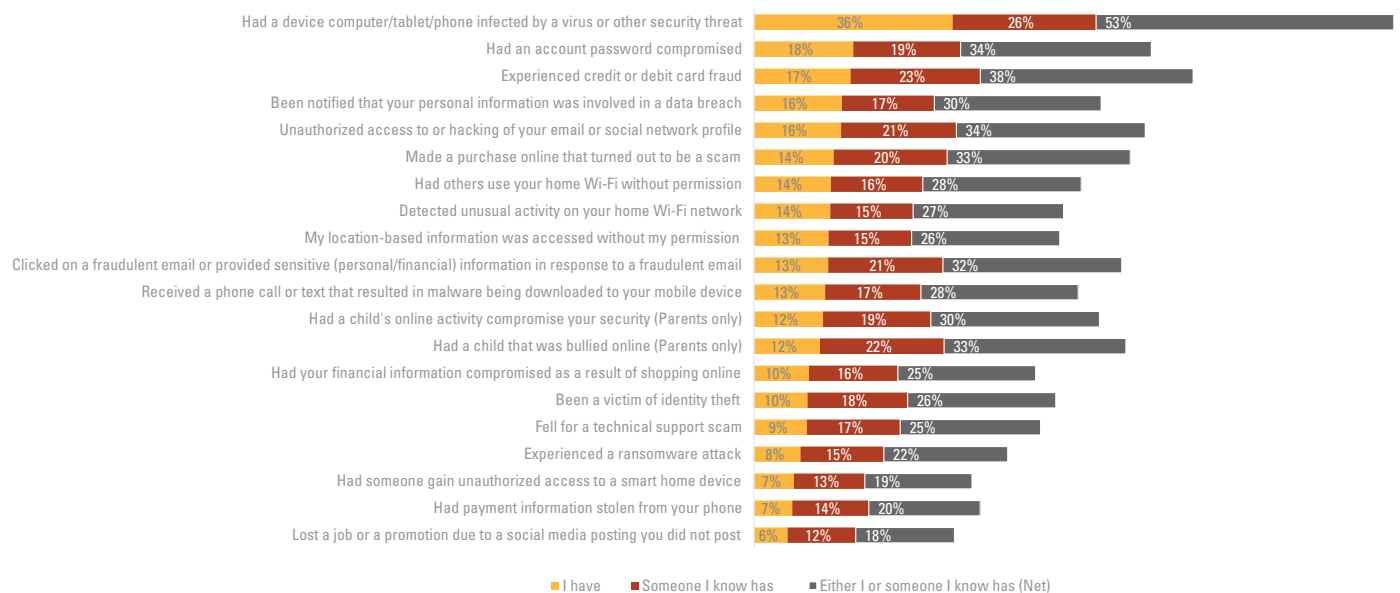


Figure 42: Cybercrime is something that many people have real-life experience with.

Source: Norton Cyber Security Insights Report, 2017

Factory automation

Robots and cobots (industrial robots that are designed to work side-by-side with human operators) need to react quickly to changes in their environment. This again means that they need to process data locally. Furthermore, they cannot always have a fast internet connection, for example when they are on remote oil platforms or in an environment with electromagnetic perturbations.

Wearable, on-body or in-body medical devices

Devices such as insulin pumps and pacemakers need to consume very little power. They do not, and cannot have, a permanent connection to a network. Yet they still need to react to changes in the environment, and act fast when they need to.

An important aspect of all these devices is that they should be fail-safe. This means that in the event that their network connection were to be interrupted, or even that (parts of) their systems were to break or fail, they should fail in a safe way so as to not injure humans.

2.3.1.2 THE SECURE COMPUTER

Now that more and more systems are connected to the internet, and more and more of our data (both personal and commercial) is digital, the security of these systems and data is becoming increasingly important to businesses. Issues range from cybercrime attacks against our systems to the protection of sensitive data.

2.3.1.2.1 Types of attack

Cybercrime is increasing and is mutating in various forms, from various origins. It seems that whereas previous cyberwar and cybercrime used to be relatively *ad hoc*, and “home-made” or “hand-made”, the situation has evolved to more systematic, professionalized, “industrial strength” activities, allowing more numerous, and larger scale and deeper attacks. Ransomware is becoming more common, and targets both individuals and companies. This is often linked to the use of crypto-currencies to make tracking financial flow more difficult.

Cyberwar has become a major concern and seems to be increasing significantly. This form of war presents a number of advantages: it remains cheap, it can be performed remotely and it is very quick. In addition, it does not always leave traces, or traces that can be unambiguously attributed, since traces can be forged to incriminate a third party. Traces of possible state-issued cyberattacks have been mentioned more often in the few past years, targeting at least the USA, UK, France, Germany and Ukraine.

In addition to “traditional” cyber-spying activity, as well as battlefield jamming or hoisting (for example of drones), attacks or pre-attacks on critical civilian infrastructures (such as power grids) have been spotted, with pre-positioning of “charges” or “dormant agents” being suspected in several cases.



Figure 43: Wannacy ransomware caused havoc around the world in 2017

Source: Screenshot of the ransom note left on an infected system



Figure 44: Ransomware is often linked to cryptocurrencies such as Bitcoin

Cyber influence is another form of cyberwar that has made headlines in recent years. It has been spotted apparently from states, who are suspected of influencing the US or French presidential elections, for example. Companies such as Cambridge Analytica [30] and Hacking Team [39] have been reported to use cyber influence and cyberwar-like strategies have also been reported.

2.3.1.2.2 Security threats to the IoT and CPS

The security challenge is increasing fast, since computing systems are more and more widespread. The proliferation of the IoT and CPS is in this respect an important issue. Indeed, as discussed in 2.2.1.5 “Cyber-physical systems and the IoT”, IoT devices are very often badly secured, because of their low cost and time-to-market requirements (for example lightbulbs: [18] and door locks [14]). In addition, they are extremely numerous, which could multiply the effects of an attack targeting them.

However, this challenge is not limited to low-cost products: more and more expensive cars are connected to the internet, but they are not necessarily well protected. Remote vulnerabilities in cars have already lead to at least one costly recall [62]. Nor is the security challenge limited to IoT devices: electronic systems in

cars also need to be protected against local attackers who have physical access to the car, or they could be hacked in that way [199].

Even without physical access, hackers can still try to wreak havoc with badly secured or implemented CPS devices. Attackers can try to corrupt the inputs to such systems in order to have them misbehave in non-obvious ways. For example, some machine learning techniques used to detect traffic signs in self-driving cars can be badly implemented in such a way that attackers can force the car to interpret them as completely different traffic signs with bogus instructions [58].



Figure 45: Putting a small, innocent-looking sticker on a STOP-sign can make it look like a speed limit sign to a machine learning algorithm trained to (un) correctly classify (US) street signs. Source: [203]

Table 1. Analogies between offensive and defensive techniques in the virtual and natural worlds.			
Technique	IT security example	Example from nature	
Offensive	Attracting the victim and fooling it into swallowing the bait (fatal)	Phishing websites or emails	Anglerfish (<i>Lophius piscatorius</i>)
	Disabling an attack's countermeasures	Worms	Bolas spiders (<i>Araneidae</i>)
	Taking control over an entity to use it for your own purposes	Botnet	<i>Ophiocordyceps unilateralis</i> (a pathogenic fungus)
	Communicating covertly	Botnet	Philippine tarsier (<i>Tarsius syrichta</i>), Richardson's ground squirrel (<i>Urocitellus richardsonii</i>)
	Generating numerous unnecessary resources	Spam	Small Balsam (<i>Impatiens parviflora</i>)
	Preventing a legitimate entity from using a resource	Distributed denial of service (DDoS)	Kudzu (<i>Pueraria montana</i>)
Defensive	Attracting the victim to achieve a designated goal (not fatal)	Honeypot	Lady's slipper orchid (<i>Cypripedium calceolus</i>)
	Preventing external threats	Firewalls	Hedgehog spines, porcupine quills, turtle shells, and acacia thorns
	Differentiating between the welcome and unwelcome	Firewalls	Allelopathy phenomenon in a Mexican shrub (<i>Leucaena leucocephala</i>)
	Detecting intruders and preventing attacks	Intrusion detection/prevention systems (IDS/IPSS)	Masked birch caterpillar—larvae of <i>Drepana arcuata</i>

Figure 46: Analogies between the natural and virtual worlds Source: [214]

BIO-INSPIRED SECURITY

Bio-inspired security goes back to the 1980s, when the term computer virus was coined. Computing and biological systems have many similarities, one of which is their vulnerability to attacks by foreign agents. Viruses, bacteria, fungi and parasites are the organisms that attack the biological system, while computing systems are attacked by hackers and malicious programs. Biological systems have capabilities that computing systems can adopt such as self-defence, self-healing, long-term (across generations) memory of attackers, decentralized defence etc.

Bio-inspired security can be classified into different areas:

- 1 Artificial immune systems
- 2 Diversity
- 3 Epidemiology
- 4 Swarm intelligence

ARTIFICIAL IMMUNE SYSTEMS

In the mid-1990s, many research groups applied bio-inspired solutions to a class of computer security problems. During this period, researchers were inspired by the immune system, epidemiology and diversity of species, and proposed solutions to address the problem of anomaly/intrusion detection, multiple versions of computer software and the model of virus spread in a network.

The immune system produces antibodies that recognize antigens or foreign agents and tags these antigens; they are later killed by phagocytes. One theory in immunology is that the immune system has the notion of self and non-self. The self and non-self model of the immune system has been applied to intrusion detection in security [171]. The system looks for odd patterns in network traffic or odd program behaviour. Essentially, the world is divided into "normal" and "abnormal". Detectors (i.e. antibodies) can be defined to recognise either the self or non-self. The detectors are the prediction engines that are trained by the data from the "normal" world using machine learning techniques, and they are randomly generated.

DIVERSITY

Biological systems are more robust to combat attacks because of their diversity, in the sense that each individual is genetically different. On the other hand, computer systems are almost entirely homogeneous, with slight variations in hardware and system software, which makes them more vulnerable to attack. Forrest et al [172] drew inspiration from biological diversity and proposed randomization techniques that can be done by compilers such as padding the stack frame by a random amount or assign a new stack frame in a random location etc.

Fulp et al [55] proposes using genetic algorithms (GA) to create resilient software configurations (e.g. Linux and Apache server settings) deployed in a network in order to protect against cyberattacks. The software configuration setting is defined as a chromosome and security is defined as its fitness function. By applying conventional GA steps like mutation, crossover and selection on the software configuration settings, new secure configurations are discovered. The main motivation is that the secure configurations revealed will thwart potential attackers.

EPIDEMIOLOGY

The behaviour of computer viruses can be better understood by comparing them to the spread of biological viruses, which is the topic of epidemiology. The virus infection model in epidemiology relies on infection and cure rates. If the infection rate is lower than the cure rate, the virus does not spread; if the infection rate is greater than the cure rate, the virus spreads. Several researchers [175] and [154] use epidemic models to understand the behaviour of computer viruses spreading in mobile and social networks.

SWARM INTELLIGENCE

Ants leave secreted hormones on trails to allow their mates to discover paths to food. Inspired by this, the researchers in [4] implemented this strategy to detect anomalies in a smart grid. They created software mobile agents to roam from one smart meter to another to observe any anomalies behaviour in the meter and its neighbour meters. Once it finds a local anomaly, it leaves a message in the meter, analogous to the hormone. When other agents visit the same meter and observe similar anomalies, the anomaly is alerted to the upper agent in the hierarchy.

A honeybee-inspired distributed intrusion detection algorithm [126] has also been developed to detect anomalies in a connected system such as a wireless network.

Bio-inspired security is on the rise because of the increasing number of computer systems (phones, wearables, sensor nodes) and the complexity of the internet thanks to IoT.

An article [214] (as shown in figure 46) points out that nature has more to offer for IT security by providing nature-equivalents to many security attacks and defence strategies used in today's IT security world. In particular, the IoT and wireless sensor networks have remarkable resemblance to insect colonies, and therefore swarm intelligence and behaviour will seem to dominate the other bio-inspired solutions (prevention and detection) and methods in the next decade.

2.3.1.2.3 Protecting user data

Of course, IoT and CPS devices are not the only ones in which the need for secure computers is critical. The increased outsourcing of both data and computations to the cloud has led to the consolidation of the software and hardware stacks of different users. However, because the infrastructure is shared with many different users, and is not located on company premises, these cloud-based systems are much more vulnerable than unconnected, locally run systems. Because businesses and customers want their cloud-based data to be secure from third-party snooping and interference, these systems need to be protected against many different kinds of attack.

Table 1: Number of medically related records breached (across top 10 breaches and those with more than one million stolen records) in the USA as collected by the Office for Civil Rights, Department of Health and Human Services (HHS) in the USA.

2009 (Oct-Dec)	134,773	18
2010	5,932,276	199
2011	13,150,298	195
2012	2,808,042	201
2013	6,939,276	265
2014	12,682,073	289
2015	113,267,174	267
2016	16,655,952	328
2017 (Jan-Apr)	1,828,956	101

In this context, it is important to stress the importance of the entire system being secure and not weakened by back doors. Some countries have argued for the presence of such back doors in operating systems and telecommunications systems, such that only "they" can (lawfully) gain access to systems and decode encrypted information. These backdoors both reduce the security of the entire system (there is no guarantee that the law enforcing agents of your own country will be the only ones with access to these backdoors; other countries and criminals might be able to use these too), and they reduce the overall trust people have in computers and telecommunication systems.

Furthermore, while the protection of personally identifiable information (PII) was important in the past (particularly post-Snowden, although there already existed (medical) privacy laws prior to this [50],[153]), its importance will skyrocket in a post-GDPR Europe. This is in direct conflict with companies collecting, processing, and storing more and more data of their users, and the aforementioned tendency towards the as a service model economy, where processing and storage of data is outsourced to third-party companies (in the cloud).

THE GENERAL DATA PROTECTION REGULATION

The General Data Protection Regulation (EU) 2016/679 (GDPR) is a regulation adopted by the EU on 27 April 2016 to protect personal data and privacy of persons. While the EU already had laws to protect privacy and data in the form of the Data Protection Directive (Directive 95/46/EC), its implementation was scattered across different national laws due to its status of a directive, rather than a regulation. The GDPR came into effect in all member states on 25 May 2018, and on 20 July 2018 in Iceland, Liechtenstein, and Norway.

RIGHTS AND OBLIGATIONS

The GDPR gives natural persons control over many different aspects of their personal data, including that

- they have the right to receive information about the processing of personal data
- they can access the personal data held about them
- they can ask that incorrect, inaccurate or incomplete personal data be corrected (the 'right to rectification')
- they can request that personal data be erased when this data no longer needs to be kept for the purposes for which they were collected (the 'right to erasure', which is also sometimes called the 'right to be forgotten')
- they can object to the processing of your personal data for marketing purposes
- they can request the restriction of the processing of your personal data in specific cases
- they can request their personal data to be delivered in a machine-readable format (the 'right to data portability')
- they can request that decisions based on automated processing about them affecting be made by natural persons, not only by computers.

This (simplified) list already makes it clear that people have clear control over their personal data. Furthermore, the GDPR makes quite a few additional requirements of entities that process or collect this personal data. Some of the more prominent requirements are that they should have data protection by design and by default and have secure processing of data. This requires among other things that companies should limit the amount of data collected to a minimum by default, and that they should take the appropriate safeguards (such as data minimisation, pseudonymisation, encryption, etc.) to protect the personal data.

IMPACT

Even though companies had more than two years to update their policies and practices, many waited up until the last moment to update their privacy policies. This resulted in a sudden wave of (sometimes even unnecessary) emails and notifications to consumers to accept updated privacy policies. Furthermore, some US companies decided to block users from the EU entirely, out of fear that they would otherwise need to comply with the GDPR.

While the GDPR puts a certain burden on companies, these should have a positive effect on the long term. Not only do consumers have more control over their own data, by having requirements on the security and protection of their data, the impact of future data leaks will hopefully be reduced. This is to the benefit of all.

However, its practical implementation can have an opposite effect: many web sites now have a pop-up menu asking the user to select a privacy policy. This is so annoying that most users click the default option without reading, explicitly giving authorisation to do what the web site prefers.

Currently, 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 by doing the data processing on encrypted data, in such a way that the PII is not known to the system performing the actual processing. Examples of such techniques are (fully) homomorphic encryption (FHE) and secure multi-party computation.

There are many fields in which homomorphic encryption would significantly increase the privacy of data in the presence 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 leaking their data to that cloud provider [56]. Similarly, we would be able to have our genome analysed by third parties without information being leaked about which genetic diseases we have or other PII such as sex, race, etc [140].

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 [207], and performing optical character recognition would no longer leak the text being processed [145]. Furthermore, if the recognized text is from licence plates that need to be queried in a database of stolen and wanted vehicles, for example, you can prevent the processing of all licence plates from leaking information about non-stolen cars [195].

Given the urgency for today's business landscape, we predict an increase in the design and use of homomorphic encryption and related techniques. Some start-ups already provide very specific applications of these techniques [393, 449], and some EU projects are trying to make these techniques more usable in practice, such as the HEAT [302] and CLOUDMAP [275] projects.

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. Currently, 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. Rather, pseudonymization of the data might be sufficient [69].

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.

For 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 can be completely broken with quantum computers. Such algorithms need to be replaced with algorithms that could withstand attacks from a quantum computer [295]. This field is called post-quantum cryptography.

However, it is not sufficient to use state-of-the-art encryption algorithms to protect PII. Software that is not secure can obviously 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 (ISA) extensions explicitly have implications on 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 [85].

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 that wants to print his 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 [6].

2.3.1.2.4 Protecting data integrity: Blockchain

In addition to protecting PII from third parties, it is also important to ensure the integrity of data. One recent trend in this area are blockchains, in which cryptographic algorithms and data structures are combined to create an immutable chain of records. Very broadly speaking, there are two categories of blockchains: public blockchains where anyone may participate anonymously, and private blockchains where participation is limited to a known set of participants [219]. The most prominent example of a public blockchain is the Bitcoin network.

While many more groups and companies promote products as featuring blockchain technology, there is currently no real consensus on what constitutes a blockchain. Many projects and persons try to cash in on the hype, either by re-branding existing technologies as block-chain related [436], or even by creating outright scams [242]. Furthermore, many blockchains which claim to be distributed (and thus "better" than alternative, centralized solutions), are in fact still centralized to a certain degree when investigated carefully [219].

Despite these shortcomings, there is a lot of interest in possible applications of blockchain-related technology, including internet domain name registries [315], prediction markets [19], land deed registries [210], share ownership [25], medical licensing registries [82] and supply chain management [79]. It remains to be seen in which domains blockchain technology will make a permanent breakthrough. In general, it seems like the concept of a distributed ledger is what will most likely remain after the hype is over.

As discussed in 2.3.2 "The energy challenge", the power consumption of some types of blockchain technology is an important point of attention. Another societal issue impacting blockchain technologies is that of privacy and the right to be forgotten. While one of the main selling points of blockchains is immutability and censorship-resistance, this has the downside that data cannot easily be erased afterwards. This is in conflict with the expectation from the GDPR that we will be able to have personal information removed when asked. Even though the GDPR and different blockchain technologies might allow for some leeway in how exactly these situations are handled [372], blockchain might face a public backlash from consumers if it turns out to operate contrary to public expectations and public interests with regard to privacy and private information.

2.3.1.2.5 Securing hardware

Evidently, the security of software and the data being processed by that software can only be guaranteed up to the correctness of the hardware on which it executes. Even perfectly safe and secure software can fail in the presence of hardware bugs. While hardware bugs have been known to be a problem for a long time (the most prominent early example is perhaps Intel's infamous floating point bug), their importance is rising dramatically. Most importantly, with the increased consolidation of hardware, many different users share the same underlying hardware. Furthermore, as more and more devices are connected, they expose this potentially buggy hardware to the outside world. This greatly increases the attack surface. Furthermore, with hardware becoming increasingly complex, bugs are becoming ever more present.

New features in modern processors take quite a while to get stable enough to be fully relied upon. This ranges from transactional memory features having to be completely disabled in hardware through microcode updates [332], to security-critical features such as hypervisor functionality not functioning as advertised [430]. In order for CPU vendors to engender trust in their new processor features, and to increase the security of the systems using them, they will have to increase the validation and the security testing of these features.

Unfortunately, the past few years have shown that even well-established processor features can be used to attack systems and lead to leaks of potentially sensitive information. While this had been known to be a problem for a while in the niche of smart cards and cryptography, it is only recently that such attacks have been extended to deal with generic user-facing software, often to dramatic effect.

The recent Spectre and Meltdown hardware vulnerabilities have led to massive changes to operating systems [134] and microcode updates that even had to be recalled due to problems with their stability [137]. These attacks share the characteristic that they are caused by resource sharing. This can either be between different physical cores, such as in the case of (L3) cache-based attacks, but also between different logical cores, or even between code that is actually executed, and code that is transiently executed but whose architectural state is never committed.

It looks as if the floodgates have been opened for these kinds of architectural-level attacks, and it appears that these will remain open for some time to come. Furthermore, as many of these attacks tend to focus on processor features used to execute code faster, and many current workarounds and fixes decrease system performance [138]. Thus, there obviously is growing tension between designing secure processors, and keeping processors fast and power-efficient.



Figure 47: Meltdown and Spectre
Source: Graz University of Technology

Even worse, these attacks also include vulnerabilities due to more 'analogue' effects. For example, Rowhammer attacks start from the fact that inter-cell coupling between rows in a DRAM chip allows attackers to flip bits in rows which are adjacent to (and thus different from) the ones they actually accessed [107]. This effect can then be abused to overwrite kernel memory from user space, and thus to compromise the security of the entire system [300].

Another interesting interaction between the digital and analogue domains that has been demonstrated to undermine the security of a system is energy management: if an attacker can trick the power management to drive the host chip outside its operating region, the resulting faults allow attackers to override and overcome hardware-enforced security boundaries [7].

A major challenge will thus be to design and fabricate hardware that is as free as possible from side channels and other attack vectors.

In the end, trustworthiness and security are properties of the system as a whole. These need to be designed into the entire system, and the entire system needs to be validated for them. This concern starts from the low-level hardware up to the user-facing software, but also includes the whole software stack in between. Furthermore, the system needs to be resilient. This means that systems need built-in ability to recover from attacks.

2.3.1.3 THE EXPLAINABLE COMPUTER

As explained in 2.2.1.2.1, explainability is also a factor of trust, mainly when the system fails: people want to have an explanation and mainly be able to assess who could be held responsible. For the time being, AI-based systems mainly mimic very low-level cognitive processes; they are a long way from understanding "ethics" or having any sense of responsibility.

However, the complexity of processes that may be managed by a machine can make detailed explainability very difficult. Finding a bug is often difficult, and who is responsible? It could be attributed to a lack of complete specifications that might open the way to the introduction of the bug, carelessness on the part of a programmer, or a failure to fully verify the complete system. The

latter is becoming increasingly difficult because, for example, use cases have introduced such a high combinatorial explosion that exhaustive simulations are no longer possible.

The problem is even worse for systems where the decision is not taken by algorithms, but is the result of data analysis (e.g. deep learning-based systems): we need to develop new approaches to be able to manage this new “way” of programming. Trust in a system is established in one of three ways, or a combination of these:

- 1 I understand the system, and, from my cognitive abilities, I am able to create trust on how it was designed or works.
- 2 I trust somebody or an organization who tells me that the system is trustworthy. This is the approach of qualification or certification.

- 3 I experience the system for a certain amount of time, and I see that it worked. I build my trust on my experience. This is typically how a child experiences life: they experience and trust gravity (or its effect) before scientifically knowing Newton’s laws.

A scientific explanation of complex processes is not always true at a particular moment of time: the movement of planets were well explained with the false assumption that Earth was at the centre of the solar system. Therefore, the key point is perhaps not being able to give the full and complete explanation of how a system works (or more importantly why it fails), but an explanation that is good enough.

AI OR AI ? (ARTIFICIAL INTELLIGENCE OR ALIEN INTELLIGENCE?)

The term “artificial intelligence” is not very well chosen to describe our current “AI” systems. Few people will disagree with the fact that computers are not “intelligent” in the human sense. What AlphaZero showed is that tasks we thought were characteristics of “intelligence” can be done with “non-intelligent” devices.

Perhaps, though, we consider intelligence only through our very human-centric view. We have a lot of biases and we believe we

make “intelligent” decisions but they are perhaps not the most optimal; however, we think we can explain them (either by logic, scientific knowledge, or tradition). AlphaGo put stones in places that were “forbidden” by centuries of human players, and finally won the game. Some of the remarks of masters after looking how AlphaGo plays Go or chess show our limitations in the understanding of the “true” nature of – in this case – the games of Go or chess. AlphaGo’s decisions seem “alien” to us, but they are the result of optimizations that haven’t followed some of our biases.

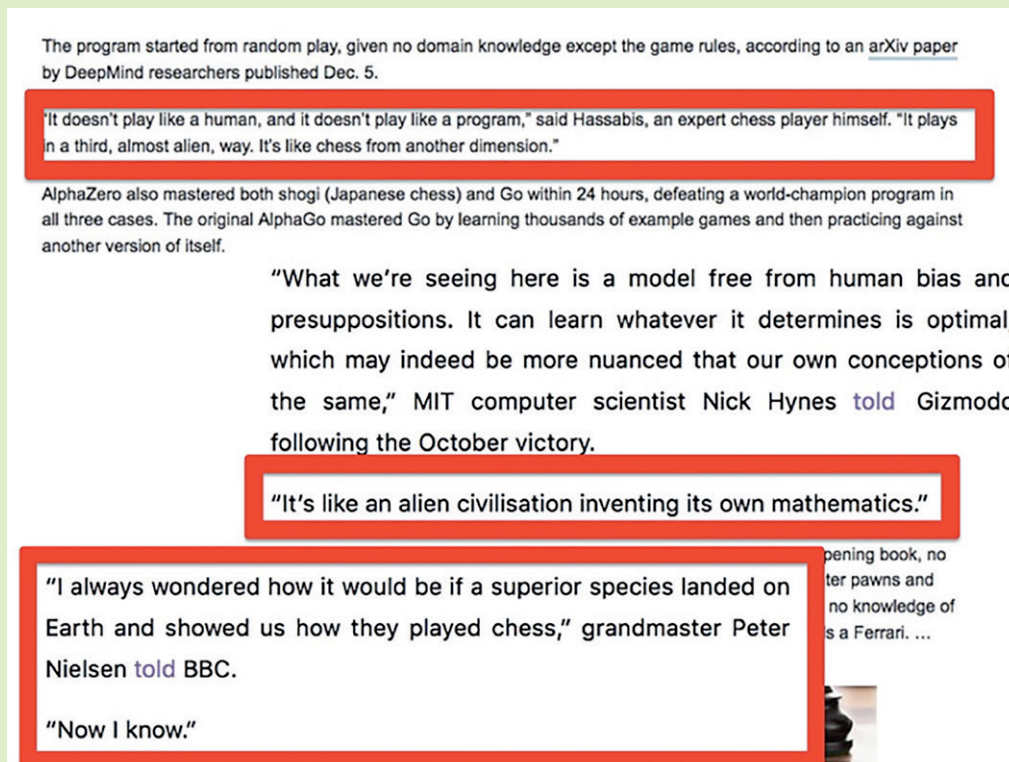


Figure 48: AlphaGo's decisions seem to alien us

2.3.2 THE ENERGY CHALLENGE

Energy considerations are an important aspect of all computer systems, from large to small.

At the largest end of the scale, we have supercomputers and data centres consuming massive amounts of power. For environmental, if not financial, reasons, their power usage cannot keep being increased. Despite of this, we are constantly planning faster and faster supercomputers, which will take massive amounts of power. That computer nodes are becoming ever more power-efficient will do little to stop this trend; an increase in power-efficiency simply means we can do more computations with the same power budget.

It is not only the power consumption of supercomputers that is problematic. Services all have a server-side that consumes power in data centres; and the ever-increasing popularity of consumer electronics devices and high-speed internet are also causing increased power consumption. While the consolidation of computing services in shared environments in data centres can decrease the power consumption of the computing tasks themselves, the power consumption of moving the data from the end-user to these data centres and back is not inconsiderable and must be taken into account. It is important to note that the cost of the energy for communication is invisible to the users.

At the smallest end of the scale, we have mobile devices and IoT devices that need to last as long as possible on a charged battery that is as small and lightweight as possible. This means both that the devices themselves should be manufactured in such a way that they can be power efficient, but also that they should be combined with power-efficient communication systems and energy-aware software.

Thus, all tasks associated with computing should consider their energy consumption. In this section, we discuss the current situation and challenges for the future in keeping the power budget in check on all these computing tasks.



Figure 49: Power use is an ever-important topic for computing devices

Source: rawpixel on Unsplash

2.3.2.1 FOR DATA CENTRES

Although (super)computers have become more power efficient, this does not mean that they consume less power. In fact, in the case of supercomputers, their average power use is increasing, as can be seen in Figure 50.

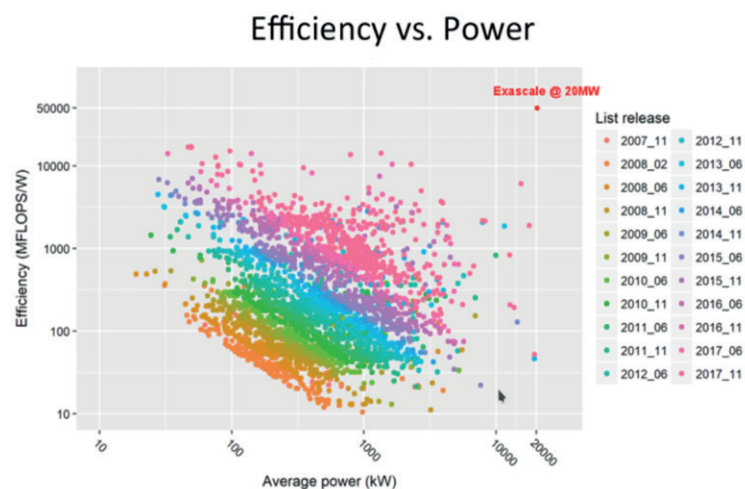


Figure 50: Efficiency vs. power use of supercomputers

Source: The 22nd Green500 List Trends and Evolutions, Nov 2017

The Japanese K supercomputer alone consumes 12.7MW of power, while its recently announced successor, the Post-K exascale supercomputer, is predicted to consume in the range of 30 to 40 MW [247]. While this is a massive amount of energy; this is a significant improvement in power efficiency over previous estimates of its power consumption, which were up to 80MW.

In 2014, data centres in the USA consumed an estimated 70 billion kWh. This represents about 1.8% of total US electricity consumption. Energy use is expected to continue increasing slightly in the near future by 4% from 2014-2020, the same rate as the past five years [17]. When zooming out further to the rest of the world, in 2015 the total electricity consumption of the world's data centres was far higher than that of the UK [368].

There are many different options to keep power use in check. The approach used by Fujitsu for the Post-K supercomputer is to use instruction set architecture extensions and to allow computations to proceed in half-precision floating point (rather than the more traditional double-precision floating point arithmetic) [247]. Half-precision floating point arithmetic allows for more power-efficient computations in application domains where precision can be sacrificed, such as machine learning [221]. Another source of efficiency improvement can come from the use of graphics processing units (typically, more specialized computing units like GPUs are more efficient). Of the top 20 of the most power-efficient supercomputers in June 2018, 17 were powered by NVIDIA Tesla or Volta GPUs [420].

2.3.2.2 FOR CONNECTIVITY

With the increase in on-demand streaming services for audio and video, and the increase in data being sent for different social networks, there is an enormous increase in the amount of data being transferred over the internet. Furthermore, a growing IoT means a growing amount of data being transported. The energy consumption of the internet is becoming a major factor in determining the total energy consumption of the IoT. All these data transfers have an associated energy cost, and thus indirectly affect global emissions of carbon dioxide. In fact, it's possible that within 10 years, internet-connected devices could produce 3.5% of global emissions, and up to 14% of global emissions by 2040, according to new research, reports Climate Home News [407].

The energy consumption of the internet is described by the energy intensity measured as energy per unit of data, e.g. kWh/GByte. It is important to clearly define the boundaries of the internet, i.e. which equipment is considered to belong to the internet and which equipment is not. Different boundaries result in energy intensity variations by as much as four orders of magnitude. For an investigation in 2014: from 136 kWh/GB [11] down to 0.0064 kWh/GB.

The fact that transferring data costs energy can already be observed on a very small scale. For example, on widely used data processing workloads, more than half the energy consumption goes on moving the data from local memory to the computer units on the device [11]. See 2.4.3.2 Near/In memory Computing for further discussion on this topic.

According to Schien et al. [183], the data chain from source to destination is broken up into the following sections:

- 1 End devices, such as smartphones, laptops, and desktops, and server and data storage devices in data centres. Internet-connected TVs are playing an increasing role in this section of the data chain.
- 2 Customer premises equipment, such as WiFi routers and cable modems, connecting the end devices to short haul communication lines to the next section of the data chain.
- 3 Access equipment giving access to metropolitan networks, and effectively to the high data volume networks.
- 4 Metropolitan networks transporting high data volumes over distances in the order of the size of cities.
- 5 Long haul networks transporting data over large distances in the order of distances between cities, or even countries.

Typically, sections 2-5 are considered part of the internet, while 1, consisting of the end user equipment and data centre equipment is considered not to be part of the internet.

2.3.2.2.1 Local wireless access: 5G, LoRa, SigFox, Zigbee, etc

Endpoint devices these days are most often connected wirelessly to the internet. Traditionally, these devices are either connected through a WiFi standard variant, or through mobile data access standards such as EDGE, 3G, 4G, and so on. However, these are not used for power-efficient applications. Standards that focus more on power efficiency have been proposed over the years for connectivity that does not necessarily involve high-bandwidth applications. Some examples of such standards are Bluetooth Low Energy (BLE), LoRa, SigFox, ZigBee, etc. See 2.4.4 "Communication and networking trends" for further discussion of networking standards.

When comparing local sensor nodes in a traditional cyclic sleep scenario (where a short-range sensor node periodically sends small packets of data to a hub), BLE is more power-efficient than other protocols [161]. For low-power long-range sensor nodes, LoRa/LoRaWAN is an attractive option with different tweakable parameters that affect its power consumption [36]. However, it is also possible to create extremely power-efficient passive 802.11b WiFi-sending devices that are even more power-efficient than Bluetooth LTE and ZigBee [157].

When looking at other standards, 5G connectivity aims to reduce energy usage by 90%, which they aim to do by having more and longer power-efficient deep sleep states that have low energy consumption and more efficient data transmissions. Another increase in energy efficiency could come from having more but smaller cells, and transmitting information as much as possible optically to/from these cells [1].

2.3.2.2.2 Local and long-haul networks

The study referenced in [183] concludes that equipment on customer premises contributes up to two orders of magnitude to the energy intensity of the internet. The reason for this large contribution is the relative inefficiency of this equipment. A typical customer on-premise set up involves between one and ten devices – which are not continuously operated – accessing the internet. This results in a relatively low amount of data being transmitted to equipment that is drawing energy 24 hours a day, seven days a week. In contrast, long-haul networks carry data concentrated from many sources over a relatively small number of physical connections. But this results in a much lower energy intensity compared to customer premises equipment, so much lower that it even results in almost negligible energy intensity for this section of the network. Recent studies estimate an average electricity intensity of 0.06 kWh/GB for transmitting data over these long-haul networks [317].

2.3.2.3 FOR SYSTEMS

If we want to reduce the power consumption of all ICT services, we need to look at systems in their entirety. Of course, one major aspect here is that the underlying hardware needs to be power efficient, and the communication protocols need to be power efficient as well. But the software running on top of these need to be able to exploit this in order to have a power-efficient *system*.

For example, consider a set-top box, which is a device that consumers leave powered the entire time (especially when they also have digital video recorder (DVR) functionality). As long as the software running on top of these boxes does not use *any* power-saving features, such a system will not be power-efficient. For example, in 2011, some digital video recorders consumed up to 31W just being idle [43].

However, even when such power-saving features *are* used, the software itself can have a considerable impact. Switching between different algorithms that take roughly the same amount of time to solve the same problem can already have a difference in performance [15].

We give three examples of application domains where the algorithms and systems used have a considerable impact on the power consumption of the applications: machine learning, IoT, and blockchains.

2.3.2.3.1 Machine learning

When training machine learning models, traditionally double-precision floating point numbers have been used. However, this precision is not necessarily needed to achieve the same results on the models. Modern hardware also supports 16 bit half-precision floating point operations.

The switch to half-precision floating point, or a combination of half-precision and single precision floating points has other beneficial effects for machine learning developers: the amount of memory required to train the networks decreases, as does the time needed for the training and inference phases [141]. With the correct techniques, the performance is the same, as can be seen in Figure 51.

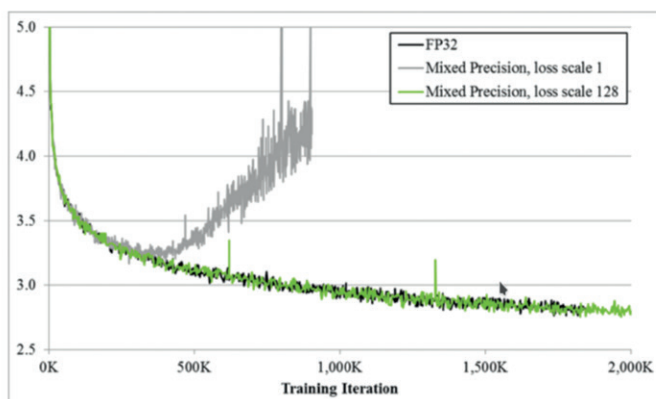


Figure 51: Training curves for the bigLSTM English language model show the benefits of the mixed-precision training techniques
Source: Nvidia

One can even go further, and look into deep learning models with only 16-bit fixed-point arithmetic, rather than floating point arithmetic. Even in this case, we can achieve the same classification accuracy as traditional approaches [48]. In fact, we can even go lower to 2-bit and 1-bit data when performing inferences with deep learning [264].

Furthermore, while until recently machine learning developers used GPUs to train their networks, they are not necessarily the most power-efficient way to do so. As discussed in Section 2.2.2.2 “Verticalization and dominance of global platforms (GAFAM + BATX)”, Google has designed a tensor processing unit (TPU) that is specialized for accelerating the operations most used in machine learning. Right now, they already have three generations of TPUs. The first generation was already 70 times more energy-efficient than GPUs and 200 times more energy-efficient than CPUs for their workloads [101].

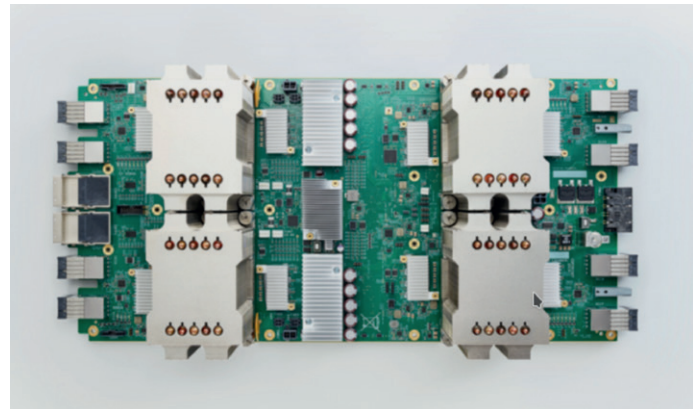


Figure 52: The 180 TFLOPS Cloud TPU card
Source: Google

2.3.2.3.2 IoT

A growing Internet of Things means a growing amount of data being transported. The energy consumption of the internet becomes a major factor in determining the total energy consumption of the IoT.

One challenge in IoT systems is the cost of moving around the data. As discussed in 2.2.1.3 “The continuum: Cloud, fog and edge computing”, depending on the data, the processing required, and the type of connection, it may be better for the power consumption of the node to either process the data locally on the node, or to offload the data to be processed remotely [110]. However, with the growth of IoT systems, we do not only need to worry about the power consumption of all nodes (of which there will be many), but also for the total power consumption worldwide. The rise in IoT systems should not increase the world’s power consumption, either through the power consumption of the nodes themselves, or through the increase in power consumption in the data centres that process the data, or even through the energy required to transport the data over the network from the local nodes to the remote servers.

2.3.2.3.3 Blockchain

One popular new technology that is especially problematic in terms of power consumption is blockchain technology. For example, Bitcoin is estimated to currently consume from 2.55 GW [100] to 5 GW of electricity. This is slightly under 1% of world electricity consumption [219]. In Figure 53, we see the energy consumption of Bitcoin during the course of a year. We clearly see

that not only all the associated mining operations consume massive amounts of energy; the energy consumption also keeps increasing over time. While private blockchains typically do not require power-inefficient mining, power consumption will definitely need to be considered for public blockchains, given that we will have to be careful with our power consumption if we want to have sustainable technologies.

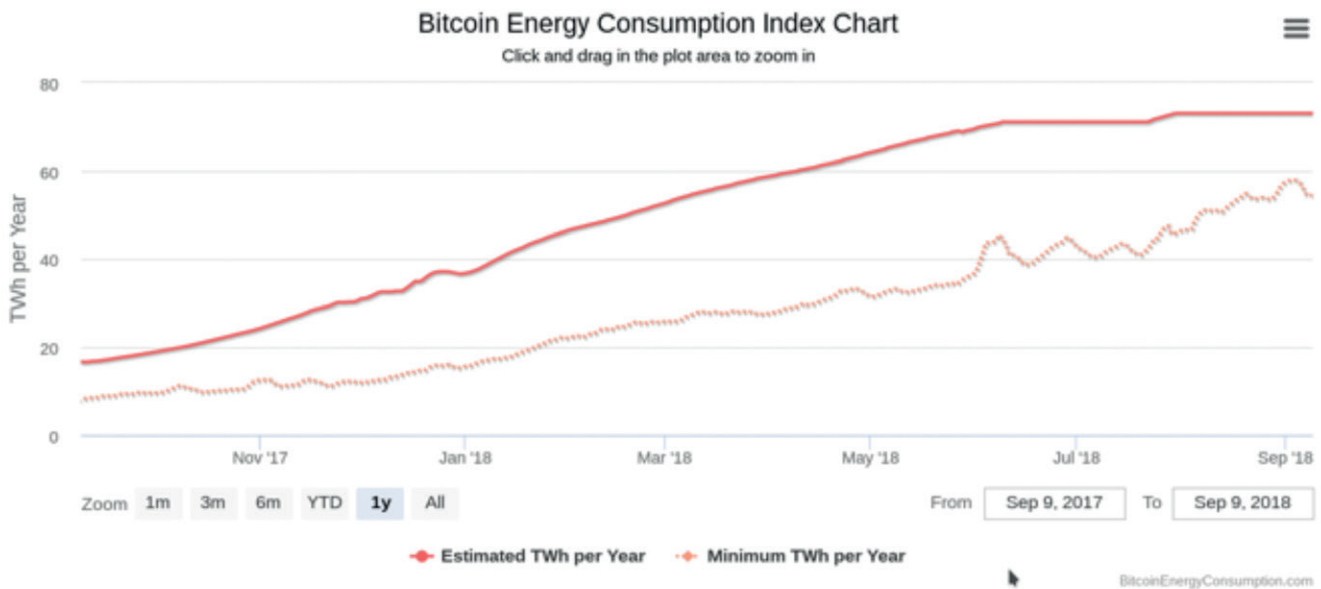


Figure 53: The energy consumption of the most popular blockchain technology, Bitcoin, has risen dramatically. It consumes more than the state of New York. – Source: Bitcoin Energy Consumption Index

ZERO POWER COMPUTING

Zero-power computing refers to the self-sufficiency of a computing device (for example, in the IoT) in terms of supplying energy to it using energy harvesters. The device may have energy storage, such as a battery or supercapacitor, that is always charged by energy harvesters and should never be recharged by human intervention. Zero power computing is also known by other names such as energy-neutral computing, intermittent computing [24]. The system needs to store sufficient energy through energy harvesters to do any useful work. When no energy is available, the system goes into a sleep mode until it stores energy again. Transient computing [72] is one step further in zero-power computing to operate the device with solely energy harvesters without any energy storage.

A typical zero-power computing system is shown in figure 54a. Figure 54b shows the energy-neutrality zone where the y-axis is the total system energy and the x-axis represents the harvested energy. The system will be self-sufficient in terms of energy as

long as the harvested energy is always larger than the system energy at any time.

Zero-power computing has unique challenges. For example, it makes the forward progress of an application unpredictable due to intermittent execution patterns. It may leave memory inconsistent, and also may not respond to sensor data in real time, while concurrency is difficult across multiple IoT devices in a collaborative environment such as a wireless sensor network.

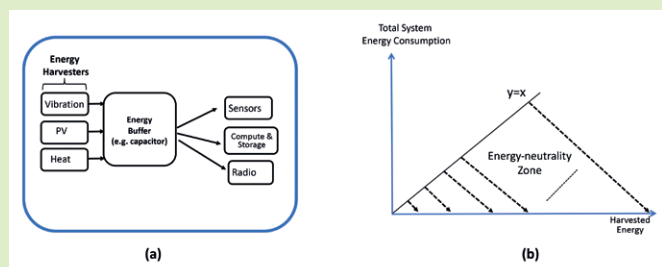


Figure 54: (a) A typical IoT device in the zero-power computing world; (b) the device has to be in the energy-neutrality zone

Also, with the increasing performance demands by the emerging applications, future IoT devices will be equipped with ambient intelligence towards becoming self-learning [54] and visual IoT [164] devices, as discussed in 2.2.1.5 “Cyber-physical systems and the IoT”. They may accommodate machine learning hardware accelerators (for example, a deep neural network) that perform compute- and memory-intensive tasks such as situational awareness, anomaly detection, activity, and pattern and emotion recognition.

Real-time response is needed because an anomaly or critical activity must be detected or recognized in situ and reported immediately, because transmitting the sensor data via radio to

its host to do this will be costly in terms of energy and latency. For example, an implantable chip must detect an abnormal condition in the organ and must take an action in real time because it cannot afford to wait for a critical decision to be made by the host.

Self-learning and visual IoT devices relying on energy harvesters will add extra possibilities to the challenges in zero-power computing. Architectural, software and system solutions must be devised to address these challenges in addition to relying on the technological advances in energy harvesters in the next decade.

In the first part of the document, we have shown the business drivers and constraints that are ingredients for an ICT product. The next part will explain that the silicon CMOS technology, which was the enabler for the ever increasing performance of our ICT devices, is struggling to keep its pace of progress. As a result, we need to seek out new solutions, at all levels – technological, architectural, hardware, software, and so on – to keep improving our ICT systems so that they will offer more features and an ever-improving level of performance to help satisfy our needs and meet societal challenges.

2.4 TECHNOLOGY DIRECTIONS

Given that Dennard scaling has ended, that Moore's law (in its original form of transistor cost) is reaching an end, and that the cost involved in developing new technology nodes is skyrocketing, new directions should be investigated to continue improving the performance of storage and communication units. This represents a brand-new opportunity for Europe to demonstrate its creativity and to invent innovative solutions that break away from current advances relying on technology improvements. It is time to revisit the basics.

The end of scaling will not only have an effect on the processor but also on communications and storage. Increasing communication bandwidth requires more sophisticated protocols and higher speeds which requires more processing power in the network nodes. Without scaling, this will translate in increased power consumption up to the point where it becomes technically and economically unfeasible to further increase the communication bandwidth.

The same holds for storage. Solid state memories are also based on lithography. When scaling ends, the scaling of SSDs ends too. Future patterned media require features that are beyond the resolution of modern lithography.

That said, disruptive innovation is not easy: in academia, researchers are locked into incremental research because of the prevailing "publish or perish" model, while in industry genuine innovation is a risk. Often, newcomers in a domain, who haven't followed the field's *de facto* rules, are able to introduce disruptive innovations. This was the case of Apple with the iPhone in the area of mobile phones and of Tesla in the area of cars.

In this section, we will not go into too much detail about "classical" silicon technologies, as these are well known and have been described in previous editions of the HIPEAC Vision. Instead, we will introduce more disruptive technologies, to provide inspiration and help trigger new activities.

At a more global level, we observe a shift within the ICT domain from compute centric to data centric, from applications (triggered by big data and also deep learning) down to hardware (we are beginning to realize the cost of moving data, hence the striving for "computing in memory" or having computing near data, which translates at all levels, from chips to systems with edge computing).

We can define the current era as "**More computing at the edge for improved safety, privacy, and cost of ownership**". Data is growing faster than Moore's law, and from the five Vs of big data (volume, velocity, variety, veracity and value - [373]), we are at time where we have volume and value, and are realizing the cost of the rest.

Computing is becoming a continuum, from sensors, data fusion, processing, storage, communication (or communication then storage) where data are progressively refined into useful information. Google proposes the following successive refinement of data from one level to another: from data, to information, to knowledge, to wisdom. Current ICT systems are efficient at the first step (from data to information); the progress of artificial intelligence will allow the step from information to knowledge to be refined; and perhaps artificial general intelligence, once developed, will do the last step?

2.4.1 TECHNOLOGY

In this section, we consider the first element of the compute-communications-storage computing systems triangle: computing units. As we will see, with CMOS scaling now reaching its absolute limits, new technologies – some more radical than others – are on the horizon, but these are more likely to exist alongside CMOS than replace it.

2.4.1.1 LIMITATIONS OF THE CURRENT CMOS TECHNOLOGY AND SILICON ROADMAP

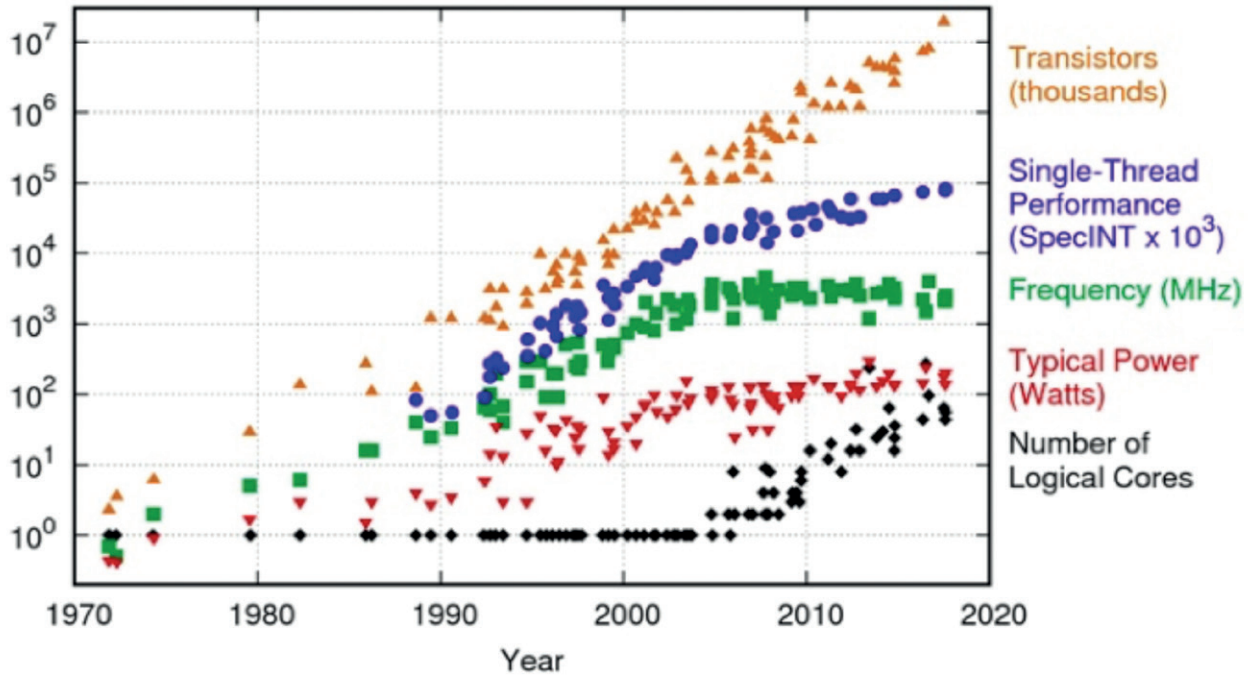
2.4.1.1.1 When the transistors stopped shrinking

News of the impending end of the silicon roadmap has reached the general public; however, the situation is more complex than it appears at first glance. The attention of the press, following Intel's lead in microelectronics for years, has mostly focused on Moore's law. In reality, this was an observation made by Gordon Moore in 1965 before complementary metal-oxide semiconductor (CMOS) device technology was even applied, which states that the **density** of the components on a chip doubles typically every 18 months. Since the sixties, Moore's law has ruled the world of chip technology, driving down the cost per transistor while at the same allowing the clock frequency and the transistor density to rise.

This statement is more a business model than a law; the real law derived from the physics of the MOS transistors was outlined by Robert H. Dennard in 1974. This law pertains to the physics of the MOS devices and makes explicit the relation between **reducing the size** of a MOS transistor and improvements in a number of other physical parameters, notably the power density (constant) and the speed.

In fact, improvements thanks to shrinking transistors had already slowed down or disappeared by 2005. In his paper, Dennard had already indicated where this scaling would start breaking down: there is a physical limit for which junctions can no longer be controlled in standard structures. This limit, related to the concentrations of dopant atoms in the silicon, is reached at around 20 to 30nm of gate length. While the power dissipation per unit area is constant, meaning that at some point it is

42 Years of Microprocessor Trend Data



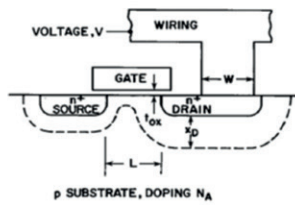
Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
 New plot and data collected for 2010-2017 by K. Rupp

<https://www.karlrupp.net/2018/02/42-years-of-microprocessor-trend-data/>

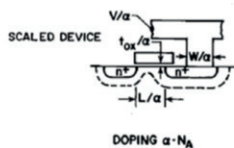
Figure 55: 42 years of microprocessor trend data

THE PHYSICS OF SCALING: DENNARD'S LAW

Device or Circuit Parameter	Scaling Factor
Device dimension t_{ox}, L, W	$1/\kappa$
Doping concentration N_A	κ
Voltage V	$1/\kappa$
Current I	$1/\kappa$
Capacitance $\epsilon A/t$	$1/\kappa$
Delay time/circuit VC/I	$1/\kappa$
Power dissipation/circuit VI	$1/\kappa^2$
Power density VI/A	1



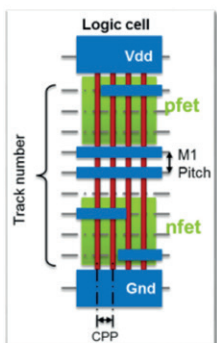
R. Dennard et al., IEEE JSSC, 1974



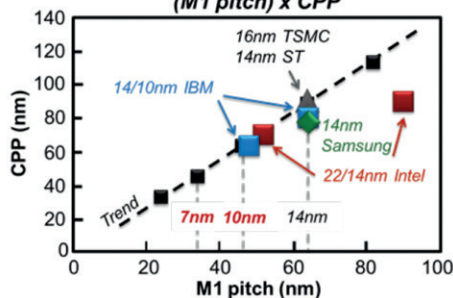
- Scaling improve performance!

Figure 56: Dennard's Law

SCALING: METAL 1 PITCH AND CONTACTED POLY PITCH



The size of logic gate is proportional to:
 $(M1 \text{ pitch}) \times CPP$



10nm node: CPP=64nm, M1 pitch=46nm
 7nm node: CPP=46nm, M1 pitch=34nm

Figure 57: Scaling: Metal 1 pitch and contacted poly pitch

impossible to have all transistors operating at the same time. The patterning at these dimensions is however complicated, as it is well below the wavelength of the light used.

2.4.1.1.2 Equivalent scaling: Compensating with complexity

The first consequences, at circuit level, were that the frequency of operation reached its limit (2-4 GHz), while the total power dissipated also hit a ceiling, as the power per unit area remains constant. The push to achieve smaller and smaller transistor sizes has continued in order to keep reducing the unit cost of the chips, but the era of pure geometrical shrinking has now been over for more than 10 years.

Instead, in order to achieve ever smaller dimensions, major changes have been made to the structure of circuit elements, with the result that they are far more complex. To control the junction after the maximum dopant density was reached, elements needed to be fabricated on thin layers of silicon, bringing about fin field-effect transistor (FinFET) and fully depleted silicon on insulator (FDSOI) technologies.

However, smaller transistors mean lower current density in the transistor channel, limiting the capacity of driving the metallic interconnects and effectively the speed. To overcome this limitation, strained layers and silicon germanium alloys have been introduced, which allow greater current mobility. In order to limit the leakage from the control gate, thicker, high permittivity gates have been introduced by adding hafnium to oxygen.

Nominal node		28nm	22nm	20nm	18nm	16nm	14nm	12nm	10nm	7nm	5nm (ITRS)
Intel	Lg		24				20		16	~12nm	10
	Fin Pitch		60 FinFET				FinFET 42		FinFET 34	FinFET	12
	CPP		90				70		54		32
	M1		80				52		36		16
	SRAM		HD 0.092µm ²				0.0588µm ²		0.0312µm ²	0.027µm ²	
	Year Publication		VLSI 2012				IEDM 2014		IEDM 2017/ISSCC2018	IEDM 2016	
Risk Prod		2011				2014		1Q18			
Samsung	Lg	32		25	25		30		~20	~16	
	Fin Pitch	BULK		BULK	FDSOI		48 FinFET		Single Fin 42	Dual thin EUV 27	
	CPP	114		86	86		78		68	54/57	
	M1	90		64	64		64		51	36	
	SRAM	0.152µm ²		0.084µm ²			0.064/0.08µm ²		0.04µm ²	HD 6T SRAM 0.026µm ²	
	Year Publication	ICSIST 2011		VLSI 2012			JSSC 2014		ISSCC/VLSI 2017	VLSI 2017/ISSCC2017-2018	
Risk Prod	2011		2013			4Q-2015		1Q2017	2H-18		
TSMC	Lg	30	30	30		33		25	~20	~16	
	Fin Pitch	BULK	BULK	BULK		FinFET 45		FinFET 45	FinFET	FinFET 4th	
	CPP	118	105	90		90/80		90/80	64	57	
	M1	90	80	64		64		64	42	40	
	SRAM	0.155µm ²	0.155µm ²			0.07µm ²			0.03µm ²	0.027µm ²	
	Year Publication	VLSI 2012	VLSI 2012	VLSI 2014		IEDM 2013		6Track	VLSI 2016	IEDM 2016	
Risk Prod	2011	2018	2013		4Q-2015		3Q2016	4Q2016	3Q-17		
GF	Lg		28				30				
	Fin Pitch		FDSOI				48 Fin FET				
	CPP		90				78				
	M1		78				67				
	SRAM		0.110µm ²				0.110µm ²				
	Year Publication		IEDM 2016				IEDM 2016				
Risk Prod		2016				2H-2016					
										C.Reita, C.Fenouillet-Beranger - CEA-LETI - 2018	

Figure 58: Nominal vs. actual node dimensions
Source: CEA Leti

To mitigate patterning issues, new lithography systems known as immersion systems were developed. For these, a higher refractive index material (water) is inserted between the wafer and the last lens in the optical path of the patterning tool (stepper).

Collectively, the use of these “tricks” has become known as “equivalent scaling”. Essentially the gate length is not scaled by a factor 0.7 to obtain a factor 2 gain in surface for the device as in the past, but these additions resulted in improved performance and the density still went up by a factor 2 through design techniques.

However, the naming convention for the “equivalent scaling” started to be somewhat misleading. Until 45-40nm, the value which indicated the node (here 40nm) was typically the gate length of the smallest devices or half of the pitch of the densest interconnect metal layer. Suddenly, Intel announced that their technology was not going to be a 28nm (the next step) but a 22nm node – yet the smallest dimension was about 30nm.

Since then, improvements in device size and in performance have slowed down by between 25% and 40% from node to node, while the density, until 7nm, kept increasing by nearly a factor 2 as predicted by Moore’s Law. However, this has been accomplished by layout techniques and lithography “tricks”, that, while very useful for low power designs, do not offer the same gains in high-performance applications.

In addition, the limitation in power also means that not all the circuit elements can work at the same time, leading to the

concept of dark silicon (areas of the chip temporarily off at any given time).

In effect, only some of the device libraries can give the density advantage due to the style of design (fewer tracks connecting to the cells, less equivalent width in the FinFET devices, fewer fingers in the circuit components and so on). For those still requiring maximum performance, like high-performance computing (HPC) designs, the effective density is far from having progressed by a factor 2 every node. Figure 58 provides some examples of node naming by the manufacturers, as opposed to the real dimensions of the design and devices. Where a reference is not provided, the data has been extrapolated from multiple sources.

2.4.1.1.3 Fables/foundry

The complexity of shrinking semiconductor fabrication process has increased the overall cost of building and running fabs (as already predicted by Gordon Moore in his original paper). In turn, this increase has induced a number of companies to drop out of the race for the development of the next generation of devices and has resulted in the beginning of a process of economic consolidation.

In parallel, the increase in capital intensity in the industry has given birth to another phenomenon: the emergence of the fables/foundry model. New players started to have fabrication facilities not for their own products but for manufacturing products for other companies wanting to reduce their investment in factories. This model was pursued very aggressively by US and EU companies which, with the exception of STMicroelectronics in

the EU, decided to stop development either at the generation of 90nm or of 45nm. The buzzword was the transition to a fab-light model in order to reduce investments and optimize use of the asset.

This in turn gave rise to new players, the fabless companies (Qualcomm, NVIDIA, Apple, and so on) that now could design circuits without having the need of worrying about where to fabricate them. Meanwhile, by consolidating demand into large volumes and taking advantage of PC and mobile consumer products, the foundry companies, mostly Taiwanese, were able to grow very quickly and maintain a very high investment rate. As an example, for the period 2005-2015 TSMC maintained an investment rate in production capabilities and research and development (R&D) of above 80% of revenue, with revenues growing at more than 20%.

With the announcement that GlobalFoundries would halt its 7nm fabrication processes [361], the consolidation of the actors has reached a turning point. There are now only two foundries pursuing processes for 7nm and below: Samsung of South Korea and TSMC from Taiwan (also aggressively using new technologies and services to lock in customers [347]), while the only integrated device manufacturer (IDM) left in this pursuit is Intel who, however, is having difficulties in producing its 10nm technology [349].

The introduction of a new lithography tool, extreme ultraviolet (EUV) at very short wavelength (13.5nm), after close to 20 years in development will probably ease some of the patterning difficulties, but its introduction cost is such that no new player has been seen to enter the race.

The Chinese government is aggressively trying to create a domestic advanced microelectronic industry. Often, the acquisition of technology by purchase of foreign companies has been blocked by opposition in the EU and USA opposition, while the investment in local players has not even started to reduce the gap between local and main players from the main players. At present an important major effort is being made to persuade TSMC, Samsung, Micron and Intel to build advanced factories in mainland China by using state financing; however, none of them have built advanced factories so far. They have only built generation n-2 or n-3 factories to serve the generic needs of the internal market. The proximity of Korea and particularly Taiwan, combined with China's political and economic ambitions in this area, may lead to strong temptations of annexation, which would complicate the independence of the supply chain for the rest of the world.

2.4.1.1.4 3nm – the end of the line?

While throughout the last sixty years of microelectronics development it was unclear what technical solution would allow development ten years down the line, there was no doubt that there was no fundamental technical limitation and that competition would guarantee that the necessary investments were made in good time. In addition, R&D provided a number of options with a degree of maturity that gave confidence in their being ready on time.

Over the last ten years the situation has dramatically changed. Technically, the specifications for the 3nm node are very close to the physical limit for transport in semiconductors (a gate length of 7 to 10nm) before stochastic phenomena introduce an intolerable degree of variability. So far, research has failed to identify materials or device architectures with the potential of behaving better than silicon [21].

	GF 3LP	Intel 7nm	Samsung 3GAA	TSMC 3
Year of Production	2021/2022	2020	2021/2022	2021/2022
Devices	HNS	HNS [1]	HNS	HNS
Contacted Poly Pitch (CPP) (nm)	45	37	45	45
Metal 2 Pitch (nm)	32	32	32	28
Track Height	5.00	5.2	5.00	5.00
CPP x M2P x Tracks (nm²)	7,200	6,157	7,200	6,300
Double/Single Diffusion Break	SDB	SDB	SDB	SDB
MTx/mm²	216.37	253.04	216.37	247.28
HD SRAM cell size (µm²)	0.0173	0.0156	0.0187	0.0164

[1] This could be a FinFET or a Horizontal Nano Sheet (HNS), we believe HNS provide a better scaling path to the required dimensions.

Figure 59: 3nm comparison

On the financial side, with the cost of an advanced fabrication facility at nearly US\$7 billion and development costs for a full technology and design platform at over US\$4 billion, it is likely that microelectronics has reached the situation we observe in aircraft manufacturing: two or three players who set the pace of development in order to maximize the return on assets in a technical environment where efficiency takes precedence over performance (“747 vs Concorde”).

With 7nm in ramp-up and 5nm in development, is likely that at 5nm or 3nm we will observe the end of the reduction of dimensions connected with devices. Structures are moving from FinFET to horizontally stacked nanowires and nanosheets. Announced by Samsung and demonstrated in the past both by IBM and CEA-LETI [400], these are really at the limit of what can be obtained in term of scaling of individual components [401]. Nothing better has been shown in the literature in the last ten years, and so we may well have to start looking for further improvements in other parts of the system, such as the assembly of multiple chips, input/output (I/O) management, packaging etc. The pace of performance development will have to rely a lot more on the assembly side of the system and architectural improvements.

2.4.1.2 3D STACKING: AN ANSWER TO CMOS SCALABILITY CHALLENGES

In the context of high-performance computing, networking, and big-data applications, the never-ending quest of computing requires very large systems which are fully scalable, easy to program, at reduced cost and with high energy efficiency. The

proximity of a huge amount of memory is one of the key challenges to address, bearing in mind the “energy cost” of large data transfers.

Until now, designers have developed their components in advanced technology nodes to meet those requirements. Nevertheless, taking all the challenges into account along with the fact that the cost of designing a chip is dramatically unaffordable for the majority of players, a move to heterogeneous integration and modular stacking is gaining more and more interest.

3D-stacking allows the density of transistors to continue increasing, not because of technology scaling, but by using the third dimension to stack dies of silicon one above another. This technology is already in use for some products, but needs to be more widely developed to enable more diversity of chips at an acceptable cost.

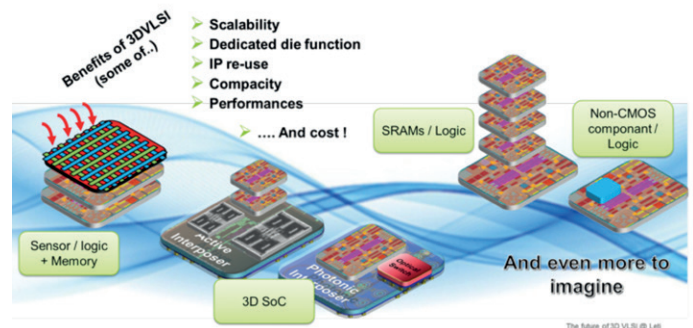


Figure 60: Benefits of 3DVLSI
Source: Leti

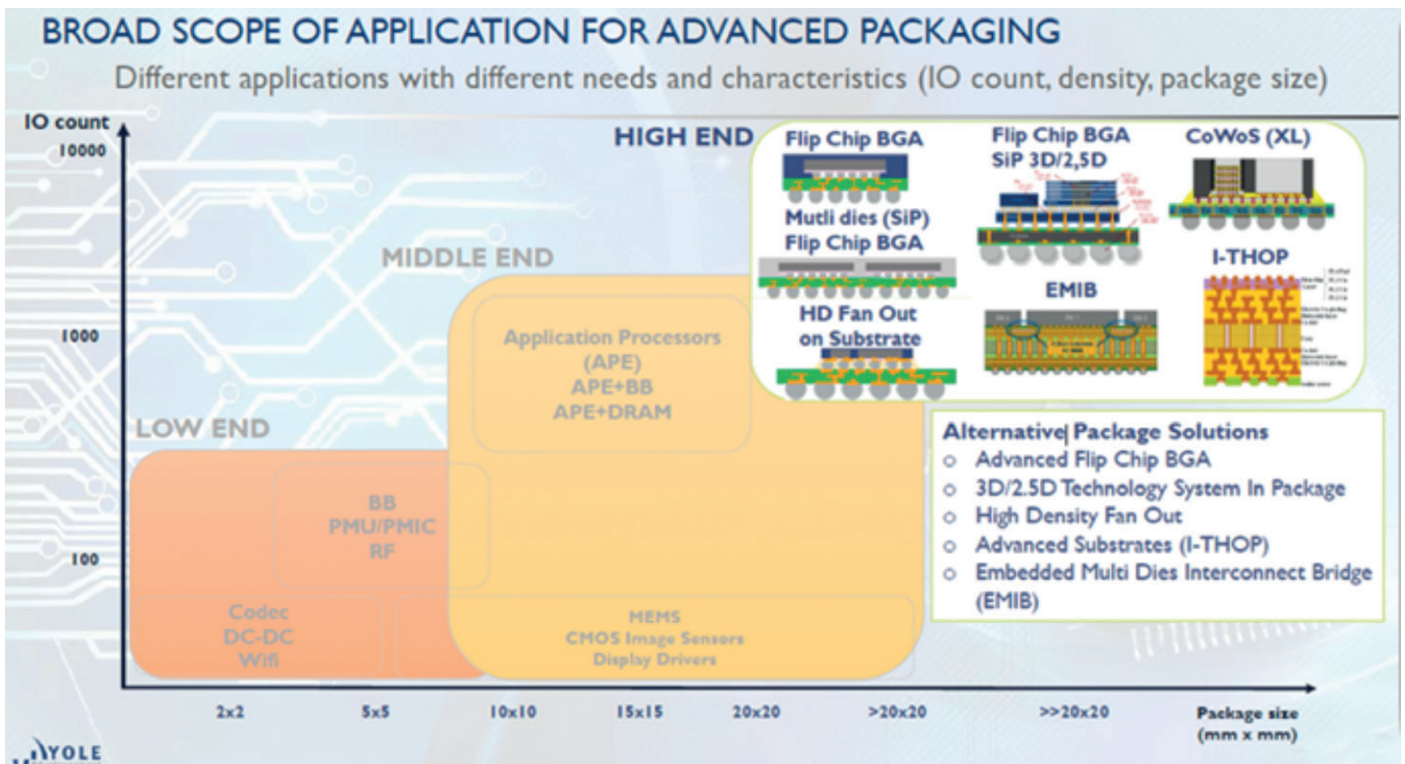


Figure 61: View of different advanced packaging solutions for high-end applications
Source: Yole, 3DTSV & 2,5D 2016 report

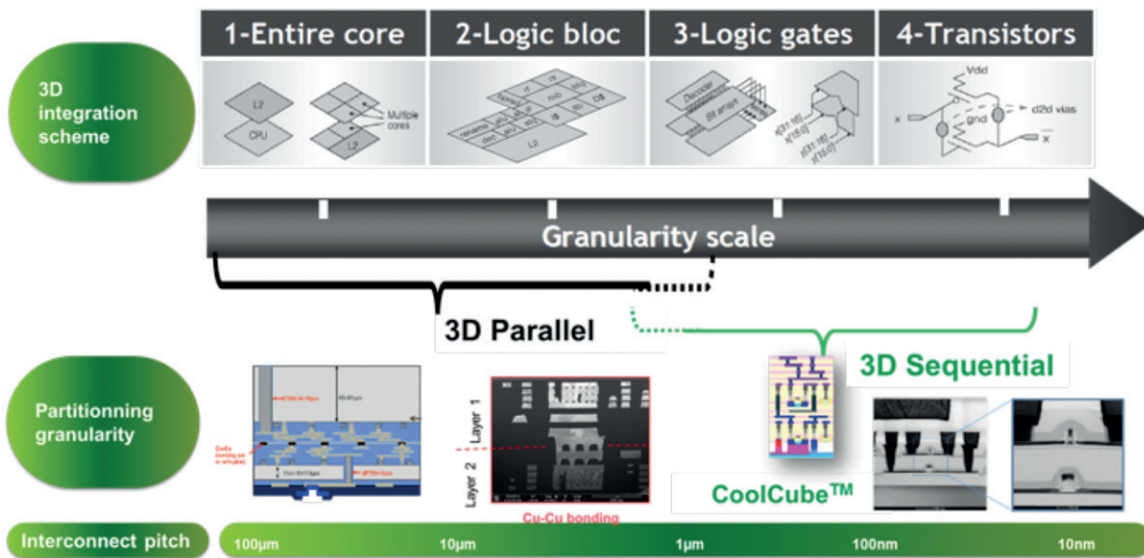


Figure 62: 3D parallel and 3D sequential positioning, depending on partitioning granularity and pitch of interconnect
Source: CEA-Leti

Some of the key advantages of heterogeneous integration compared to classical planar architecture are the following:

- Transistors, or function by surface unit increase.
- Possible re-use of advanced intellectual property (IP), which allows faster time-to-market as well as cost decrease.
- Use the right technology node for the right function.

The first challenge of heterogeneous integration in this context is to maintain or even decrease the energy efficiency of the global system. Several solutions are already available in foundries or outsourced semiconductor assembly and test providers (OSAT) to go in that direction, mainly driven by TSMC with their integrated fan-out (InFO) advanced packaging, or Intel with its EMIB.

Nevertheless, current solutions may meet some tough challenges concerning overall energy consumption: high density of interconnect between separate functions (mainly logic to memory) is absolutely required, with an objective of less than 1pJ/bit per vertical link.

This figure of merit naturally leads to two different but complementary integrations: 3DIC stack (also called 3D parallel) and 3D monolithic (also called 3D sequential). Here also, the technology will be chosen in regard to the architecture and the density required, as shown in figure 62. We will briefly discuss both integrations next.

1 – 3DIC stack

3DIC stack with through-silicon via (TSV) and μ bumps are well known, mainly for field-programmable gate array (FPGA) applications (partitioned dies on a passive silicon interposer), and for high-bandwidth memory (HBM) and hybrid memory cube (HMC) stacking. The pitch between 3D features is in the range of

40 μ m, while a pitch of less 10 μ m is evaluated to reach <1pJ/bit for the consumption of each vertical links.

That's the reason why advanced 3D technology with very small TSV (a diameter in the range of the μ m) and very fine-pitch chip-to-chip interconnects are required. The chip-to-chip fine-pitch interconnect may be based on the μ bumps (copper and solder) interconnect or direct hybrid bonding. Wafer-to-wafer by direct hybrid bonding is also a potential solution, already famous for CMOS image sensor (CIS), by partitioning the sensing layer from the logic layer, and more recently by embedding in the stack a third layer with dynamic random access memory (DRAM).

As an example, DARPA's Common Heterogeneous Integration and IP Reuse Strategies (CHIPS) program demonstrates work in this area, while AMD and CEA-Leti have published work and launched some initiatives towards advanced silicon interposers and chiplets.

In several application domains, advanced 3D integration may advantageously replace a multi-core monolithic planar die.

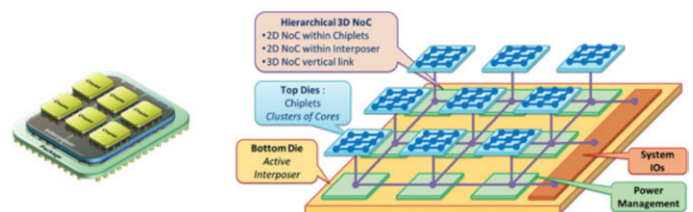


Figure 63: One possible architecture partitioning between chiplets and interposer – Source: Leti

A further advantage of die stacking is the innovative potential it provides to introduce novel materials, such as III-V chemical compounds such as gallium nitride (GaN), onto silicon CMOS wafers. This is another example of “using the right function, the right technology, the right material” for the right function. Additionally, this would save some rare or expensive materials by limiting its use.

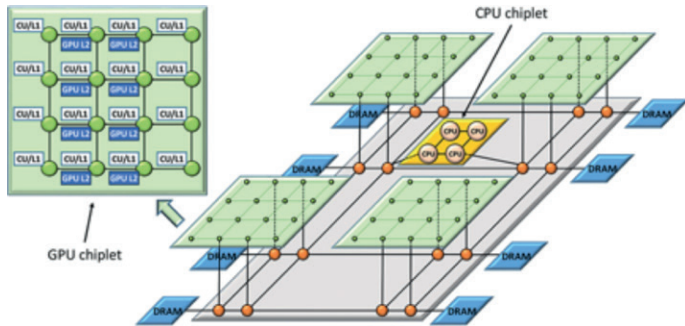


Figure 64: A future system might contain a CPU chiplet and several GPUs all attached to the same piece of network-enabled silicon – Source: AMD

The roadmap of alignment accuracy depends on the technology and the integration, but the objective of reaching under 10µm of pitch has already been achieved, and advanced proofs-of-concept from laboratories have delivered a 1µm pitch for wafer-to-wafer hybrid bonding (Leti, 2017), or 3µm for a µbumps interconnect (Imec, 2017).

On the industrial side, the new AMD “Rome” Epyc processors use this principle: all of the I/O and memory controllers are set into an 14nm “interposer” that sits at the centre of the Rome package. The compute chiplets are 7nm.

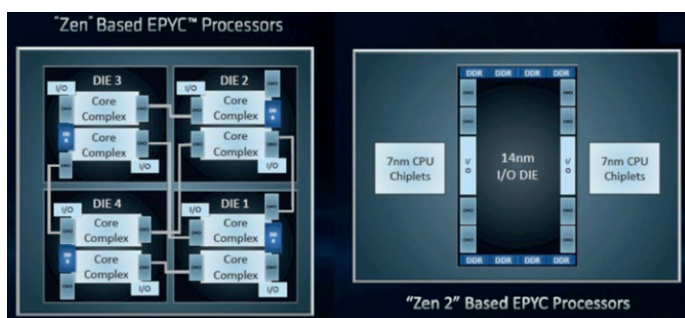


Figure 65: Zen Based EPYC Processors Source: The Next Platform

2- 3D sequential

3D sequential, which consists of the wafer-level manufacturing of a low-temperature device layer on top of a standard Front-End-Of-Line, introduces the notion of very high density on the transistor-to-transistor interconnect, opening up a range of new architecture (sensor on top of CMOS, low-energy CMOS such as FDSOI on top of high performance FinFET devices etc..).

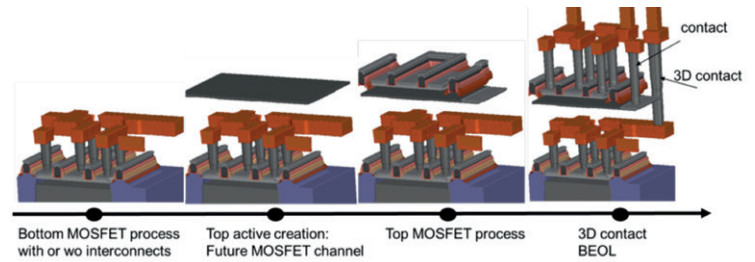


Figure 66: Principle of 3D sequential integration Source: Leti

The alignment of the two layers is performed by lithography (rather than by bonding) which means a possible sub-10nm alignment accuracy between the layers.

The various 3D integrations outlined above are not at the same level of maturity. While wafer-to-wafer stack is already in production in a ≈5µm pitch for image sensors, other solutions may require a learning curve of between 10 to 15 years. In addition to integration challenges, design and computer aided design (CAD) tools may need some disruptive innovations in order to fully exploit such technologies.

2.4.1.3 CRYOGENIC COMPUTING

An alternative to try and increase the speed and power efficiency of computing devices is to try and create a superconducting computer. The idea is to cool the device to such a low temperature that it could take advantage of superconduction effects. For example, superconducting switching devices (Josephson junctions) can switch very quickly and do so with very little energy cost per switch. Furthermore, communication could also take advantage of superconducting transmission lines.

While these approaches are still in their infancy, the end goal here is to create exascale supercomputers at a fraction of the energy cost that would normally be associated with such devices using traditional approaches. Currently, both the US and Chinese governments are backing projects to create a superconducting supercomputer [31, 37].

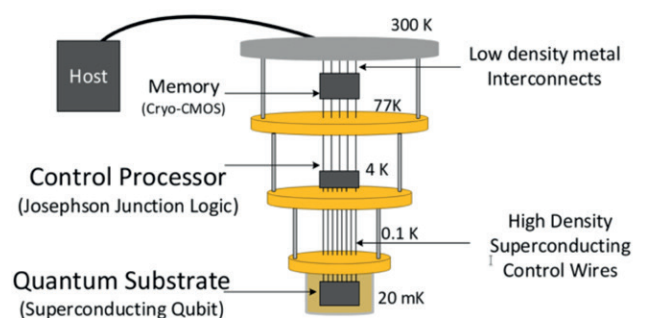


Figure 67: Proposed organization of a scalable superconducting quantum computer. Source: [38]

Another area of computing where cryogenic temperature is being investigated is quantum computing, as discussed in 2.4.1.5 “Quantum computing”. As qubits are very sensitive to noise, quantum computers need to be operated at as low temperatures as possible (in the orders of milliKelvin). These qubits and quantum circuits need to be controlled and programmed. In a proposed architecture, this is done using a superconducting computer operating at 4 Kelvin [162]. If these cryogenic computers then need memory, they can interface with regular DRAM that is cooled to 77 Kelvin [38].

2.4.1.4 PHOTONICS FOR COMPUTING

Using light in computing technology is not a new thing. In fact, photons are used every day in high bandwidth communication links and have contributed to the growth of the internet and our hyperconnected world. 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 also faced major engineering issues: power, photonic/electronic conversion, co-integration with CMOS, etc. that has hindered its adoption in core computing functions.

Nevertheless, the last decade has witnessed a rapid growth of photonics technologies for computing. In particular silicon photonics is nowadays considered a technology that would impact applications such as high-performance computing, data centres and sensing. Silicon photonics is also considered as one of the key technologies to enable novel paradigms such as **neuromorphic computing** (which we will consider in section 2.4.6.1) or **quantum computing** (as discussed in 2.4.1.5 “Quantum computing”).

Photonics computing can also benefit from the steady progress in basic optical components: integrated lights sources such as light-emitting diodes (LEDs) and coherent light sources (lasers) of various wavelength and power, digital micro-mirror devices (DMD), display technologies like liquid crystal display (LCD) or organic LED and integrated camera technologies, high-speed photodiodes arrays and time-of-flight sensors (TOF).

The availability of high-performance photonics components often developed for other applications (e.g. DMDs and LCDs for video projectors, integrated cameras for smart-phones) allows the exploration of various concepts of optical computing. In fact, the basics of optical computing are pretty simple: set up a controlled light source and form a beam, use the light beam to illuminate a 2D or 3D substrate for which you can modify optical properties by applying an electrical or another optical signal, make several beams interfere and finally observe the resulting light beam using an array of photodetectors (such as a camera) that represent the output of a particular computational task.

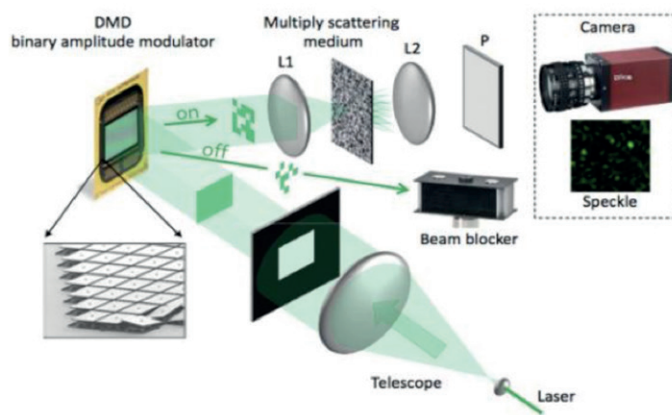


Figure 68: The Lighton optical processing unit concept
Source: Arxiv

As an example of this trend, a small company, Lighton, has designed a photonic system for speeding up matrix-vector products [226] using laser light and off-the-shelf cameras and DMDs. The optical processing unit (OPU) is designed to replace power-hungry GPUs and is to be integrated in data centres.

The company Optalysys has proposed another architecture in which they use low-power lasers passing light through spatial light modulators implemented with LCDs. The company claims to tackle important applications like genetic searches, weather forecasting or high throughput mathematical processing more generally.

Another approach developed at UCLA [241] introduces a physical mechanism to perform machine learning by demonstrating an all-optical diffractive deep neural network (D2NN) architecture that can implement various functions following the deep learning-based design of passive diffractive layers that work collectively. After a neural network has been trained using classical means, the interlayer connections are translated into patterns on layers of diffractive optical elements by 3D printing. Once the stacks of diffractive layers are aligned and set up, the system can perform simple classification tasks such as MNIST at the speed of light.

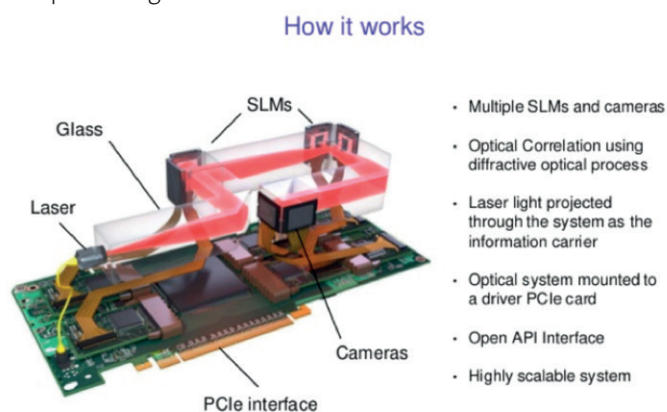


Figure 69: An optical computing setup
Source: Optalys Ltd

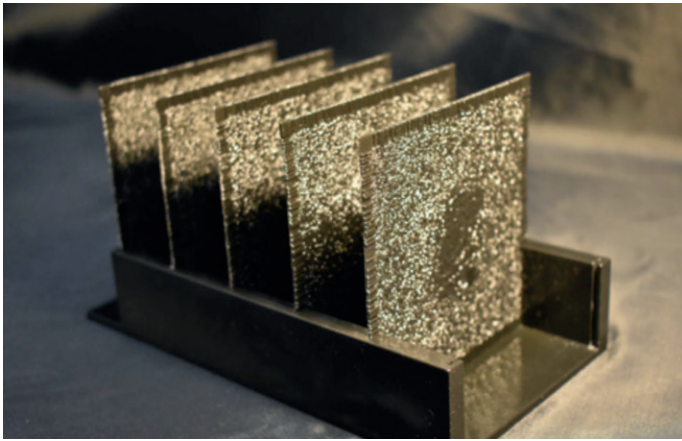


Figure 70: Diffractive deep neural network (D2NN)
Source: [220]

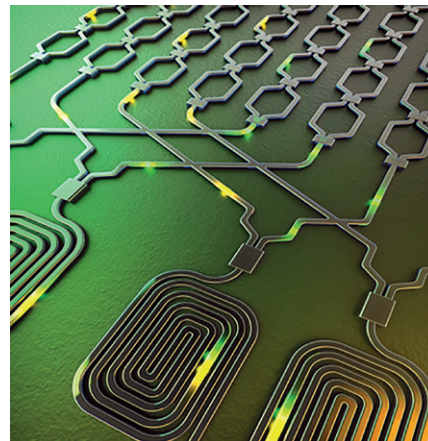


Figure 72: Spirals of waveguides generate photons that are then routed around the processor circuit to perform different tasks
Source: [381]

The concept of reservoir computing (see section 2.4.6.2.)” for more details) has also recently been demonstrated using photonics hardware for a dynamic system, which opens up the path towards ultrafast brain-inspired computing. One implementation involves electro-optic phase-delay dynamics designed with off-the-shelf optoelectronic telecom devices, thus providing the wide bandwidth required. The efficiency of the implementation has been demonstrated experimentally on speech-recognition tasks [306]. Another work investigates the possibility of designing a fully photonics analogue reservoir computer even removing the need for pre- or post-digital processing [382].

In another field, it is well known that photons are pretty good qubits. In fact, most experimental proofs of the reality of quantum mechanics have been done using light [307], even before the idea of quantum computing became popular. Even now, the most successful examples of quantum computation that have been demonstrated involve photons in order to prepare or measure qubits. See for example the recent work of [309] observing entanglements in a 20 trapped calcium ions qubits system. However, implementing photonic quantum computing practically

gets very difficult due to the engineering complexity of photon sources, channels, operators and detectors. Things might change with recent advances in integrated photonics. A team recently designed a 2 qubit circuit based on gallium arsenide (GaAs) material quantum dots for which they observed entanglement of photons on the chip [381].

Teams led by QuTech (NL) and Princeton have published two independent works demonstrating the first evidence of strong coupling between the electron spin in qudot and a microwave photon in a resonator. The authors suggest that this would open up ways to realize interacting qubits on a silicon substrate without having them in neighbouring locations [240]. Indeed, coupling lots of qubits on a silicon wafer is currently the main challenge of silicon-based spin qubits technology. By providing a way to couple qubits via photonic lines on a chip (i.e. transport quantum state on chip) this could open up new architecture and design possibilities [378].

Silicon photonics has made tremendous progresses in the last decades. Although not yet ready for generalized use in computing systems (for core computing tasks) it will be a key technology for

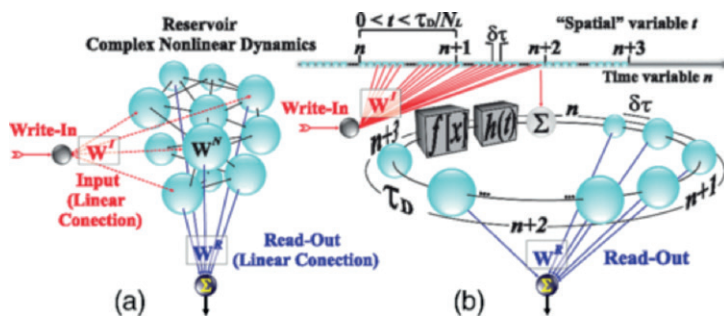


Figure 71: Principles of phase-delay optical Reservoir Computing used for speech recognition
Source: [118]

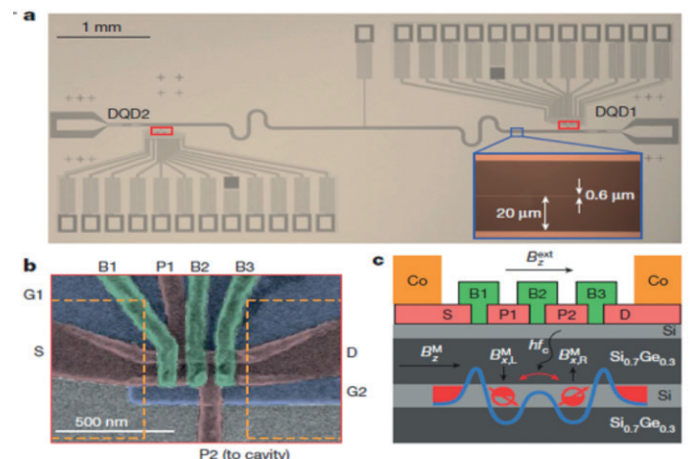


Figure 73: A spin-photon interface in Silicon
Source: [378]

the success of novel computing paradigms that are being developed. This is particularly the case for high-throughput machine learning tasks for artificial intelligence (AI) and in quantum computing engineering.

The reader interested in exploring the topic deeper might find reference [235] useful.

2.4.1.5 QUANTUM COMPUTING

While 3D stacking is already in use, cryogenic computing relies on electronics and photonics and needs to strengthen its presence in computing. Quantum computing is still in an emerging state, but it is promising.

2.4.1.5.1 The various flavours of quantum computing

There are different ways of envisioning a quantum computer (QC). We can roughly classify QC architectures in three main categories, from the most achievable to the most complicated to build.

Quantum emulators are classical computers for which elements of the architecture have been specialized in order to better execute operations needed for solving the Schrödinger equations. In general, most of the enhancements are found in the processor memory subsystem: size, connections and speed. Indeed, memory size and management are their main limiting factor. An example of this type is the Quantum Learning Machine from ATOS [265]. QC emulators can also benefit from custom computing accelerators such as GPUs or FPGAs very often used to speed up the computation of vector, matrix and tensor operations. As an example, Fujitsu presented an annealing accelerator [248] developed together with University of Toronto, a classically designed CMOS chip that can speed up such QC emulators.

Simulated QC implements a collection of qubits on a physical substrate for which we are interested in their collective (hopefully quantum-assisted) behaviour. Introduced at the turn of the 21st century as adiabatic quantum computation [228], a good example of such a concept is a quantum annealer as proposed by the Canadian company D-Wave [346]. Although this type of QC seems readily feasible with their programming style relying on



Figure 75: A D-Wave 2000 qubits machine
Source: D-Wave Systems Inc.

constraint programming paradigms, their potential acceleration factor is not very clear yet.

Universal QC also known as “unitary QC” in which each physical qubit is precisely controlled through a sequence of quantum operations. The universal QC is the “holy grail” of Quantum computing and can be considered the quantum version of a general digital instruction set architecture-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 technology used for implementing qubits is therefore critical for achieving those goals.

Superconducting qubits is one of the most popular technologies for achieving large-scale universal QC. As examples IBM announced a 50 Qubits chip in 2017 [327] and Google announced the Bristecone 72 Qubits chip in 2018 [259]. Among the technologies being considered, each with their share of pros and cons, are trapped ions technology, photonic, cold atoms, etc.



Figure 76: multiple IBM Q systems in the IBM Q computation center. – Source: IBM



Figure 74: The ATOS Quantum Learning Machine

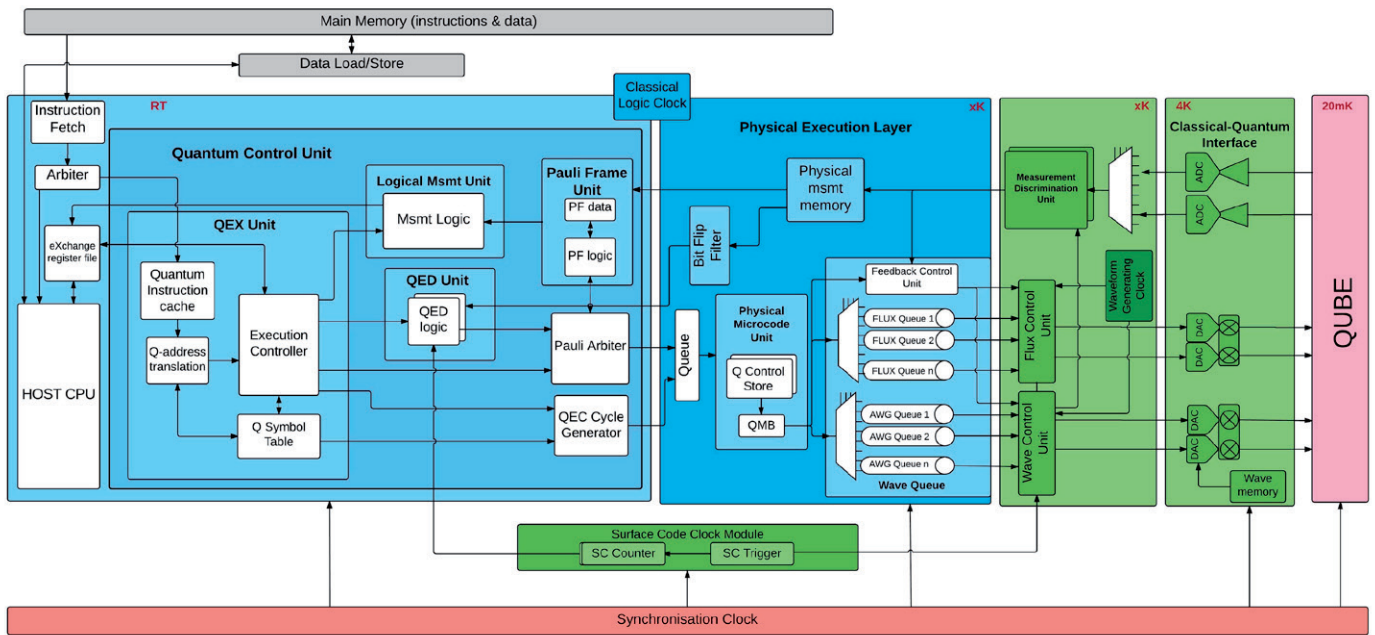


Figure 77: Hardware circuit architecture of a heterogeneous quantum computer
Source: [233]

One promising path is silicon-based qubits, also known as spin qubits technology. Circuits based on Si-28 implementing a two-qubit logic gate were demonstrated by [375] at the University of South Wales. The compatibility of such circuits with industrial grade CMOS process has been demonstrated by CEA-LETI [379]. More recently teams of researchers at Qutech [240] and Princeton [378] independently demonstrated spin-photo coupling in silicon that could pave the way for coherent interconnections of silicon qubits on a single chip.

2.4.1.5.2 Programming the Quantum computer

In the event that QC hardware becomes available, it is clear that the resulting machine will be **hybrid**. It will combine a quantum engine closely coupled with a classical digital computer. This can be pictured by the work of Qutech [233] in which the architecture of such a heterogeneous QC architecture is proposed. In the diagram of the physical architecture (see figure below), the quantum part is coloured pink, the interface circuits in green and the conventional computing (by far the largest) part in blue.

In order to operate such a complex machine, a system stack needs to be designed that links low-level quantum hardware to high-level programming. The system architecture will implement a number of layers ranging from qubits control, quantum interfaces, analogue parts, microarchitecture, quantum and classical instruction sets as pictured in Figure 78.

Programs running on such a machine will obviously 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 [232], but there is still a lot of research and development to be done.

For the time being, a programming model for a quantum computer could be summarized by the following 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 lots of things can go wrong!) the final measurement of the system state will yield a solution to the problem with a probability even greater as the fidelity of the

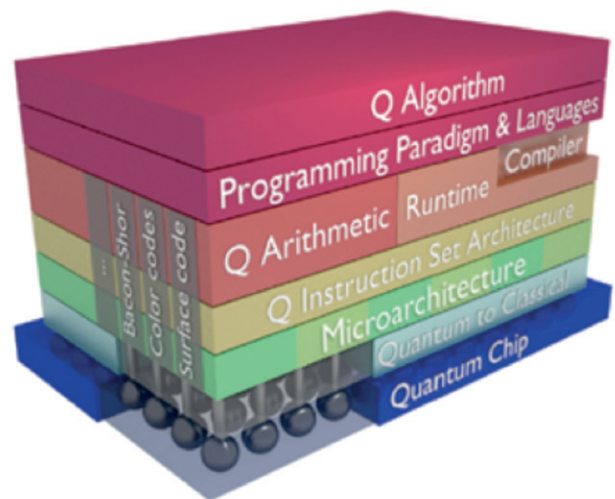


Figure 78: The Quantum Computer system stack
Source: [233]

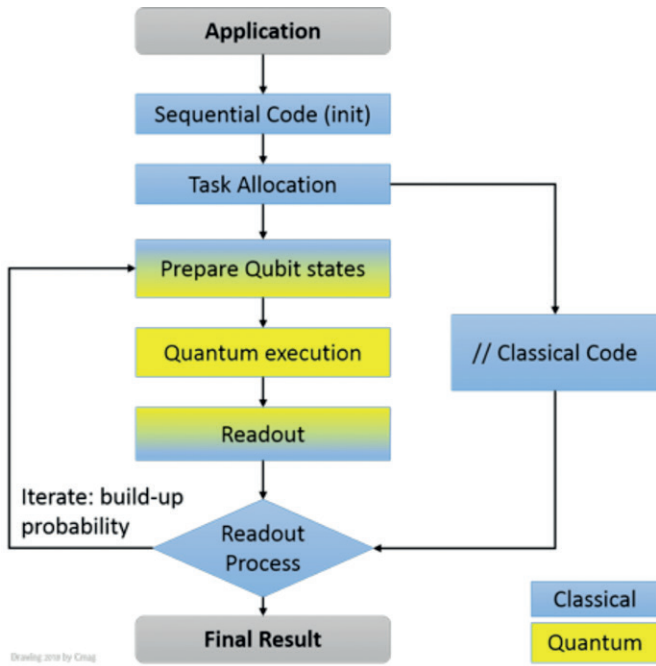


Figure 79: Execution flow of a typical quantum program: Quantum stages (yellow) are integrated into a classical execution flow (blue). Source: CEA

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 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 the figure below. 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 [71].

The way a quantum computer works is radically different from von Neumann's classical computer. As an illustration, on a classical computer the result is deterministic whereas it is probabilistic for 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 problems are solved in order to adapt them to quantum computation, we are also advancing classical algorithms [426] and pushing their limits a little further.

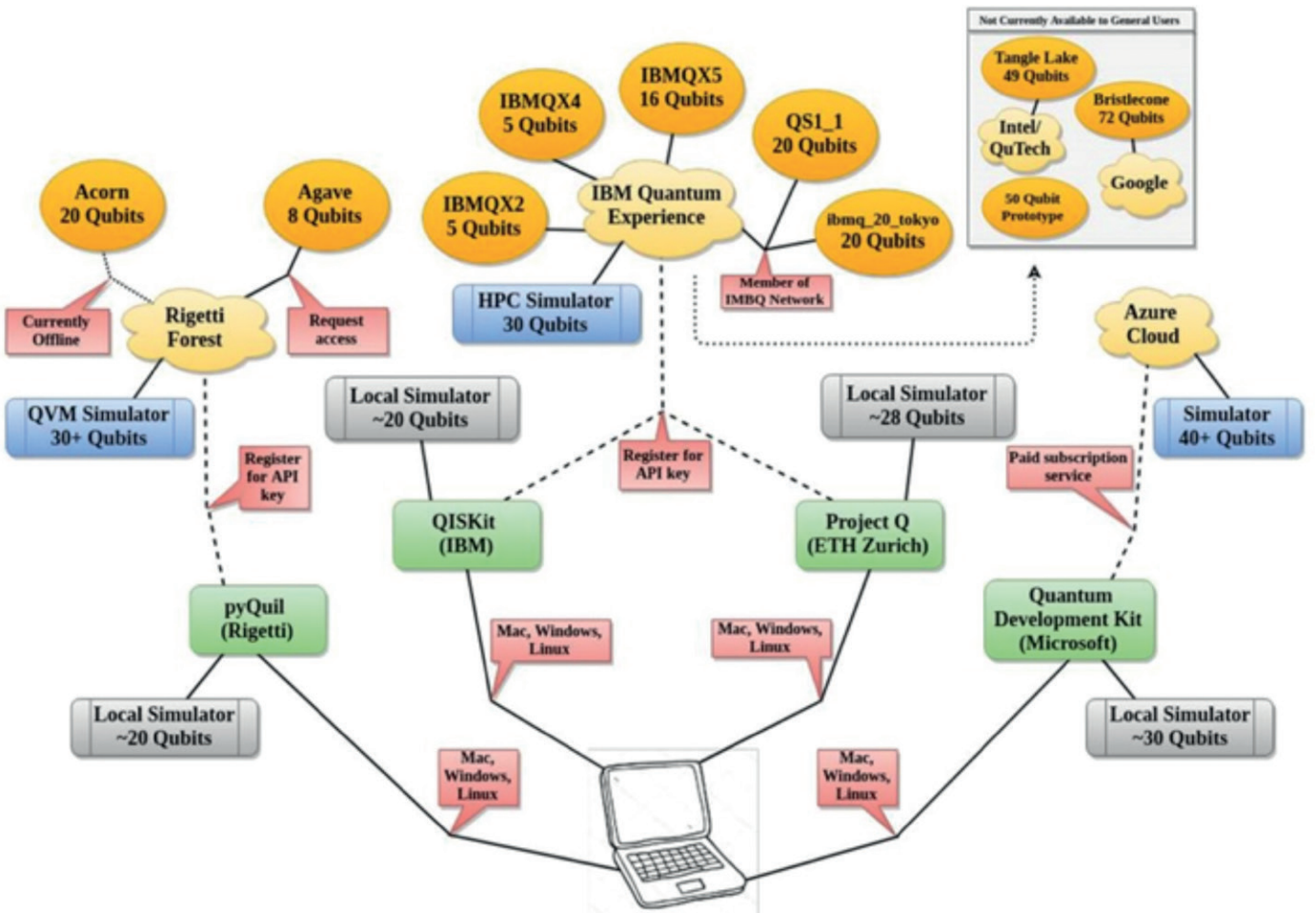









Figure 80: How to connect a personal computer to a (cloud based) Quantum computer? Source: [227]

	Visual Design	Platform	Library	High-level Language	Runtime language	ISA	Cloud
	N/A	Quadrant (ML)	Qsage, ToQ, ...	Qbsolv	QMASM	QMI	yes
	Quantum Playground	?	OpenFermion (chemistry)	Cirq	?	?	yes
	Q Experience	QisKit	QisKit	QisKit	OpenQASM	?	Q Experience
	N/A	Forest	OpenFermion (chemistry)	pyQuil	N/A	Quil	QVM
	N/A	LIQUI >	?	Q#	?	N/A	yes
	oui	?	?	?	?	?	yes
	N/A	QML	QLIB	pyAQASM	AQASM	CIRC QPU	

(cc) Olivier Ezratty, 2018

Figure 81: Quantum programming frameworks issued by the main industrial actors
Source: adapted from [388]

2.4.1.5.3 Languages and programming frameworks

Although a real QC has not yet been developed, there is no shortage of quantum programming languages, software simulators and development tools. For example, the quantiki website [397] lists over 116 quantum simulators written in more than 17 different languages. Similarly, most companies with active research in QC (IBM, Google, Microsoft, etc.) often provide software simulators in order for users to get their hands on quantum programming. However, downloading and installing such packages is not as straightforward as it is for usual digital applications. In a recent paper [227] we find a diagram showing the possible paths to connecting a personal computer to a usable gate-level quantum computer allowing anyone to experiment with quantum computing.

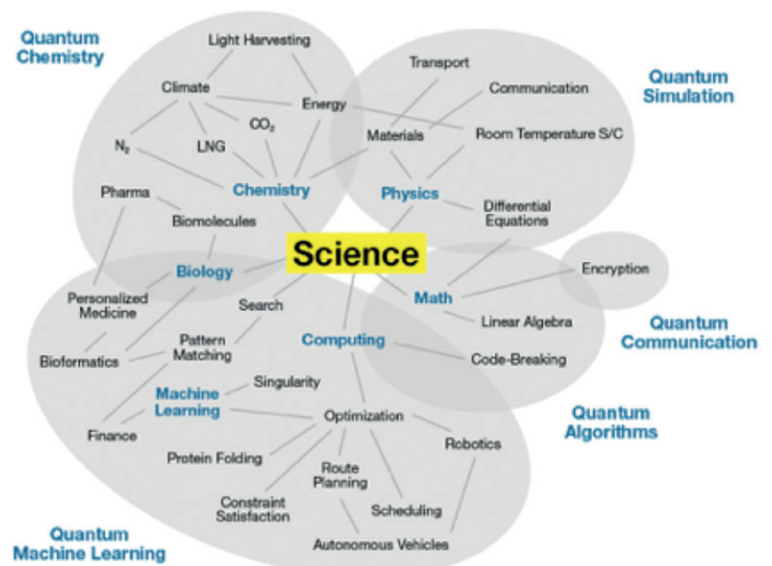
The table above gives an overview of the main quantum programming frameworks proposed by commercial companies.

2.4.1.5.4 Applications of QC

Quantum chemistry is the most cited application field. Computational chemistry problems involve determining molecular orbitals, calculating their spatial and energetic distributions, and the properties of a molecule's fundamental states. From the calculation of these quantities we can deduce and predict the stability, the reactivity and other key properties of a molecule that are critical for understanding its behaviour for diverse application fields: pharmacology, catalysis, etc. The quantum wave function (Schrödinger function) makes it possible to obtain this in-

formation, but it is a calculation that can only be performed using a classical approach by making a certain number of approximations. The bigger the problem (the more complex the molecule), the more important these approximations become.

Quantum Computing Use Cases



gartner.com/SmarterWithGartner

Source: Adapted from Peter Shadbolt and Jeremy O'Brien © 2017 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner is a registered trademark of Gartner, Inc. or its affiliates. PPL_338249

Gartner

Figure 82: Quantum Computing applications as seen by a Gartner Report

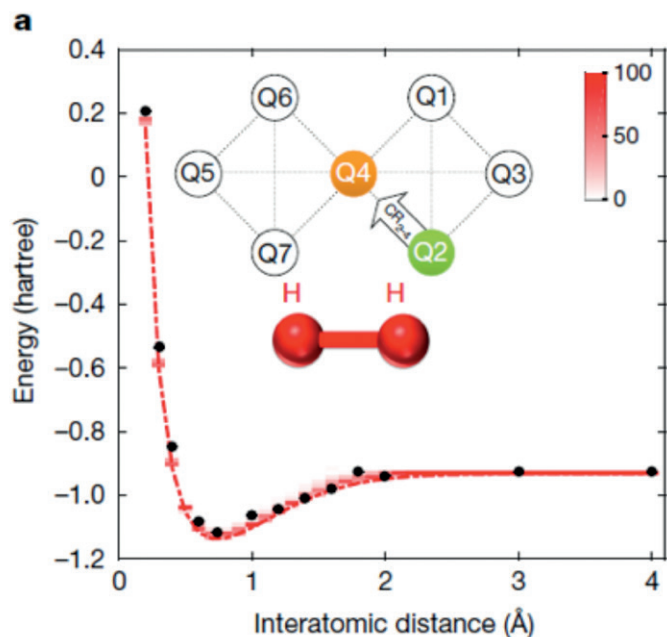


Figure 83: Computation of an elementary molecule (H_2) with a quantum computer: Energy as a function of inter-atomic distance. Black dots: experimental data.; red plot: Quantum simulation Source: [83]

One is typically in a classical modelling approach, in which the real problem cannot be solved in all its complexity; it is necessary to go through a simplified representation phase of the problem (the model) to find an approximate solution. Quantum computation, due to its computational power potential and exponential capacity to store information, makes it possible to envisage a direct simulation of these complex molecular systems. The situation is very similar for nuclear physics and generally speaking for all systems based on quantum physics.

Quantum cryptography was first proposed in 1970 [174] and developed in the 1980s [224]. It proposes a secure way for the problem of exchanging keys: quantum key distribution. A method for secure communication, called BB84, has been developed from this work and is now used in commercial products (ID Quantique, MagiQ). The announcement in 2015 by the US National Security Agency (NSA) that it will initiate a transition to quantum resistant algorithms in a not too distant future [367] has boosted a number

of research studies into what is known as post-quantum cryptography.

When Peter Shor published an article showing how it would be possible to factorize prime integers in a polynomial time with a quantum computer [229], the idea of quantum computing suddenly became popular. Indeed, the principle of decomposing a sufficiently large integer into its prime factors is the basis of the Rivest, Shamir and Adelman (RSA) encryption algorithm used to exchange confidential data over the Internet. This way, the owner of a quantum computer could potentially “break” the RSA code in a reasonable time using the algorithm discovered by Shor, threatening numerous kinds of confidential information in areas like e-commerce, among others. The quantum computer thus made a fanfare entry into the world of computing and cybersecurity. In response to this potential risk, in early 2017 the US National Institute of Standards and Technology (NIST) launched an initiative to solicit studies on cryptography algorithms that would be capable of resisting attacks by future quantum machines known as “post-quantum algorithms”, as discussed in 2.3.1.2 “The secure computer”. In fact, such algorithms do exist, showing that the quantum computer also has its limits.

Blockchain mining has become popular thanks to the emergence of crypto-currencies such as bitcoin, but the technology can have many applications in various fields requiring a trusted third party. QC might have a role to play in providing trust in computing. However, if QC can speed-up mining the block-chain it could be used to crack existing ones as well.

Machine learning and in particular deep learning have seen rapid progress in the last two years. However, to tackle bigger and more realistic cases, deep-learning algorithms face a hard-computing constraint on the learning side. High-performance computers with the help of GPU accelerators often run for days or weeks with very deep networks and extremely large learning sets. Since learning algorithms are highly dimensional optimization problems requiring careful descent in very complex phase spaces, quantum computing could theoretically solve this much more efficiently. In a review published in 2017 [377] the potential speedup of several machine learning algorithm has been evaluated.

Method	Speedup	AA	HHL	Adiabatic	QRAM
Bayesian Inference [107, 108]	$O(\sqrt{N})$	Y	Y	N	N
Online Perceptron [109]	$O(\sqrt{N})$	Y	N	N	optional
Least squares fitting [9]	$O(\log N^{(*)})$	Y	Y	N	Y
Classical BM [20]	$O(\sqrt{N})$	Y/N	optional/N	N/Y	optional
Quantum BM [22, 62]	$O(\log N^{(*)})$	optional/N	N	N/Y	N
Quantum PCA [11]	$O(\log N^{(*)})$	N	Y	N	optional
Quantum SVM [13]	$O(\log N^{(*)})$	N	Y	N	Y
Quantum reinforcement learning [30]	$O(\sqrt{N})$	Y	N	N	N

Figure 84: Speedup techniques for given quantum machine learning subroutines Source: [377]

Indeed, learning techniques often consist of minimizing an error function by mean of a gradient descent heuristic in a very large state space. This is achieved by making iterative adjustments to a considerable number of parameters. Moreover, this error minimization must be done on a large number of patterns to be learned. This results into very iterative learning algorithms whose execution times can be calculated in days or even weeks depending on the complexity of the problems, even on very powerful conventional machines. In this type of problem, the quantum approach could allow a more efficient exploration of state space by exploiting quantum superposition and quantum parallelism.

Combinatorial optimization: Machine learning is a particular optimization problem trying to minimize an error function on the desired output. However, there are a number of very useful applications which are sometimes used on a daily basis and which can be expressed as an optimization problem on a set of constraints: traffic schedules, traffic management, timetables, or scheduling and planning problems more generally. As with other applications, the quantum approach may prove useful when the size of the problem to be addressed becomes very large.

The **quantum algorithm zoo** [237], is a website that maintains a list of all known quantum algorithms so far. At the time of writing (mid-2018), the quantum zoo listed 60 classes of algorithms, with most of them being refinements of the Shor or Grover algorithms.

Commercial applications of QC are still further up the road, however some companies like D-Wave and IBM [421] have attracted paid customers for their quantum technologies. These “customers” are major companies that don’t want to lose ground should QC become mainstream. Although it is true that the few million dollars needed to acquire a D-Wave machine is a serious cost, the odds of witnessing the emergence of a competitor mastering quantum technology might well justify the investment for some businesses.

2.4.1.5.5 The moving horizon of quantum supremacy

Quantum supremacy is “the potential ability of quantum devices to solve problems that classical computers practically cannot” [293] and was initially introduced by John Preskill [225].

If quantum supremacy ever comes to light, it is unlikely to be a ground-shaking event. It might come about in very specific, mostly unnoticeable, areas of computing and will be very progressive. The fact is that QC will first need to overcome enormous engineering challenges both at the hardware and software levels; at the same time, non-quantum information processing is also progressing.

Quantum supremacy is a moving target. Not only does it move according to progresses in non-quantum information processing but also because of our better understanding of hard problems and when and how to tackle them. QC is a very brute-force

approach to what we define as computationally hard problems. Some results suggest that other ways exist to tackle hard problems where brute-force numerical approaches such as QC fail [376].

Paradoxically we might face a situation in which we solve most of QC challenges, build the QC, program the QC and finally end up in a situation where the horizon of “quantum supremacy” has been pushed away by advanced knowledge and progress in problem solving.

QUANTUM SIDE EFFECTS

Research in quantum may have some good side effects: the low-temperature electronics for controlling and interfacing qubits could lead to cryogenic computing, and research in quantum algorithms may lead to improvements in 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, 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.” From [395]

This example shows that regardless of the actual availability of QC hardware, rethinking algorithms and programming in the light of quantum information brings a new vision on the actual limits of classical computing. Quantum supremacy might well be a moving target.

2.4.2 EMERGING TECHNOLOGIES: BEYOND SILICON

In this section, we will discuss the emerging technologies that are not based on the element silicon (Si) on which today’s electronics and computing mainly rely. However, this should not imply that these are alternative technologies to Si-based technologies and computing. They are instead technologies that will address the particular requirements of a specific domain.

For example, flexible electronics will make almost every object smart by embedding a flexible, conformable, disposable and low-cost IC into an object or the package of an object. Similarly, although it may seem far-fetched, synthetic biology will enable the programming of living beings such as cells to do tasks that they normally would not do. Cells will eventually turn into

programmable compute engines that are capable of communicating with us. Programmable cells will be domain-specific devices that will have significant impact on healthcare, agriculture, bioenergy and environment in the next decade or so.

2.4.2.1 FLEXIBLE ELECTRONICS

Flexible electronics is a term that covers a wide range of electronics manufactured on a low-cost flexible substrate such as thin glass, plastic, metal foil or paper. The process to create electronics on a large flexible substrate can take the form of roll-to-roll printing, inkjet printing of solution based inks for organic electronics, low temperature and cost photolithography to thin-film ICs [112]. It has found applications in sensors, radio frequency identification (RFID) tags, solar cells, batteries and displays in the fields of medical, automotive, human-machine interfaces, mobile computing platforms and embedded systems.

The Organic and Printed Electronics Association (OE-A) roadmap [252], as shown in figure 85, identifies five key application areas:

- 1 OLED lighting
- 2 Organic photovoltaics (OPV)
- 3 Flexible and OLED displays
- 4 Electronics and components
- 5 Integrated smart systems (ISS)

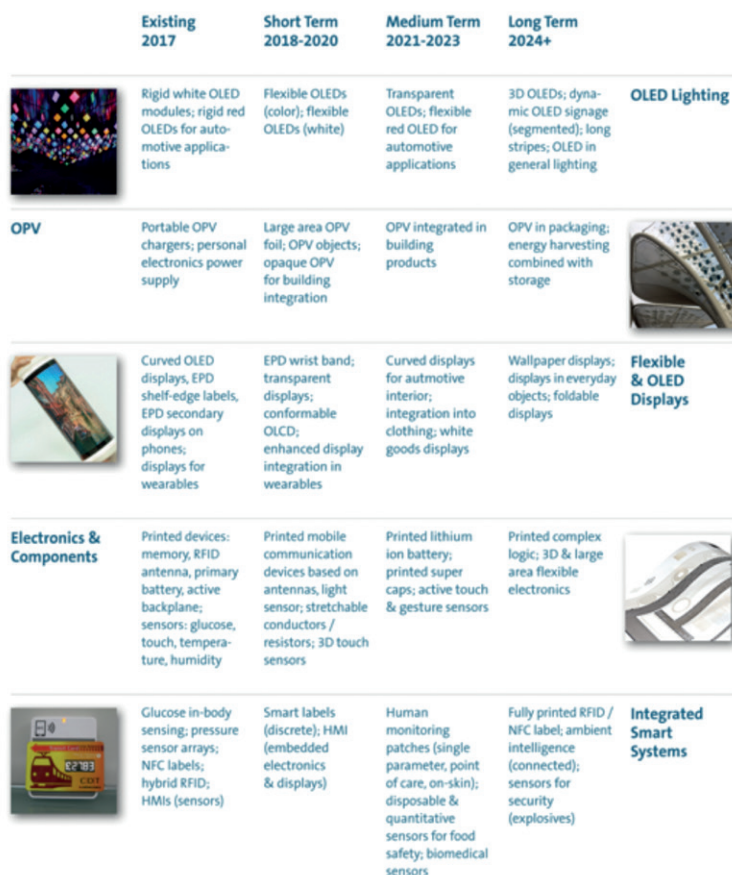


Figure 85: The 5 key application areas and their evolution from short term to medium term and then to long term
Source: OE-A Roadmap for Organic and Printed Electronic Applications 2017

The roadmap predicts the evolution of these application areas from short term to medium term and then to long term.

Figure 86 shows the potential applications of printed electronics. There are six key driver markets for flexible electronics applications: automotive, consumer electronics, healthcare, IoT, smart packaging and smart buildings. Although not explicitly mentioned, fast-moving consumer goods (FMCG) covers the IoT and packaging in this spectrum, in particular with smart labels, flexible sensors and displays. The main pull for flexible integrated smart systems (ISS) comes from consumer electronics, healthcare, IoT and packaging.

2.4.2.1.1 Impact on computing

The Organic and Printed Electronics Association (OE-A) roadmap predicts that flexible integrated smart systems (ISS) consisting of memory, logic and sensors will appear in the medium term (2019-22). More complex flexible ISS such as smart standalone body monitoring systems and wireless sensor systems are predicted in the long term beyond 2023. Flexible ISS will put ambient intelligence into billions of disposable things creating a flexible intelligent IoT market.



Figure 1. Organic and printed electronics solutions in important industry sectors

© OE-A 2017

Figure 86: Potential applications of flexible electronics
Source: OE-A Roadmap for Organic and Printed Electronic Applications 2017

MARKET FORECAST FOR PRINTED, FLEXIBLE AND ORGANIC ELECTRONICS

IDTechEx [250] forecasts that the total market for printed, flexible and organic electronics will grow from £21 billion in 2016 to £55 billion in 2026 with a compound annual growth rate (CAGR) of 10%. Although the majority of the market share is OLED displays, it is expected that logic, memory and thin film sensors will have huge growth potential by 2026. Currently, logic/memory take up about 5% of the market share.

Another IDTechEx report [249] estimates that the printed sensor market will be worth £6.4bn in 2025 with a CAGR of about 42%. Biosensors and pressure sensors take up the largest share in the market but some sensors like photodetectors, temperature and gas sensors are coming out of research and development (R&D) to production and will grow fast over the next 10 years.

Thin-film transistors (TFT) can be used to build ICs on flexible surfaces to enable new wearable IoT devices. Today, thin-film RFID and near-field communication (NFC) tags and toy microprocessors have been demonstrated in the literature. TFT technologies are potentially low cost due to simpler lithography process. At present, the mainstream TFT technologies available in consumer electronics products are amorphous silicon (a-Si), low temperature polycrystalline silicon (LTPS) and amorphous metal-oxide semiconductors (mainly indium–gallium–zinc-oxide, IGZO). Metal-oxide TFT is a promising n-type-only technology for flexible IC circuits, as it can be manufactured at process temperatures within the thermal budget of flexible substrates.

Flexible ICs are not in competition with Si-based ICs but rather target a market space where low-cost, flexibility, conformability, biocompatibility and disposability are desirable properties in applications.

There have been a few attempts to build a simple microprocessor in plastic. For example, the first thin-film flexible microprocessor was published in 2005 [106]. It comprises about 32,000 transistors based on flexible complementary low temperature polysilicon transistors. The first organic microprocessor was fabricated directly on flexible substrates exhibited a clock frequency of 40 Hz to execute 8-bit operations [143]. Most recently, ARM announced a research prototype, building a 32-bit Cortex-M0 based microcontroller in plastic called “PlasticARM” in 2µm metal oxide semiconductor [453].

As computing becomes pervasively intelligent, many applications in market segments in which flexible ICs can be deployed such as packaging, healthcare, FMCG etc. need low-cost flexible and disposable integrated smart systems. The type of computing

required in such a system can be a custom processing engine that is designed to perform a fixed function such as detection and recognition of a modality (such as audio, image or odour). Essentially, this is a hardware customization or specialization, and flexible electronics is better suited to allow low-cost customization than expensive Si fabrication technologies. For example, a project led by ARM called “PlasticArmpit” attempts to achieve this goal with Plastic Neural Networks [435].

2.4.2.1.2 Impact on low-cost fabs

Today, flexible electronics manufacturing uses moderately expensive and large equipment, and we predict that flexible electronics technology, particularly for integrated circuits (ICs), will track a similar trend as optical disc manufacturing which has evolved from manual, batch-based cleanroom production to fully-automated production in a self-contained module. These modules cost three orders of magnitude less than a Si fab. Self-contained fab modules will be affordable enough to be owned by smaller businesses, research institutes and even university consortia. These fab modules could be located in the EU (as opposed to Si fabs in Asia) due to the fully automated nature that reduces the operating costs. The projected production times of future flexible ICs will be under an hour compared to seven days today, as opposed to the 8-12 weeks required for Si (excluding design effort).

Over the next decade, end users will personalize the sensing and intelligence of wearable/IoT devices by selecting sensors, their interfaces, and customizing flexible systems to applications. This will have unprecedented effects on the industry, research communities enabling rapid prototyping of custom flexible products, faster time-to-market for SMEs, low-cost research prototyping/testing of novel ideas. It will also have an impact on education where students in universities and in secondary education can demonstrate their creative skills by building custom flexible ISS prototypes (e.g. wearables) at a cost that can be affordable by universities and schools.

2.4.2.2 SYNTHETIC BIOLOGY

Synthetic biology is a cross-disciplinary field that applies engineering principles of specification, design, modelling, testing and validation so that new biological systems can be produced. It is relatively a new field that started at the beginning of 21st century, thanks to advances in DNA sequencing (i.e. reading), synthesis (i.e. editing) and computer technology (i.e. compute power, storage, computational modelling etc.).

Synthetic biology aims to build novel and artificial biological parts, devices and systems. A biological part or biopart is a module designed to build larger components such as biological devices and systems. Such bioparts must be characterized so that input and output functional characteristics are documented and can be stored in an inventory or registry (e.g. similar to the standard cell libraries in chip design). The goal is that well-characterized bioparts can be re-used in many applications.

PROGRAMMABLE MATTER

Programmable matter [204] refers to an intelligent material that contains all the elements that a compute unit has such as sensors, actuators, memory, processor, and communication. Programmable matter is also known as claytronics [303]. The ultimate goal in programming matter is to change the physical properties of the material so that its shape and functionality can be programmed through software.

Metamaterials [217], which are engineered materials that do not exist in nature, have the potential to achieve the vision of programmable materials. Depending on the material, the envisaged granularity of programming varies from atom to molecules all the way to visibly tiny devices. It has been predicted that programmable matter will lead to disruptive shape-shifting objects around us and its impact on our lives will be equivalent to the impact of the Internet today. The following excerpt from [303] is a good example of how it can impact our lives:

“... Matter can be transformed into any shape for any purpose. Furniture could change shape; blank walls could grow doors or windows. Catoms (i.e., tiny micro robots) could form into people that we would find difficult to discern from the real person. They would appear as an actual physical being, not a hologram.

...

For example, should we be at risk; programmable clothing would become stronger than steel, while still maintaining its light weight. Sensing danger, these ‘smart clothes’ could form an impenetrable shield to stop bullets and knives from piercing our skin; or become cushion-like to protect us from auto accidents.

On command, walls in our homes could light up with a radiant glow; TVs would look less like moving pictures and more like 3-D windows; and as wild as this may sound, we could actually move doors and windows to different walls. There is almost no end to the magic that this technology could create.

Claytronics would reduce the number of furniture pieces required in a home. A dinner table might be changed to a poker table for a party, then into a bed at night. In addition, a single room could be used as living-room, dining area and bedroom, simply by morphing furniture at different times.”

A biosystem designer can use the bioparts in the inventory to build biodevices and biosystems. Tolerances are built into the design of any engineering biopart, device or system to compensate for imperfections in the manufacturing. This is very similar to the design margins used in circuit design within semi-conductor IPs.

Figure 87 shows the synthetic biology design library hierarchy from DNA to systems. The bioparts consist of promoters, activators, repressors and terminators; devices are made of bioparts and encode man-made functions such as logic gates, protein generator; systems are built with devices to perform tasks such as counters, switches, oscillators etc.

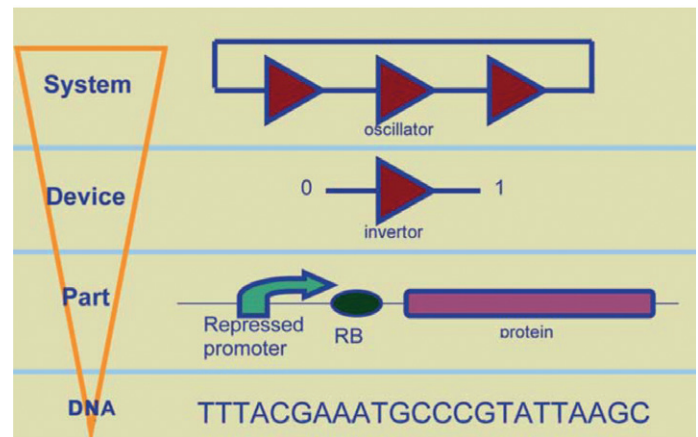


Figure 87: Synthetic biology design library hierarchy

Source: [201]

Synthetic biology has many commercial applications with an impact on diverse sectors. Figure 88 shows the potential commercial applications in associated sectors. In particular, the healthcare and environment segments offer great opportunities for synthetic biology.

Health	Energy	Environment	Agriculture	Other Industry
Cell counter	Bio power units	Emissions sensors	Starch synthesis	Biological computers
Biological sensors	Biofuels	Spill/chemical/radiation detection	New seed products	Digital/bio converters
Disease diagnosis	Enzymes	Artificial leaf	Bioenergy feedstock	Logic gates
Disease fighting		Biodegradable packaging	Agro-fuels	Switches/oscillators
Controlling signs of ageing		Stronger/lighter materials	Optimised food production	Cleansing biofilms
Custom drugs				Responsive materials, eg oil
Tissue engineering				Nano particle production
				Bioremediation
				Biofabrication

Figure 88: Synthetic Biology commercial applications

Source: [201]

The key enabling technologies in synthetic biology are synthetic gene circuits and CRISPR gene editing. Synthetic genetic circuits are engineered gene circuits that perform user-defined functions in a predictable and reliable manner.

A typical gene circuit has three elements: a) sensor that accepts inputs, b) processor that computes the response and c) actuator that produces the corresponding output. The inputs to gene

circuits are regulatory molecules to which cells respond by activating or repressing gene expression. The outputs of synthetic gene circuits are mainly fluorescent reporter proteins that can be detected using optical measurement techniques. Figure 89 shows the logic gates that can be built with different regulatory molecules.

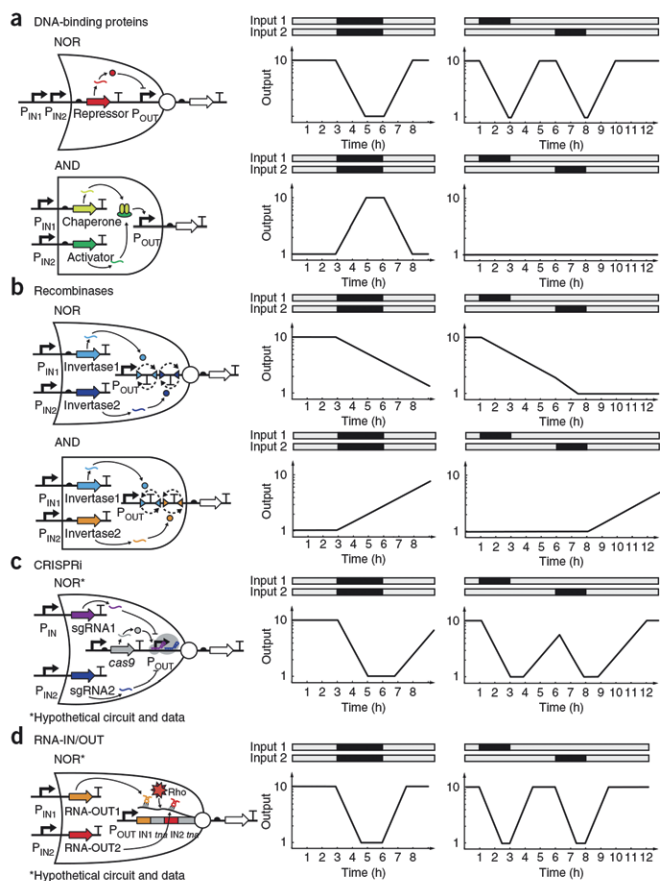


Figure 89: All 2-input logic gates can be built through gene regulation
Source: [89]

A typical use-case scenario is to build a synthetic gene circuit and embed the circuit into a biological organism to do useful work. For example, synthetic gene circuits are built by manipulating DNA, RNA and proteins and injected into mammalian cells for gene therapy and drug delivery as shown in figure 90.

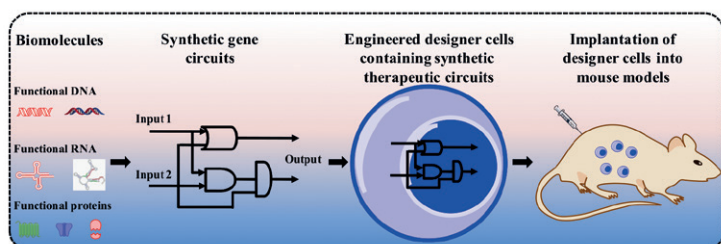


Figure 90: Synthetic gene circuits used in gene therapy to combat diseases
Source: [76]

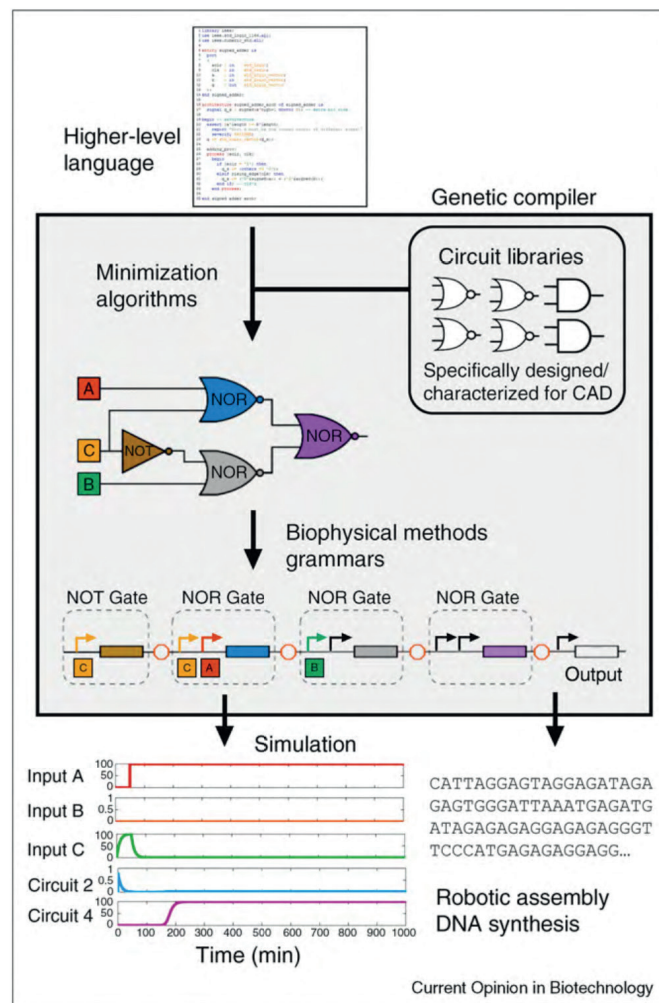


Figure 91: Automated genetic compiler [104]

CRISPR technology [116] is a gene-editing technology that has been used to fix the bases in the DNA more precisely than other gene-editing technologies.

Many design automation tools have been developed for synthetic biology to automate the design cycle from a specification to the biolayout. BioCAD tools enable to design synthetic genetic circuits from a high-level language specification using an automated genetic compiler. Essentially, the input specification is made by a high-level language similar to RTL in electronics and the output is the DNA sequence. The automated genetic compiler as shown in Figure 91 abstracts away the details of molecular biology and biochemistry. The user can design genetic circuits using standard library parts, simulate it and finally send the designed DNA to a bio-foundry for DNA synthesis and assembly. There are several emerging CAD tools to automate the genetic circuit design such as GenoCAD, TinkerCell, ClothoCAD, GenomeCompiler, CelloCAD and many others.

There is a trend in synthetic biology to separate the design from the fabrication similar to the silicon world. For example, the biofab concept is offered by companies such as *Twist Bioscience*, *Integrated DNA Technologies* and *Ginkgo Bioworks* who advertise themselves as “a foundry for designing living cells” [360].

SynBio is a young field with a lot of applications, some of which will take many years to become reality such as tissue engineering or biocomputing. The following are the key challenges for synthetic biology [67]:

- Safe delivery of synthetic gene circuits into mammalian organisms for therapy.
- Engineer cells acting as programmable devices and functioning safely in human body.
- High throughput and non-invasive measurement technologies beyond fluorescent reporter proteins.
- Build complex and multiple synthetic gene circuits.
- Well-characterized, reliable and robust complex synthetic gene circuits that can be used across multiple applications.
- Synthetic gene circuits that function in a truly multi-cellular fashion.

These challenges also offer opportunities for the computing community to contribute to synthetic biology such as in building automation tools, complex custom synthetic gene circuits using digital design principles and system design, and eventually exploring how to develop biocomputing devices.

2.4.2.3 OTHER NEW MATERIALS

2.4.2.3.1 Carbon-based

For some time, new materials have been investigated as replacements for silicon. These materials have the characteristics allowing them to be made into very thin films, fixing the thickness of a layer to just a few atoms. It is these thin layers of just a few atoms thick that exhibit the desired characteristics, comparable to those found in stacks of silicon oxide and polysilicon to construct transistor used as switches.

One of these promising materials is graphene. Known to exist for a long time, it was not until 2005 that a method was found to produce it. Graphene is such a promising material because it has a high electron mobility that would make the construction of fast devices possible. However, graphene lacks the bandgap in its electron energy level which is characteristic of semiconductors and which makes the construction of electronic switches possible. Graphene can be constructed in narrow strips which induces bandgap-like behaviour, but such graphene layouts have a lower electron mobility, lowering the potential for high speed devices.

2.4.2.3.2 Other new devices

A number of directions are being exploited to develop new switching devices. Current developments can roughly be divided into two groups: transistor-like devices and majority gate devices.

Transistor-like devices switch a charge current through an electric force. The traditional CMOS transistor, which switches a current by modulating the conductance of a channel through electric charge, falls in this category. New devices based on charge current switching, the tunnel field-effect transistor or TFET, switch a charge current by modulating electron tunnelling through an energy barrier. In the ferroelectric device, the regulating charge is stored in a ferroelectric material instead of in a capacitor. Some ten such TFET devices of various construction are under investigation. Some of these TFETS use carbon nanotubes as a channel, others use graphene sheets embedded in monolayers of insulating material.

Spintronic devices drive the magnetization of a material through either an electronic current or an electric force. Note that magnetization is an intrinsic particle property: the magnitude is fixed, and cannot be amplified (comparable to the charge of an electron). Information is not stored by the absence or presence of magnetization but by its orientation. Switching in spintronic devices is accomplished by combining three or more inputs in a majority gate circuit. Note that spintronic devices require an interface to go from spintronic to electric and vice versa.

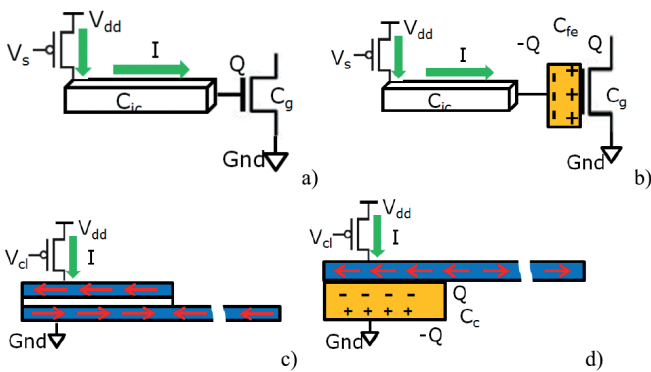
Most devices are still at an experimental stage, in which it is infeasible to construct larger circuits from gates. To make fair comparisons, however, multi-device circuits have been modelled using the characteristics of the new devices. Such circuits then

allow comparison of the devices in terms of power and switching characteristics. In [51] a set of new devices is compared with 28 and 15 nm CMOS devices in an arithmetic logic unit (ALU) circuit.

This comparison shows that TFETS, in particular van der Waals TFETS, have switching speed capabilities comparable to or even slightly faster than CMOS FETs. Ferroelectric devices come close, but devices based on magnetization are two orders of magnitude slower. The picture is different in the case of power, both active and standby: devices based on magnetisation have markedly lower power dissipation. These devices are thus better suited for

mobile and IoT applications. Also note that because in principle a device based on magnetization retains its state, and thus its information, when powered down, it is possible to construct instantly on circuits.

There is no clear winner among the new devices now under investigation as CMOS alternatives, and they are often not drop-in replacements. As an example, magnetics-based devices lend themselves especially well to in-memory computational architectures. This is a trend visible in many of the new device technologies, and in quantum computing, for example: the new



Scheme of driving switching of (a) electric device, (b) ferroelectric device, (c) ferromagnetic device, and (d) magnoelectric device.

Figure 92: Scheme of driving switching of (a) electric device, (b) ferroelectric device, (c) ferromagnetic device and (d) magnoelectric device

Source: [21]

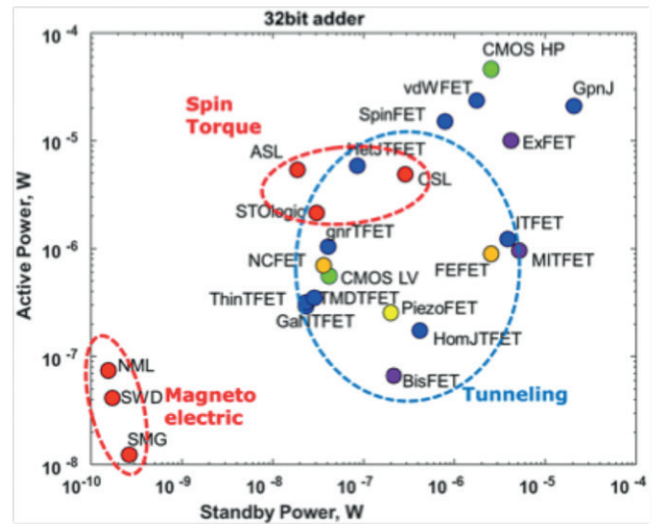


Figure 94: Active power versus standby power of a 32-bit adder Source: [51]

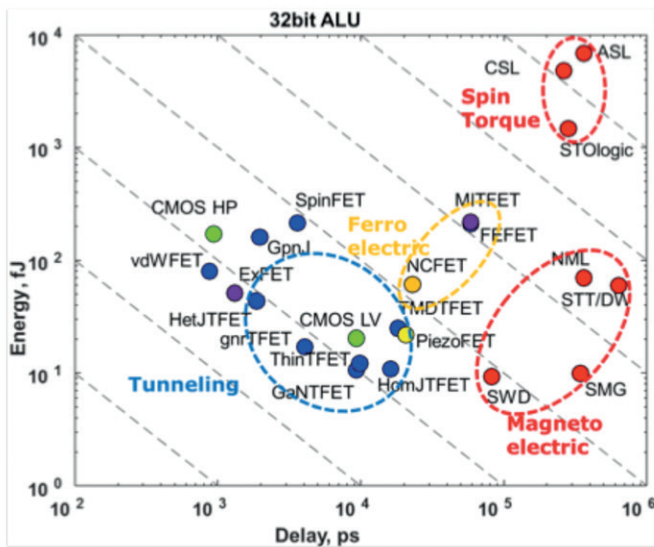


Figure 93: Switching energy versus delay of a 32-bit ALU Source: [51]

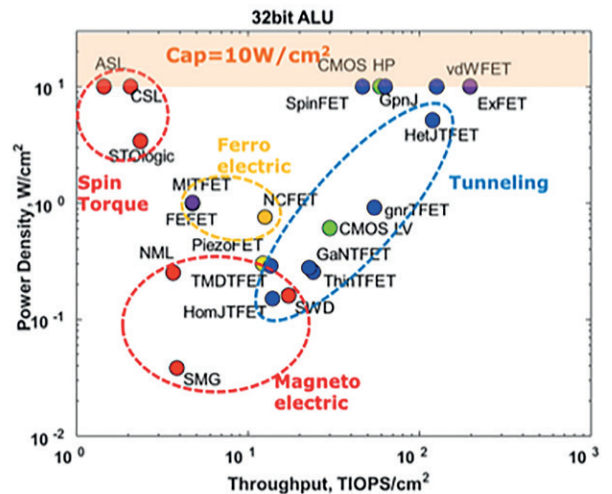
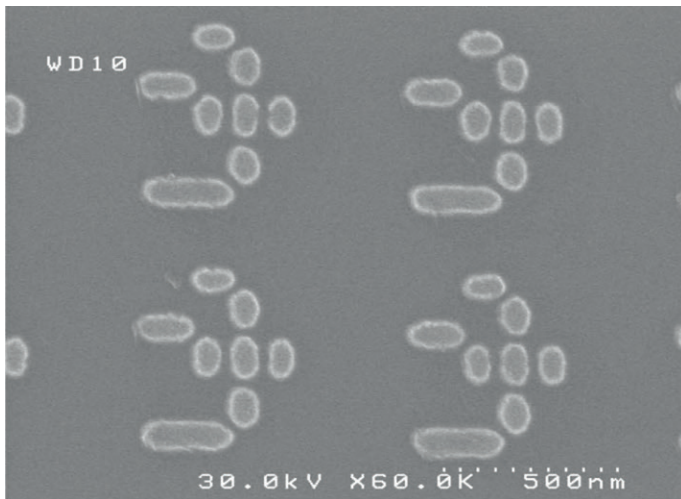


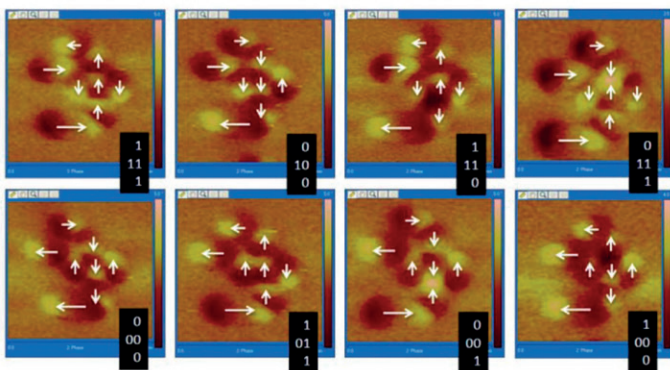
Figure 95: Dissipated power versus computational throughput (capped at 10W/cm²) related to a 32-bit ALU Source: [51]

switching devices will lead to different architectures and with that to new models of computation.

With no clear winner, it is very likely that future applications will contain a mix of interfaced architectures. New computational paradigms/models and the particular characteristics of the emerging devices have to be considered holistically. Thus, the HiPEAC community should therefore work in close cooperation with the device community to design efficient models of computation.



(a)



(b)

Figure 96: Fig. 4. Scanning electron micrographs of Nanomagnet logic quantum cellular automata (NML) (a) SEM photo of NAND2 (b) magnetic force micrograph (MFM) of NAND2. Source: [51]

2.4.3 ARCHITECTURE: HETEROGENEITY, ACCELERATORS AND IN-MEMORY COMPUTING

2.4.3.1 MORE SPECIALIZATION THROUGH ACCELERATORS

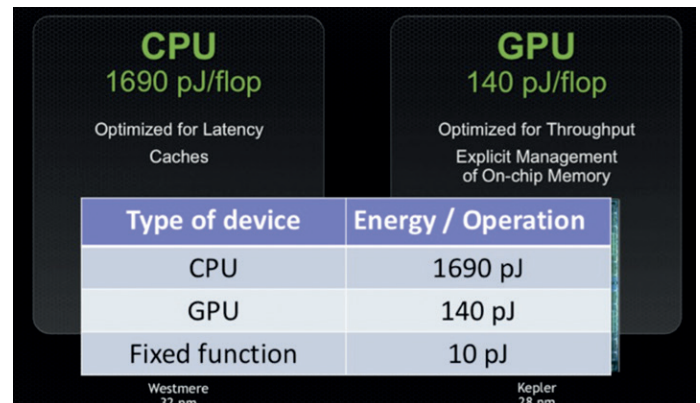


Figure 97: Optimized CPU and GPU Source: Bill Dally (Nvidia)

While the architecture of processors makes them very versatile, they are not necessarily optimized for all functions. Accelerators, on the other hand, are tuned for a set of operations, meaning that they are more energy efficient or require fewer transistors to carry out a similar functionality.

Indeed, as we saw in 2.2.2.2 “Verticalization and dominance of global platforms (GAFAM + BATX)”, there is a general trend towards vertical companies entering the field of chip and accelerator design in order to have more efficient systems, tuned to their ecosystems or needs. Mastering the hardware also allows better control of costs, as there is no third party involved, and of the availability. Little wonder that the major technology companies – GAFAM and BATX – are moving towards making their own chips like fabless companies: [342].

New architectures are also emerging which exploit new potentialities of technology for innovative applications and algorithms. For example, 3D stacking, which we discussed in another, allows memory and computing to be stacked on top of one another, increasing the bandwidth between the two and reducing energy.

Stacking sensors and processor arrays also unlock new potential: for example, CEA’s intelligent retina [117] stacked an array of image sensors with processors, allowing parts of an “image” to be processed independently, at their own frequency, freeing the algorithms from the notion of images where all pixels are sampled at the same time.

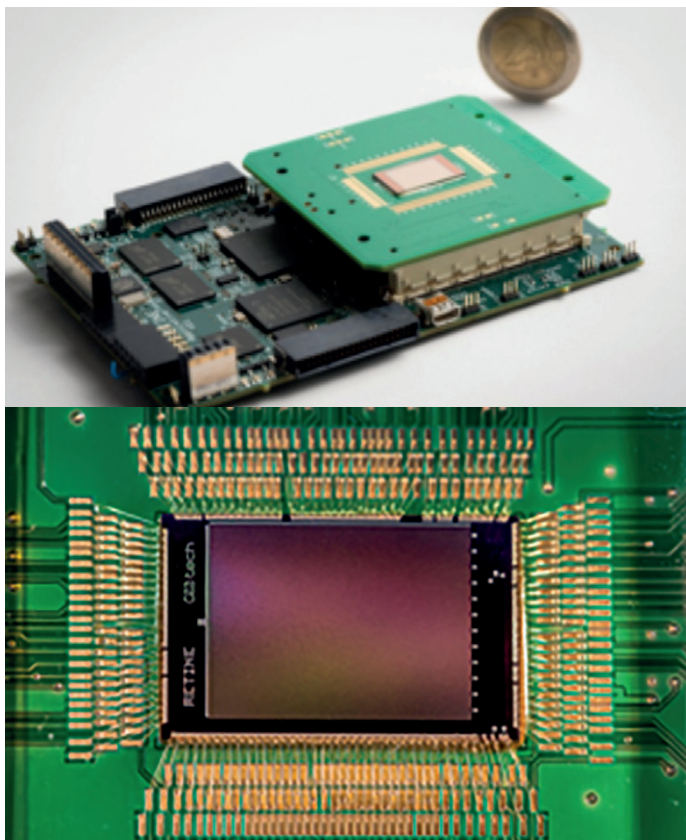


Figure 98: CEA's intelligent retina 3D stacked chip and development board
Source: CEA

2.4.3.1.1 Heterogeneity of computing elements

One clear trend here is that we are moving towards combining general-purpose CPUs with application-specific accelerators. One application domain that is currently dominating these CPU-accelerator combinations is machine learning. As mentioned in 2.2.1.1 “The AI bandwagon”, Google are producing their own TPU chips that are specialized towards accelerating machine learning applications. Furthermore, Intel is adding support for accelerating machine learning in their upcoming Cascade Lake Xeons [202].

These accelerators may either be external accelerators, as in the case of TPUs, or fixed functionality on the CPU, but other

possibilities exist. For example, Intel is working on a Configurable Spatial Accelerator, which is a kind of dataflow engine or coarse-grained reconfigurable-style architecture [86].

We have already seen in section 2.2.3.3 “Gaming: testbed for consumer advanced technologies” that GPUs are very efficient because of their high parallelism: as in single instruction, multiple data (SIMD) architectures, a large number of operations on data are executed per program instruction. In addition to GPUs, a number of functions, such as cryptographic functions, can be accelerated by optimized hardware.

2.4.3.2 NEAR/IN MEMORY COMPUTING

While enormous progress has been made on optimizing computational units over the last 70 years, the same cannot be said for data storage and movement [458]. This has led to unbalanced, inefficient computing systems, with as much as 95-99% of the “real estate” being dedicated to units that simply store and move data.

As a result, systems have become excessively complex in response to the need to get data to the processors quickly, using workarounds such as out-of-order and speculative execution engines, many levels of cache hierarchy, complex pre-fetching mechanisms and large amounts of multithreading. These workarounds both complicate the design process and have an adverse impact on predictability, reliability and energy efficiency.

A single memory access costs 2-3 orders of magnitude more energy than a complex arithmetic operation, reduces performance and increases security vulnerabilities by exposing data to the outside world for longer. Hence, methods to reduce data movement could help create systems that are more energy efficient, higher performance, more reliable and secure.

Facilitated by the decreasing cost of RAM memory and increasingly common 64-bit operating systems which allow a much larger memory set to be addressed, *in-memory computing*, or *in-memory processing* [181], is one approach to reducing data

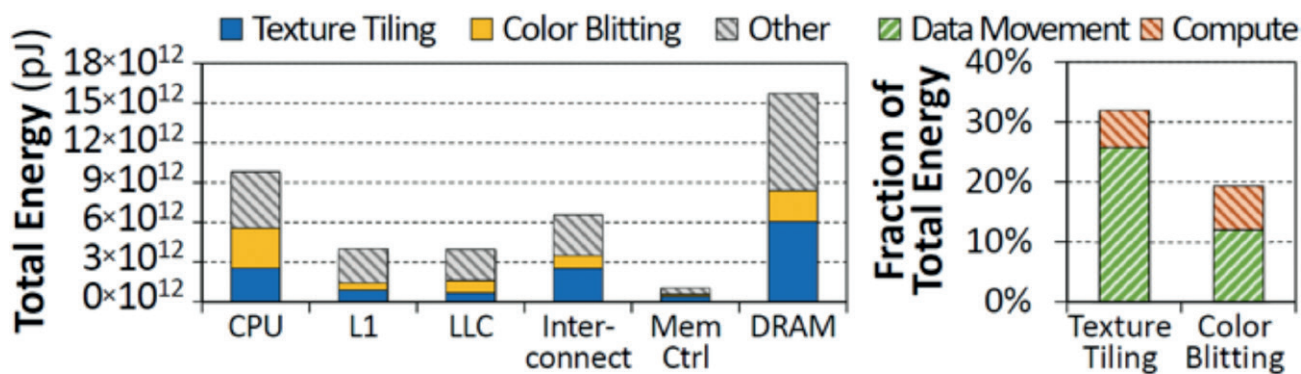


Figure 99: Browser energy breakdown
Source: SAFARI Research Group, ETH Zurich and Carnegie Mellon University

movement and accelerating computation. It consists of processing data in the main memory (RAM) rather than only in the CPU, thus benefiting from much lower access latencies and higher transfer speeds.

In-memory is therefore especially useful where large amounts of data are to be processed, such as in bioinformatics applications, business intelligence applications and graph processing (the technology underlying web searches such as PageRank, social media and information networks). Using in-memory processing to help provide artificial intelligence at the edge may also power many applications, including for self-driving cars [180], databases [213], or genome analysis [98].

Research on Google workloads on mobile devices has shown that more than 62% of the total system energy, on average, is spent on data movement between main memory and the compute units. By processing data close to memory using either very simple cores or specialized accelerators in the logic layer of 3D-stacked memory, it is possible to halve the energy used while doubling system performance in a state-of-the-art mobile device [11].

Similarly, designing an in-memory graph-processing accelerator led to 13.8x better performance and 8x less energy on graph processing [103]. Even light-touch modifications to existing systems allowing users to take advantage of processing in memory have been shown to deliver up to 50% better performance and reduce energy consumption by 25% [102].

Consequently, developing in-memory or near-memory computing through specific applications, frameworks and strategies is an efficient and cost-effective way of improving information and communication systems performance, since most of the infrastructure investment has already been made.

2.4.3.3 HW/SW CODESIGN

HW/SW codesign has been around for quite some time now (see for example [66]).

Google's TPUs – which are discussed earlier in this document – can be seen as substantive evidence that HW/SW co-design may be the most practical way to sustain a radical change of paradigm in compute-intensive algorithms. Choosing this direction is a bold lateral step, which – renouncing tradition and consolidated solutions – requires capacity for massive investment and a very clear overall system concept as the ultimate goal.

Another direction of interest for dynamic, long-lived systems that include specialized parts whose deployment may make replacement or refurbishment not practical (due to remoteness, such as in space, or due to the number of units, such as in smart city infrastructures, especially in edge nodes) are reprogrammable field-programmable gate arrays (FPGAs). FPGAs are attractive for two direct and tangible reasons:

- i reduced weight and board space, due to decrease in devices required;
- ii increased flexibility, allowing design changes after the board layout is complete.

They are also attractive for two indirect repercussions:

- i increased reliability, owing to reduced solder connections;
- ii lower cost of ownership, owing to there being fewer vendors to qualify for use in critical systems.

An interesting ramification is that FPGAs are designed with hardware description languages, which make their development more similar to software (with its same risks and threats, and the corresponding mitigation practices) than to hardware. See [354] for a discussion of this technology ambit. However, the improvement of High-Level Synthesis allows now to program FPGAs using a programming language such as C.

The essential message here is that, increasingly in the future, the requirements of the application-level algorithm where most of the added value is and where performance, accuracy and efficiency are paramount, will determine the architecture of the processors that will run them. This will reverse what we know from the history of computing, where processor architectures have determined the characteristics of algorithms.

2.4.4 COMMUNICATION AND NETWORKING TRENDS

The second piece of the compute-communications-storage triangle, communications are central at all levels of integration in computing systems, from interconnections on a single die to exchanging data between continents. In the sections which follow, we discuss different kinds of communications networks.

2.4.4.1 WIRED: FROM BETWEEN DIES TO BETWEEN RACKS

Wired connections require cable infrastructure, and have the advantage of being relatively secure: it is not easy to eavesdrop on a wired connection. One wired connection does not experience interference from other wired connections (if shielded properly). They are also more energy efficient than wireless connections.

2.4.4.1.1 On-chip communication

Since the beginning, buses have been the favourite choice for interconnecting devices on a chip. A bus is a simple, fast point to point communication mechanism. However, it is also large and power hungry, and when a bus spans a “large” distance, the long wires suffer from signal degradation unless measures are taken on each wire. To overcome some of these negative effects, networks on chip (NoC) were developed. A NoC solves the real estate, signal degradation and energy problem, but it is more complicated in design, and is in general slower than a bus for small distances.

A bus transports information in parallel or semi-parallel mode: all the bits of the information word are carried next to each other from source to destination. Sometimes the information word is divided into large chunks: the chunks are transported serially, but the bits of each chunk are still transported in parallel. In this way, the number of interconnecting wires is traded for communication speed.

Often, a mix of buses and NoCs are found on a chip: buses for shorter and NoCs for longer on-chip distances. Buses and NoCs are almost always electrical interconnects, because this means that no interfaces between different transport technologies are required.

As chip dimensions decrease, wire diameters decrease as well, increasing resistance, and thereby reducing the distance an electric signal can travel before becoming too attenuated. This has already been recognized for a long time in long-distance communication, hence the development of fibre optical communication. Photons travel for large distances: a fibre can carry an optical signal for about 100 km before amplification is required.

Optical communication technology is, in principle, also usable for on-chip interconnections. Optical communications allow for much higher data rates: at 1000nm, the theoretical data rate is in the order of 100 Tbit/s. But there is a catch: the dimensions for an optical connection must be about one half the wavelength of the light used. For the most popular wavelength used in optical communication, around 800 nm, this would amount to 400 nm waveguides, about 40 times larger than the 10 nm technology node now aimed for by high-end chip manufacturers. This means that, while optical NoCs would be just about feasible, optical buses are simply ruled out for now because of required chip real estate.

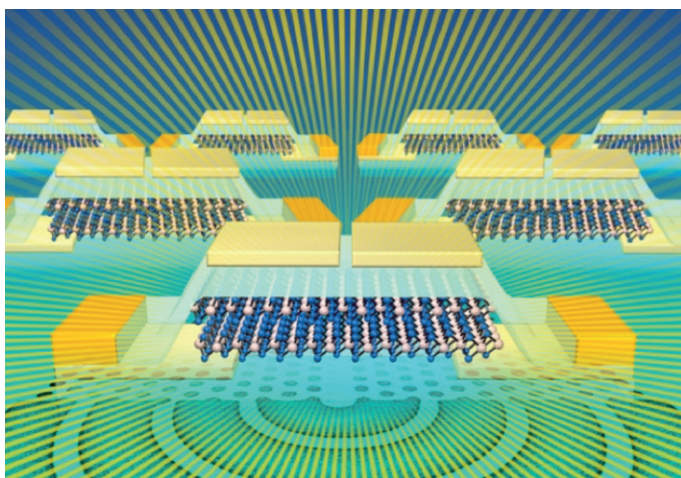


Figure 100: Molybdenum ditelluride light source for silicon photonics

Source: Sampson Wilcox

Going to a smaller wavelength poses problems in producing photons (the light source), and is also challenging because of the optical properties of silicon at smaller wavelengths. At a 200 nm wavelength, light is attenuated to 36% after an optical path length of 1 mm. Introducing other-than-silicon materials (such as glass or quartz) for optical communication between different parts of chips is a challenge for the chip production process. However, in [316] a new semiconductor material, molybdenum ditelluride seems to hold promise for use in CMOS devices to construct on-chip optical NoCs.

A further disadvantage of optical communication is the conversion cost: the electrical signal has to be converted to photons and back to an electrical signal.

To make the choice between buses and NoCs, future challenges will relate to balancing these technologies between speed and on-chip distances.

2.4.4.1.2 Between-chip communication

Originally, wire wrap was the preferred method of computer building, used some 30-odd years ago for implementing the Cray, which at the time was the-fastest computer in the world. However, for a long time now, circuit boards using printed copper wires have been the choice for interchip connections as they are relatively easy, and thus cheap, to mass produce. But as communication speeds increase, even the shortest wires start to act as delay lines. Interconnecting devices with photonics is therefore a logical choice. Although more complicated than copper wires, this is a technology that is slowly gaining ground.

A way to increase circuit integration efficiency is to combine chips – or chiplets – on interposers to form circuits with a larger functionality. The interconnect between the chiplets can be done using metal (copper or gold) wires, but optical communication is an excellent alternative for this application. The same holds for 3D chip technology, discussed in 2.4.1.2 “3D stacking: an answer to CMOS scalability challenges”, where chiplets are stacked, with through-chip vias for interconnect: these interconnects can be either metal or optical interconnects, although metal is still prevalent. The architectural consequences depend on the type of connection: bus type (many wires, fast, usually metal) or network type (single wire, slower, optical or metal).

2.4.4.1.3 Interboard communication, intra-rack communication

In the past implemented as flat cables, comparable to computer buses, or network-like coaxial cables, for the fastest interconnects nowadays optical fibres are being employed between circuit boards. Connections between units are typically implemented as optical connections as well.

2.4.4.1.4 Inter-rack communication

Unshielded twisted pair (UTP) cable connections between racks have been replaced by optical fibres at a rapid pace over the past few years.



Figure 101: Fibre optic cable

Source: Chaitawat Pawapowadon on Pixabay

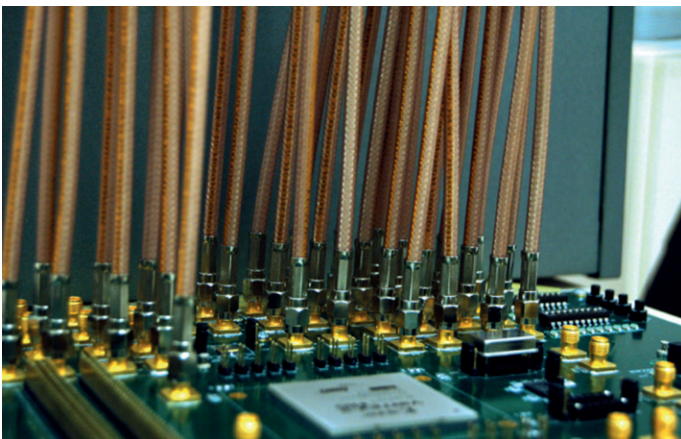


Figure 102: Interconnect prototypes from the EU-funded ExaNeSt project – Source: ExaNeSt

2.4.4.2 WIRELESS

In some circumstances, a cable infrastructure is impractical (mobile devices), not cost effective, or simply impossible. In such cases, wireless connections are used.

Wireless connections can in principle reach many devices in one transmission. Yet, in many cases, wireless connections are used for point-to-point connections. Certainly, with the advent of the internet of things, wireless communication is playing an increasingly important role, as it is infeasible to connect billions of devices by wires.

2.4.4.2.1 From short range to long range

A number of standards exist for wireless connections, each covering a particular range. Probably the most well-known among consumer users are Bluetooth and cellular networks.

Bluetooth and Zigbee, the standards developed for short range, offer data rates of up to 1 Mb/s and interconnect distances

ranging from a few centimetres to 100 m. These standards have the capability of trading speed for range, or for connection quality; see for example the Bluetooth 5 standard, now offered in several smartphone devices. Bluetooth has limited networking capabilities: up to seven devices can create a local network, and each device can be part of more than one network. Both Bluetooth and Zigbee have standards for low energy connections designed to accommodate communication for the IoT.



Figure 103: Wireless internet is a staple of modern life

Source: Bernard-Hermant on Unsplash

Other standards, such as wifi and different cellular standards have a significantly larger range, limited by its line of sight and transmit power. One way to overcome this, is to use a network of a set of interconnected non-hierarchical devices called a meshnet. Meshnets are self-organizing, and by their setup have high interconnect reliability. Meshnets can and have been constructed based on existing network technology, such as wifi, Bluetooth (BLE, Bluetooth Low Energy), and ZigBee (there is even an Ethernet mode, shortest path bridging, allowing switches to be connected in a mesh network).

It goes without saying that wireless interconnections, especially for self-driving cars and the IoT, require rigorous security measures, to which providers are paying an increasing amount of attention.

2.4.4.2.2 Wireless 5G will change the landscape

The mobile phone networks of the 1990s evolved into 4G data connection networks in the 2010s. The next generation, 5G, will evolve mobile networks into general-purpose high data rate networks, connecting devices with demanding applications such as driverless cars. The demand for high data rates can be met by increasing the baseband frequencies, which is equivalent to lowering the wavelength.

As a consequence, the range of these cm radio waves is smaller than for 4G networks, which employed between 30 and 15 cm radio waves. This entails a denser network of radio access points, which increases the total energy consumption of the internet.

The main challenge that 5G networks impose on the hardware and software community is to optimize the energy efficiency of these access points, as they impose the highest demand for energy. However, solutions may be implemented not only at the access points but at all levels on the internet communication network.

The current standard, 4G, has data communication fully integrated: watching streamed video on a smartphone is a common sight even on the street; a laptop with a wired internet connection is no longer required. The next generation of wireless networks, 5G, will shift the balance even further towards a computer network. IMT-2020, the mobile communications standard for 2020 and beyond produced by the International Telecommunications Union (ITU), provides for speeds of up to 20 gigabits per second at a frequency of 15 GHz (2 cm wavelength). The 3rd Generation Partnership Project (3GPP) standard, which allows for frequencies up to 6 GHz (5 cm wavelength), permits speeds which are 15% to 50% higher than 4G.

5G divides communications into three categories: enhanced mobile broadband, ultra-reliable low-latency communications, and massive machine-type communications. An example of ultra-reliable low latency communications are the connections to self-driving cars. The connections of such devices must for safety reasons be very reliable. In addition, since self-driving cars must react fast to real time events, the communication latency must be low. 5G is an important illustration of the computing spectrum: from edge through fog to cloud.

2.4.4.2.3 What's next? 6G!

5G networks are just being rolled out, but planning for the next generation of systems for the connected world is already ongoing. Should this be called 5G LTE, or is it really a step up?

As 5G is introduced, with little consumer demand as yet, technological experience and business models still need to be developed. However, 6G development represents a further step in projected technological requirements in order to deliver data-driven, almost instant, virtually unlimited connectivity.

Future applications might require communication latencies below 1 millisecond and data rates in the range of terabits per second. That means several things:

- 1 Data processing must be evenly distributed over the spectrum, from the edge to the cloud. Mobile devices will take up some of the data processing to guarantee low data latency.
- 2 New applications are anticipated to be data hungry, requiring data rates in the Tb/s range, and thus requiring terahertz communication technology. That will spur the development of new communication devices to open up this mm wavelength regime.

- 3 As the communication range at these wavelengths is reduced to line of sight, the number of radio access points will have to be increased, with some estimates even pointing to about 1,000 radio transmitters per person. This spreading and multiplication of access points will challenge the communication architecture as well, even more than 5G.

Hence, even though it is still early days even for 5G, the challenges 6G is posing are already keeping researchers in all fields of computing and communication busy.

2.4.4.3 SATELLITE COMMUNICATIONS

For decades, communications using satellites have been the realm of government agencies such as the US National Aeronautics and Space Administration (NASA) and the European Space Agency (ESA). This all changed when the Motorola-founded company Iridium launched a constellation of communication satellites for a worldwide satellite telephone network. Although not a commercial success, the Iridium network is undergoing its first major overhaul in 2017-2018, replacing its first-generation satellites.

During the past ten years, universities have been using small satellites (the size of milk cartons) to allow students to gain experience in building, launching, and operating satellites. Some of these so called nano-satellites were simply used to experiment with satellite communications; others were used as a low-cost experimental platform for new space-born technologies such as innovative solar cell technology, or wireless intra-satellite communication.

This was made possible through the price reduction that came from standardizing the nano-satellite design. This same price reduction, and a new generation of students experienced with the nano-satellite design coming out of universities, has sparked a number of companies experimenting with nano-satellite communication [336].

A relatively low-cost network with a global coverage such as this of course holds enormous potential for the IoT to penetrate even the remotest locations on Earth. However, this comes at a cost, and not only because of the still much higher network costs. Building and launching a nano-satellite still costs in the order of several €100K, but that price may go down through the advent of commercial launch services such as SpaceX. The other cost is in the energy consumption required to communicate with a satellite, which is still in the order of several watts. So, again, for this technology to succeed, intelligent energy consumption is the key to success.

2.4.5 STORAGE TRENDS

Storage is the third element of the computing systems triangle (with compute and communications forming the other two). In this section, we review the state of the art and forthcoming technologies.

The field of memory devices can be separated into two major categories:

- 1 Volatile memories, which require the presence of a power supply and which lose data at power-off in a certain, relatively short, amount of time. These are represented mainly by dynamic random access memory (DRAM) – specific cells, very dense, fast access but requiring regular data refreshes – and static random access memory (SRAM) – logic devices cells, less dense than DRAM, fast access, refresh not required.
- 2 Non-volatile memories, which retain data after power-off for a long time, ideally tens of years. These are today mostly represented by NAND flash memories, as NOR flash is now limited to a few small niche areas; similarly, older generation magnetic memories, excluding disks and tape, are confined to off board system use, while new magnetic technologies like magnetic tunnel junction (MTJ) devices are still in their infancy (see Section 2.4.5.3, 2.4.5.2.2).

Both volatile and non-volatile memories can either be fabricated in standalone chips or embedded in logic chips. The latter approach, however, requires a number of compromises in speed, density and performances that (with the notable exception of SRAM), means they are not as good as the corresponding standalone version, although they are often extremely useful.

2.4.5.1 VOLATILE MEMORIES

As noted above, DRAM and SRAM are the main forms of volatile memory currently available.

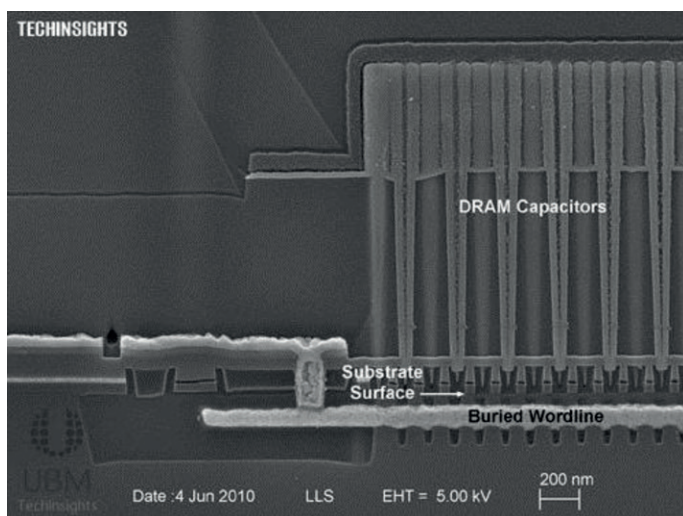


Figure 104: Cross section of a typical DRAM memory.
Source: Aspencore, Inc.

DRAM memory still works on the principle of a charge stored in a capacitor accessed by a transistor. To increase the density of the storage, the capacitor is increasingly a high cylindrical structure with a high permittivity insulator material allowing access to the transistors below the capacitor cell. An example of the structure is shown in figure 104, although the state of the art is now much smaller with unit cell having a surface of around 0.002-0.0026 μm^2 (as witnessed in the latest LPDDR4 products from Samsung and Micron), giving a density of about 0.1-0.15Gb/mm². Over the next five to ten years, material improvements and the use of more complex lithographic techniques should be able to deliver density improvements of between 30 and 70% along an evolving path.

The biggest progress, however will come from high-bandwidth memory (HBM) components. They increase the “volume” density, hence capacity, and bandwidth by 3D assembly of chips using through-silicon via (TSV) technology. Such assembly is schematically illustrated in Figure 105 along with comparison with some other type of DRAM implementations. For high-performance applications, in particular, HBMs are the best solution to attain the data transfer rates necessary to alleviate the bottleneck between the memory and computing units.

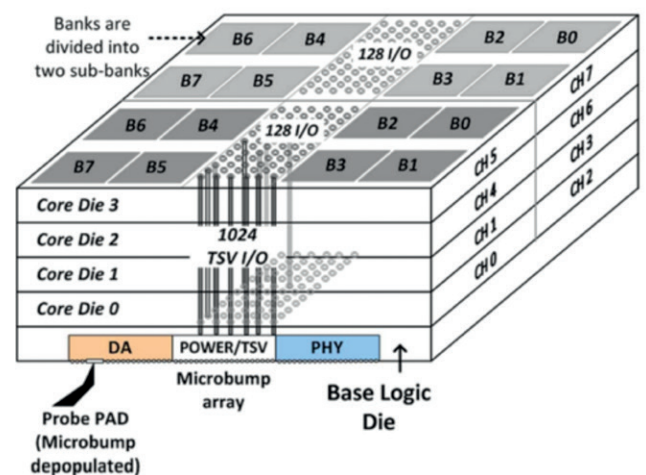


Figure 105: HBM memory with 4 DRAM chips
Source: [52]

2.4.5.2 NON-VOLATILE MEMORIES

2.4.5.2.1 NAND flash

As noted above, the architecture that has come to dominate the non-volatile memory market is the NAND flash. Its major advantage is the high density that can be achieved thanks to the compactness of its unit cell. As with all flash memories, information is represented by a charge stored into an insulator or a polysilicon layer acting as a secondary gate on a transistor. This limits the minimum size achievable as, below a certain size the number of electrons stored becomes too small for a reliable operation.

However, the structure of the cell is very regular and lends itself to a 3D implementation. This type of integration is very complex,

as it involves very high aspect ratio hole etching and filling and this has been very difficult in production. Fortunately, in this type of integration, the lithographic constraint and those on the size of the storage gate are strongly reduced. As soon as the number of superposed layers (hence number of bits per cell) goes above about 80, the cost equation becomes favourable compared to a planar solution. Above 80 the cost per bit rapidly decreases and now 3D flash memories are progressively replacing 2D flash in very high density chips and applications.

Over the next few years we will see an increase of the density due to an increase of the number of layers until the limit of the access resistance of the vertical structure is reached. Subsequently, EUV lithography should be able to allow the unit cell to be further reduced in size.

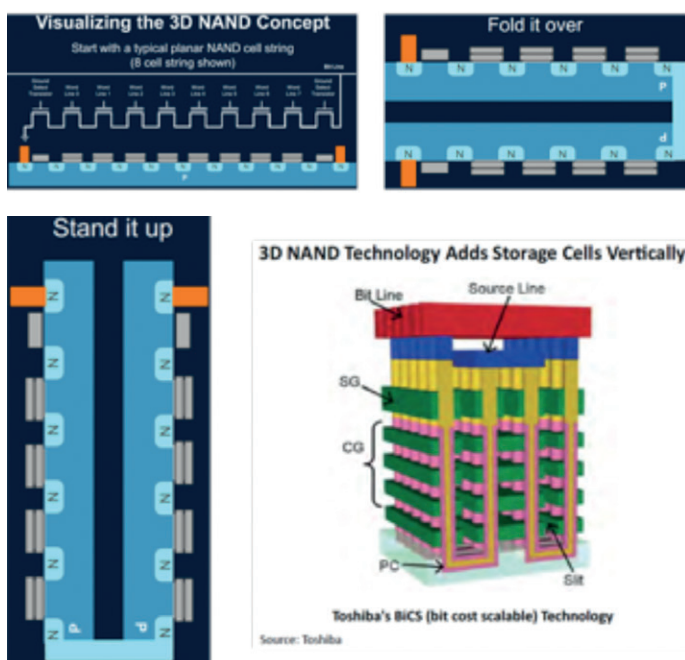


Figure 106: Principle of 3D NAND flash
Source: Toshiba

2.4.5.2.2 New non-volatile concepts based on material resistivity

In recent years, limitations of non-volatile memories have spurred research on new concepts based not on charge but on semi-permanent variations of the resistivity of a material or of a stack of materials. Different effects have been studied with a view to exploitation, in particular:

- the thermally induced change of the phase from crystalline to amorphous and the reverse in a chalcogenide material (phase change memories, PCM or PCRAM);
- the formation of a metallic conductive path between electrodes on a chalcogenide or insulating materials (conductive bridge RAM, or CBRAM);
- the formation of a conductive path due to ions or vacancies in a metallic oxide (oxide-based RAM, or Ox RAM);

- and the tunnelling through a magnetic junction (Magnetic RAM, MRAM).

With the exception of the CBRAMs, all the others are currently very actively researched and demonstrations of systems with over 1Mb have been obtained. However, only the embedded version of some of them (notably the MRAM integrated by Samsung on its 28nm FDSOI technology) has been taken close to market. One other exception, which is difficult to comment on due to lack of scientific papers, is the 3D XPoint technology from Intel/Micron, which is supposed to be a form of PCM and which has been introduced as a standalone memory.

While their characteristics are quite different from one to the other with very different values for the major parameters (on/off resistance level, on/off ratio, endurance, retention time, access time, density, read/write voltage levels and so on), they all have in common that they can be integrated at a relatively low temperature between the metal layer of the backend process and, potentially, to be realized in stacked planes of cells.

The possibility of integrating these memory cells in the backend also opens up the possibility of integrating them within a logic process and obtaining memory planes much closer to the computational element. Ideally, this would open up the possibility of doing some computation directly within the memory planes, as discussed in section 2.4.3.2 “Near/In memory Computing”.

2.4.5.2.3 Carbon nanotube based memories

A more futuristic-sounding approach to memory is the nanotube-based memory technology being developed by Nantero. The idea is to form a film of carbon nanotubes on a silicon substrate containing logic to select and index the memory. Depending on the state of the nanotubes (either touching one another or not), they can represent either a 0 or 1 bit [32]. One of the proposed advantages would be that it has very low power costs, should scale to extremely low feature sizes, and is compatible with existing CMOS fabs.

2.4.5.2.4 Magnetic storage

Today, many consumer devices (smartphones, tablet, laptops) start out equipped with solid state disks due to their superior performance (faster, more energy efficient). Desktop computers are following. In the future, magnetic storage might become a niche market for data centres. Given the large price difference between magnetic storage and solid-state storage, there is a good reason to assume that hard disks will still be used in the foreseeable future in data centres, in combination with magnetic tape storage. The fact that manufacturing companies keep innovating the storage technology for hard disks also confirms this.

Bits on a hard disk are stored in sets of magnetic grains. A magnetic grain is about 8nm, and it cannot be made much

smaller because super-paramagnetism will cause random flips of the magnetic grains under the influence of temperature. One stored bit consists of 20-30 grains and has a bit width of 75nm and a bit length of 14nm. The number of grains cannot be reduced much if we want to keep a sufficient signal-to-noise ratio. Therefore, the maximal density of perpendicular recording is about 1 Tb/in². Today, hard disks with a density of 1 Tb/in² are commercially available.

The bit density can be further increased by reducing the bit (track) width. The idea is that a track is written full-width, but the next track partially overwrites the previously written track (just like shingles on a roof, hence the name 'shingled magnetic recording'). The remaining strip of the track is wide enough to be read, but it can no longer be written without destroying the data in the neighbouring tracks. This leads to disks where data must be stored in bands. These hard disks have to be used like solid-state disks; bands must be written sequentially and cannot be changed, they can only be overwritten. However, since much contemporary data is write-once (like images, movies, audio files), the fact that rewriting requires more work is not that problematic.

Shingled magnetic recording increases areal density about 25% [184]. In 2016, major hard disk vendors introduced helium-filled hard drives. Helium is seven times lighter than air, and creates less friction and less turbulence inside the hard disk, and hence less heat. This allows for higher rotational speeds (10,000 rpm) and 50% more platters in the same volume, increasing both the bandwidth and the capacity of the hard disk.

Narrower tracks lead to more interference from adjacent tracks when reading. Two-dimensional magnetic recording improves the signal-to-noise ratio by using multiple read heads: one to read the central track, and two heads to measure the interference from neighbouring tracks. By combining the three signals, the signal-to-noise ratio can be improved, and the track density can be further increased.

Beyond shingled magnetic recording, other approaches are needed. One approach is energy-assisted magnetic recording, of which heat-assisted magnetic recording is the best known. It uses heat in combination with a magnetic field to record the bits. This, however, requires that a heat spot be localized on a single track and that the rise and fall times be in the sub-nanosecond range. Designing such a head is challenging. Heat-assisted magnetic recording could eventually lead to an areal density of 4 TB/in².

The next approach is to make use of patterned media. In patterned media, each bit is recorded on a small island of magnetic material, surrounded by a non-magnetic material. In this case, a bit can be made as small as a single magnetic grain (instead of 20-30 grains for perpendicular recording). In order to reach 1 Tb/in² in patterned media, it is necessary to etch islands of 12 nm, which is beyond

the resolution of current lithographic systems. That means that patterned media will have to rely on self-ordering. Densities of up to 10 Tb/in² by 2025 seems to be theoretically possible with patterned media, if combined with heat-assisted magnetic recording. However, today, bit patterned media is not yet ready for the market.

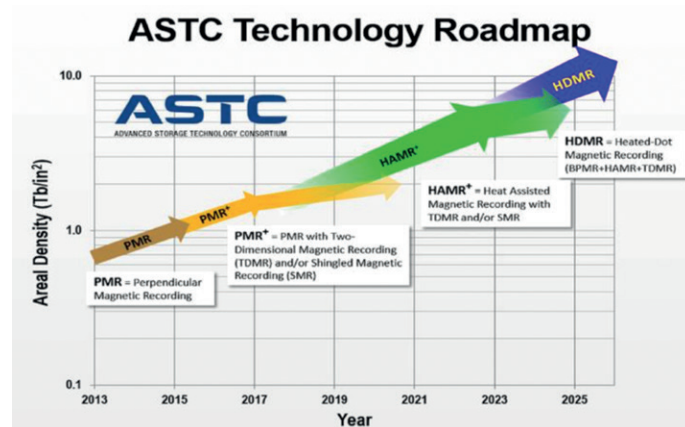


Figure 107: Evolution of areal density according to the ASTC Technology Roadmap
Source: ASTC

Hence, it does not seem that the magnetic storage technology is currently running out of steam. There is a roadmap for at least seven more years.

Another magnetic storage technology has an even more promising roadmap [131]: tape. Tape has survived for a very long time, for two main reasons: it is cheap, and its growth capacity is 33% per year (compare to 15% for hard disks), or a doubling of the capacity every two to three years. The areal density of tapes is lower than that of hard disks (up to 100x), but this is compensated by a much larger surface (modern tapes can be as long as 1 km). Since the areal density is so much lower than that of hard disks, there is still a lot of headroom before the superparamagnetic limit will be reached.

The difference in growth rates means that tape storage is getting cheaper faster than disk storage, and that offline storage in a robotic library becomes a valid alternative for data that do not need to be online all the time [87]. Data stored in such a library has a number of unique advantages:

- It is the cheapest form of mass storage.
- It is more energy efficient because a tape in the library does not need to be powered. This is very important for long-term storage.
- It is more secure, because it cannot be accessed if it is not sitting in a drive. The air gap protection in the robot library is very reliable.
- It is more reliable than hard disk storage (five orders of magnitude, which is important for archival applications).

2.4.5.3 FUTURISTIC STORAGE

In the future, human and/or synthetic DNA could be used to store and retrieve data in an extremely dense and efficient manner. Data is stored by manipulating the base-pairing mechanism in the DNA using DNA synthesis methods, and can be read from it by using DNA sequencing methods.

Microsoft in collaboration with University of Washington [90] built a DNA-based storage archival storage system on a synthetic DNA and demonstrated its feasibility, robustness and random access to the storage with wet lab experiments. Most recently, they stored 35 files, equivalent to 200MB of data, on a synthetic DNA and have been able to recover each one without any errors [152]. Recently, Church et al [173] in the Wyss Institute for Biologically Inspired Engineering and Harvard Medical School have stored a digital movie in a living bacterial cell, and retrieved it with 90% accuracy using the CRISPR-Cas gene editing system.

2.4.6 COMPUTATIONAL MODELS

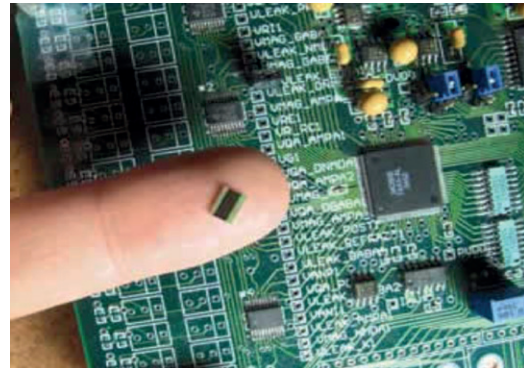
Another approach to enhancing the performance and efficiency of ICT devices is by changing the computing paradigm, of course matching the right paradigm to the right application. The following section gives a few examples of interesting approaches, without offering an exhaustive summary.

2.4.6.1 NEUROMORPHIC COMPUTING

The idea of mimicking how the brain works in order to design better (more intelligent) computing machines is not a new idea. As touched upon in section 2.2.1.1 “The AI bandwagon”, from the early effort of modelling neurons and synapses by McCulloch and Pitts in 1949, the idea of the “perceptron” by Rosenblatt in 1950, works on multi-layers neural networks by Fukushima in the 1970s to the modelling of the dynamics of recurrent neural networks in the early 1980s by Hopfield, the topic really took its modern form in the late 1980s with the work of Carver Mead. See [291] for an overview.

Since then we have seen a variety of computing approaches. Some are very loosely related to the initial idea of neuromorphism. In fact, they are implementations of vector matrix products accelerators often known as neural processing units (NPU) or neural engines, and are finding their way into almost all new “AI powered” devices, such as smartphones, cameras, washing machines et so forth.

On the true neuromorphic side of the spectrum the work is still mostly academic due to the exploratory nature of the field, with some industrial companies also investigating the possibility of getting real inspiration from the brain. For the purposes of this discussion we consider efforts that try to mimic the human brain by imitating how the architecture of biological networks processes information with streams of spikes (or event data) as “neuromorphic”.

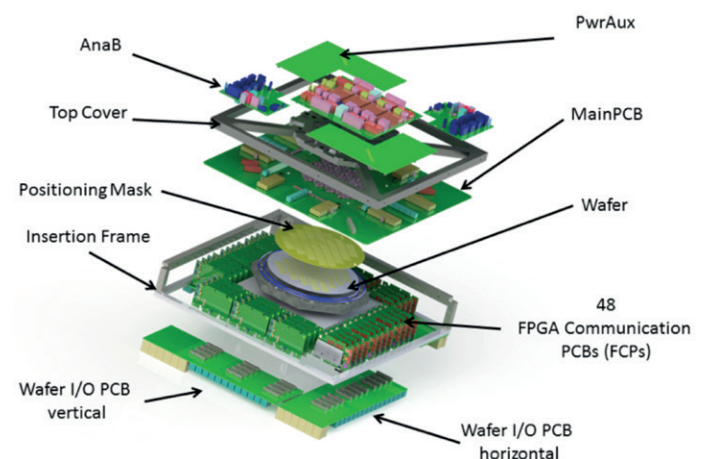


The Neurogrid chip models 1,024 excitatory pyramidal cells and 256 inhibitory basket cells. – Source: Emily Nathan 2007

The **Neurogrid** system [22] from Stanford’s team “Brains in Silicon” is designed for fast simulation of biological neural circuits. It is based on mixed digital and analogue circuits aiming at simulating the behaviour of various parts of the biological brains such as synapses, membrane potential and ion-channels, in order to reproduce the shape of brains signals (spikes). The Neurogrid system is made of several custom chips (Neurocore) implementing 65,536 artificial neurons, each with up to 256 connected synapses per neurons and several shared parameters that can be tuned to adjust the neuron simulation.

The **True North** chip originated from the contribution of IBM Research to the DARPA synapse project. The purpose of the IBM’s TrueNorth [9] architecture is to provide a generic platform for computational applications. The architecture includes 4,096 cores on a single CMOS chip. Each core contains 256 “leaky integrate and fire” (LIF) digital neurons and each neuron is connected to 256 synapses. All cores are interconnected by a network-on-a-chip (NOC) and it is possible to interconnect several cores together to form a larger network. In this case, the architecture time is discretized and cores are synchronized by a global signal of 1kHz. The average power consumption is estimated at around 68mW.

Brain Scales [182] is a neuromorphic architecture developed at the University of Heidelberg, Germany. The circuits of BrainScales



The Wafer-scale structure of Brainscales – Source: [365]



The Brainscales system – Source: [365]

mimics the structure of biological neurons and synapses and try to retain their biological functions as far as possible. It is based on a mixed-signal design (analogue/digital) and is implemented as a set of wafer scale circuits running up to ten thousand times faster than real time.

The BrainScaleS system [365] (NM-PM-1) contains 20 8-inch silicon wafers in 180 nm process technology. Each wafer incorporates 50 x 106 plastic synapses and 200,000 biologically realistic neurons. The system does not execute pre-programmed code but instead evolves according to the physical properties of the electronic devices. The estimated power consumption of the architecture on a wafer is 1 kW. The BrainscaleS system have been developed and deployed as part of the Human Brain Flagship (HBP) project.

While writing this document, the European neuromorphic computing community was saddened by the sudden death of Prof. Karlheinz Meier, 24 Oct. 2018, coordinator of the Human Brain Flagship and leader of the BrainScaleS project.

The Spinnaker [65] neuromorphic system is a based on ARM cores and aims at speeding up simulation in computational neurosciences. In contrast to BrainScaleS, Spinnaker is a fully digital design build upon industry standard ARM processors. Nevertheless, the hardware and software architecture of Spinnaker is really targeted at speeding-up computational neurosciences problems, with some impressive results. The SpiNNaker system (NM-MC-1) provides almost 30,000 custom digital chips, each with eighteen cores and a shared local 128 MB RAM, giving a total of over 500,000 cores. A single chip can simulate 16,000 neurons with eight million plastic synapses running in real time with an energy budget of 1W. It has been developed at University of Manchester, UK, in a team led by Prof. Steve Furber.

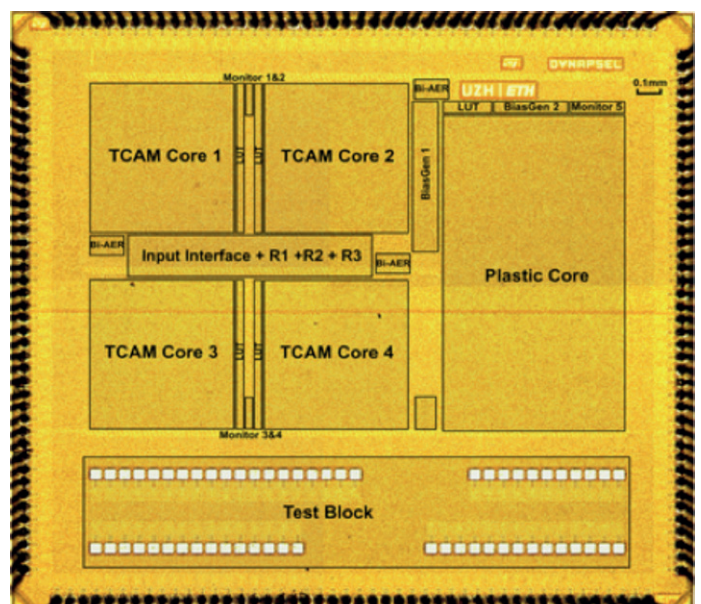


The SpiNNaker system – Source: [65]

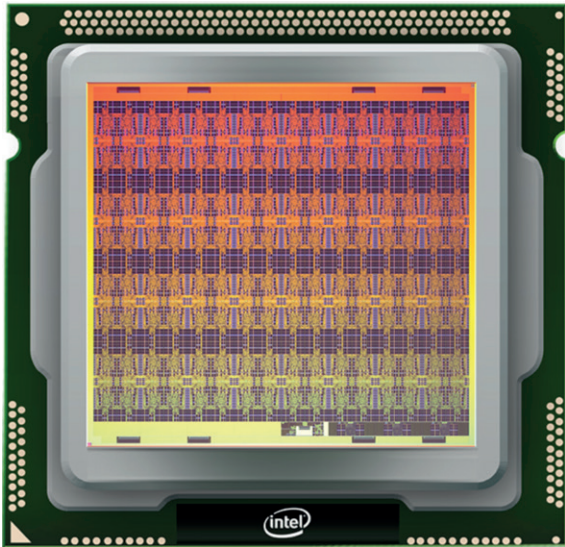
Both the BrainScaleS and Spinnaker systems are part of the HBP Neuromorphic Computing Platform [366]. The platform provides a number of documentation resources, software tools, ways to request a resource allocation and help on running neuromorphic simulations onto the physical machines.

Dynapsel [433] is a five-core fully-asynchronous mixed-signal spiking neural network chip with on-chip learning (STDP) fabricated in 28nm FDSOI process with a silicon area of 2.8mm x 2.6mm.

The chip comprises four TCAM cores with 1k analogue LIF neurons and 64k 15-bit TCAM synapses sub-divided in four cores, one learning core with 64 analogue leaky I&F neurons, 8k digital 4-bit plastic synapses, and 8k 4-bit digital configurable synapses. The architecture of the chip is based on the work of the team Prof. Giacomo Indiveri at the University of Zurich [83]. The estimated power for the Dynapsel is of 2.8 pJ per synaptic event, giving a total power efficiency of 320 giga synaptic operations per watt



Die photo of the Dynapsel chip – Source: [433]



Intel's Loihi neuromorphic chip – Source: Intel

(GSOP/W) compared 46 GSOP/W for the TrueNorth circuit. The DynapSel circuit was designed as part of the European collaboration NEURAM3 [432].

Intel's new Loihi chip [47] is a neuromorphic manycore processor with on-chip learning fabricated in Intel's 14 nm process. In contrast to the other chips mentioned here, which mostly aim at brain simulations, the Intel effort is clearly targeting the AI market. The aim of Loihi is to offer a self-learning solution based on neuromorphic principles with spike-based computation.

By attempting to mimic the biological brain as closely as possible, both its architecture and its way of coding information, neuromorphic circuits are primarily targeted toward computational neuroscience. Indeed, by designing hardware circuits reproducing detailed functions of the brain and scaling their number using state of the art microelectronics technologies, neuromorphic chips can speed-up simulations faster than by using a pure software approach. This contribution of neuromorphic computing to a better understanding of the brain (how it works, how it encodes its environment, how it is so power efficient) is certainly very important. However, it would not be fair to only consider neuromorphic chips as simple accelerators of brain simulations.

In reality, most of what we learn in designing neuromorphic computing chips and systems, could be used to vastly improve current digital or analogue designs based on conventional neural networks for deep-learning. Spike (or event)-based coding is naturally adapted for processing data collected from ever-changing environments, such as the flow of information from various sensors in an autonomous vehicle or robot. Spike-based coding is also a promising way to reduce the power consumption of computing systems since it results in very sparse activities [70].

Indeed, spike coding and neuromorphic computing architectures modelled on the biological brain is a very active field of research. While sometimes overshadowed by deep-learning fever, it is

definitely on a track for a bright future. For more references on the field of neuromorphic computing hardware, the reader could read the extensive review on the works in the field proposed by Schulman *et al.* in 2017 [459].

2.4.6.2 RESERVOIR COMPUTING

The concept of reservoir computing is often considered as a special case of recurrent neural networks. It was proposed independently by Wolfgang Maas [374] as the liquid state machine and Herbert Jaeger [239] as echo state networks. It consists of an input layer fed into a recurrent network of nonlinear neurons with randomly fixed weights (the reservoir) and a readout layer with trained weights.

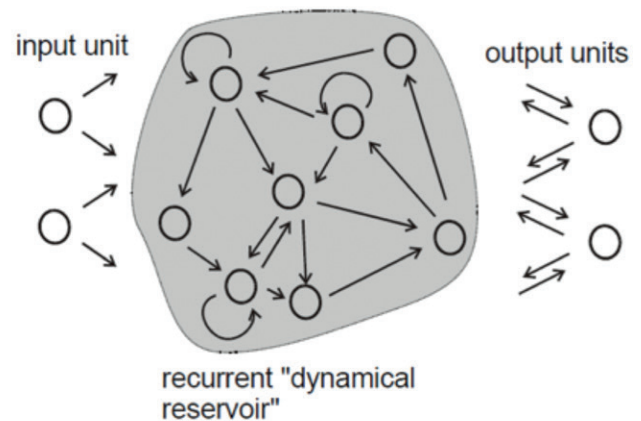


Figure 108: The concept of reservoir computing as sketched by Herbert Jaeger

The basic idea is to project an input vector onto a higher dimensionality space implemented in the reservoir before reducing the information by training the readout layer. The main benefit is that learning is restricted to the readout layer and can be very fast in principle. Another benefit of reservoir computing comes from the delayed dynamics of the reservoir which makes it very well adapted to process timed data series inputs for prediction purposes.

Several projects have explored the concept of reservoir computing implemented in various technologies, for example using photonics technology as in the Phocus and BioPhoProc projects [308], or even directly embedded into a ionic sensor as shown in the RECORD-IT [341] project. This latter example is an interesting case of using a computing paradigm (reservoir computing) in a sensing device: the non-linear reservoir is implemented by a biological fluid. It was observed that the variation of ionic species concentration of the fluid induced changes in its non-linear properties that could be sensed with a simple readout layer.

Another approach by the group of Wei Lu at Michigan University demonstrates how memristors could be used to implement a non-linear dynamic reservoir [380]. They implemented a fairly small reservoir with 88 memristors and fed it with MNIST data

(transformed to spikes) to perform the classification operation using a simple readout layer with a success rate of 88%. This way we could imagine that a reservoir network could be implemented using memristor arrays with randomly set weights using their natural dispersion and randomness. Therefore, the dispersion of memristor characteristics, a major weakness for their use in traditional deep networks, becomes a key advantage when implementing a reservoir. A reservoir computing architecture could thus be composed of layers of memristor-based reservoirs combined with simple readout layers implemented with standards circuits.

Reservoir computing techniques can also be simulated using software packages. An open source version of the Oger Reservoir Computing_simulation toolbox is available on GitHub. Oger was introduced by a group at Ghent University in 2012 [251] after experiments previously performed by that same group in 2007 [255].

Although reservoir computing techniques have been proposed and developed for some time and show promising properties in terms of learning time [333], they still have not garnered the level of interest of deep learning techniques. Reservoir computing techniques are still in their research phase, much like other unsupervised approaches.

2.4.6.3 AI BEYOND DEEP LEARNING

In spite of the success of deep learning in AI, it is not without its critics, who believe that deep learning is just a tool in AI instead of a panacea in machine intelligence. The deficiencies of deep learning identified by critics are threefold:

- 1 Deep neural networks are black boxes whose outputs cannot be explained [40].
- 2 They need huge amounts of labelled training data to learn while humans can learn and generalize with a single example via “one-shot learning” [122].
- 3 They are “naïve” and can be easily fooled because they are programmed with no common sense [424]. A good example is shown in the figure 109 below [268] where a picture of a panda is taken and a “gibbon” gradient is added to it, then a deep neural network classifies it as a gibbon. As discussed in section 2.3.1.2.2 “Security threats to the IoT and CPS”, malicious attempts to fool deep learning networks in this way could have serious repercussions.

Of course, there are examples of research showing potential solutions to those problems, such as self-supervised learning [301]. However, many of the critics of deep learning argue that causality and learning with uncertainty are missing in deep learning-based approaches. There is some consensus among these critics that the next big wave in machine learning and AI

will be combination of different techniques also using probabilistic and Bayesian learning.

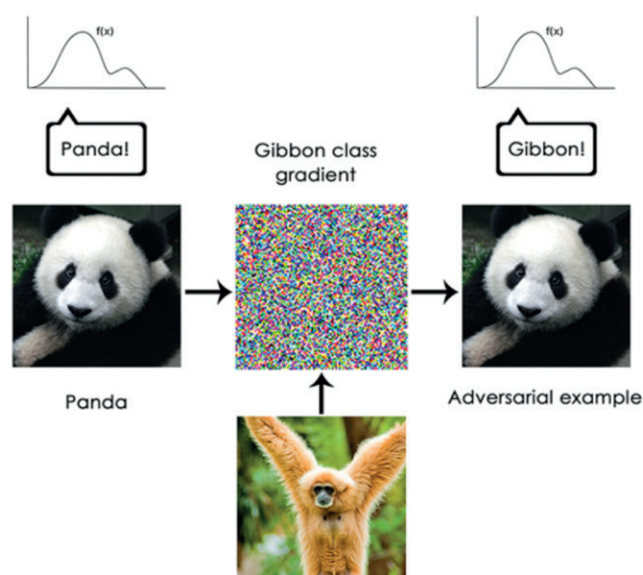


Figure 109: Limitations of deep learning
Source: Francois Chollet, ‘Deep Learning with Python’, 2017

For example, Lee et al, [444] proposes Bayesian program learning capable of learning a large class of concepts from a single example in which concepts are represented as simple probabilistic programs. Ghahramani says in [450] Bayesian non-parametric models (e.g. Gaussian processes) are flexible models that do not need a large amount of training data and are dynamic to learn better as they observe more data. He argues that probabilistic programming is essential for representing probabilistic models. Similarly, Intel Chief Technology Officer Mike Mayberry thinks probabilistic computing will be the third wave of AI after rule-based and deep learning-based AI waves [323].

Judea Pearl, a Turing Award winner and inventor of Bayesian networks, goes beyond deep learning and probabilistic approaches in his new book *The Book of Why?*, arguing that causal reasoning could provide machines with human-level intelligence where machines will communicate with humans more effectively [396]. He proposes an alternative approach, “reasoning with cause and effect”, rather than “reasoning with uncertainty”. According to Pearl: “If we want machines to reason about interventions (‘What if we ban cigarettes?’) and introspection (‘What if I had finished high school?’), we must invoke causal models. Associations are not enough — and this is a mathematical fact, not opinion.” He predicts that “reasoning with cause and effect” will eventually lead to “machine free will”.

EVOLUTIONARY COMPUTING

Evolutionary computing or algorithms have been around for 30 years and are mainly used in optimization and search problems. More recently, they have emerged as a viable toolset for AI applications, in particular to evolve neural networks, a concept called “neuroevolution” [390].

Evolutionary algorithms (EA) mimic natural evolutionary mechanisms such as mutation, selection, recombination and reproduction. EAs are used in many applications for searching, optimization, bioinformatics, VLSI chip implementation, hardware verification and testing.

EAs generate a population of random solutions to a problem, and each solution is evaluated by a fitness function measuring the goodness of the solution. A subset of the solutions from this population is selected by the fitness function as parents. Then, the parents reproduce by mutation to generate an offspring from which the survivor selection is made. The solutions in the offspring become the new population of solutions. This process is iterated until a particular termination condition is met. The flow of an EA is shown below:

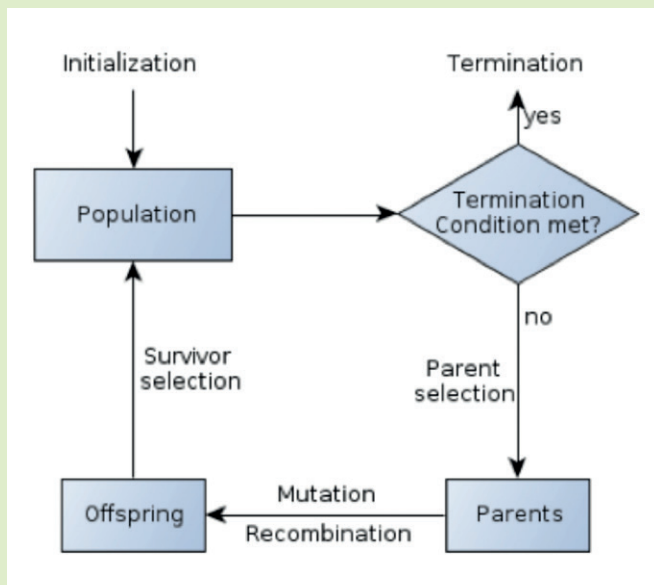


Figure 110: Flow of an EA

Source: 2017 Sentient Technologies Holdings Limited

Most recently, it has been used in learning how to play video games; in some video games, it has even outperformed deep learning [405].

Uber has recently demonstrated that EAs can be used as a non-gradient based training algorithm for deep reinforcement neural networks, which can be a good alternative to backpropagation training algorithms [263].

2.4.6.3.1 Statistical and probabilistic reasoning

The convergence between traditional embedded real-time systems and the cognitive CPS envisioned in this document will require worst-case analysis of non-functional properties for cognitive CPS. Interestingly, however, the analysis techniques to use for the emerging systems will have to adopt other strategies than classic static analysis.

As long as certain basic preconditions are met, statistical and probabilistic reasoning may prove a useful approach, or a combination of various approaches including symbolic ones. This is because, as processors become ever more complex and heterogeneous by including various accelerators, it is increasingly difficult for static analysis techniques to cope with the explosion in the state space when attempting to model the inner operation of all relevant components. As a result, conservative assumptions are made to compensate for unknown details.

The benefit of statistical and probabilistic techniques is their ability to reason on black-box observations, which significantly easier to obtain than white-box knowledge on the relevant internals of an execution. A string of research works recently conducted in Europe (such as [60], [61] and [80]) span issues and challenges ranging from tailoring statistical techniques used in other domains to fit the timing (and energy) analysis problem, to dressing the hardware or the software runtime of the system to match the premises of probabilistically analysable behaviour, via revisiting the way such analyses should work.

This body of work and the evidence collected from representative use cases show potential worth of industrial consideration as well as further research. The biggest research challenge still ahead of this novel branch is to provide confirmatory arguments that the observations collected during analysis are sufficient to capture all of the significant contributions to worst-case scenarios. This is often called “the representativeness problem”. Injecting randomization in the non-functional behaviour of selected hardware components, those for which best-case and worst-case behaviour span a large distance, has been explored as a valid technique to yield statistical representativeness to measurement observations.

Interestingly, these hardware modifications have attractive repercussions on security, in that the observable non-functional behaviour becomes non-deterministic and therefore less apt to use by attackers, and assurance, in that the risk of pathological (deterministic) behaviour is averted by construction. Bayesian reasoning and machine learning techniques are also beginning to be explored to address the variability caused by software execution taking different program paths across observations.

2.5 SYSTEM-LEVEL DIRECTIONS

Having explored the problems and potentials opportunities of novel solutions for the “physical” part of ICT systems, in this section we first discuss how system organization is expected to evolve and the consequences for software composition. Subsequently, we discuss software implementation for those systems, considering programming and compilation.

2.5.1 THE CONTINUUM OF COMPUTING

The notion of “the continuum of computing”, encompassing “edge”, “fog” and “cloud” computing to support data-driven applications and business intelligence more generally, has been around for some years; see for example [124].

As discussed in 2.2.1.3 “The continuum: Cloud, fog and edge computing”, at one end of the continuum are pervasive devices located as near as possible to the user or to the physical target of interest, known as the “edge”. These in turn are linked to specialized services which can be run either on centralized servers, as in the “cloud” model, offering maximum scalability, or closer to the original device(s), as in “fog” computing, which provides greater responsiveness.

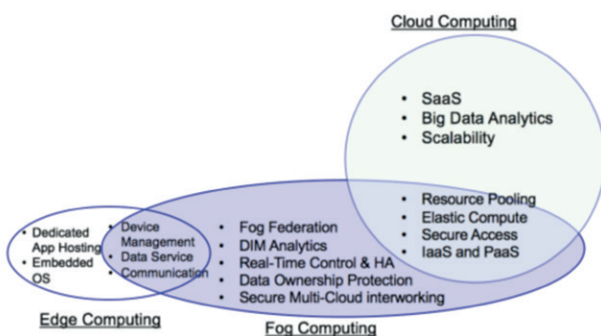


Figure 111: A pictorial view of the edge-fog-cloud continuum of computing

Source: [222]

Initially, the concept of the continuum was brought to public attention through attempts by hardware providers such as Intel [84] and software providers (Microsoft’s attempt to enter the mobile phone market and join its application space to its desktop platforms) alike to cover the full spectrum, in addition to being quietly but steadily pursued by Google. Nearly a decade later, the same notion has risen to a fully acknowledged and attractive prospect, now that it has embraced the need to accommodate heterogeneity.

The boundaries of this continuum are flexible and can cut a thin or thick slice of the fog, edge and cloud space, according to application need. Similarly fluid is the functional apportionment, as evoked in figure 111. Yet the internal architecture of each instance of a continuum system is bound to have very similar characteristics. It are these characteristics, arguably of high

strategic importance for future applications, that we discuss in this section.

The software fit for “continuum” systems exhibits two distinguishing and closely intertwined traits that are vectors of high value-added potential.

- First, they are designed to be provided *as-a-service* (i.e. through internet connectivity), which means that they require very little in the way of installation and execution on the target device. As a consequence, the need and extent of embedding are reduced to a minimum.
- Second, they are designed to inter-operate at the *highest level* of the internet protocol stack, which allows them to realize value-added functions via natural distribution (and possibly even decentralization) by functionally aggregating components regardless of their physical location – without the limitations in the addressing capabilities of lower-level protocols – and of the technology stack in which they reside.

The union of these two traits yields a very powerful combination, which sets a clear trend for all other software that aims to make lasting impact.

Individual or professional users alike are becoming increasingly familiar with software applications being provided as-a-service; indeed, as discussed in 2.2.2.1 “Renting instead of buying”, the trend to “everything-as-a-service” is gaining traction in modern economy, as shown in the infographic in figure 112 [429].

The fruition of those applications via web browsers – which are becoming known as “progressive web apps” for the mobile app market segment (see Google’s views on this at [281]) – has prompted significant advances in web technologies, although there are still plenty of possibilities to explore. What used to be a rather simple, near-dumb, client merely tasked to represent server-side contents, has become a full and yet comparatively light-weight run-time environment where client-side value-added computation takes place at a different time to server-side activity. This asynchrony is essential to assuring acceptable user experience, as well as being attractive in that it makes use of compute cycles that would otherwise be wasted in synchronous waiting. This observation explains the massive emergence of asynchronous programming languages, whose role is vital to the client side of as-a-service applications, and has consequently established asynchronous programming as an important paradigm.

Incidentally, but also importantly, the fact that the client side of web-based applications may have a larger share of responsibility in the overall computation makes it more evident that privacy requirements can and should be addressed on the client side too. In this regard, see for example [414], which urges the deprecation of server-controlled cookies.

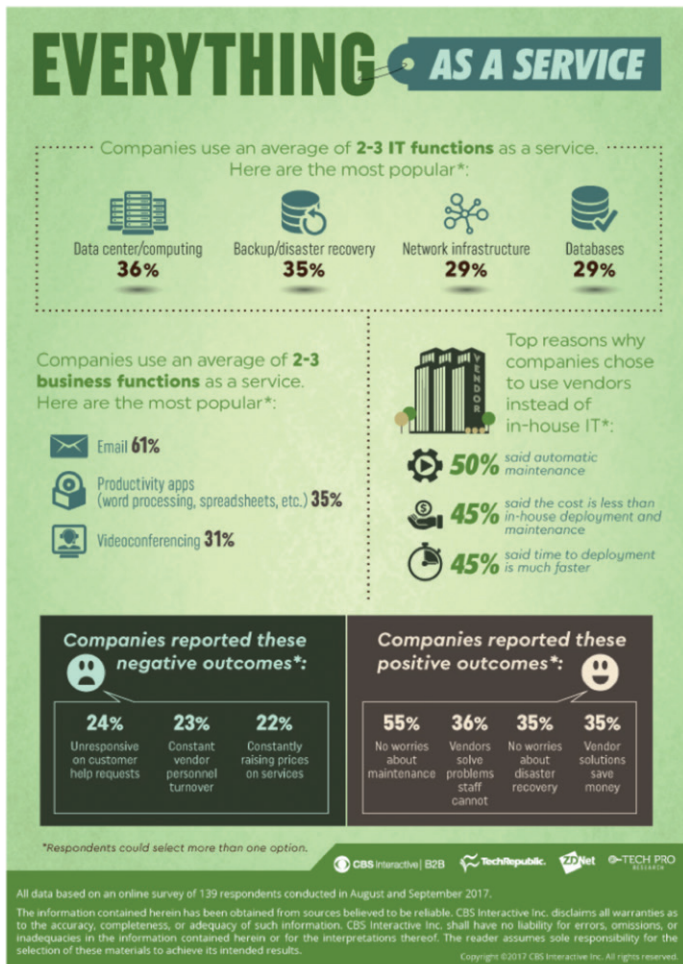


Figure 112: A business-oriented infographic on the “everything-as-a-service” trend in the IT market
Source: CBS Interactive Inc.

Another major implication of the rise of the *as-a-service* style of software provisioning is that this modality is attractive to all devices – and all applications – that are intended for connectivity, for user-side reachability, system-side operation or both. As the client-side device has sufficient computational capabilities, it becomes practical and desirable to transfer part of the overall computation to it, with the aim of reducing the response-time latency that would be incurred if the bulk of computation were to occur on the server side.

Taking this observation in account together with the second characteristic itemized at the start of this section (the desire to cooperate at the highest level of the protocol stack), we see that “the client” and “the server” are increasingly less self-contained monoliths, destined to fully and permanently reside at one place. Instead, they are increasingly becoming *aggregates of distributed components*, which may not even need to have a fixed physical residence and can instead move around (also known as roaming).

This architecture concept gives rise to the notion of “**the continuum of computing**”, which allows a single software system to be comprised of interconnected parts that span across the edge, the fog and the cloud, and that may individually, transparently and

Computing Continuum

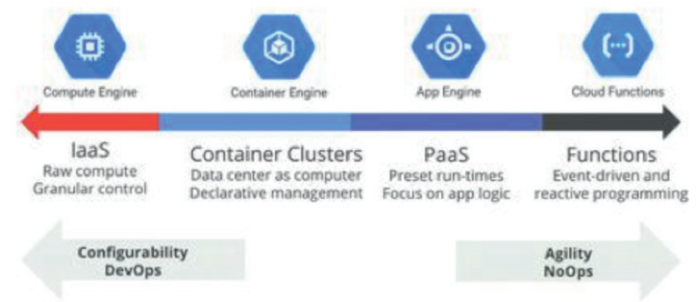


Figure 113: A pictorial Cloud-centric view of the continuum of computing
Source: [136]

dynamically change location and configuration of interconnect, for performance, fault tolerance, and security reasons.

The components of software applications designed in accord with this vision (which can be called *services*, or better yet *microservices*, to stress that they are intrinsically other than traditional monoliths disguised within containers) have a number of relevant characteristics:

To earn maximum reachability (that is, to be able to potentially span all nodes of the internet and, within them, to reach out to any place in their local storage and to any service attached to it), the interfaces they expose to the outside are programmed against HTTP(S), rather than against arbitrary APIs. Note that this span cannot be achieved with TCP-level solutions, which have no uniform way to reach out to local-storage places owing to the limited span of port-based services.

The state-of-the-art paradigm for that purpose is REST, short for “representational state transfer”. The primary intent of REST is to “transfer, access, and manipulate textual data representations in a stateless manner” [42].

Elevating information into a first-class element of software architectures allows decoupling (i.e. the outcome of drawing value from data) and processing (i.e. the act of doing something with information, typically producing more data for more information processing down the line). When realized correctly, RESTfulness is an “in-the-large” architectural style that provides for **uniform interoperability** between different services (as part of applications) or whole applications on the internet. In addition to relying on textual representation, this interoperability descends from **statelessness**, which allows application services to communicate agnostically and therefore be able to accommodate **heterogeneity**.

Statelessness is also a prerequisite to scalability; its opposite, statefulness, instead nails the software asset where its state is persisted, which cannot be copied elsewhere, short of slowing execution down dramatically in order to assure transactional

consistency. It is worth noting that horizontal scaling (scaling out) has more strategic value in future systems, as it lives on roaming across existing execution platforms, without incurring the total cost of ownership and the added complexity of having to scale up to more powerful dedicated infrastructures.

RESTful interaction is very simple to comprehend, much simpler than with arbitrary APIs, which is a good omen for verification. Admittedly however, using a simple style of interaction to build complex systems needs a very profound understanding of how to break a complex whole down into simpler parts that can be individually mapped to RESTful simplicity. Later in this section we return to what should be done to elevate RESTful interaction to the needs we envision in this document.

2.5.1.1 OPEN SOFTWARE ARCHITECTURE

Over the last 20 years or so, the evolution of the architecture of large software applications (large in terms of feature set, size and volume of users) has followed a course that – in retrospect – appears very evident. In fact, software architectures have progressed as follows:

- 1 From single-location monoliths, with variable room for modular and recurrent organization, to
- 2 TCP-level static client-server pairs (now nearly extinct, due to the limitations and lack of flexibility explained above), to
- 3 HTTP-level *pull-based* static client-server aggregates, to
- 4 HTTP-level *push-based*, publish-subscribe dynamic compositions.

Figure 114 gives a pictorial representation of this course of evolution.

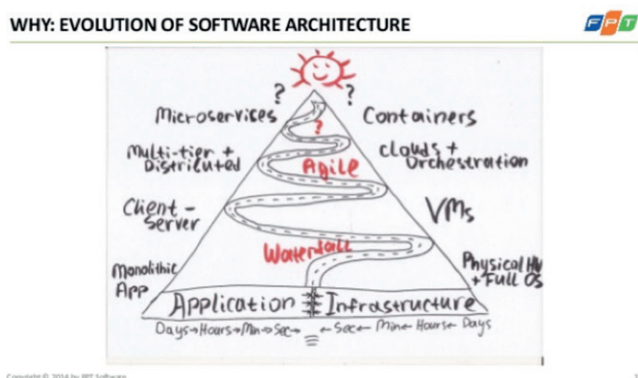


Figure 114: A pictorial view of the evolution of software architectures – Sourced: FPT Software

The step from 1 to 2. responded to the massive demand for more uniformity, reachability and interoperability, but it did not correspond to efficient, lean, modular, evolvable, and robust architectures. Very few systems were internally reorganized to

better fit the new architecture, because in software production, conservatism tends to prevail over decisive change, leading to effects that becomes particularly noticeable in systems that outlive their anticipated lifespan and therefore are not designed for long-term evolution. Those applications frequently become “the land of legacy”, where evolutive maintenance is rarer (because costlier) than corrective and adaptive maintenance.

The step from 3 to 4 responded to the advent of software-defined social networks, whose information flow was fully inverse to the request-reply paradigm ingrained in 3. Interestingly, however, in spite of this fundamental difference, 4 was built on top of 3, as a token of pragmatic conservatism (hence the retention) of the pre-existing network infrastructure. Yet the software architecture of the client and server sides in 4-type systems had to undergo sufficiently large changes to motivate the use and the furthering of pattern-based architectural design. This trend is being accompanied by growing attention to *software frameworks* – which embody precise architectural styles – in preference to simple libraries, which may have no conscious architecture underneath their API.

4-type architectures, which have a publish-subscribe core at their centre, are intrinsically more open than their predecessors as they allow the transparent addition or removal of end-points (the former requiring advertising to potential users; the latter requirement client-side handling of access failure), as well as transparent replication or relocation of components, without any modifications to the rest of the system.

All the types of architecture cited above in some ways are variants of the same client-server paradigm. However, the advent of the blockchain concept and technology has shown the new frontier of *decentralization*, where an arbitrarily complex system can be constructed without requiring server components as well as without fully connected distribution, in which all nodes are logically connected among them. Lately, technology assets are beginning to emerge that help construct full-stack blockchain-based systems, which preserve the web-based nature of 3- and 4-type systems.

TOOLS FOR BUILDING DECENTRALIZED SYSTEMS

DappRadar [277] offers a snapshot of existing and advertised decentralized applications. Technology-wise, Ethereum [355], with its MetaMask browser adapter [313], and the InterPlanetary File System (IPFS) protocol [305], are enablers worthy of mention. Here, smart contracts are the service methods exposed to the application in the same way as normal web applications, front-ended with React or some such, and, thanks to IPFS, storage is either on chain (where secure persistency incurs variable costs), or off chain, depending on mutability needs and cost considerations.

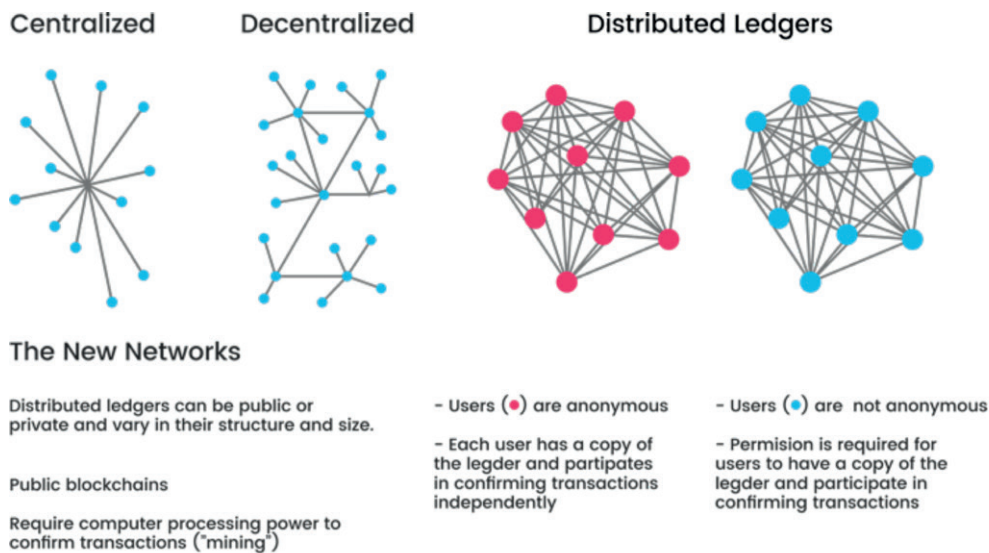


Figure 115: A pictorial view of how blockchain's decentralization relates to other forms of networked organizations

Recently, decentralization has gained traction and interest in the numerous application domains (digital conservation most notable among them), where trusted persistence is paramount, and doing away with the costs of traditional centralization solutions is attractive.

There are, however, two main limitations to blockchain technology, with further study needed to overcome them. First, the energy costs of persisting securely (what came to be known as "mining" in the bitcoin world) need to be drastically reduced, which may be done using more elegant and sophisticated solutions like those of Algorand [166]. Second, scalability needs to be ensured, possibly by federating private and public resources, in all of three dimensions of the so-called "scale cube" [128], that is, functional replication, functional composition, and data sharding, without breaking the principle of decentralization.

2.5.1.2 SOFTWARE COMPOSITION

The pursuit of software composition acknowledges the attractiveness of constructing software systems as aggregates of pre-existing components, hence in a *bottom-up* fashion. This notion contrasts with waterfall-style top-down development, which is losing traction in current practice owing to the combination of a number of phenomena, including the increasing availability of potentially reusable software assets and its poor disposition for compressed time to market.

Software composition at its fullest should embrace heterogeneity, striving to use software technologies for what they are good at, in a context where they fit. Heterogeneity used to be a complex challenge for software development, enjoying limited support from programming languages, as supporting interoperability may be arduous and is more likely to offer little return on investment than to be a marketable asset.

By targeting a single runtime model, compiled programming languages indeed have a hard time contemplating lasting interoperability solutions towards other languages that evolve independently and sometimes also very rapidly. Scripting languages, which do not have a runtime model of their own, are much more versatile in that respect (which adds to their attractiveness for compositional development), but at the cost of offering no guarantees on program semantics (which is an extremely serious shortcoming with regard to reliable systems).

The trend that has emerged to address these challenges is comprised of two complementary principles: *containerization* and *microservices*. Containerization, which was initially born as a lightweight alternative to virtual machines for seeking resource isolation, has more recently joined the microservices architecture paradigm, forming a formidable enabler of modern, heterogeneous, software composition.

The "microservices" architectural style yields a single application from the coordination of a suite of unitary services [132], each of which exposes an application programming interface (API) *outside* of their codebase (central to the composition style), which the user invokes using *asynchronous* (crucial to loose coupling) *web-based* service requests (key to reachability).

An individual microservice is a small self-contained application that has a single responsibility (which gives it a clear and distinct role in a composition), a fully-self-contained and preferably lightweight stack (which allows its software dependencies to be always fully satisfied), and can be deployed, scaled and tested independently (which facilitates software evolution) [91]. In fact, at the present state of the art, these attractive traits can only be achieved with containerization. A RESTful software architecture, based on publish-subscribe aggregation of containerized microservices is the most natural evolution of iv.-type architectures, embodying characteristics that respond very well to the challenges discussed in this section.

In this section we have discussed two high-level principles that should guide software composition for the system architectures of the future. First, designing software to be provided *as-a-service*; second, designing software for inter-operation at the highest level of the internet protocol stack. The combination of these two principles – each of which has numerous ramifications – allow for reachability, openness, flexibility, mobility, agility, heterogeneity. We have seen how those principles imply the use of containerization and RESTfulness.

In order to explore this direction of evolution further and use it to develop systems that in addition to being modern, open, heterogeneous, interoperable, evolvable, are also **reliable**, however, requires the addressing of two crucial needs that are currently *not* satisfied by state-of-the-art technology:

Developing solutions to specify *programmatically* – and not solely declaratively – and execute the orchestration of the individual parts and of selected aggregates of containerized microservices.

Orchestration specifies the lifecycle (deployment, scaling, upgrade, retirement) of the individual parts and the logical interconnect among them (binding between the in/out ports of the parts). Orchestration technology must assure that all lifecycle operations on individual parts can be kept transparent to the other side of the interconnect.

Augmenting the (REST) APIs of individual components with *non-functional* contracts, so that component binding not only responds to functional needs but also allows assessment and assurance of the transitive satisfaction of all assume-guarantee pairs stipulated at the point of binding.

2.5.2 SOFTWARE IMPLEMENTATION: THE LIMITATIONS OF TRADITIONAL PROGRAMMING

2.5.2.1 CONTEXT AND INTRODUCTION

In this section, we discuss requirements that consolidated programming technology does *not* address satisfactorily at the present state of the art. In the subsequent section, we make recommendations on ways to reinvent programming for a new era of this discipline.

The areas in which the main limitations:

- assuring correctness, both functional and non-functional, especially the latter, which is largely neglected in almost all programming languages;
- longevity, with the ability to accommodate legacy, reuse, adaptation and evolution;
- predictability, safety and security, efficiency (with the issue of compiler-mediated optimal use of hardware and its negative impact on portability).

We discuss each of these shortcomings in dedicated clauses.

We also elaborate on how erroneous (short-sighted) the current interpretation of productivity is, and what we should do to revisit it.

2.5.2.2 TECHNOLOGY TRENDS

Before discussing the future of programming as we envision it, a few observations are in order on how the pace of application development relates with the relative intensity of use of specific programming languages and technologies.

One trait that can be discerned very clearly in software production strategies is the desire to use *as few technologies as possible* for building the full application, often trading how well a technology fits and/or how long-lasting it is for shorter time to market. Arguably in fact, the race to positioning software products in the market, both commercial and free, as soon as possible, without paying sufficient attention to their quality attributes is becoming so dominant as to give rise to growing concerns.

What is sorely missing, evidently, is a “*culture of quality*” that can help the customer, private or institutional alike, to *demand* quality and to be able to appreciate the presence or the absence of it. History shows that public bodies have a decisive role in instigating quality-oriented production policies. Entities with that capacity are largely absent when it comes to software products.

In the case of the web-enabled applications discussed above, for example, the attitude of taking a superficial view of productivity (where *faster* obscures *better*) has prompted programming environments to emerge that enable and favour the use of client-side asynchronous programming languages for the server side too, regardless of the difference in the respective architecture needs. Whereas the cover plate of those programming environments warns users that they are intended for fast prototyping, the current economy has very few users able to scratch the surface and tell a prototype apart from a solid long-term solution.

One particularly serious consequence is that numerous software products (applications, libraries, utilities) increasingly often exceed their span and scope of use, and are spread through opportunistic use. The 2018 Top Ten Programming Languages ranking recently published by IEEE Spectrum (Figure 116) is quite revealing of that trend. Indeed, it shows that, in terms of (public) use-based popularity, so called public-oriented, “easy to use” newer languages either have overcome (in the case of Python) or are about to overcome (in the case of PHP and JavaScript) older more established infrastructure-oriented, enterprise-solid languages such as C, C++, Java and others.

The ranking reported in Figure 116 only reflects public (e.g. posted on GitHub), measurable use, which cannot be directly transposed

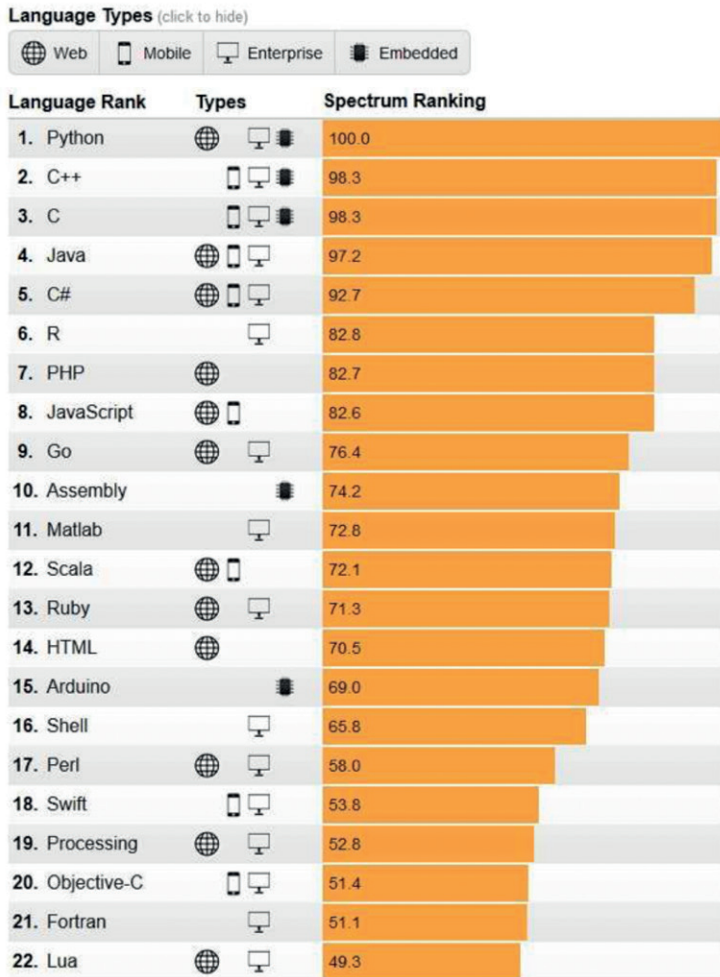


Figure 116: The IEEE 2018 programming language survey shows Python topping C and C++
 Source: The IEEE 2018 programming language survey

to enterprise trends, but certainly has profound implications on the programming skills on offer and therefore indirect influence on enterprise orientation.

One reason behind this phenomenon, on the software production side, is that professional programmers capable of mastering *multiple* languages with equal aptitude are a very small contingent, too small to meet market-driven productivity requirements (incidentally, this scarcity of available expertise, combined with the technical and cultural difficulty of interoperability among programming languages, explains why few organizations see this trait as a requirement). In general, therefore, once an enterprise – or indeed a group, project or lab – has acquired a software programmer sufficiently versed in a programming language in high demand at the time, such a language becomes the “hammer” and *all* types of programs become “nails”, to borrow Mark Twain’s famous aphorism. In this way, programmers condemn themselves to a single-tool development situation: “I program it in X because X is what I know best.”

It is worth noting that the “demand” of a programming language also negatively correlates with the complications of interoperability. In other words, a language that has gained use in a spotlight application sector will attract the creation of libraries and utilities written in and for it, which will increase geometrically the use factor of that language and therefore the “demand” for it. This is very evidently behind the massive rise in the use of Python for machine learning applications and beyond, for instance.

2.5.2.3 THE OVERARCHING CHALLENGE: MASTERING COMPLEXITY

Despite all attempts to master it, software complexity continues to grow, and defies our understanding of the systems that we design and use. An increasing number of systems are already regarded as *no longer completely understandable* [3]. This situation, which shows no sign of abating, constitutes a new software crisis.

When Dennard scaling stopped, processor systems became tightly-interconnected multi-core (exposing parallelism with and without concurrency), and fitted an increasing number of accelerators (exposing heterogeneity), which then in turn were aggregated in variably deployable units (exposing statelessness), and networked (exposing geographical distribution and decentralization), for access via the web (exposing asynchrony). Programmers and programming models are struggling to adjust to all of these vectors of evolution.

In addition, with the expanding pervasiveness of the use of computer systems in virtually every aspect of our daily life, the production side of the IT community is faced with additional complexity factors — energy, time and other resource constraints, ever-advanced human-computer interaction, the weaving of cyberspace into physical reality, continuous delivery within continuous operation — that further deepen the complexity of programming.

While mainstream programming languages incorporate abstractions for data and control, capabilities that distinctly matter in contemporary and even more in future information systems lag behind. For example, parallelism at the application level still has to be expressed explicitly, that is, in fine-grained, low-level programmatic detail, since most programming languages lack adequate facilities to specify and stipulate actions that have to occur in parallel, such as sensing and control of the physical world, at higher levels of abstraction. Even less prevalent is the ability to attach non-functional properties, for example in the areas of power, energy or time, to units of execution.



The variety of programming languages currently in use, although helping to address specific issues and therefore master complexity at some level, also introduces an extra complexity factor in itself. As noted earlier, no programmer can be fully proficient in *all* programming languages, and the average programmer has difficulties mastering well just one or a very few of them. Yet for reasons that often pertain more to the pragmatic preservation of legacy than to the search for proper fit, no single programming language can be conceivably expected to sweep all others away. Accordingly, as use paradigms and deployment opportunities evolve, software systems are increasingly comprised of multiple heterogeneous components, written in various languages, reused and glued together, often in distributed aggregates.

The componentization solutions that we discussed earlier, notably container technology, offers a way to address this rising complexity, using modular assets to encapsulate local complexity and hide it from the outside. Whereas this solution has proven effective for encapsulation, however, it has expanded the issue of complexity to encompass the challenge of creating sound, trusted, reliable and fully interoperable assemblies of components.

Figure 17: the maze of currently hyped programming languages
 Source: Patrick O'Neill, 'The Most In-Demand Programming Languages', 2018

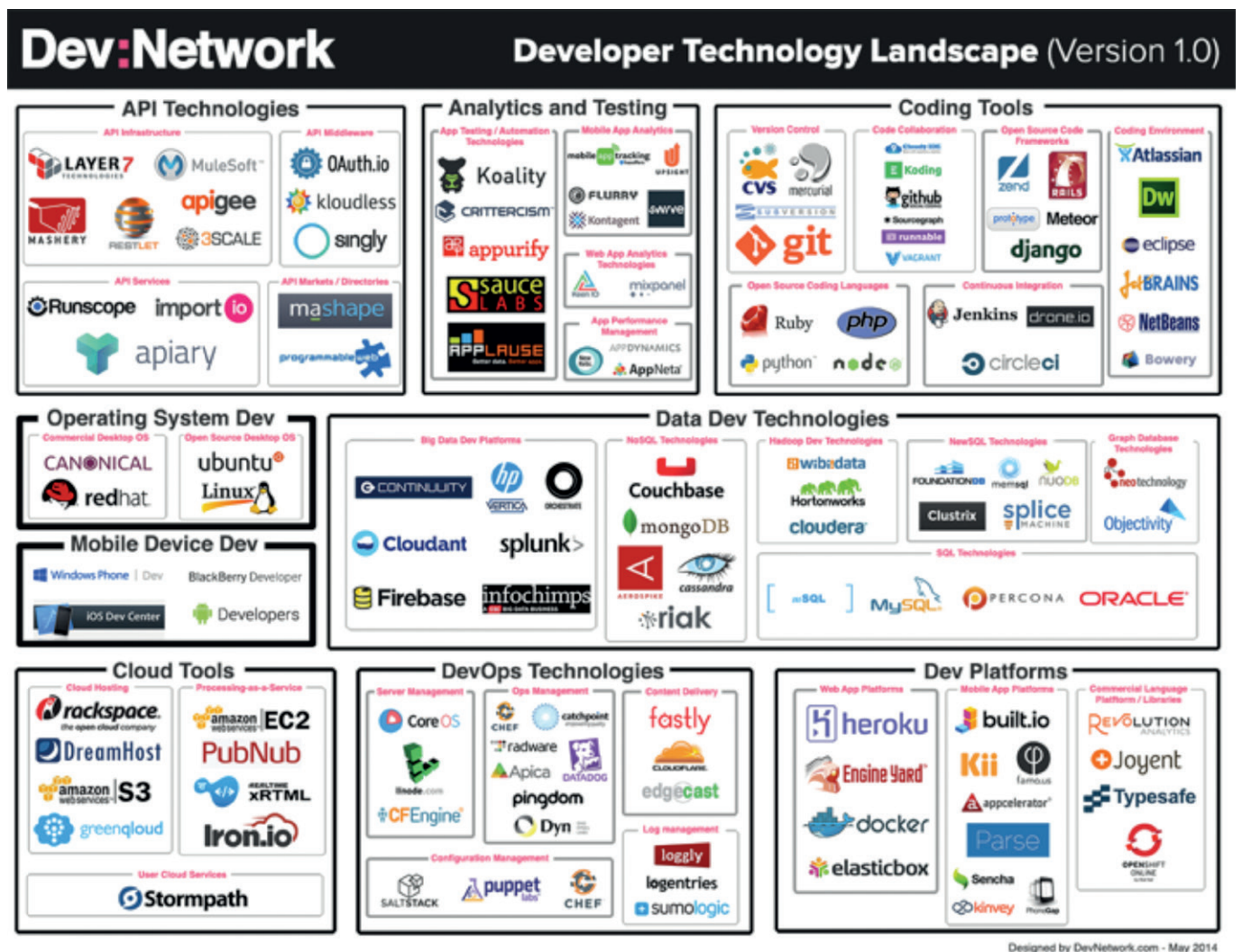


Figure 18: A sample of the developer's technology landscape
 Source: DevNetworks Sought enhancements: asserting correctness

When software components are written in a single language and follow one and the same convention, the interfaces which become standard are therefore pretty simple. With a large, possibly huge, number of components of various origins, written in various languages, with various coding styles and often non-matching conventions, instead, the complexity of their respective interfaces becomes a critical hurdle. It is this interface complexity that must be addressed, with solutions that are language-independent, hence fully interoperable, low-overhead, hence resource efficient, and supportive of build-time verification as well as of run-time enforcement. In section 2.5.3 “Software implementation: time to reinvent programming”, we discuss how we envision this challenge should be addressed.

2.5.2.4 SOUGHT ENHANCEMENTS: ASSERTING CORRECTNESS

Although it has always been a paramount aspect of programming, correctness – the assurance that that a computing system exhibits the specified behaviour, both functional and non-functional – is still relatively poorly attended to and mastered concern.

Indeed, business constraints (time to market, cost of production) and the programmers’ mindset have generally focused on delivering functionalities to customers, since this is what sells and what is perceived as the creative act. Integral correctness is rarely pursued by design; more often it is sought as a product of quality assurance activities, either performed retrospectively or in parallel to development, but not sufficiently ingrained in it.

In the general market, integral correctness is not a visible trait, and its presence earns no distinction, success, or fame. Not surprisingly, therefore, efforts aimed at supporting and achieving correctness have been scarce. While some enterprises do specialize in providing tools that help seek correctness, their success has never even remotely approached that of organizations providing functionalities to the end user, such as the likes of Facebook or Twitter.

In addition, the traditional approach to seeking correctness suffers from an initial flaw. Indeed, correctness has traditionally *solely* focused on *functional* concerns, that is, that the program performs the actions it is supposed to. Other aspects of correctness, now collectively captured under the umbrella term “non-functional properties”, instead have been almost regularly neglected. In the general acceptance, non-functional properties include timing, power/energy consumption, and other resource usage, security, and safety. This omission is so imperceptible in the notion of most business actors that, even if the system fails to meet some of its non-functional requirements while carrying out its intended function – hence being slow, late, wasteful, or leaking memory or energy – it is still (deludedly) regarded as correct.

In certain domain-specific, demanding markets, such as critical embedded systems (for example aeronautics, nuclear), which live in a quality-aware culture where correctness has always been seen as paramount, this omission does not occur. In the systems developed for those domains, correctness by design is normally sought and methods that help achieve it are devised, with active support from research (see, for example: [215]).

The products of these development practices are normally so trustworthy that they are intended for deployment in services and infrastructures that have impact on people’s lives and wellbeing. Yet, they are less understood, more prosaic and less fashionable than general-purpose apps, which makes them much harder, costlier and less attractive to imitate, unable to capture the imagination of the general public.

Not surprisingly, therefore, the mainstream of computing system development is not in those privileged domains. The bulk of new and trend-setting applications is for mobile communications, including social networks, and for their superficial manifestation in CPS and IoT systems (predominantly via the user-oriented utilities of new-generation automotive), and more occasionally in smart-everything-everywhere contexts. These reach billions of people every day and impact the surface of our daily lives far more obviously and quickly than the former kind.

Sadly, in the latter type of systems, non-functional properties are for the most part neglected – due to a lack of good role models to set authoritative trends, and strict-enough assessors – or just injected as an *afterthought*, owing to the emergence of evident and intolerable flaws (as we have seen recently in the case of security). The potentially negative impact of this situation is huge, for loss of value, increase of risk, and spread of threats, and should be acted upon with a more vigorous quest for quality.

There is no doubt that, beneath the surface of user-driven applications, the hardware and software infrastructure of new-generation cognitive CPS will have to confront unprecedented demands of correctness under very strong constraints of competition, economy and market pressure. To address those needs adequately, new solutions will have to be devised that allow to achieve the assurance of correctness without imposing unsustainable rigidity and slowness to the development process.

2.5.2.5 SOUGHT ENHANCEMENTS: ACCOMMODATING LEGACY, REUSABILITY AND EVOLUTION

IT is a domain in which systems evolve constantly, and do so at a very rapid pace, especially for software. Indeed the immaterial aspect of software (bits of information) makes it possible to update software very easily as compared to hardware. This becomes even easier when remote updates become possible, thanks to connectivity, without anything or anyone having to physically move.

The same immateriality of software makes it also more amenable to reuse, as part of new software parts, either through extension or duplication.

However, managing software evolution and the reuse of existing software parts, which involves the integration of legacy software components into newer programs and systems, are increasingly complex challenges that need innovative solutions to be mastered effectively. A lot of the outcome depends on the developer’s ability to understand the parts that are being reused or evolved. Having helpful and above all up-to-date, documentation, throughout development and maintenance, is certainly essential, but it is not sufficient. Yet, even this very basic need is frequently neglected, because documentation and program source are often seen and treated as separate and disjointed artefacts.

It is true that programming languages have for a long time allowed comments to be embedded in the program sources; some of them, like Java, putting particular emphasis on the automatic generation of program documentation from source code. However, it is well known that developers tend to be reluctant to put effort into comments that provide no additional program functionality for themselves. Other example-setting, niche languages such as Eiffel (<https://www.eiffel.com/>), Ada (<https://www.ada2012.org/>), and SPARK (<https://www.adacore.com/sparkpro>) embed *design-by-contract* information in the source code, as preconditions, post-conditions and other assertions that help build, debug, and document the source code, while also involving various extensions of run-time semantics, to aid programmers to keep this information up-to-date.



Figure 119: A word-cloud that evokes the prominence of the notion of design by contract
 Source: 123RF

A flurry of complementary specialized tools helps statically analyse software programs and extract structure information from them, generating structured diagrams (in UML or other fitting formalisms) in a semi-automated way. Those tools however are good for comparatively small software units, and their usefulness degrades as the program size grows, which leaves unsolved the problem of maintaining large software aggregates.

Advanced visualization of software through metaphors is offered by some tools, which may be better fit at providing a quick understanding of very large legacy software [155]. However, much remains to be done towards assuring *all* of the expected qualities of large software assemblies, throughout building, debugging and execution.

Attaching richer semantics to the interfaces of software modules, beyond functional APIs and state-of-the-art contracts, which for the most part continue to focus on function, promises to be of great help in mastering the complexity of software integration. Yet, pursuing this vision requires devising solutions that do not explode the complexity of compilers and do not oppress runtimes, while providing the desired assurance.

It must be noted, however, that the premises of agile development (partial releases, frequent increment) are antagonistic to design-by-contract practices, which – at the current state of the art – require *all* interface contracts to be fully defined before verification activities may start on their binding. This requirement marries well with a top-down style of development, but is ill-fit for continuous integration, which is the agile connotation of most modern systems. Research on this topic should devise solutions that support reasoning and assurance making on incomplete (contract) specifications

2.5.2.6 SOUGHT ENHANCEMENTS: SECURITY, RESILIENCE, TRUST

As already mentioned in Section 2.3.1.1, software should be trustable and secure. A major issue with software development is that developers scarcely appreciate that, without taking adequate precautions, their programs may be insecure, and thus not trustable. This attitude reflects the fact that, often enough, security is often neglected; it is not a prime requirement, and is rarely tested against. This deficiency is not helped by the fact that many programming languages tolerate sloppy programming, where code that looks reasonable at first sight may in fact contain major vulnerabilities. (For a comprehensive summary of programming language vulnerabilities, see [81].)

Different solutions are possible to help developers create more secure software. One example is to create programming languages that enforce (more) secure coding patterns, and that cause developers to make their assumptions on unsafe parts of the code more explicit. For example, looking at recent endeavours, the Rust programming language, originally developed by Mozilla,

makes many aspects of memory management and memory safety explicit in its language constructs.

Another approach is to prove the correctness and security of the developed programs with programming languages such as Coq and Ada's SPARK. While taking correctness into account during development may seem to incur a high overhead, it is worth bearing in mind that the cost/benefit ratio should extend beyond produce release to include maintenance. Outages further down the line may be massively damaging to critical systems and infrastructure.

Recent research-level successes, which include the verified parts of the seL4 microkernel (<https://sel4.systems/>) and the CompCert compiler (<http://compcert.inria.fr/>), prove this point well. More large-scale studies should be encouraged that consolidate this quantification and invert the commonly held belief that *quicker may obscure better*.

Other notable examples in the line of program verification include Frama-C (<https://frama-c.com/features.html>) and SPARK's Discovery toolset (<https://www.adacore.com/sparkpro>): those tools operate on the premise that the source code must conform to some formal specification. This notion of conformance is essential and prerequisite to software products that are categorized critical from the outset, but it is unheard-of in the vast majority of the existing software base, some of which does begin to approach critical use (e.g. in the millions of lines of code embedded in self-driving cars, where assurance is more sought by isolation than by assurance). With current-generation tools, such specifications concentrate on functional traits. Future work will expand the capability set to non-functional concerns.

Of course, as long as security vulnerabilities exist, customers expect these to be fixed. First of all, this means that manufacturers now have a burden to keep their software secure and up-to-date long after they started (and maybe even stopped) shipping it. Furthermore, another aspect is that, once such a security vulnerability has been found and fixed by the software vendor, the patch needs to be delivered to the users, and the users need to install it. In order for such updates to be delivered to the users, they have to be secure, that is to say, authenticated to come from the original software vendor, in such a way that they cannot have been tampered with. Otherwise, attackers could latch on to the software update mechanism, and substitute their own, malicious updates. A similar concept is that of code signing, where the operating system tries to ensure that only software that is signed by known and trusted software vendors, can get installed.

2.5.2.7 Sought enhancements: predictability, safety, and conformance with specifications

The proportion of command-and-control software infrastructures is rapidly expanding beyond its more traditional domains of application, enveloping industrial plants, transport and service

networks, as well as other commodities. Regardless of the differences in ambit of use, those infrastructures have (at least) two distinguishing traits in common: they have to constantly acquire possibly large amounts of data from an increasing variety of sources, and they have to draw intelligence from them in order to decide time-bounded actuation operations.

The increase in the type, quantity, throughput, and heterogeneity of the data sources, and in the computational intensiveness of the intelligence-gathering algorithms that have to be run on them present unprecedented challenges for non-functional requirements, which the current software production practices are scarcely prepared to face. Those impending requirements concern:

- rising demands for time-predictable execution behaviour: what used to be a very specialized and niche trait of real-time systems, now becomes a common need, transversal to data sensing, data fusion and algorithmic computation, revealing fundamental shortcomings in programming language notions, constructs and capabilities to address execution-time behaviour as a first-class citizen;
- similarly critical needs for the assurance of safe behaviour, in the face of missing data, late or erroneous computation, hardware or mechanical failures. The traditional approach to dealing with these needs was such that the more stringent the requirements, the more restricted the programming capabilities, with the aim of reducing the complexity of designing, implementing and verifying the contingency strategies. This simplification conflicts with the nature and needs of the emerging systems, which will therefore require programming capabilities much beyond what current technology can do in this regard.

Interestingly, the increase in importance of such non-functional concerns equally raises the importance of specifications against which conformance can be ascertained, and without which nothing final can be said about the fitness of the system.

The solution that we have envisioned in this particular regard is to expand the expressive power of interface contracts and the support for them, so that they can become the place where non-functional requirements are specified, checked for soundness, assured at build and deployment time, and preserved during execution.

2.5.2.8 SOUGHT ENHANCEMENTS: BALANCING EFFICIENCY AND PERFORMANCE WITH PORTABILITY

With computer platforms becoming increasingly heterogeneous, there will be increasing tension between the quest for optimization and the preservation of portability. The former makes software code tightly coupled to the specificities of the hardware target, and its effectiveness strictly depends on the particularity of the adaptations. The latter aims to preserve the

capital investment in the code development in the face of mobility, which requires different deployment (preferably without re-compilation), or evolution.

The optimization of compilers for large and feature-rich programming languages is a very complex and costly endeavour, which speaks against diversification and suggests convergence to a few, common and open-source back-ends where target-specific vertical optimisations can be concentrated and benefit multiple language frontends by improvements, enhancements and feedback from use. This trend is silently happening, but more as a matter of pragmatism (where bootstrapping a programming language on a hardware target is seen as vastly more complex, costly and risky than piggybacking on an existing language base) than as an organized coherent front. A lot more can be done and achieved in this respect thanks to plug-and-play compilation systems such as LLVM.

Solutions to this end, which go beyond purely technical challenges, need to be investigated.

In complement to this, the adoption of common runtimes for interpreted languages can benefit the return on investment in their optimisation and the maturation of the corresponding codebase.

2.5.2.9 SOUGHT ENHANCEMENTS: INCREASING PRODUCTIVITY FOR FASTER, CHEAPER AND BETTER PRODUCTS

As noted earlier in this section, in relation to software development, arguably the dominant notion of productivity means that “faster” is obscuring “better” in too many areas.

As software infrastructures permeate and sustain different aspects of our professional, social and even personal lives, agility in the development, operation and feedback-based maintenance cycle – whose acknowledgement has given rise to the DevOps movement – is going to increase, pushing the “faster” dimension even more. The rising criticality in most aspects of those software assets will however require returning attention to their overall quality attributes, including the whole spectrum of non-functional requirements that we discussed above.

Over the last decade (if not more), a lot of energy has been deployed to serve the “faster” side of the challenge. This has resulted in productivity enhancements (an increasing number of libraries covering various needs with sufficient recurrence to draw attention, “intelligent” program editors, etc.), being devised in programming language environments.

This growth however has not been accompanied by a proportionate rise in the support for the “better” quality of the software product. The ultimate response to this situation is to reinvent programming, as we discuss next.

2.5.3 SOFTWARE IMPLEMENTATION: TIME TO REINVENT PROGRAMMING

2.5.3.1 INTRODUCTION

Sadly, the limitations of current programming are numerous and are having a major impact. Although continuous progress is being made in various domains and traits of programming, we believe we are to some extent reaching dead-ends, a swan song from an overall perspective, where software no longer is a solution, but a part of the problem. Extricating programming from these dead ends requires a radical change of perspective and mindset. It is time to reinvent programming. In the following, we discuss what features programming languages and their runtimes should support.

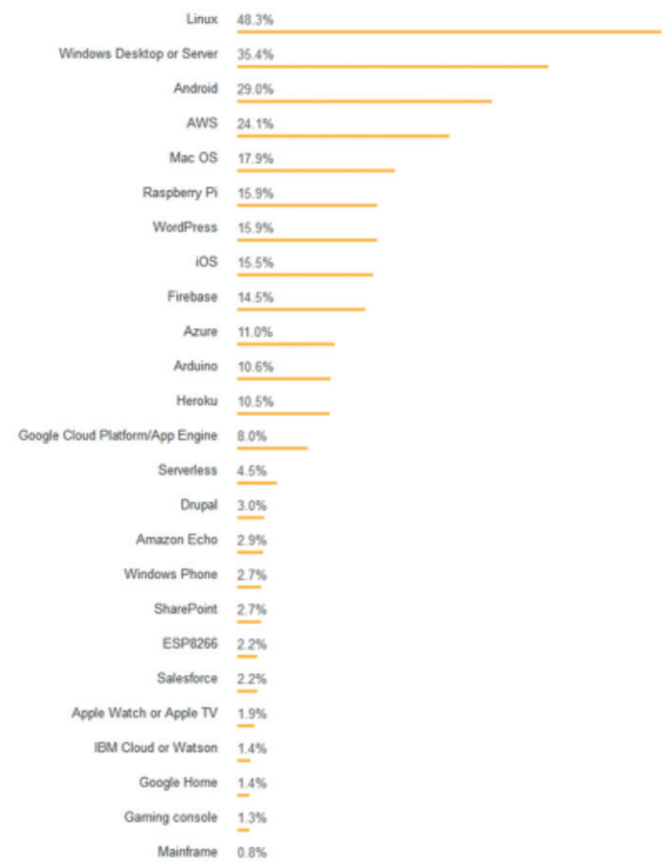


Figure 120: A large variety of deployment platforms exists, which shows no sign of shrinking. Pressure must be exercised on the makers of those platforms for them to be inclusive, and technical solutions must be invented to ease trusted mobility among them. Source: Stack Overflow, Developer Survey Results 2018

Before delving in the particulars, some general observations are in order, which we draw and elaborate from [331]:

- Finding trusted programmers capable of handling security concerns satisfactorily will be a critical challenge. The same can be said of several other non-functional properties. One radically-different way to seek a solution to this problem looks into computer-aided programming, where (a comparatively

small number of) trusted computer programs produce (an infinitely large number of) high-quality programs, elevating the human role to producing declarative specifications of needs, wishes, preferences, and constraints instead of source lines of code in any particular syntax.



Figure 121: one way of highlighting the human factor in software programming

Source: Stack Overflow, Developer Survey Results 2018

- Public pressure works for creating inclusive technology environments. Numerous deployment platforms exist (as shown in Figure 120).
- Instead of dreaming of a single solution for all needs (in other words, undesirable monopoly), it is more opportune to devise technical solutions to ease source- or object-level mobility across them. As noted in [370], several languages – e.g., Groovy, Scala, Clojure, Kotlin, etc. – exist that run on the JVM, but there is only one JVM. By the same token, one can run many languages on .Net’s VM as well. This shows that the JVM – and by extension .Net – is a very convenient base to build upon, achieving at one time robustness, interoperability, and portability.
- At long last, pressure is rising on programming languages for them to adopt modularity (see for example Java and JavaScript efforts at refining their support for modules). This is essential to constructing more robust software, but it only one step if the long road to system-level modularity, which needs componentization and containerization to become programmable too.

2.5.3.2 NON-FUNCTIONAL PROPERTIES AS FIRST CLASS CITIZENS

An urgent ingredient in the process of reinventing programming is to devise ways to address all aspects of a computing system *together* in the act of programming, contemplating functional and non-functional properties simultaneously, from the outset, at the same level of prominence. Programmers and support tools should become to express, manipulate, and reason about non-functional properties, to make runtime decisions based on them, to yield static proofs of correctness, to support runtime assertions to check that the necessary properties hold during execution, and adequate semantics to handle violations so that safety conditions are restored. Prominent non-functional properties that need special attention for new-generation cognitive CPS are timing and reactivity, power and energy, security, safety.

Indeed, CPS have evident reactivity requirements. Yet, reactive programs for the most part still rely on low-level techniques such as call-back functions and explicit task handling. Developing higher-level abstractions for reactive systems with explicit non-functional properties will improve productivity and scalability, paving the way for higher-level, higher-impact resource optimization.

CPS also have timing constraints, which include completing units of execution within a certain time interval or deadline, but also, for example changing semantics based on events. Time must thus be a first-class citizen in the programming languages destined for CPS, which promote coding styles that facilitate (worst-case) execution-time analysis, and optimization.

The continuous contact of CPS with external systems places a heavy obligation for safety and security on the programming languages and tools used to implement them. Part of the system must be able to continue running in partial or full isolation. The system must be able to detect intrusion (attempts), and take countermeasures to guarantee safety and security. This hinges on the hardly investigated semantic properties of programming language constructs in terms of safety and security. There again, abstractions of security and safety must be integral parts of the languages and design methods.

CPS have power, energy and even for some of them thermal constraints, as part of their requirements. Energy is thus another physical dimension that must be visible as a first-class non-functional property in programming languages and tools to allow the programmer to design energy-efficient systems. Indeed, the energy awareness of such systems is crucial, and most of the time has to be dynamic, so that the system can react and adapt to the changing environment. Many CPS are autonomous, relying on batteries with limited autonomy and peak power. Some are able to harvest energy from their environment, to extend their lifespan. Correct modelling and explicit handling of all these aspects is necessary to ensure appropriate operation of such systems.

Research should thus be encouraged to devise new programming concepts, styles, methods and tools that help capture non-functional needs – most notably time, power, energy, safety, security, and privacy – conveniently and aid their assurance at build and execution in manners that do not hinder agility.

2.5.3.3 BETTER ABSTRACTION AT BOUNDARIES

Software components and containers, enhanced with interface contracts that express assume-guarantee pairs on the functional and non-functional behaviour of the internals and their intended domain of use in the continuum of computing, and are enforced at build and execution, are central assets to the programming of the future. They help tackle a large fraction of the various issues that (current and) future software systems face. Their primary benefit is that they promote a practical, agile, and sound way to

envelope legacy software into well-defined parts, and afford quick time-to-market to the construction of products that integrate novel and reused parts in a trusted and reliable whole.

Interesting and promising work (e.g.: FP7-ICT projects COMBEST [27], SATURN [177], ServFace [185], NEXOF-RA [144], PROWESS [156], CONTREX [29]; FP7-JTI projects nSafeCer [431], SESAMO [178], CONCERTO [28], and H2020 SAFURE [176] and AMASS [8]) has been carried out in Europe around this particular subject in the last two framework programs. This wealth of work should be furthered by new research efforts and accompanied by industrial assessment.

2.5.3.4 THE NEW PROGRAMMING LANGUAGES



Figure 122: A toolbox

As noted earlier, the current situation with software quality is not good; see for example [428]. Current software is for the most part of low quality and absorbs huge resources for bug-fixing and corrective maintenance. The shortage of trained and qualified programmers pulls in the workplace people from various other lines of profession, with insufficient training, which can hardly be acquired on the job given the shortage of qualification and the production pressure. This situation poses questions on what should we do to counter this trend effectively.

Programming is becoming multi-paradigm: imperative vs functional, synchronous vs asynchronous, strictly vs selectively object oriented, sequential vs concurrent or reactive, parallel vs data flow, homogenous vs heterogeneous, centralized (shared-memory) vs distributed or decentralized, transactional vs eventual consistency, monolithic vs componentized. It is unlikely that a single programming language will be able to support all such paradigms into a consistent, manageable and efficient whole. It is more plausible that a single software system will result from the integration of a collaborating collection of software parts each of which internally adopts some of those paradigms. Given this diversification, emphasis should be placed on devising programmatic solutions to specify the orchestration of those, possibly heterogeneous, parts, that is, how they are to interact, where they are to be deployed, how they should transparently scale and how their life cycle should be managed.

Greater attention should be placed on correctness-by-construction development practices, which promote active and

preventive enforcement of restrictions, application of fitting patterns, and automated generation of trusted, proved code. Different solutions may be required for in-the-small and in-the-large programming scenarios.

A style of programming that aimed at *revealing intentions* – which could be assured by verification – was one of the highlights of late 1990s extreme programming, XP, (see: [310]). Despite the hype surrounding this trend, it was not equally well followed up in the practice.

Arguably, this situation happened for two reasons: programming language syntax is so stylistically varied and diverse that is hardly always objectively revealing, and certainly not in one and the same way: this nature causes the XP practice to slip into the subjective, thereby becoming less attractive and less effective. Second, systematically tracing back code parts to their specified intent is a major brake to the rate with which source code is committed: this tension is often resolved by loosening the obligations on quality assurance. Once again, the human factor gets in the way. It would be much easier, more reliable and systematic if the code artefacts were the product of computer-aided automation, and the human contribution were the declaration of the intent.

DevOps should be brought to the next level. It is acknowledged that future systems will be always and continually evolving. In the same way as DevOps practices have embraced the attention for security, they should be augmented with solid support for the other non-functional concerns.

2.5.3.5 NEW DOMAIN SPECIFIC LANGUAGES

Besides general-purpose languages (GPL) such as C, C++, Haskell, Java, etc. the past decade has seen increasing interest for domain specific languages (DSL). DSLs are designed to keep as close to the problem domain as is possible, and thereby bridge the gap between GPLs and a specific problem domain. VHDL for example bridges the gap between traditional GPLs such as Ada and the hardware design domain.

The division between GPLs and DSLs is not sharp: it has been argued that COBOL is a DSL for the business domain, and some DSLs are sufficiently rich to allow to program problems from other domains as well as GPLs. It is important that a DSL is designed with a specific problem domain in mind. As a general observation, DSLs are often small languages, more declarative than imperative, and have focused expressiveness. DSLs are intended to be used by non-programmers.

DSLs can be implemented in several ways. Libraries, with a well-defined API, can be seen as a form of DSL, where the DSL is then integrated in an existing GPL. The many libraries developed for Python for specific purposes can be seen as DSLs of this form. The term embedded DSL is often used for this form.

However, often a DSL is literally a language, with its own syntax and semantics tuned for a specific problem area. It allows domain experts to describe knowledge or express problems in a notation that is close to the expert domain. It can be a standalone language, an extension of an existing language, a restricted version (subset), or a combination of the last two. In these cases, separate tooling is required in the form of a compiler or interpreter.

Embedded DSLs are the most efficient in terms of implementation effort. If the experts using the DSL also have programming experience, then the user efficiency is also quite high.

If a DSL is in the form of an actual language, then a compiler or interpreter has to be developed for that language, or an existing compiler or interpreter has to be extended. In general, this implementation work is relatively extensive: it requires scanning, parsing, and semantic analysis in addition to code generation and the development of the functionality. That last part can be skipped, or kept to a minimum if the goal of the DSL compiler is limited to analysis and/or the checking of consistency of the model.

DSLs should be defined as precise as required. Experience shows that informally defined languages cause problems when the semantics of the language is implemented. Also, the effort spent on defining a language pays off through a decrease in implementation effort. A precise definition of a DSL allows the use of tools to implement the language.

The cost of a DSL has to be weighed against its benefits. Factors such as development costs, expected life, (non-)programmer efficiency have to be taken in to account.

On the user side of DSLs, if an expert has little programming experience, then the use of a DSL, which is a formal language, can still be difficult for the expert. Experience [130] shows that the acceptance of DSLs is limited. The validation and reasons for acceptance have hardly been studied.

In this part, we have seen that DSLs are widespread and important tools. Acceptance of DSLs varies greatly, end user productivity varies as well. Much emphasis has been paid to the languages themselves, which explains their widespread use.

Boosting acceptance of DSLs will require a better tuning of the languages to the way of working and the application domain. Being able to experiment with a DSL is therefore an important aspect in their development, e.g. development environments allowing the domain experts to experiment with the DSL design. Such environments should move away from the traditional text-only and towards a graphical environment, allowing the domain expert to design a DSL through examples, assisted by AI techniques to transform the examples into specification. Existing compiler technology will still play an important role in such environments, but much more behind the scenes.

2.5.3.6 ATTENUATING THE HUMAN FACTOR: COMPUTER PROGRAMS GENERATING PROGRAMS

We mentioned earlier in this document that the distorted notion of productivity has a role in the hype around certain programming languages, whose proportion of use and presence much exceeds the language design's intent. That intent hardly is to "conquer" the world, but – more prosaically – to do some things, which happen to be recurrent in a given domain of application or in some part, component or layer of some systems, better.

When the recurrence occurs in spots which are very visible to the programmer community – which acts as a sort of vertical social network – and the public opinion around it, the programming language that appears to best serve that recurrence gets traction and earns attention. At that point, the quest for social credits yields a flood of apps, utilities and libraries, which attract opportunistic re-use as long as they help cut some corners here-and-now and achieve rapidity of development. One distinct consequence of that phenomenon is that short-sighted software artefacts become part – for the mere reason of being at hand in the moment of need – of system and application infrastructures where they would not really belong if quality barriers were in place and policed consistently, and that, when those systems or infrastructures happen to become mission- or business-critical, will pose the difficult question of replacing them with proper-quality software.

One reason for the lack of stringent attention to quality in software production is the very large variety of developer types as illustrated by figure 123, which tends to fragment vertically and to follow the market economy – call it "the dominating culture" – of the sector that they serve.

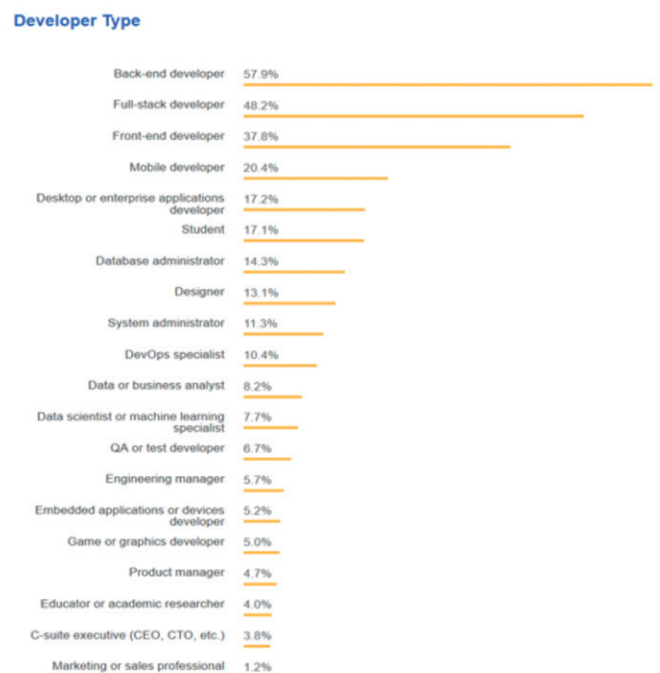


Figure 123: There is a large variety of developer types, which gives rise to rapid fall into verticality and loss of common, transversal practices. – Source: Stack Overflow, Developer Survey Results 2018

Another interesting factor of influence on the lack of balance between productivity and quality in software is the demographic of the programmers' population. Figure 124 provides a view into that demographic, focusing on the gender ratio and the developer kind.

Another angle of the same demographic, explored in Figure 125 shows that the level of education may not be as high as the current and future impact of the products from that community requires. In this very regard, it is also interesting to notice that the same 2018 survey of the Stack Overflow community reports that a considerable proportion (more than 1/3) of the programmers' community members do **not** have computer science, computer engineering or software engineering education.

Humans writing software may therefore be more a problem than a solution, for their exposure to – and generation of – idiosyncrasies, fashion trends, tribalism, subjectivity, and short sightedness. As a testimony to this problem, there currently appears to be more attention to programming language syntax – which is required to be (subjectively) cool, intuitive, clever, and ultimately tribally idiomatic – than to the depth, assurance and run-time overhead of their semantics, and more proclivity in practitioners and educational agencies to follow and replicate the mounting trends than to scratch the surface and take a deeper understanding of what is actually needed.

With current trends in the offer and access to education, it is difficult to imagine that university curricula will manage to reverse the trend and create a sufficiently diffuse quality culture that can meet the massive demand for the software programs that are going to feed and drive our social, professional, and service applications and infrastructures. An increase in the “mechanisation” of programming, certainly much more attainable today that it used to be in the recent past, may be a practical evolution.

It can be expected, however, that programming will change from human-made hand-writing narrative text – no matter how far assisted by program editors – into **automated translation** of design and programming intents, in a sort of next-generation model-driven development. This translation, whose outcome should be commented source code fully traced back to requirements, should **not** tell the programmer what to do (lest the programmer's responsibility is lost), or even scrap the human actor altogether, but rather produce the code corresponding to the programmer's intent and, with it, help the programmer ascertain the soundness of the original intent. Incidentally, this trend will lessen the programmer's sensitivity to the popular perception of coolness of language syntax, and consequently create more room for “less cool” programming languages, more apt at addressing non-functional requirements more soundly, from the ground up.

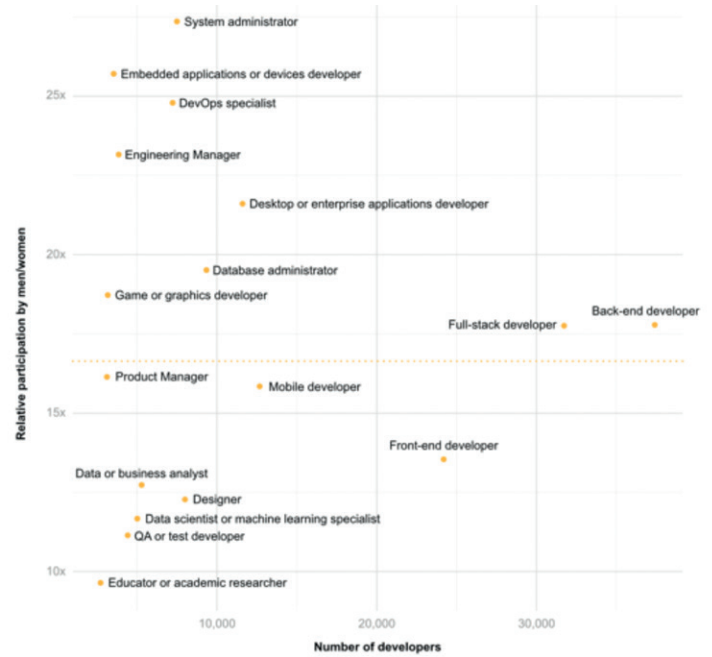
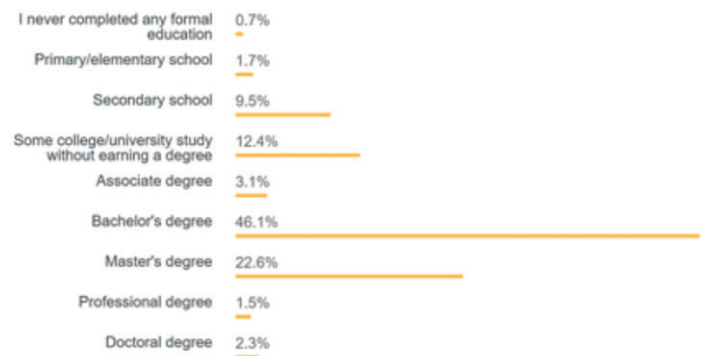


Figure 124: Density of developers per type, plotted against the distribution of gender

Source: Stack Overflow, Developer Survey Results 2018



94,703 responses

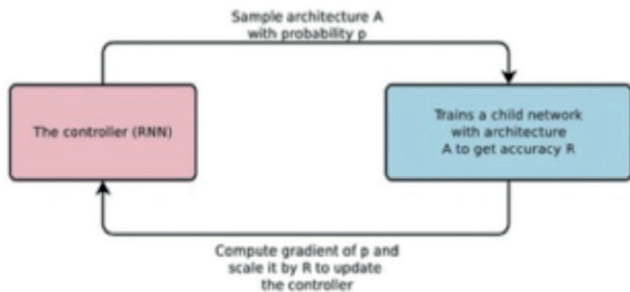
Worldwide, about three-fourths of professional developers have the equivalent of a bachelor's degree or higher. It is not that rare to find accomplished professional developers who have not completed a degree.

Figure 125: Distribution of education in the software programmers' community– Source: Stack Overflow, Developer Survey Results 2018

The **tool support** that conforms to this vision should be capable of continually learning rules, styles, and patterns from **good** practices and resources submitted to it, as well as of proposing, exploring and evaluating alternative solutions against weighted criteria, in addition to tracing all code fragments being released to the quality requirements and constraints that apply to it.

The emphasis placed on the need to learn from “good practices” sets this direction in a different course from that pioneered by Alpha Zero (as evoked for example by: <https://www.futurity.org/artificial-intelligence-bayou-coding-1740702/>). In the latter case, in fact, game rules exist, which could be actively enforced, and outcomes whose goodness could be measured objectively. There is no such thing for software production, instead, and least of all

“Neural Architecture Search”, using a recurrent neural network to compose neural network architectures using reinforcement learning on CIFAR-10 (character recognition)



Model	Depth	Parameters	Error rate (%)
Network in Network (Lin et al., 2013)	-	-	8.81
All-CNN (Springenberg et al., 2014)	-	-	7.25
Deeply Supervised Net (Lee et al., 2015)	-	-	7.97
Highway Network (Srivastava et al., 2015)	-	-	7.72
Scalable Bayesian Optimization (Snoek et al., 2015)	-	-	6.37
FractalNet (Larsson et al., 2016)	21	38.6M	5.22
with Dropout/Drop-path	21	38.6M	4.60
ResNet (He et al., 2016a)	110	1.7M	6.61
ResNet (reported by Huang et al. (2016c))	110	1.7M	6.41
ResNet with Stochastic Depth (Huang et al., 2016c)	110	1.7M	5.23
	1202	10.2M	4.91
Wide ResNet (Zagoruyko & Komodakis, 2016)	16	11.0M	4.81
	28	36.5M	4.17
ResNet (pre-activation) (He et al., 2016b)	164	1.7M	5.46
	1001	10.2M	4.62
DenseNet ($L = 40, k = 12$) Huang et al. (2016a)	40	1.0M	5.24
DenseNet ($L = 100, k = 12$) Huang et al. (2016a)	100	7.0M	4.10
DenseNet ($L = 100, k = 24$) Huang et al. (2016a)	100	27.2M	3.74
Neural Architecture Search v1 no stride or pooling	15	4.2M	5.50
Neural Architecture Search v2 predicting strides	20	2.5M	6.01
Neural Architecture Search v3 max pooling	39	7.1M	4.47
Neural Architecture Search v3 max pooling + more filters	39	37.4M	3.65

Figure 126: Neural Architecture Search – Source: Barret Zoph, Quoc V. Le ; Google Brain

in indistinct lumps of software artefacts, such as online source-code repositories. Before a machine-learning agent can learn useful knowledge, there must be a good base of data to learn from, which includes the “rules of the game”, that is to say, the traits that “good” software should have, in functional and non-functional terms. Codifying this rule base comprehensively in a form that can guide deep learning, and collecting quality programs expressed in **polyglot** source artefacts, which solve categorised problems, is a massive prerequisite effort, which require software engineering and not just machine learning experts, lest the learned knowledge has biases, quirks, and holes, which add to the problem instead of solving it.

In some way, the generative vision that we have evoked here elevates the so-called **low-code development** [26] to the next level for grander and wider goals.

2.5.4 SMART DESIGN TOOLS

The complexity of hardware and software developments for systems has become so large that humans are finding it harder and harder to generate efficient solutions. Complexity is managed by abstracting or clustering, but at the cost of extra layers that are generally decreasing overall performances: a sum of local optimizations is less efficient than a global optimization.

Current hardware and software are composed of various parts or layers, with interfaces, allowing to manage (for human) the overall complexity. But computers were invented to manage complex problems, and there is an emerging trend to use progress in computing power and optimization algorithms, or even using techniques derived from artificial intelligence, to help optimize systems and software. Computers are good at optimizing problems with a very large number of parameters, which is very

difficult for humans. Compilers are already using advanced optimizations techniques and place and route systems as well.

To cope with the ever-increasing complexity of today’s and tomorrow’s systems, we need to have better tools. For example, DARPA in the US is launching a call for project in the fields of automated design tools [326], [328].

Solutions are currently designed allowing to find good meta-parameters for deep learning solutions: these solutions explore the space of parameters to find a good topology for the neural networks and parameters used during learning. This is generally called auto-ML. Google is launching its cloud auto ML, allowing its users to develop rapidly deep learning solutions to their problems [273].

Software development can also be helped with new techniques, in what it is sometimes called “programming 2.0”. The aim is to

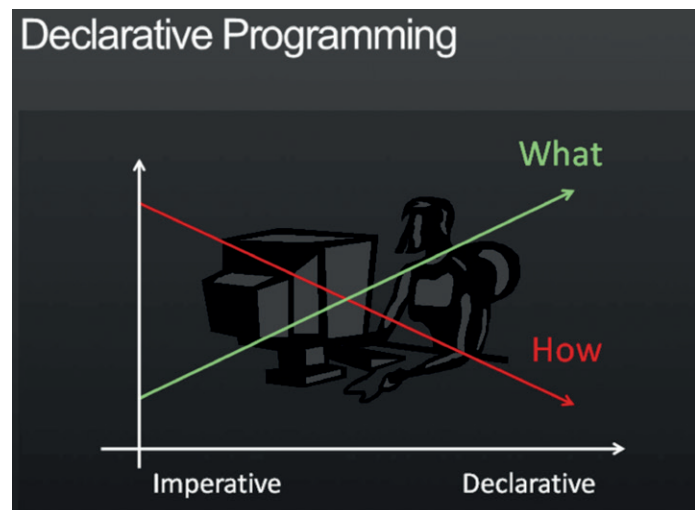


Figure 128: Declarative Programming

Avoid premature commitment, seek design alternatives, and automatically generate performance-optimized software.

Programming by Optimization

When creating software, developers usually explore different ways of achieving certain tasks. These alternatives are often abandoned or abandoned early in the process, based on the idea that the flexibility they afford would be difficult or impossible to exploit later. This article challenges this view, advocating an approach that encourages developers to not only avoid premature commitment to certain design choices but to actively explore promising alternatives for parts of the design. In this approach, called Programming by Optimization, or PBO, developers specify a potentially large design space of programs that accomplish a given task, from which versions of the program are selected for various use scenarios and generated automatically, including possible variants derived from the same sequential sources. We outline a simple, generic programming language mechanism that supports the specification of such design spaces and discuss ways specific programs

are generated and how they are used to find the best solution. This approach is based on the idea that the flexibility they afford would be difficult or impossible to exploit later. This article challenges this view, advocating an approach that encourages developers to not only avoid premature commitment to certain design choices but to actively explore promising alternatives for parts of the design. In this approach, called Programming by Optimization, or PBO, developers specify a potentially large design space of programs that accomplish a given task, from which versions of the program are selected for various use scenarios and generated automatically, including possible variants derived from the same sequential sources. We outline a simple, generic programming language mechanism that supports the specification of such design spaces and discuss ways specific programs

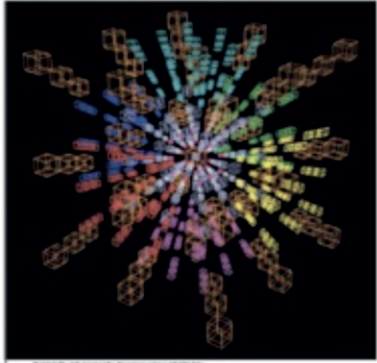


Figure 127: Article on Microsoft's AI is learning to write code by itself

Microsoft's AI is learning to write code by itself, not steal it

Written by Dave Gershgorin

What if instead of searching through menus within programs like Microsoft Excel, our computers could understand the problem we're trying to solve and write the software to solve it? It's a hyper-futuristic idea, but one that has recently seen progress from Microsoft Research and the University of Cambridge.

In a November 2016 paper (pdf), which gained notoriety after being accepted into one of the year's largest artificial intelligence conferences, Microsoft and Cambridge built an algorithm capable of writing code that would solve simple math problems. The algorithm, named DeepCoder, would be able to augment its own ability by also looking at potential combinations of code for how a problem could be solved. (It's a bit complicated; we'll break it down later.) However, this doesn't mean it steals code, or copy and pastes it from existing software, or searches the internet for solutions, as some reports have claimed.

Communications of the ACM, 55(2), pp. 70–80, February 2012
www.prog-by-opt.net

Figure 127: Article on Microsoft's AI is learning to write code by itself

develop systems where the “programmer” is describing *what* the program should accomplish, rather than describing *how* to accomplish it as a sequence of the programming language primitives. The “what”, which can be explicitly given to the system, or by examples, it is where this approach and AI techniques meet.

For example, the designer of an application should describe the *concurrency* of an application, not how to parallelize the code for it. It is already evident that (good) compilers know better about efficiently using architecture than humans, they are better at optimizing code with the multiple constraints of modern architectures.

These ideas are not new – they were already presented in the HiPEAC Vision 2010, for example – but the recent increase of performance of AI techniques makes them a more realistic in the short term.

In the domain of hardware, multicriteria optimization techniques can be used to help define an efficient architecture.

It is expected that automated techniques will allow the design of more efficient computing systems and their software, and also the integration or the creation of both simultaneously, leading to a more optimal co-design approach.

There are plenty of domains within ICT where AI-related techniques could be used to improve efficiency:

- Automatic generation of user interface (UI) from sketch. Deep learning is trained from UI layout (and associated code), and in the inference phase, the user sketches what she wants as UI, and it generates code.

- Debugging: the AI system find similar piece of code in repository such as Stack Overflow [329] and finds out if other people have problems with this piece of code.
- Adaptability: Using AI technology such as reinforcement learning to self-improve the software while it is in operation (e.g. in cloud computing).
- Security: Using AI technique to analyse and detect abnormal behaviour, even weak signals, to detect intrusions or other malware.

2.5.5 THE OPPORTUNITIES AHEAD: THE SOFTWARE ROADMAP

From the above, a few key recommendations can be derived, laying out a roadmap for software in EU, that provide significant opportunities for the future of computing.

- 1 Non-functional properties (e.g time, power and energy, security and safety, etc.) will become a central focus of the cognitive CPS of the future. Non-functional properties should therefore be recognized and integrated as first class-citizens in software tooling, from programming languages to compilers, runtimes and libraries.
- 2 Software applications and infrastructures will increasingly be aggregates of heterogeneous artefacts with a variety of deployment requirements. Controlling them can hardly be done in a merely declarative way or scattered in a maze of uncorrelated and independent scripts. Languages and tools for orchestrating collaborative distributed and decentralized components are thus needed.

- Ne-XVP project – Follow-up of the TriMedia VLIW (<https://en.wikipedia.org/wiki/Ne-XVP>)
- 1,105,747,200 heterogeneous multicores in the design space
- 2 millions years to evaluate all design points
- Optimization techniques allowed to reduce the induction time to only few days

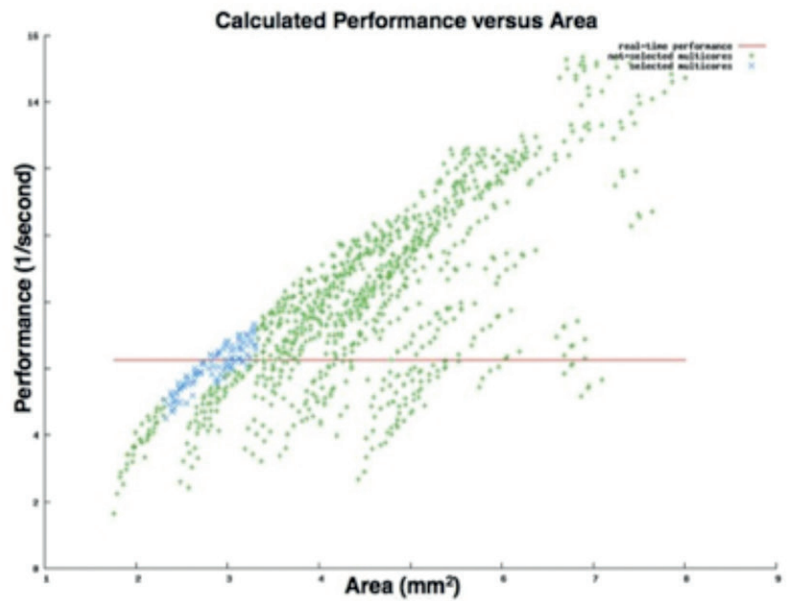


Figure 129: Ne-XVP project, calculated performance versus area – Source: [125]

- 3 The legacy problem is best addressed by containerization, but emphasis in containers must be shifted to enhancing their interface specifications, so that they help assess semantic conformance at build, integration, deployment and execution time. The required enhancements should take the form of enforceable contracts, which need to be codified in manners that afford agility, performance, and assurance. Hence containerization must be enhanced to support components augmented with interface contracts (covering both functional and non-functional properties).
- 4 Human programming is less than ideal, especially in the face of the complexity of the upcoming systems. Programmers' education, whether professional or practitioners, is insufficient; that of users, private or institutional, is even less, which frustrates the quest for and the assurance of quality. At current trends, the throughput and rate of education agencies is unable to meet the demand for software programs to be developed. To solve this crisis, computer programs should help write software programs by learning from the "quality rules of the (programming) game" as well as from a base of selected "good" examples in polyglot source languages (since no single programming language is good at everything).

2.6 THE SOCIETAL DIMENSION

Computing is a disruptive technology, which means that it introduces fundamental changes into existing systems. Over the last five years, many people have become aware that the impact of computing and the internet is so profound that it is changing society as a whole too. In this section, we look into the effects on society, people, the job market, Europe, Planet Earth, education and ethics.

2.6.1 IMPACT OF COMPUTING TECHNOLOGY ON SOCIETY

The use of computing technology is changing society in unprecedented ways. Examples are abundant:

- It has changed the nature of information storage and processing. Everything that can be digitized (text, pictures, audio, video, and so on) has been digitized and made available 24 hours a day, seven days a week. Since digitization eliminates the use of a physical medium, distribution becomes immediate, protection of copyright becomes more difficult, and archiving content for the future generations becomes a challenge [23]. Digitization has also had a huge impact on the music industry [147].
- It has changed relationships between people. Face-to-face contact based on geographical proximity has been replaced in many cases by remote contact. Even dating is increasingly taking place via the internet, leading to more diversity and more social integration [358], but also to more matches within the same socio-economic group and hence less social mobility.
- It is having a major impact on the job market. Globally, millions of jobs are disappearing due to automation, while at the same time millions of new jobs are being created. The content of the

jobs that remain is continuously changing to keep up with technological evolution. The effect of this change seems to be more inequality, a shrinking middle class and the emergence of a dual economy [159]. See 2.6.3, “Computing technology and the future job market”, for a full discussion of this phenomenon.

- Governments and companies are collecting billions of records on their citizens and customers. This information is used to optimize their processes. There is a growing concern that big data analytics is encoding historical biases, driving positive feedback loops, and leading to unwanted outcomes [33, 212].
- It has changed the nature of politics. Political parties now use social media to build constituencies, while governments use it to interact with citizens. As a consequence, politics has become faster, more personalized, and more direct. Unfortunately, this process can also be hijacked by third parties trying to influence this process by spreading false information, especially during elections when governments are at their weakest [179, 296].

These examples demonstrate a non-negligible and growing impact on society. Few people fully understand internet companies’ business models. Facebook is a free platform with around 2 billion active users. In 2017, its revenue was 40 billion USD – an average of 20 USD per user. That is the average value in 2017 of the information we share on our Facebook accounts. Facebook’s real customers are the companies and organizations paying for marketing campaigns. The goal of a marketing campaign is to change the behaviour of the target group (for example by convincing them to buy a particular product, to sign up for a service or to vote for a political party).

For companies like Facebook or Google, *the users are the product*, and as any other company Facebook and Google *try to adapt their*

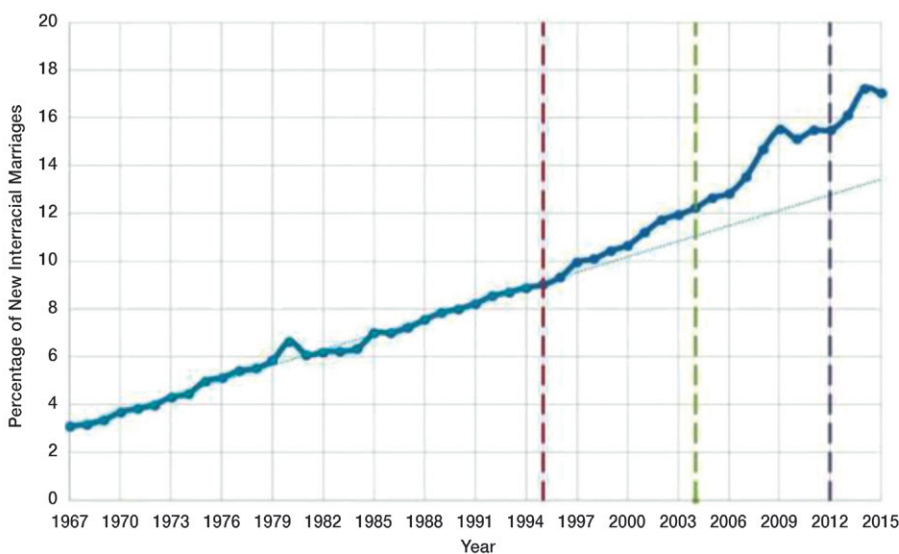


Figure 130: Percentage of Interracial marriages among newlyweds in the U.S.

Source: Pew Research Center analysis of 2008-2015 American Community Survey and 1980, 1990 and 2000 decennial censuses (IPIMS). The red, green and purple lines represent the creation of Match.com, OKCupid and Tinder, three of the largest dating websites. The blue line represents a linear prediction for 1996-2015 using the data from 1967 to 1995.

product (i.e. us!) to the needs of their customers. The perfect product is a user who spends a lot of time on the platform, and reacts in ways intended by the (paying) customers (i.e. buying goods and services, voting and so on). The more information the platform has about its users, the more targeted and the more effective the marketing campaigns can be made, the more the platform can charge for them, and the bigger its revenue will be. The longer a user spend on the platform, the more advertisements can be shown, and, again, the bigger the revenue of the platform will be. The more features the platform offers (face recognition, language translation, video, games, and so on), the more time users will spend on it, and the more frequently they will return.

Platforms deliberately use mechanisms to make them addictive, or at least habitual. These include likes, automatic notifications, clickbait and scoring. This has been called *brain hacking* [167]. Addicted users come back frequently, which translates into higher revenue. Finally, the number of users has to grow fast for start-up internet companies and this influences the content. On one hand, platforms try to ensure that nobody will be offended by content on the platform, so they censor all content that might offend valuable groups of users. On the other hand, viral content is welcomed because it means more people spending more time on the platform, and hence generates extra revenue.

“One of the core things going on is that they have incentives to get people to use their service as much as they possibly can, so that has driven them to create a product that is built to be addictive. Facebook is a fundamentally addictive product that is designed to capture as much of your attention as possible without any regard for the consequences. Tech addiction has a negative impact on your health and on your children’s health. It enables bad actors to do new bad things, from electoral meddling to sex trafficking. It increases narcissism and people’s desire to be famous on Instagram. And all of those consequences ladder up to the business model of getting people to use the product as much as possible through addictive, intentional-design tactics, and then monetizing their users’ attention through advertising.”

Sandy Parakilas, former Facebook platform operations manager, currently Chief Strategy Officer at Center for Humane Technology [200]

Manually analysing the actions of billions of users is not feasible, which is why these companies are making huge investments in artificial intelligence – including the development of custom hardware to accelerate their algorithms – to extract more information from the raw data. It is no coincidence that companies like Google and Facebook are leading in this area [139]. The better their big data analytics, the higher their revenue. There is an arms race between (social) media companies for the attention of the

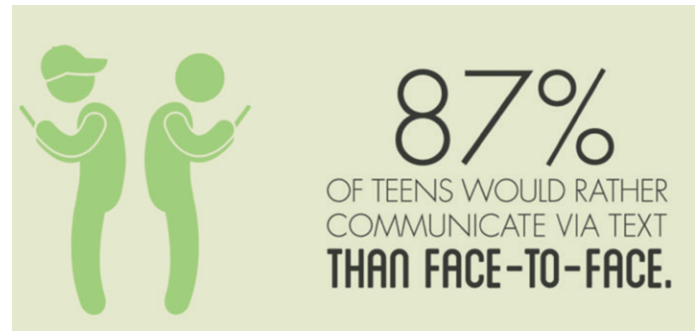


Figure 131: Phubbing
Source: stopphubbing.com

user. However, a user cannot spend more than 24 hours a day on any one social network, search engine or streaming service. All these companies are thus competing against each other to get more attention: by making their platforms more attractive, more addictive, easier to use and so on.

And they are successful: in the younger generations, social media has almost completely outcompeted traditional media like television and newspapers [95]. In their competition for more attention, they are also monopolizing people’s time, in both their professional and private lives. Active professionals believe they have to have a presence on the social media, and to amass large numbers of followers. This leads to loss of productivity and mental absence at meetings, etc. In many people’s private lives, screens have replaced face-to-face interactions at home, at the dining table, at the pub, in restaurants and on public transportation. This leads to a phenomenon known as “phubbing”, or phone snubbing: checking your smartphone during social events instead of giving your full attention to the people who are physically there [113].

The final frontier is competing with people’s sleep. Studies show that millions of people suffer from sleep deprivation resulting from excessive use of smartphones and tablets [96, 391].

In the sections which follow, we examine a limited number of particular societal effects.

2.6.1.1 PRIVACY EROSION

There are multiple definitions of privacy. In the 19th century, privacy was defined as the “right to be left alone”. A more modern definition is that privacy is the “control one has over the information about oneself”. It is necessary that doctors maintain medical records about their patients, but nobody expects the doctor to share this information with third parties (medical privacy) unless this were to be required for medical treatment. We expect the same behaviour from financial institutions (financial privacy), websites (internet privacy) and voting systems (political privacy). We do not expect an email service to use the content of our messages to influence the advertisements we see on websites, or a booking website to use the type of rental car we prefer to result in seeing advertisements for that particular type of car.

Gathering information about users is crucial to the business model of internet companies. That is why many websites nudge users to complete their profiles, thereby collecting additional monetizable information. Some companies, like the now notorious Cambridge Analytica, have made a business model out of collecting information, analysing it, and selling it to whoever is willing to pay for it.

Many people are largely unaware about the cost of convenience in terms of lost privacy; or if they are aware, they willing to give up some of their privacy in return for convenience:

- Booking websites collect numerous details about every single trip their users book. This is crucial marketing information for hotels, airlines, rental car companies.
- Streaming music applications have data on when and where users listen to music, as well as what their musical preferences are. The better streaming music providers can profile their users, the better suggestions they can make and the more frequently and longer people use the service.
- Companies selling e-books know the identity of every single reader of a book, when they are reading a book, which parts they actually read and so on. In a sense, they know what a buyer learned from the books they bought. The more they know, the better suggestions they can make; it is not difficult to guess the interests of somebody buying books on classic cars, cookery, political history, or travel guides, for example. By (not) making particular suggestions, they can even steer what their users read and even think.
- Social media networks monitor all the private details users share with their most intimate friends, and use this data to infer information (for example, that the person feels depressed), in order to send them targeted advertisements they know work well (such as make-up or medication for depressed people). Their aim is not to help people, but to sell and to influence. The people in social media control rooms are not medical staff; they do not have to comply with professional codes and they do not care about whether the advertised drugs are effective or safe.

- News websites track which articles users read, and adapt their content offering (news and advertisements) to their interests. They basically decide what their users will read, which might lead to a biased perception of the world.
- Satellite navigation systems detect where the navigation system (and, by extension, probably its owner too) is at any time. It is comparable to being shadowed by somebody wherever you go.
- Voice controlled devices keep track of what goes on in a house or office, and they can be hacked to eavesdrop on conversations. Few people would appreciate a stranger sitting in their house all the time.

In addition to the examples above, people are already under surveillance for a large part of the day, through access control systems in companies and hotels, numerous cameras in public places, licence-plate recognition, Google Street View filming the street, tourists taking pictures with people in the background and posting them on social media, and so on. Most people do not protest about this surveillance because they believe that it helps the government to enhance their safety and prevent terrorist attacks.

Irrespective of the application, the fact is that (i) all our actions in *cyber space*, and an increasing number of actions in *physical space* are being recorded and stored in huge databases, (ii) that an increasing number of such databases are being linked (often through acquisition, or by linking government databases to facilitate e-government), and (iii) that that there is no guarantee that this data is only used for the purpose it was collected for.

It is clear that there is an urgent need for a (global) legal privacy framework and that computing systems will have to support better privacy mechanisms. The EU General Data Protection Regulation (GDPR) is the most important change in data privacy regulation in 20 years. Besides the automatic protection of privacy, it also helps to create awareness about the importance of privacy. It protects citizens against illegal use of their data owned by companies or governments. Although GDPR was meant to protect the privacy of Europeans, it has had a much bigger impact and today the privacy of all global internet users is better protected. Surprisingly, California recently adopted a law similar to GDPR in the home state of Google and Facebook [35]. It shows that Europe – even without hosting one of these major internet companies – can still influence policies that affect them.

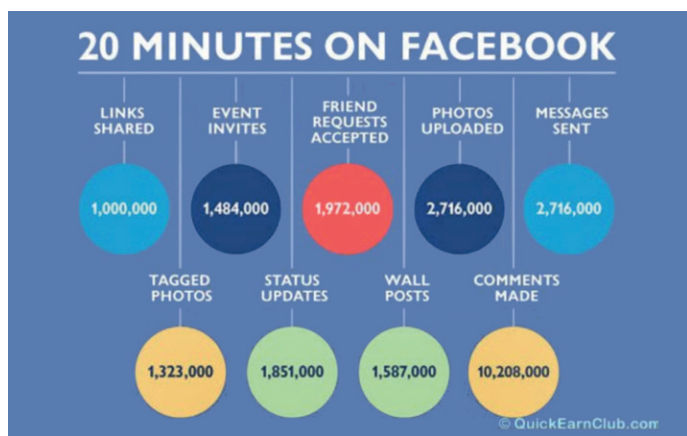


Figure 132: 20 minutes on Facebook
Source: QuickEarnClub

2.6.1.2 FAKE INFORMATION

Whereas traditional media have built-in filters that require journalists to verify their sources, there is no such thing in social media. Anybody can post anything, and as soon as it passes social media companies' decency filters, it becomes public. The social media reviewers censor particular content (child abuse, sexual content, hate speech, ...) but not fake information. The higher the number of people reading and liking the fake information, the

better it is for the business results of the platform. In response to public concern over the spread of fake news and hate speech on social media, major companies such as Facebook have employed editors to try and monitor the content, but these resources are often inadequate for the task of checking content posted by two billion users.

Over the last few years, there has been a surge in false or misleading information such as fake news, fake science and deep fake videos. Fake information is information which is presented as a reliable piece of information, but is either completely made up or highly misleading. Such messages are like hoaxes on steroids. Popular genres are the launch of conspiracy theories (such as those about the condensation trails behind airplanes), and the spreading of pseudo-science (such as the dangers of

vaccination). The motives of people spreading such information range from making money (mostly from advertisements alongside stories that go viral) to political objectives (influencing elections, creating unrest, destabilizing societies).

Recent years have also seen an increase in the spreading of fake news by *professional internet trolls*. Unlike individual internet trolls – unpaid people who deliberately comment on online posts to generate a specific reaction – these are paid by states to broadcast propaganda in general online media (alternative news channels for example) as well as in specific communities such as the defence and military community. Especially in the latter, acting in groups, they target chosen topics, start rumours and launch alternative narratives, repeating them and getting them quoted by fellow propagandists to make them more credible, in order to advance their country's ideas and weaken their adversaries'. As such, these professional internet trolls are in fact members of cyber armies, waging hybrid warfare [187] on the internet.

The most recent technical evolution of fake information are the so-called deep fakes, a successful application of face swapping technology to video. Originally designed to put the face of celebrities on pornography actors in action, the technology has been used to create credible fake interviews [417]. For the naïve viewer, these interviews are hard to distinguish from the real thing. In combination with video call services like Skype, Facetime

HOW TO SPOT FAKE NEWS

- CONSIDER THE SOURCE**
Click away from the story to investigate the site, its mission and its contact info.
- READ BEYOND**
Headlines can be outrageous in an effort to get clicks. What's the whole story?
- CHECK THE AUTHOR**
Do a quick search on the author. Are they credible? Are they real?
- SUPPORTING SOURCES?**
Click on those links. Determine if the info given actually supports the story.
- CHECK THE DATE**
Reposting old news stories doesn't mean they're relevant to current events.
- IS IT A JOKE?**
If it is too outlandish, it might be satire. Research the site and author to be sure.
- CHECK YOUR BIASES**
Consider if your own beliefs could affect your judgement.
- ASK THE EXPERTS**
Ask a librarian, or consult a fact-checking site.

International Federation of Library Associations and Institutions

Figure 134: How to spot fake news
Source: IFLA

HOW THE WEB WAS LOST
Hacks, viruses, leaks, and surveillance have been part of online life since the beginning

- 1971**
The world's first computer worm, "the Creeper," is created.
- 1981**
Ian Murphy becomes the first American convicted of a cyber-crime after hacking into AT&T's system to give customers discounted calling rates.
- 1982-83**
A teenage cyber-gang, who refer to themselves as "the 414s," hack into the Los Alamos National Laboratory and Memorial Sloan Kettering Cancer Center.
- 1996**
Hackers break into Web sites for the United States Department of Justice, the C.I.A., and the U.S. Air Force.
- 2001**
President Bush signs an order initiating the N.S.A.'s domestic-spying program.
- 2006**
As part of its News Feed launch, Facebook posts personal details of users.
- 2009**
In a speech, Berners-Lee warns that the power of online information "is so great that the commercial incentive for companies or individuals to misuse it will be huge."
- 2013**
Google acknowledges that Street View, its photo-mapping program, used its technology to collect data from home networks without people's knowledge.
- 2014**
The New York Times reports that the N.S.A. is using facial-recognition software to store the images of millions of people.
- 2016**
BuzzFeed News uncovers at least 140 fake-political-news sites designed to generate shares on Facebook by using false information, such as a claim that the Pope endorsed Trump for president.
- 2017**
Facebook acknowledges that Cambridge Analytica collected data on more than 80 million users.

Figure 133: How the web was lost
Source: [108]

or Google Hangouts, this technology can easily be used for phishing purposes. Imagine being called by (the voice of) your financial advisor with some advice about the management of your retirement portfolio. Eventually, this may mean that we are no longer able to trust phone conversations or even video conferencing sessions. Fake, yet very convincing, political or religious speeches could also be produced and distributed throughout the world almost instantaneously, quickly creating large opinion movements for specific purposes.

The term virtual reality was used for the first time in 1987, swiftly followed by head-mounted virtual reality devices. Thanks to increased computing performance, the virtual environments that can be created are becoming more realistic than ever, including very realistic avatars. However, when users put on a head-mounted device, they know that they are entering a virtual world. Fake information is much more dangerous because it invades our world disguised as real information. Its appearance is so realistic that it has the power to change people's opinion and behaviour. It gives a totally new definition to the term "virtual reality".

In conclusion, too many people trust the internet as they would trust a newspaper. An alarmingly high number of people take fake information seriously. Fake information can now be spread more quickly and more convincingly than ever, multiplying its power. The only antidote seems to be better education, specifically targeted to help people distinguish fake from real information.

2.6.1.3 DIVIDE AND CONQUER

What sets social media apart from traditional media is that traditional media broadcast their messages publicly so that everybody can receive them and, ideally, learn about the arguments of a range of stakeholders by watching their channels. In contrast, the combination of advanced big data analytics and significant computing power hosted in large data centres has enabled social media platforms to create a personalized experience for each individual user. That means that every user gets to see a different stream of messages and that users cannot see the message streams of other users. Users can share messages in their own network, but since networks tend to be clustered, users tend to see more of the same messages rather than different points of view.

In so doing, social networks create information silos or filter bubbles and act as echo chambers which reinforce the values of the members of the network. Awkward facts – like a mistake made by a member of the network, for example – will not garner a large number of "likes", and will quickly disappear from timelines. Hence it is very difficult for information in one information silo to make it into another. The following figure illustrates three different communities living in Israel: pro-Palestinian, pro-Israel and religious/Muslim. There are very few links between the pro-Palestinian and pro-Israel communities. Most links are shared via the religious/Muslim community. There



Figure 135: Israel, Gaza, War & Data - Social networks and the art of personalizing propaganda – Source: [75]

is little chance that messages from the pro-Israel network will ever make it into the pro-Palestinian network and vice versa.

What is worrying is that a handful of private global companies and their proprietary algorithms decide who gets to see what messages, in which order, and when. They can even gradually modify the user's preferences by proposing only a limited set of items and removing items that are old, in low demand or not in accordance with the ideas of the providers, for example. In the past, opinion-shaping messages came in hard copies, which were harder to remove – it was necessary to physically find them in the house of customers and burn them, as in *Fahrenheit 451* – compared to digital media on private servers and streamed to people who are not using local backups. Already, a number of classic films are not included on streaming services.

"1984" REMOTELY REMOVED FROM ALL AMAZON EBOOK READERS

"In July 2009, Amazon remotely wiped Orwell's "1984" and "Animal Farm" from all Kindle e-readers, because the publisher of the e-books didn't have the rights to sell them in the United States. The move was seen as Orwellian in itself, and raised questions of whether the consumer really owns digital content that is downloaded and paid for." From [462]

All this means that social media companies are in a sense helping to create a worldview per user, formed by purely business decisions – i.e. decisions that will optimize the profitability of the company – mostly unregulated by governments.

The fact that traditional media such as newspapers and television news have declined in popularity among "digital natives" strengthens the impact of social media on the world view of young people.

This explains to a certain extent why traditional media outlets anticipated neither Brexit nor the election of Donald Trump. They were simply unaware of messages shared in circles they did not belong to [179].

2.6.2 IMPACT OF COMPUTING TECHNOLOGY ON PEOPLE

Sixty years ago, the distance travelled by car per year was considered an indicator of progress; today, for environmental and health reasons, this is no longer the case. Similarly, consuming lots of energy-rich food and drink was viewed positively by people who suffered from a lack of food when they were young, whereas today it is no longer considered healthy as it leads to obesity. Consuming too much digital information leads to what is known as “digital obesity”. It seems likely that we will need to develop a healthy, balanced digital lifestyle, avoiding the negative effects of the technology described in this section.

The effects of digital technology on humans has been studied extensively, and there are both positive and negative effects. Customers have access to online information, they can make online appointments and buy goods and services without having to queue, physical meetings can be replaced by virtual meetings, collaboration tools allow people to work together efficiently and form the basis of the paperless office. On a personal level, it is now easier to keep in touch with friends and family members via social media. Many disabled and older people can also participate in social networks because their participation is not constrained by their limited mobility; this, in turn, helps them maintain or develop cognitive abilities.

People from poor countries who cannot afford to travel can access high-quality learning resources such as online courses (MOOCs) developed in wealthy countries. Children can get access to a virtually unlimited source of information about a huge range of topics, leading to a lot more informal learning, including learning foreign languages [93].

However, there are also some side effects [92]. In some cases, people have become dependent on their smartphones. The smartphone does to the brain what using a lift does to the body, compared to the stairs. Rather than memorizing information, people constantly refer to the internet, which can lead to digital amnesia [170]. Skills like mental arithmetic, memorizing numbers (mathematical constants, phone numbers) and driving without a navigation system are disappearing in young people.

Perhaps even more disturbing is the fact that the web is full of texts that fit on just one or two screens, and that this has been linked to losing the ability of “deep reading”, that is to say, the ability to focus on a long text for an extended period of time. Research suggests that the disappearance of this skill, which is needed to read a book or to study [146], can lead to lower academic performance [192].

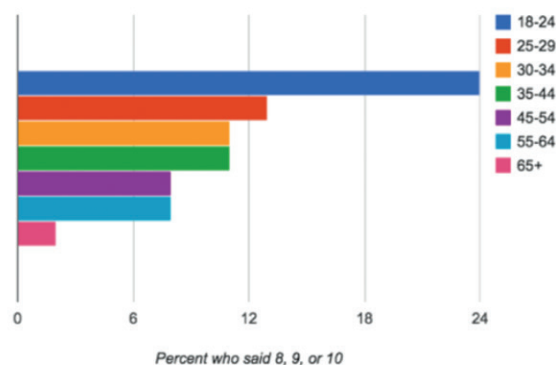
For many teenagers, their smartphone has become part of their personality. Without their smartphone within reach, they feel incomplete. They would rather give up owning a car, a television, or a private swimming pool than give up a smartphone. Some

people would never date a person with a crack in their smartphone screen because they believe that somebody who does not care for their smartphone will not care for people either [254]. The fear of missing an important message can lead to an overload of digital information [412].

Information technology has made sharing information so easy and cheap that it has become epidemic. Many modern workers receive hundreds of messages per day; reading and responding to these messages takes up a significant part of their time, without being explicitly mentioned in their job description. Processing emails has become a struggle, putting people’s bodies in fight mode for extended periods of time, and leading to exhaustion, burnout and faster ageing [46].

Mobile devices invite users to engage in multitasking, i.e. to use their device while performing other activities. Using mobile devices while driving is now forbidden in most countries, but, unfortunately, it still happens all too often. Using the internet during meetings is a very common practice even though it reduces the effectiveness of the meeting as people are often mentally absent in the meeting and therefore not really part of what is going on. Many people believe that multitasking increases their productivity, but there is clear scientific evidence that it is detrimental for productivity and for the quality of work [196].

There is plenty of evidence that the use of technology has an impact of the amount of sleep we get. A survey from 2015 shows that the sleep of young adults is impacted most by technology.



Percent of respondents, by age bracket, who said they agreed with the statement “I don’t sleep as well as I used to because I am connected to technology all the time” at a level of 8, 9, or 10, with 10 points meaning it describes them exactly (Time/Qualcomm)

Figure 136: Percent of people who don’t sleep well because of technology

Source: [150]

More recent studies show that that the problem is at least as severe in teenagers [78, 97], who practise late-night socializing, called vamping, in some extreme cases at any time during the night. Teenagers need around nine hours of sleep, but in 2015, 43% of US adolescents reported less than seven hours on most nights which means that half of teenagers in the USA are seriously sleep deprived. The 18-year-olds were the most affected.

Adults face similar problems. According to the Centers for Disease Control and Prevention in the USA, 35% of American adults are not sleeping enough, an increase from 29% ten years ago. According to the same study, an estimated 70 million US adults sleep fewer than six hours a night. This leads to concentration problems and a number of health issues.

Causes of disturbed sleep include (i) the use of social media which is both mentally and emotionally stimulating [119], and (ii) the blue light emitted by smartphones and tablets which simulates daylight, inhibiting the brain's production of melatonin, the hormone that regulates sleep.

According to some researchers, heavy smartphone use and the consequent sleep deprivation is one of the biggest unaddressed public health issues of our time [168]. Potential consequences include lower academic performance, obesity, and mental health issues including anxiety, "nomophobia" or the fear of being without one's mobile phone, depression, and low self-esteem. Some people even suffer from phantom vibration syndrome, also called ringxiety or fauxcellarm: a perception that a phone is ringing or vibrating when it is not.

Slowly, awareness about the negative effects of heavy smartphone usage is growing and even technology companies have started to offer tools to measure or restrict screen time, such as Apple's "Screen Time" and Google's "Digital Wellbeing". These tools inform the user about the time spent on the different platforms. They are positioned as tools to help users to control



Figure 138: Screen time
Source: Apple/Victor Tangermann

their social media usage, but according to [211], they are not very effective. The pop-ups are a nuisance, comparable to the frustration children experience when their screen time is constrained by parental control apps, and temporarily being shut out of a platform is frustrating. A much more effective solution would be to make the platforms less addictive, but internet companies are unlikely to take action that has a negative impact on their bottom line.

A number of former employees at the larger internet companies have started regretting what they built [149]. Some of them founded the Center for Humane Technology (<http://humanetech.com>) and give advice on how to take back control. The most extreme suggestion is to go "cold turkey" and delete all one's social media accounts [94]. It has been claimed that this simple action will increase productivity, reduce stress and improve overall wellbeing.

A more balanced approach is to advocate a healthy digital lifestyle by consciously avoiding the excessive use of a smartphone, by avoiding using screens before going to bed, by turning off addictive features like notifications, by not checking work-related messages outside working hours, by exercising the GDPR right to be forgotten or to be left undisturbed, etc.

Some companies have introduced a policy not to allow their workers on the corporate network to check emails outside working hours. Sometimes it is useful to observe what insiders do; a number of high-profile executives at internet companies have admitted that they put serious restrictions the use of social media and mobile devices for their own children.

However, at the same time, many schools are intensifying the use of technology as part of the learning process, for example by introducing MOOCs and flipped classroom courses, by using learning platforms that need to be used by children and students for their homework in the evening. According to an OECD study [194], the results are mixed at best. Students who use computers moderately at school tend to have somewhat better learning outcomes than students who use computers rarely. But students who use computers very frequently at school do a lot worse in

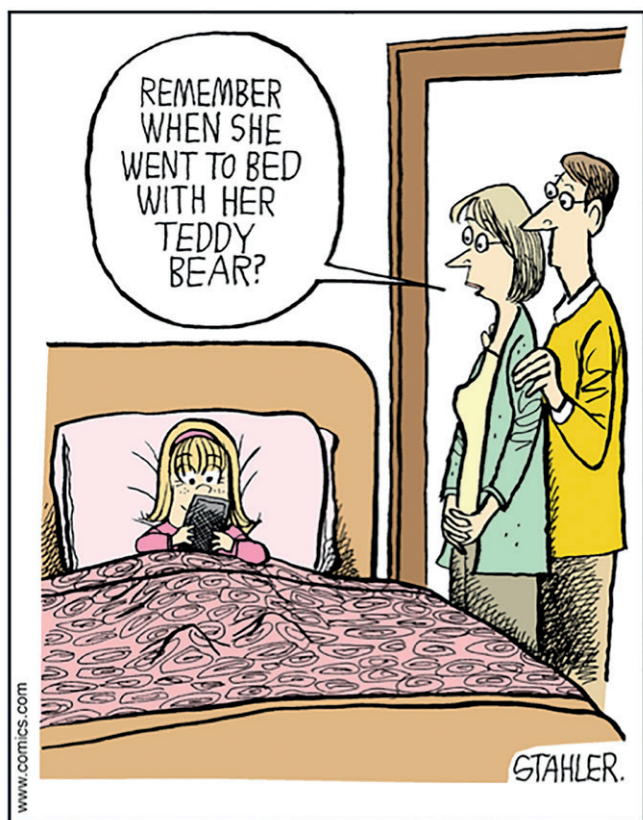


Figure 137: Cartoon by Jeff Stahl

© 2011 Jeff Stahl/ Dist. by UFS, Inc.

most learning outcomes, even after accounting for social background and student demographics. Time will tell whether the benefits of technology outweigh the side effects on the development of the children.

“I don’t have a kid, but I have a nephew that I put some boundaries on. There are some things that I won’t allow; I don’t want them on a social network.”

Tim Cook, CEO of Apple



Figure 139: Logo Centre for Humane Technology

“Technology is hijacking our minds and society.

Our world-class team of deeply concerned former tech insiders and CEOs intimately understands the culture, business incentives, design techniques, and organizational structures driving how technology hijacks our minds. Since 2013, we’ve raised awareness of the problem within tech companies and for millions of people through broad media attention, convened top industry executives, and advised political leaders. Building on this start, we are advancing thoughtful solutions to change the system.

Why is this problem so urgent?

Technology that tears apart our common reality and truth, constantly shreds our attention, or causes us to feel isolated makes it impossible to solve the world’s other pressing problems like climate change, poverty, and polarization.

No one wants technology like that. Which means we’re all actually on the same team: Team Humanity, to realign technology with humanity’s best interests.”

Center for Humane Technology

2.6.3 COMPUTING TECHNOLOGY AND THE FUTURE JOB MARKET

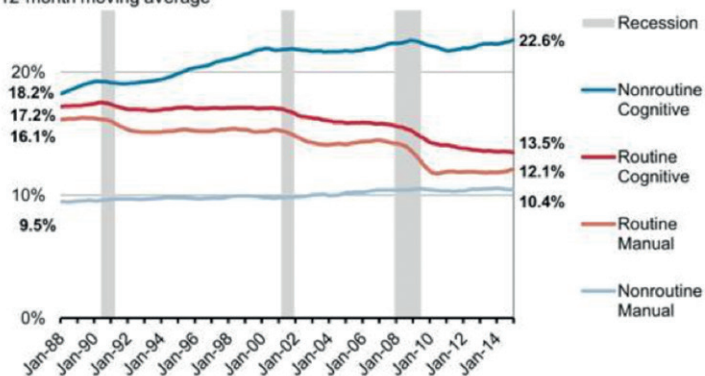
Computing, by definition, has an impact on the job market. The introduction of automation destroys jobs, creates new ones and changes the content of the remaining jobs. This has always been the case since automation was invented. The key question many people have been focusing on is whether the current wave of automation fuelled by artificial intelligence and robotics will create more or fewer jobs than it destroys. As of today, there are no signs that there are fewer jobs than, for example, 20 years ago. On the contrary, there have never been more people employed than today – which can only be partially explained by the fact that there have never been more people/consumers than today.

In addition, several countries have reached the point where there are more open positions than available candidates who are qualified to fill them. The techno-optimists see this as a sign that the fourth industrial revolution is creating more jobs than it destroys – as was the case for the previous industrial revolutions [115]. Studies indeed show that the western economies have by now recovered from the great recession in 2008, in terms of number of lost jobs that have been recreated [49].

Techno-pessimists argue that the labour market is complex, and that although the numbers look promising, a deeper analysis reveals that the jobs created are quite different from the jobs that were destroyed [12, 13] and that computing is transforming the job market in fundamental ways.

Decline of Routine

Percentage of the population in jobs that have been identified as routine and nonroutine, 12-month moving average



Source: Henry Siu and Nir Jaimovich for Third Way | WSJ.com

Figure 140: Percentage of people in jobs identified as routine and nonroutine

Source: Henry Siu and Nir Jaimovich, WSJ.com

The jobs that are destroyed are mostly routine jobs (manual and cognitive). The newly created jobs are mostly non-routine cognitive jobs, and, to a lesser extent, non-routine manual jobs. Routine jobs are jobs that are standardized, and that need a specialized but limited skillset. The typical routine manual job is a factory worker job. The typical routine cognitive job is an administrative job. Many such *medium-skilled* jobs were destroyed

in the 2008 recession. Some of the routine manual jobs came back when the economy revived, and manufacturing needed to increase production volume. The routine cognitive jobs did not come back: once administrative processes have been digitized, the associated jobs are gone forever. This trend is seen worldwide. Even in low-income countries (countries in which the average income per person is less than \$1.90 per day), medium-skilled jobs are being replaced by low-skilled and high-skilled jobs.

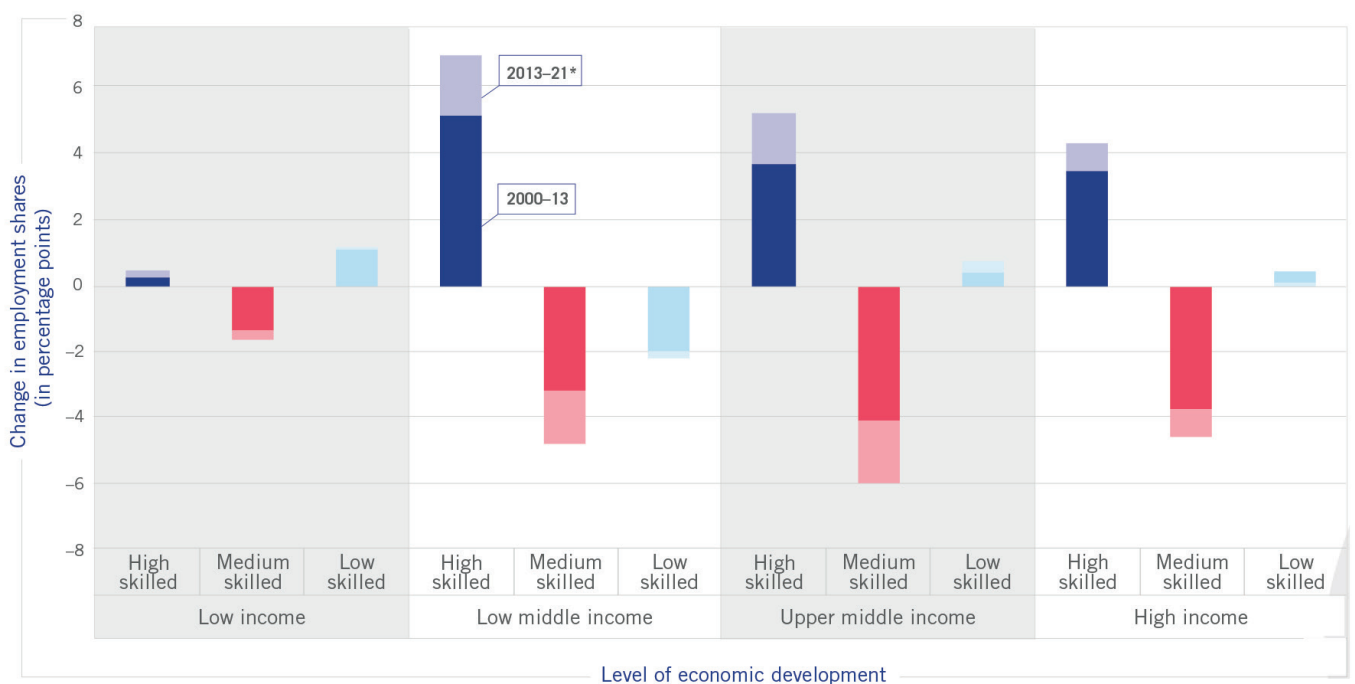
The loss of jobs is, however, largest in middle income countries. This is no surprise. These countries manufacture a lot of goods for the global economy; as they develop, their labour becomes more expensive, and automation is used to stay competitive with the cheaper labour in the low income countries. Comparing the number of destroyed and created jobs is a simplistic way to assess the impact of computing.

What happens in reality is that routine tasks within jobs are being digitized, and this is something that happens in all jobs. It is only when the remaining part no longer justifies the cost of an extra worker that the job disappears. In many cases, the workers will be given other tasks within the organization, or they might not be replaced after leaving the organization. This incremental process explains why 73% of Americans believe that artificial intelligence will eliminate more jobs than it creates, but 72% thought it was “not likely” or “not likely at all” they would lose their own job in the next 20 years [165].

It turns out that this destruction and creation of jobs also generates demographic shifts. Studies in the USA show that in the last decade, many jobs that were traditionally held by male workers (such as factory jobs) have been replaced by jobs taken by female workers (such as those in healthcare), and that the traditional white male worker has found it harder to overcome the effects of the big recession of 2008. This might however change in the future. Many non-routine cognitive jobs require science, technology, engineering and mathematics (STEM) competences. Since women are underrepresented in STEM studies, their participation in highly paid non-routine cognitive jobs might shrink in the future.

It also turns out that medium-skilled jobs that are lost in the rural areas are often replaced by new jobs in the cities [63].

The growth of high-skilled jobs also has an impact on the required level of education. The more advanced problem solving skills one has, the easier it is to find a job. The best guarantee for securing employment is a university degree. University degree holders form the only group that have fully recovered from the 2008 recession and have even seen an increase in their income [353]. Jobs that were traditionally done by middle-class workers without university degrees are now done by workers with one or more university degrees. There is no reason to assume that this trend is going to change in the future.

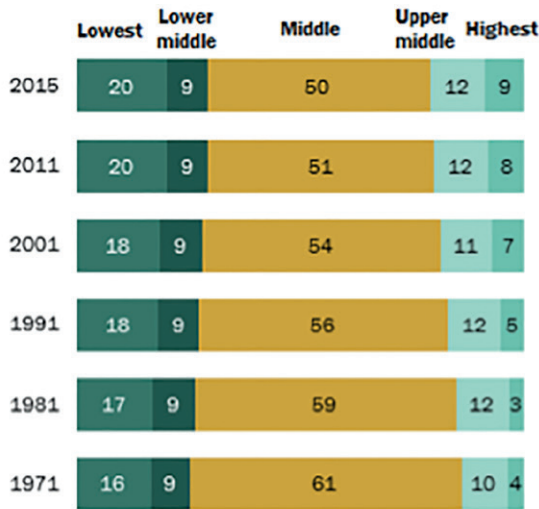


Notes: Change in employment shares, in percentage points. * Forecasts after 2016.

Figure 141: The impact of technology on the quality and quantity of jobs
Source: ILO Trends Economic Models

Share of adults living in middle-income households is falling

% of adults in each income tier



Note: Adults are assigned to income tiers based on their size-adjusted household income in the calendar year prior to the survey year. Figures may not add to 100% due to rounding.

Source: Pew Research Center analysis of the Current Population Survey, Annual Social and Economic Supplements

PEW RESEARCH CENTER

Figure 142: Pew social trends: the American middle class

But even a university degree is no guarantee for job security. Stable, long-term employment with a single employer is no longer the norm and temporary unemployment or underemployment is no longer exceptional. Future workers might be jobless, freelance, employed or entrepreneurs at different stages of their career. The current education system does not adequately prepare the next generation to deal with this future because it is still training millions of people for routine jobs in large organizations.

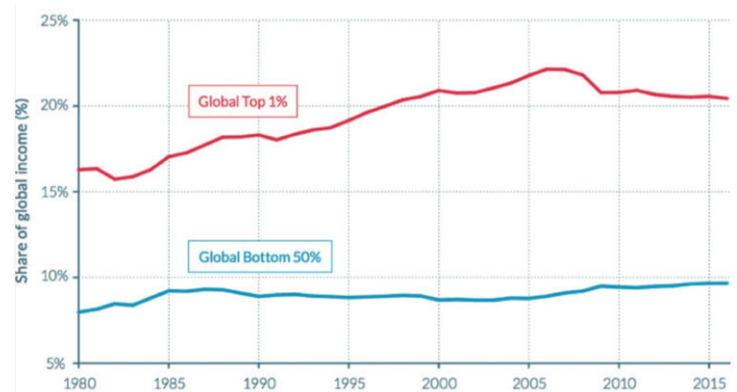
An increasing number of people are being employed in the so-called "gig economy" [148], sometimes also called the platform economy because the work is distributed piece-by-piece via a platform (Uber, AirBnB, Lyft, Blabla Car, Nubelo, Amazon Mechanical Turk, Task Rabbit, YoupiJob, Frizbiz, etc) or work on zero-hour contracts. The gig economy is growing faster than the traditional economy, which means that an increasing number of people do this kind of precarious work in which they have no protection at all.

In 2017, the Employment Appeal Tribunal in the UK decided that "Uber drivers are considered to be workers when they have the Uber app switched on due to the level of control exerted by the company over its drivers" [456]. As a result, Uber drivers are entitled to receive (i) the national minimum wage, (ii) protection from unlawful deduction from wages, (iii) paid annual leave, (iv) a working week of at most 48h, (v) protection to make disclosures

under the whistleblower legislation. At the time of writing, Uber was appealing the decision. The expectation is that more companies with dependent self-employed workers will face similar claims.

The biggest losers in this transition are middle-class workers. According to [455], the number of people living in middle-income households has been steadily declining since 1970.

This is leading to growing inequality and a polarized job market, which is a trend which is being noticed across the world and is attributed to the introduction of automation. Today, 1% of the richest people own 50% of the global wealth and 20% of the income. The eight richest people in the world own as much wealth as the poorest 50% [410, 411].

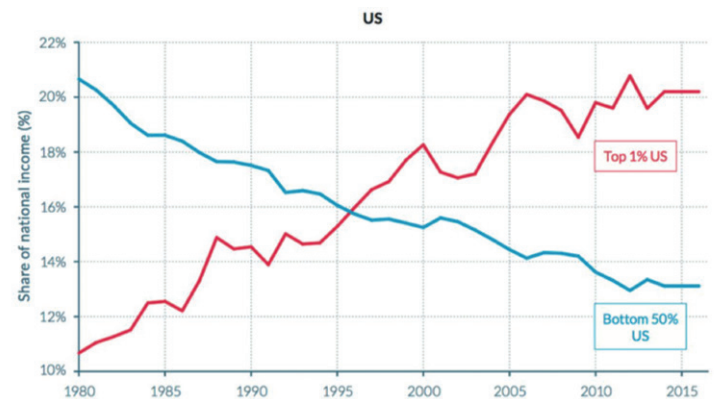


Source: WID.world (2017). See wii2018.wid.world for data series and notes. In 2016, 22% of global income was received by the Top 1% against 10% for the Bottom 50%. In 1980, 16% of global income was received by the Top 1% against 8% for the Bottom 50%.

Figure 143: Share of global income of top 1% and bottom 50% of the world population

Source: WID.world

The increase in income inequality is particularly pronounced in the USA. The bottom 50% has experienced a steady decrease in their income over the last 35 years.



Source: WID.world (2017). See wii2018.wid.world for data series and notes. In 2016, 12% of national income was received by the top 1% in Western Europe, compared to 20% in the United States. In 1980, 10% of national income was received by the top 1% in Western Europe, compared to 11% in the United States.

Figure 144: Share of national U.S. income of top 1% and bottom 50% of the world population

Source: WID.world

In combination with a shrinking middle class, this growing inequality might lead to societal polarization (since the challenges for the wealthy and for the poor are quite different), political problems, and economic stagnation [34, 57, 158, 160].

2.6.4 COMPUTING TECHNOLOGY AND FUTURE OF EDUCATION

Today's globalized world is being described as VUCA: volatile, uncertain, complex and ambiguous. Change is accelerating, there are no longer lifelong guarantees, especially when it comes to employment. Furthermore, there are a number of global societal challenges that need to be solved in the coming decades like (i) how to support 10 billion middle class people on one planet by 2075, (ii) how to completely decarbonize the economy by 2075, (iii) how to support an ageing global population, (iv) how to come up with economic models that take long term sustainability into account and so on. See 2.6.6, "Computing technology and Planet Earth", for a fuller discussion of such challenges and how ICT can help to address them.

The children and young people that are at school, college or university today are the ones who will have to support a family in this VUCA world, and who will have to find solutions for the associated challenges. This immediately raises the question of what they should learn at school in order to be ready for this world, and to tackle these challenges. Since the future is uncertain, it is hard to make predictions, but for education there are a number of things we know.

- 1 At the competence level, study programs should focus on the eight key competences for lifelong learning [285] as adopted by the European Parliament in 2018.

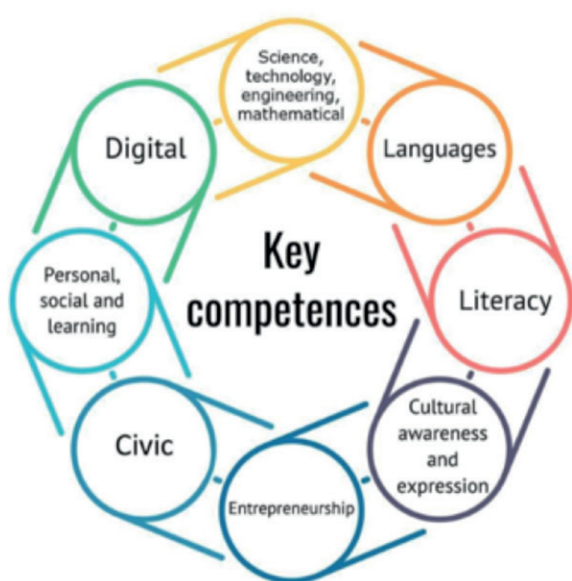


Figure 145: Key competences for lifelong learning
Source: European Commission

Notable 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 must make Europe more competitive. The combination of cultural awareness and expression competence with civic competences should provide all Europeans with a common framework for values, democracy, globalization, multi-culturalism. Finally, literacy and (foreign) languages are important as a means to learn, listen, and express ideas. These eight key competences are essential 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, i.e. the dichotomy between the humanities and the sciences, and the dichotomy between pure and applied training [114].

- 2 At the content level, it is clear that formal education will not be able to provide all the knowledge that one needs for a whole life (especially since we cannot train students for jobs that still need to be invented). Furthermore, all knowledge has a half-life (facts, business models, even secrets). According to [443], the half-life of an engineering degree, for example, is at most five years. This means that some of the engineering knowledge students acquire in the first year of a five-year engineering course is already obsolete by the time they graduate. It is not a coincidence that many technology companies have a median worker age below 35. Therefore, future study programs should not focus too much on teaching solutions (which are by definition changing), but instead focus on the basic principles of the discipline, which have a much longer half-life. Furthermore, it are the basic principles that will be needed to come up with future outside-the-box solutions. In addition, workers will have to compensate for the decay of their knowledge by continuing to learn throughout their lifetimes (lifelong learning), and to keep working on the development of their competences in order stay attractive in the job market until retirement age and beyond.
- 3 Graduates should be at least T-shaped. This means that they 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, coordinating with others, emotional intelligence, judgement and decision-making, service orientation, negotiation and cognitive flexibility [10]; that is, competences that set humans apart from computers and robots.

The deep knowledge and skills part must encourage the student to learn how to take 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

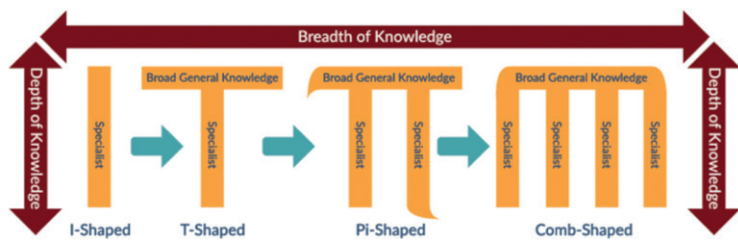


Figure 146: I-Shaped vs T-Shaped Professionals

Source: DevOps Institute

part, the more they will learn, and the better adapted they will be to tackle the technical challenges of the 21st century.

One thing is for sure: there is little value in specialist training to do routine tasks (so-called I-shaped profiles), as routine jobs are disappearing. T-shaped education offers a best guarantee for self-fulfilment, happiness and a good life.

- To help digital society develop, all students should get a basic understanding of computing, big data analytics and artificial intelligence (on a par with a basic understanding of sciences, history, one or more foreign languages). Globally, there is a shortage of millions of ICT workers to tackle all the challenges ahead of us (digitizing industry, securing ICT systems, designing smart grids for the transport of renewable energy, the development of precision agriculture to reduce the use of pesticides and irrigation, and so on). At this moment, there is a big deficit in this area in many study programmes in higher education in Europe.

Finally, education should also educate the next generations about the grand challenges of the 21st century, and provide hope that these can be tackled if we are willing to change our unsustainable habits and collaborate to find sustainable solutions that benefit the global population.

2.6.5 COMPUTING TECHNOLOGY AND THE FUTURE OF EUROPE

In relation to the rest of the world, Europe as a continent is generally depicted and perceived as “the old world”. This connotation was coined in contrast to what Europeans perceived as a “new world”: the Americas, especially North America. Today, the old world could also be understood literally: Europe is the continent with the oldest population and today faces challenges that other continents will face later in the 21st century.

When this “old-new” contrast started applying to information technology and its ramifications, being “old” started to be associated with not being ready, willing, able, bold, and visionary enough to be at the forefront of innovation. Ironically, the example of China, which has recently earned an international reputation of being and wanting to be a fast-paced innovator, shows that one can be old (in terms of history and wisdom) and new (in terms of vision and energy) at the same time.

Most people in the West still have the 20th century world views they learned in school. They do not realize how drastically the world has changed over the last fifty years. Europe is still one of the best places on earth to be born, but many countries are catching up quickly. Relatively recently, Europe started to realize that it is losing ground compared to the rest of the world in a number of technology domains, including high-performance computing, artificial intelligence, cybersecurity and renewable energy.

In order to stay ahead, we need more than funding programmes. It is necessary to inculcate a completely different view of entrepreneurship combined with more curiosity in, openness to and perhaps even thirst for innovation in the public opinion. This change requires favouring and promoting the perception that technology-enabled innovation can be a powerful vector for improving our collective and individual wellbeing.

This section discusses some domains in which Europe needs to invest in order to keep or regain a leading global position.

2.6.5.1 HIGH-PERFORMANCE COMPUTING

High-performance computing (HPC) allows the simulation of military devices and planes, cars, pharmaceutical products and many more things. Many scientific discoveries are made through the so called “fourth paradigm” [206]. Rather than physical experiments or mathematical models, the fourth paradigm starts from massive datasets. Many of the innovations made by Google, Facebook, Amazon and the like are made possible by the combination of access to massive datasets and to powerful computing. Countries need a powerful computing infrastructure to be able to compete in science and research. This is as important as access to raw materials for the manufacturing industry.

The USA is currently the dominant provider of computing solutions with CPUs (Intel) and GPUs (NVIDIA), which are used to build high-performance computing and servers, as well as increasingly being used for developing solutions based on artificial intelligence. Components can be banned from export under a number of US regulations, such as the International Traffic in Arms Regulations (ITAR), which controls the export and import of defence-related articles and services. The US Department of Commerce prevented Intel and NVIDIA (but also AMD and IBM for their processors and HP for its optoelectronic devices) from shipping the processors required for the upgrade of the Chinese Tianhe-2 supercomputer, citing concerns over nuclear weapons-related research [105].

As a result, all major countries want to control a large part of their ICT infrastructure to avoid being blocked in their development by other countries. Consequently, China developed, over the span of only three years, a completely new system, including a very energy-efficient computing chip. The resulting supercomputer, the Sunway TaihuLight, reached the top of the TOP500 list of most powerful supercomputers on the LINPACK benchmark in



Figure 147: Sunway TaihuLight
Source: Xinhua

June 2016 with 93 petaflop/s (quadrillions of calculations per second) [197]. It superseded the Tianhe-2, which was the first-placed supercomputer in the previous six TOP500 lists.

Japan is also aiming for exaflop computing, and the “post-K” computer, designed by Fujitsu, will similarly use a processor made in Japan, based on Arm architecture (the previous architecture supported by Fujitsu was based on the SPARC architecture).

In June 2018, the US Department of Energy’s Oak Ridge National Laboratory (ORNL) announced that the US supercomputer Summit would have a peak performance of 200,000 trillion calculations per second, or 200 petaflops peak performance, meaning that the USA regained the top spot on the top500. The Summit system is built around an IBM Power9 22C at 3.07GHz and NVIDIA Volta GV100 GPU. The Chinese supercomputer Sunway TaihuLight took the second spot on the list, while third place went to a reduced version of Summit (1,572,480 cores instead of 2,282,544 for Summit).

It is interesting to see that the Sunway was built around custom processors, while Summit was built around a processor which is very efficient for data processing and management and a lot of GPUs as accelerators. It is foreseen that the future HPC machine will not only have simulation loads, but also more loads based on high-performance data analytics (HPDA), and also that applications will use more and more artificial intelligence-based solutions.

We observe that in the short time since the last HiPEAC Vision, some countries have gone from having intentions to having real plans and fully operational systems. Their architecture is either based on brand-new designs (like the Chinese ShenWei SW260), on MIPS (Russian Baikal-T1) or on Arm (Japanese future Fujitsu chip for HPC or Chinese FT-2000/64). China, Russia, Japan and



Figure 148: Summit Supercomputer
Source: IBM

India are actively developing processors either for desktop computers, servers, HPC, or even embedded devices. Open-source hardware processors like RISC-V are also attracting a lot of interest. Regardless of whether this is related to the revelations of Edward Snowden or not, there is a growing movement away from well-established US computing platforms, such as those of Intel, Google, Apple and Microsoft, either to avoid bans on accessing core components, or because of fears that hardware and software might have spyware deeply implanted.

AMERICA'S MOST POWERFUL SUPERCOMPUTER IS A MACHINE FOR SCIENTIFIC DISCOVERY.

- The US Department of Energy's Summit supercomputer enables scientists to simulate complex physical systems and make predictions critical to advancing research and development.
- Summit's "smart" architecture merges GPU acceleration and dense local memory to support expanding applications in data science and artificial intelligence.

<p>A 200-petaflop machine, Summit can perform 200 quadrillion (peta) floating point operations per second (flops). If every person on Earth completed one calculation per second, it would take 305 days to do what Summit can do in 1 second.</p>	<p>At over 340 tons, Summit's cabinets, file system, and overhead infrastructure weigh more than a large commercial aircraft.</p>
<p>For some AI applications, researchers can use less precise calculations than flops, potentially quadrupling Summit's performance to exascale levels, or more than a billion billion calculations per second.</p>	<p>Occupying 5,600 sq. ft. of floor space, Summit could fill two tennis courts.</p>
<p>Summit is connected by 185 miles of fiber optic cables—or the distance from Knoxville to Nashville, Tennessee.</p>	<p>Summit's file system can store 250 petabytes of data, or the equivalent of 74 years of high-definition video.</p>
<p>More than 4,000 gallons of water pump through Summit's cooling system every minute, carrying away about 13 megawatts of heat.</p>	

Figure 149: Summit Supercomputer
Source: ORNL

2.6.5.1.1 European Processor Initiative

To meet the global HPC challenge, Europe has begun a strategic initiative to support the next generation of computing and data infrastructures with a European project of the size of Airbus in the 1990s and of Galileo in the 2000s. EU efforts are synchronized in the establishment of the EuroHPC Joint Undertaking, a legal and funding entity which will enable the pooling of EU and national resources on high-performance computing to acquire, build and deploy in Europe the most powerful supercomputers in the world.

The European Processor Initiative (EPI) is one of the cornerstones of this EU HPC strategic plan. EPI brings together 23 partners from 10 European countries with the aim of bringing to market a low-power microprocessor. EPI will ensure that the key competence of high-end chip design remains in Europe, a critical point for many application areas. Thanks to these new European technologies, European scientists and industry will be able to access exceptional levels of energy-efficient computing performance. EPI aims to benefit Europe's scientific leadership, industrial competitiveness, engineering skills and knowhow, and society as a whole.



Figure 150: European Processor Initiative
Source: BSC

The design of a novel HPC processor family would not be sustainable without thinking about possible additional markets that could support such long-term activities. Thus, EPI will cover other areas such as the automotive sector, ensuring the overall economic viability of the initiative. One specific objective for the automotive sector is to develop customized processors able to meet the performance needed for autonomous cars.

EPI brings together experts from the HPC research community, major supercomputing centres, and the computing and silicon industry as well as the potential scientific and industrial users. Through a co-design approach, it will design and develop the first European HPC systems-on-chip (SoCs) and accelerators. Both elements will be implemented and validated in a prototype system that will become the basis for a full exascale machine based on European technology.

EPI will provide European industry and research with a world-class, competitive HPC platform and data processing solutions

which consider the interests of data security and ownership. The initiative aims to achieve unprecedented levels of performance at very low power, and EPI's HPC and automotive industrial partners are already considering the EPI platform for their product roadmaps.

2.6.5.2 SECURITY

The situation in countries bordering the European Union is definitely less stable now than it was a decade ago. Whereas until a couple of years ago, the European Union acted as if it could ignore problems outside of its borders, recent history shows that they are increasingly affecting internal European affairs; for example, an unstoppable stream of refugees from the Middle East and Africa trying to enter the European Union, and an unstable political situation in Turkey that is being exported to some European countries. Some political parties use war rhetoric in order to mask their inability to address these issues directly, and to build support for more investments in internal security.

In parallel with the increase in physical threats, there has also been a surge in cyber-attacks [5]. This is a logical consequence of the fact that a large part of modern society has critical dependence on its cyber infrastructure (banking, communication, businesses and utilities to name but a few). Stealing information is now as lucrative as robbing a bank (68% of funds lost as a result of a cyberattack turn out to be unrecoverable), only less dangerous for a robber because it can be done at a distance. Disrupting a global cyber infrastructure can have a serious impact on society and on the economy. Disclosing classified information can have serious political consequences as demonstrated by the multiple *-leaks incidents like Wikileaks.

2017 COST OF CYBER CRIME STUDY FROM ACCENTURE AND PONEMON INSTITUTE



Figure 151: 2017 cost of cybercrime study
Source: Accenture and Ponemon Institute

Many people are amazed at how apparently easy it is to hack the email servers of political parties, and to bring down government and company websites. Governments are increasingly worried about attacks by organized crime (including terrorists), and state-sponsored attacks. The FBI even keeps a list of "most wanted" cybercriminals [356]. The US Navy receives more than 100,000 cyber-attacks per hour. Cybercrime incurs a cost of several million

euros per year for major corporations and governments, and hence it weighs on the economy. The website of the Norse Corporation [236] has a good visualisation of the global cyber war which is taking place 24/7.

According to Quora [209], the top 10 cyber army superpowers are: the USA, Iran, China, Israel, North-Korea, Russia, Canada, the UK, Germany and India. Different countries specialize in different areas: the USA focuses on defending its own infrastructure and attacking its enemies; China specializes in spying; Israel exports more cybersecurity products than all other countries in the world combined; North Korea specializes in financial hacking; Russia in political hacking; etc. The fact that the top six countries are not particular friends is not reassuring.

After 30 years of cutting down military investments in Europe, it has become clear that this trend will come to an end. The USA is demanding higher European contributions to NATO, and European countries are starting to realize that they will have to invest more in a European defence system. Currently, many European countries do not support military research with their funding instruments. While this is a principled stance to take, the question is whether such a position is in the interest of Europe. Instead of buying American weapons systems, Europe could also develop and buy its own systems.

2.6.6 COMPUTING TECHNOLOGY AND PLANET EARTH

There is overwhelming evidence that sustainability is the mother of all long-term societal challenges. Sustainability has many definitions, but a very concrete one is “all that needs to be done to make sure that *homo sapiens* can continue to thrive on Earth for the next 10,000 years”, as it has done since the introduction of agriculture 10,000 years ago.

There are a few boundary conditions for sustainable development. The first is population. At the global level, the average number of children per woman was 2.5 in 2015. Even in the most populated continent, Asia, it is now as low as 2.17, which is the replacement rate. Only Africa reports a much higher fertility level, but this is now dropping as fast as it did in the rest of the developing world in the 1980s. In a few decades, Africa will probably join the rest of the world with a fertility rate of two children per woman or fewer.

This evolution has taken place over the last 50 years and can be attributed to improved living conditions and lower infant mortality rates. As soon as a country transitions from a poor country to a (lower) middle-income country, the fertility rate automatically drops to around the replacement rate, irrespective of culture, religion, political system, or other factors.

However, in the 21st century, the population will continue to grow, but only due to the so-called “fill-up” from increased life expectancy. That means that there is currently nothing that can be done to immediately halt population growth.

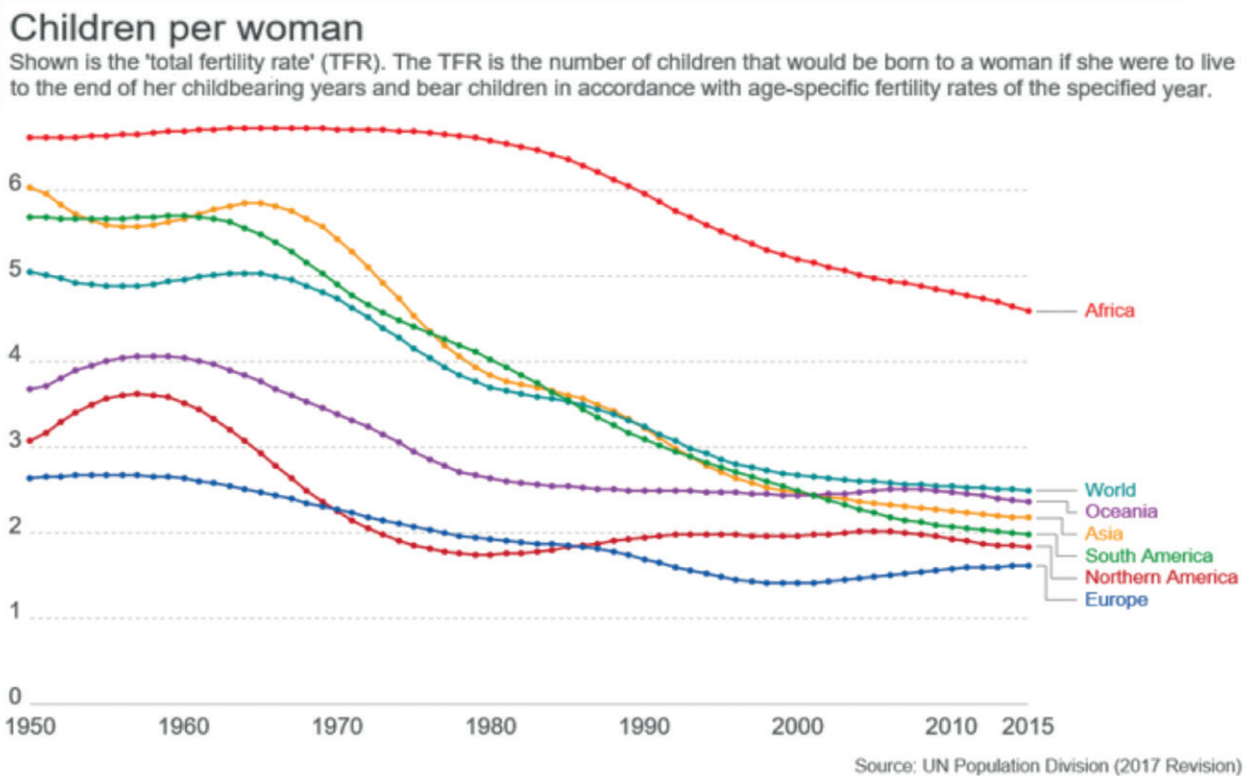


Figure 152: Evolution of the number of children per woman – Source: UN Population Division 2017

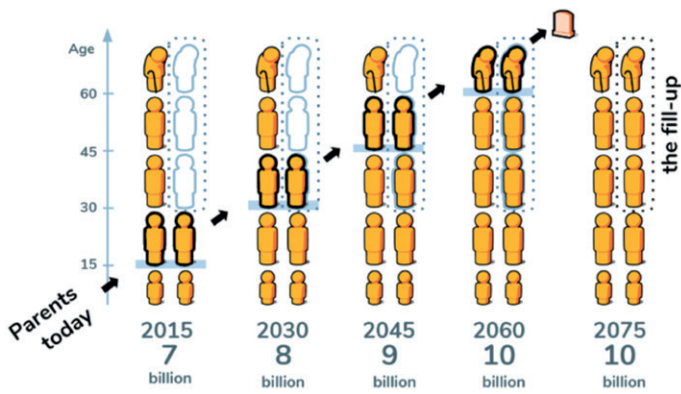


Figure 153: Evolution of the world's population

Source: [gapminder.org/factfulness](https://www.gapminder.org/factfulness)

With current life expectancy, the peak global population will be 10 billion. But every extra year in life expectancy will eventually add around 100 million people to the global population. In the longer term, the world's population can only be reduced peacefully by bringing down the average number of children per family to fewer than two. This will only happen when more countries have few to no people living in extreme poverty.

In 2009, 25% of the global population was said to belong to the middle class, half of them living in the USA and in Europe. In 2030, over half of the world's population is predicted to belong to the middle class, and two thirds of them will live in Asia [121]. That means that between 2009 and 2030, more than three billion people will join the middle class. This will have a profound impact on the distribution of global gross domestic product (GDP). One day, the membership of the G7 might have to be reconsidered.

A second boundary condition is the ecological footprint of modern society. Today, the European Union (EU) has an ecological footprint that is about twice the biocapacity of its surface area [439]. This means that the EU currently uses two Europes to support its lifestyle. It also means that Europe depends on solid trade and a good relationship with a sufficient number of countries willing to share their resources with us, even if they are scarce. Some might one day decide to keep them for their own population, or create an artificial shortage in order to increase prices. Hence, it is in the interest of Europe to stay within the biocapacity of the continent.

At the global level, 1 August 2018 was the "Earth Overshoot Day" of 2018, which means that the world population had consumed all renewable resources of Planet Earth on that day (for example, all the wood that will grow in 2018, or all the rainwater that will be captured in 2018), and that for the rest of the year, we are using historical reserves (for example by clearing longstanding forests, or by pumping fossil water).

We currently consume the renewable resources of 1.7 Earths per year; in budgetary terms, we have a deficit of 70% on the yearly ecological budget. This is obviously not sustainable; indeed, it is considered the root cause of all environmental problems, of which climate change is only one (loss of biodiversity being another) [169]. A large part of the ecological footprint is due to the use of fossil fuels. Recent efforts to stimulate green energy and recycling seem to have had an effect on the evolution of Earth Overshoot Day (even with a growing global population). The goal is to push it back to 1 January.

Share of world GDP

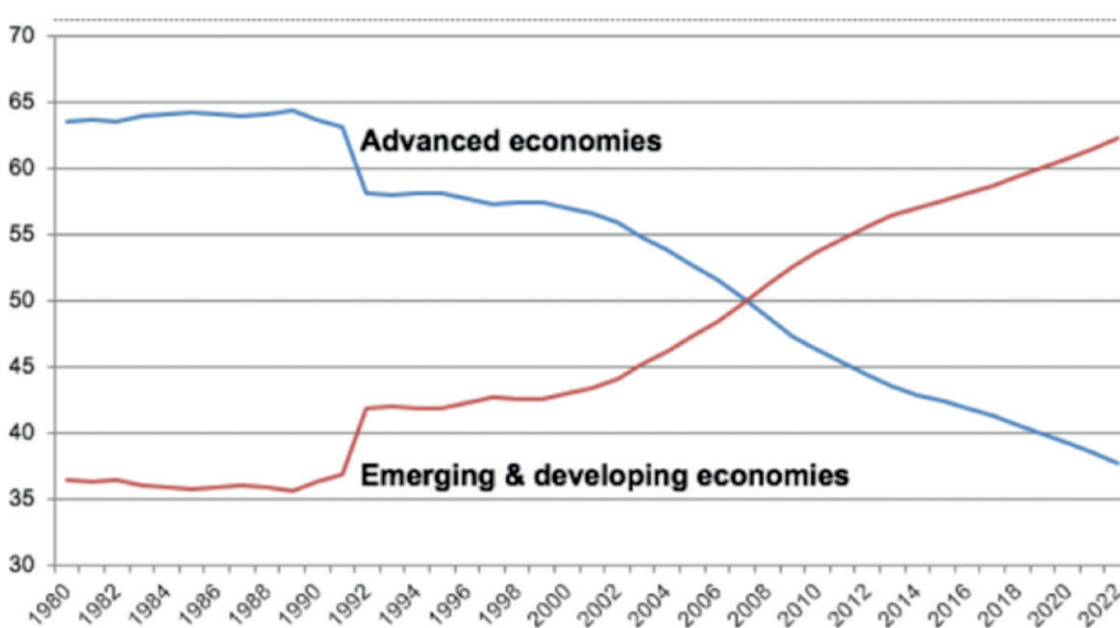


Image: VoxEU

Figure 154: Evolution of World GDP – Source VoxEU

Past Earth Overshoot Days

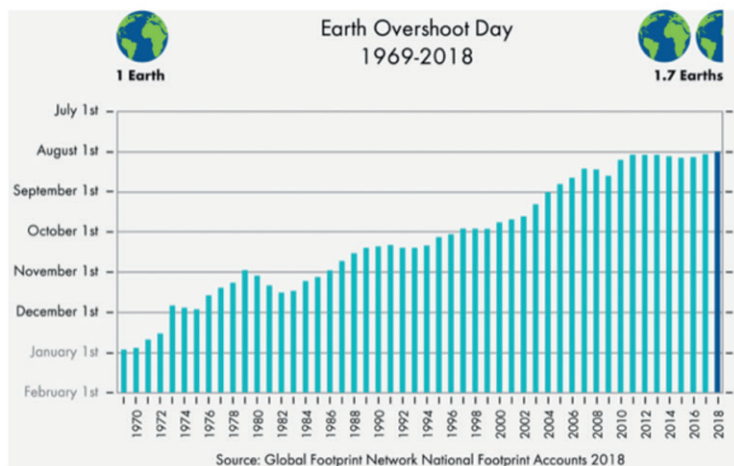


Figure 155: Earth Overshoot Days

If we want to keep the planet inhabitable for 10 billion people, we must control climate change [120]. United Nations Secretary-General António Guterres summarized the current situation very well in his 2018 New Year's address when he said that "climate change is moving faster than we are". The only known solution at this moment is to completely decarbonize the economy and start capturing carbon from the air by the end of the century.

The graph below shows who is emitting the CO₂ emissions worldwide. The richest 10% of people – mainly living in advanced economies – are responsible for 49% of total CO₂ emissions. The mechanism is clear: the higher the income, the higher the level of consumption, and the larger the carbon footprint.

The carbon footprint in this graph is based on consumption, not on production. That means that the carbon footprint of products manufactured in poor countries but consumed in rich countries are part of the carbon footprint of the rich countries. It is clear that rich and middle-income countries have a large responsibility to cut down their carbon footprint. At the current CO₂ production

Percentage of CO₂ emissions by world population

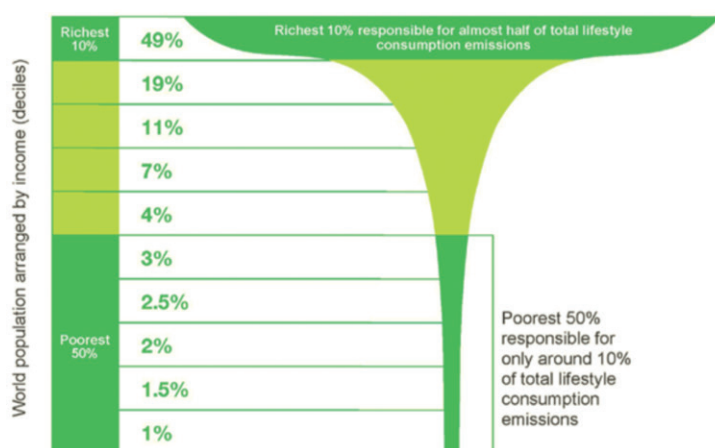


Figure 156: Percentage of CO₂ emission by world population

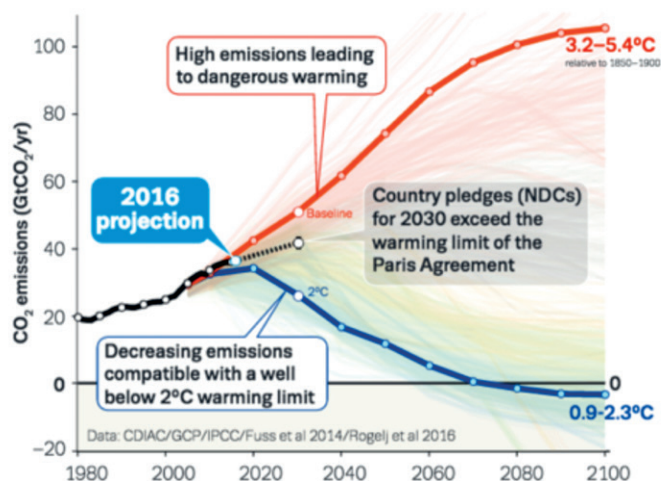


Figure 157: Emissions reductions will need to pick up momentum everywhere to meet the goal of limiting warming to the internationally agreed goal of staying "well below" 2°C above pre-industrial levels.

rate we can still go on for another 18 years before we will reach a 2°C increase in the global temperature [409]. Keeping global warming to 1.5°C, a difference which would have a serious impact on many biosystems and extreme weather events, would require drastic action within 12 years [408].

In order to keep the temperature increase below 2°C, carbon emissions must be reduced by 80% by 2050. In addition, it will be necessary to capture and store a staggering 810 billion tons of carbon from the air by 2100, or the equivalent of 20 years of burning fossil fuels at the current rate [45, 457].

However, there is no cheap way to capture carbon yet, and given the current growth in renewable energy production, it is unlikely that renewable energy sources will be able to replace even 50% of all fossil fuel consumption in the next two decades.

World Energy Consumption, 1965-2016

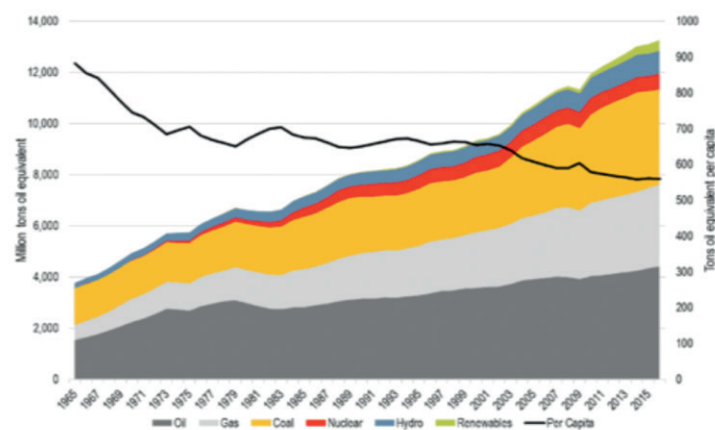


Figure 158: World Energy Consumption
Source: BP Statistical Review of World Energy. Population data from World Bank

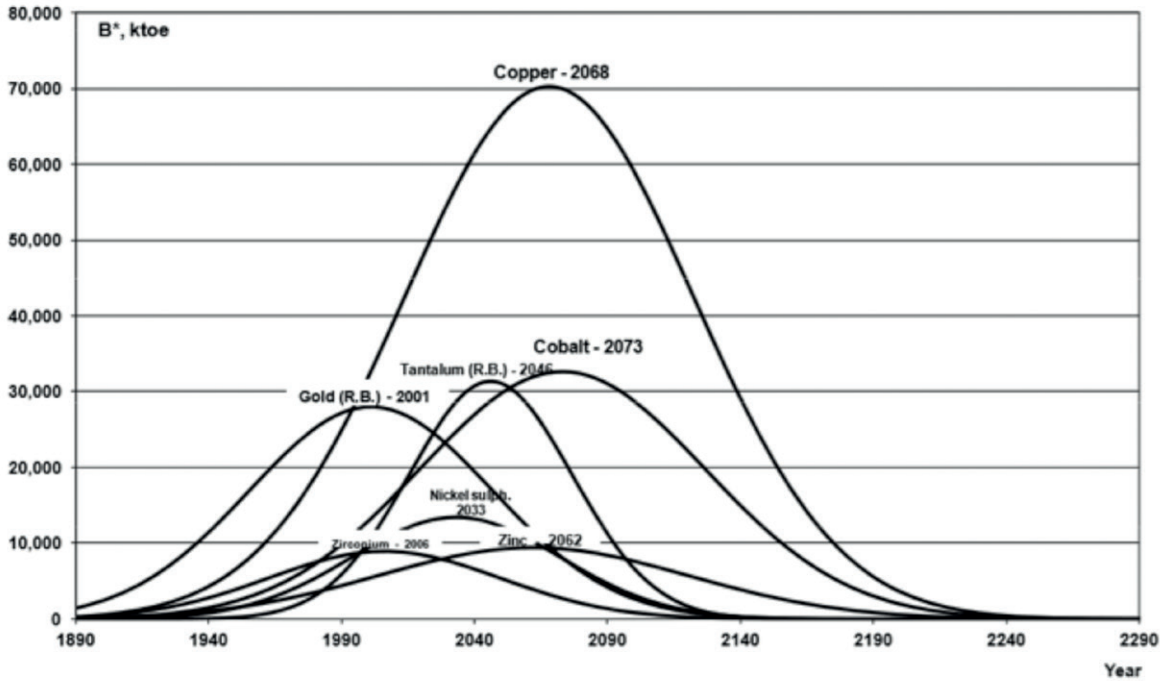
The best option for the moment is to consume less energy [44]. The sooner we start, the cheaper it will be. The richest 10% of the world has to start; the poorest 50% does not have the resources to reduce their carbon emissions.

The third boundary condition is natural resources. If we continue to run Earth as we do today, in 100 years many deposits of natural resources (minerals, fossil fuels, historical ground water deposits,

and so on) will be depleted. So the question is how *homo sapiens* will continue for 9,900 years after that.

Today's electronics rely on elements of almost the entire periodic table, including the rarest elements on earth, such as iridium [135].

ICT relies on rare materials making the European ICT supply chain very fragile and sustainability questionable if no dedicated research is developed [448].



Source: A. Valero and A. Valero (2014). *Thanatia: the Destiny of the Earth's mineral resources*. World Scientific Publishing



Figure 159: Depletion of natural resources

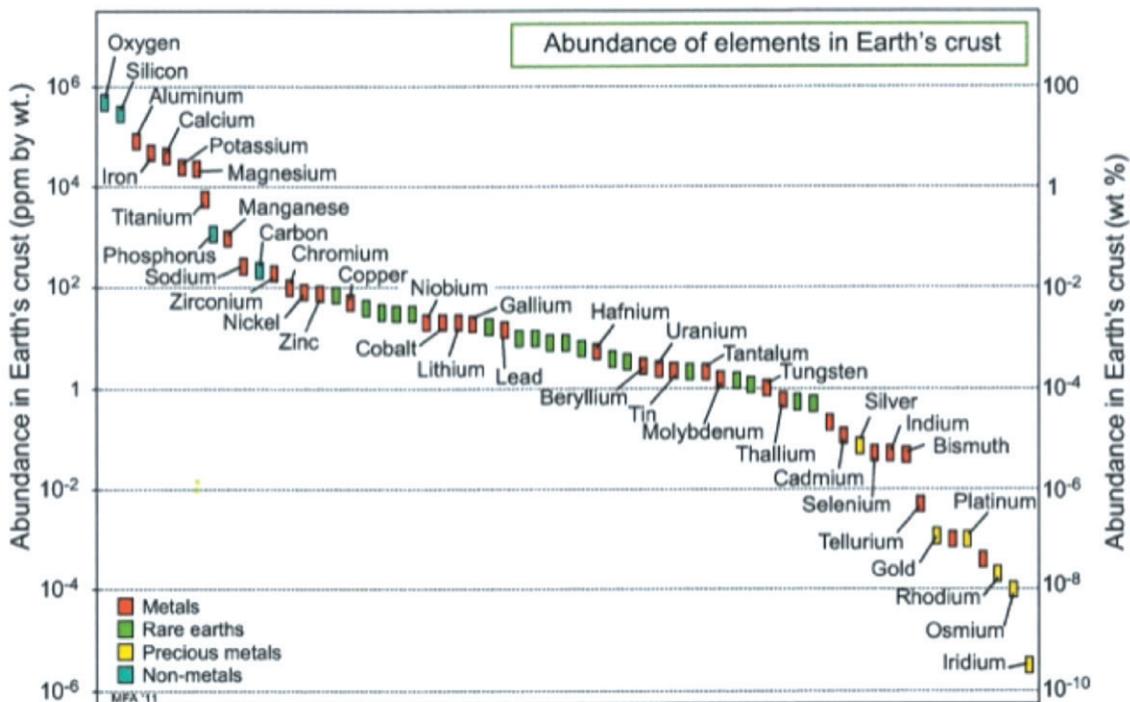
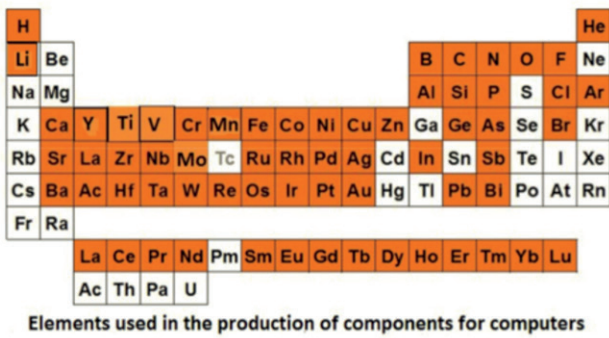


Figure 160: Elements in the Earth's crust – Source: *Materials and the environment*, Michael F. Ashby, 2011

Almost the entire Periodic Table is being used



Elements used in the production of components for computers

Source: Adapted from different sources



Figure 161: Elements used to produce electronics

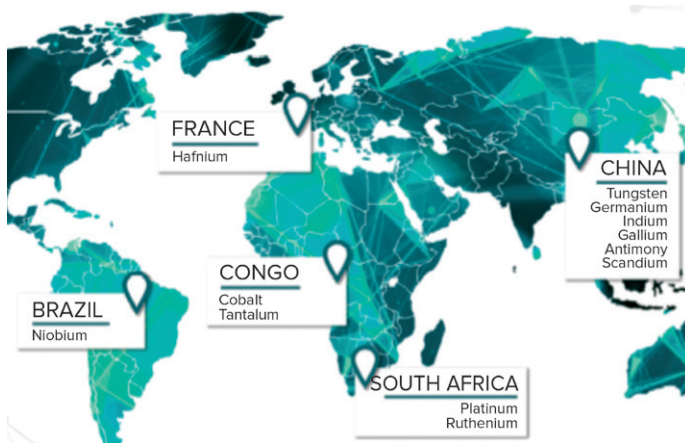


Figure 162: Rare materials that ICT relies on
Source: [448]

Everything seems to be fast, clean, reconfigurable and so on, but behind our screens there is an industry which requires, more than ever, space, energy and matter. There is an urgent need to revisit the economic, technological, and societal models to develop a sustainable electronic industry which cares about its impact right from the initial design of these objects. For example, China provides 95% of the production of rare-earth elements to Occidental countries and this monopolistic situation is a major stake for coming years. Another example is the dangerous conditions faced by African workers in cobalt mines that reinforces the attention paid to human rights issues in global supply chains.

However, no clear methodology exists today to design, manufacture and deploy the IoT in a “sustainable way” that preserves enough resources to avoid political, economic, and environmental tensions in the next twenty years. The European microelectronics industry depends on rare raw-material sourcing. European technology supplies are very fragile, and economic, strategic independence, ethical and environmental considerations are converging into a common requirement: design our technologies differently in the early stages of the research process.

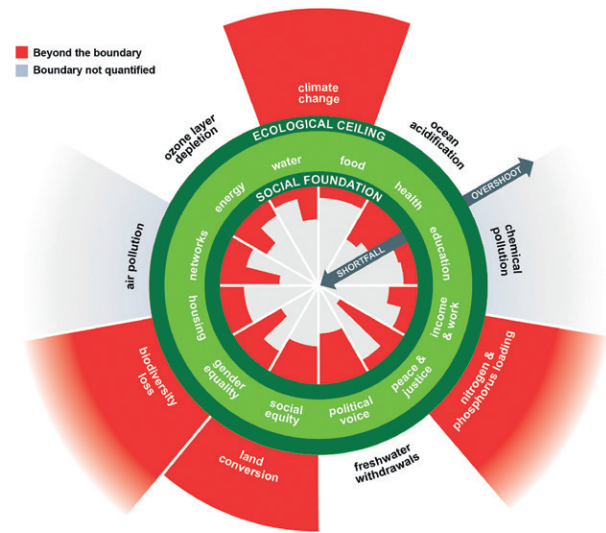


Image: Kate Raworth and Christian Guthrie/The Lancet Planetary Health

Figure 163: Doughnut economics

Source: Kate Raworth and Christian Guthrie/The Lancet Planetary Health

Sustainable design of computing technologies, from materials to systems, and corresponding sustainable business models are needed, with joint efforts from research and industry. In particular, the sustainable use of rare raw materials is an economic, social, environmental and major geopolitical stake for the current and next generations. Some key elements considered today in the emerging electronic devices for expanding markets, such as IoT, transport, connected medicine or so-called “green” energy production and storage, must be drastically substituted or saved in the near future [438].

The challenge for the 21st century is clear: how to support 10 billion middle-class people within the renewable resources of the planet we live on. This will require new economic models, such as so-called “doughnut economics”, for instance [109] which tries to balance between the social foundation (decent living for 10 billion people), and the ecological ceiling (the biocapacity of Planet Earth).



Figure 164: Sustainable Development Goals
Source: United Nations

The 21st century urgently needs visionaries who can show society a clear path towards the 22nd century. It is unclear what this path will look like, but it is clear that it needs to lead to sustainability. In 2015, more than 190 world leaders committed to 17 Sustainable Development Goals (SDGs) to help end extreme poverty, fight inequality and injustice, and fix climate change. Achieving these goals should lead to a more prosperous, equitable, and sustainable world.

Computing should contribute towards the realization of these goals, and it should not work against any of them. It is worth pointing out that technology of any kind is simply an enhancement of human capabilities, which can then be deployed to achieve different objectives. If ICTs are to contribute to the Sustainable Development Goals, conscious policy decisions will be required to steer the implementation of ICT solutions in the direction of these goals.

A detailed analysis of how ICTs could contribute to the sustainable development goals is provided in the report [441]. Drawing on this report, below are a few examples of the role ICTs will play in achieving the sustainable development goals, along with challenges that they pose.

1. No poverty

How ICTs can contribute: Expanding the ICT infrastructure and making use of digital technologies can help fight poverty by, for example, connecting poor communities with the rest of the world, providing mobile training, increasing productivity, fighting fraud, creating financial services through mobile banking, supporting free elections and enabling new business models. Investing in the digital skills of young people empowers them to build their own digital society and help their extended families.

Challenges: To avoid entrenching existing inequalities, the so-called “digital divide” needs to be overcome, with all members of society having access to high-quality digital technologies.

2. Zero hunger

How ICTs can contribute: ICTs can help achieve sustainable and inclusive rural transformation, and help increase food production. Access to the internet brings information about weather, financial services, market information, agricultural advice, and disease control information to rural areas. Satellite imagery helps to monitor land use and water resources, and hence to optimize the agricultural production of a country and to create food security.

Challenges: Adequate ICT infrastructure will be necessary for people to take advantage of the benefits of ICTs for agriculture and distribution in poor countries.

3. Good health and wellbeing

How ICTs can contribute: E-health is becoming an increasingly important aspect of health and wellbeing. For instance, ICTs

allow information to be collected, analysed and managed more easily in all areas of healthcare. Investing in technologies such as remote diagnostics, accessible medical imaging, affordable implants and labs-on-a-chip will help provide high-quality accessible and affordable healthcare for 10 billion people. Computing is an important enabling technology for e-health.

Challenges: Deploying and maintaining advanced medical devices in environmentally harsh conditions (humidity, heat, dust, unreliable power supply and so on) is a challenge and might require customized devices.

4. Quality education

How ICTs can contribute: Online courses can be used in places where there is a lack of qualified teachers, for teacher training, or for teaching children in emergency situations. ICTs can also facilitate access to quality educational materials. The internet is a powerful tool for expanding access to knowledge, reducing learning divides, and supporting lifelong learning, and the impact of getting children online will be felt in the extended family and wider community.

Challenges: It will be necessary to equip public educational institutions with good ICT infrastructure. Special emphasis should be placed on giving women and girls access to the internet and teaching them digital skills, as they are currently underrepresented. To empower students and help them fully participate in the digital economy, ICT education should focus on producing as well as consuming ICTs.

5. Gender equality

How ICTs can contribute: Information and communication technology and access to the internet is a key enabling technology to emancipate and empower women. With increasing automation, gender equality can be promoted in society if women are better equipped to get jobs requiring technology and engineering skills, which are likely to be better paid and with better conditions. Bringing more women into the field will also contribute to reducing bias, creating technology which meets the needs of more people and increasing the ICT workforce (see 2.6.4, “Computing technology and the field of education”).

Challenges: The field of ICT is well-known for its gender imbalance, in countries of all income levels.

6. Clean water and sanitation

How ICTs can contribute: As the demand for water will grow in the future, but the supply cannot be controlled, we will have to use it more efficiently (in agriculture, in manufacturing and in the home). Technological solutions which rely on advanced computing will be crucial in this respect. They can also make wastewater treatment more efficient, and provide monitoring technology to promote water saving.

Challenges: ICT production and recycling requires lots of water.

7. Affordable and clean energy

How ICTs can contribute: Producing energy to meet human requirements without fuelling climate change is a major challenge for the 21st century. Since renewable energy such as solar and wind cannot be produced on demand, countries will need smart grids and storage facilities to balance the supply and demand on a country scale or beyond. This will require sophisticated distributed energy management systems.

Challenges: ICTs themselves are a major and growing cause of energy consumption. Their energy consumption should therefore be reduced as much as possible, and the energy savings they deliver in other domains (like transportation) should be at least equivalent to the energy required to power them. See 2.3.2, “The energy challenge”.

8. Decent work and economic growth

How ICTs can contribute: ICTs have enabled new ways of doing business, relieved workers from many tedious and repetitive tasks, and contributed to benefits such as flexible working hours. The ICT industry also offers many stimulating, rewarding and well-paid jobs, and the number of these is rising.

Challenges: Decent work is understood as employment with adequate earnings, social protection, freedom of association, etc. A growing polarization of the workforce has been attributed to the introduction of ICTs, with more high-skilled and low-skilled jobs and fewer jobs requiring mid-level skills (see 2.6.3, “Computing technology and the future job market”).

Moreover, automation does not always have a positive effect on the quality and the quantity of jobs, and the speed of technological change is likely to have a disruptive effect on the job market. The challenge is to use computing to improve the quality of jobs, and to increase their number (within the boundaries of the planet). Globally, more than 400 million jobs are needed by 2030 for new entrants into the labour market. This is a challenge.

9. Industry, innovation and infrastructure

How ICTs can contribute: The world needs to build infrastructure to host 10 billion people by 2075: transport, irrigation, energy and communication technology are crucial to build a sustainable middle-class future for all these people. Technology and innovation are essential for industrialization, and industrialization is a requirement for development and a middle-class society. Computing will be an essential part of future (sustainable) industry. It is clear that ICT, including of course the internet, has facilitated an explosion in research, collaboration, globalization and innovation. There is no reason to doubt that it will continue to do so.

Challenges: With large multinational companies dominating the ICT industry, there is a danger that some regions will not reap the full benefits of ICT innovations, as well as being dependent on ICT imports. ICT infrastructure also needs to be improved in many regions in order for them to benefit.

10. Reduced inequalities

How ICTs can contribute: ICTs, not least connectivity, can be used to promote integration, and empower disadvantaged and excluded communities to join a global, networked society. They can help local entrepreneurs compete in the same arena as the largest international companies.

Challenges: Computing has created (i) a small but thriving upper class in finance, technology and electronics that controls the economy and (ii) a large group of workers in the low-wage sector that usually struggle. This divide leads to societal tensions and political difficulties. The challenge is to find ways to let everybody benefit from the productivity gains technology produces. Bill Gates calls it the robot tax [111]. See 2.6.3, “Computing technology and the future job market”.

11. Sustainable cities and communities

How ICTs can contribute: By 2030 60% of the world population will live in cities (from 50% today). 95% of urban expansion will take place in the developing world. Cities account for 60% of energy consumption, 70% of carbon emissions and 70-80% of global GDP while only taking 3-4% of the world's available landmass. The challenge is to build large healthy, safe, carbon-neutral cities. Through the smart city model, advanced computing technologies will be needed to reduce energy consumption and carbon emissions.

Challenges: Transparency and the needs of different groups of citizens will be important factors when designing the smart cities of tomorrow. People should be made aware of how their data is used and why, and should be empowered to participate fully in the digitalization of their societies. Introducing new technologies to smart cities should also reduce rather than increase energy consumption and resource use.

12. Responsible consumption and production

As noted in the introduction to this section above, we currently consume the renewable resources of 1.7 Earths per year. Ensuring a more sustainable model of consumption and production will be essential to ensuring that future generations have the resources needed to thrive.

How ICTs can contribute: ICTs can contribute to more efficient production practices. Moreover, by allowing data to be collected on themes such as pollution, ICTs can provide the evidence required to prompt changes in policy and citizen behaviour. They can also contribute to the circular economy, for example through car sharing networks or networks to identify new users for unwanted goods.

Challenges: To contribute to this goal, the computing industry needs to abolish programmed obsolescence and produce goods which last for longer. By adding semiconductors to products that previously had none, and by shortening the life of such devices

(they are almost never repairable and in some cases, the battery cannot even be replaced), the internet of things will create an explosion of e-waste [19]. Of the roughly two billion smartphones, tablets and computers sold every year, 80% (around 35 million tonnes per year) end up in landfills, where they will eventually contaminate the environment with hazardous chemicals [437]. Instead, these devices could be recycled in urban mining projects.

13. Climate action

How ICTs can contribute: ICTs can play a significant part in the fight against climate change by providing data and climate models which can be used to justify policy directions. They can also be used to improve energy efficiency in industrial processes, in transportation, and in the home, as well as enabling the creation of smart grids for renewable energy.

Challenges: ICTs currently account for a significant and increasing amount of global energy consumption, as well as being used to support activities which contribute to climate change, such as fossil fuel extraction. ICTs need to become more energy efficient and should be used to promote renewable energy production rather than prolonging the use of fossil fuels. See 2.3.2, “The energy challenge” for a detailed discussion of ICT and energy.

14. Life below water

Oceans provide food for billions of people, add three trillion USD to the global economy, regulate the climate, produce 50% of global oxygen and store carbon, among many other benefits. But the seas suffer from pollution and overfishing, with over 600 officially recorded “dead zones” in the oceans.

How ICTs can contribute: ICTs can help in protecting the life below water by supporting a fine grained global monitoring system that can alert us when changes take place, and that will allow us to develop more detailed models of the oceans. They can also provide people with the mechanisms to raise awareness and campaign for ocean conservation.

15. Life on land

Since 1970, there has been a nearly 60% decline in wildlife populations across land, sea and freshwater due to loss or degradation of natural habitat, invasive species or pollution. Climate change adds to that. Decreasing biodiversity weakens the earth’s regenerative capacity, which we all depend on.

How ICTs can contribute: ICT is crucial to protect remaining biodiversity and to manage the production resources (land, sea, fresh water) more sustainably, for example by providing data through remote sensing technologies.

16. Peace, justice and strong institutions

How ICTs can contribute: Digital forensics is essential to extract digital evidence stored on computers, tablets, smartphones. Data analytics is another discipline that helps identify criminal

organizations, for example by tracking financial transactions and money laundering. Computers and the internet are instrumental in creating efficient public services, fomenting trust between the government and citizens by enabling greater transparency, and empowering citizens. For the 1.5 billion people who currently lack state-recognized identification, providing them with this would give them access to social security, healthcare and voting.

Challenges: Enforcing the law requires strong institutions that are not corrupt, and are subject to democratic accountability. ICTs can also be used to abuse states’ power, compromising privacy rights by enabling global surveillance. There is also a danger that ICTs will be used for new kinds of warfare, from state-sponsored cyberattacks to automated weapons such as drones.

17. Partnerships for goals

From the above is clear that ICT will play a major enabling role in the realization of the sustainable development goals. However, it is self-evident that ICT cannot solve all the challenges standing in the way, nor are governments and traditional non-governmental organizations able to develop, install and maintain the ICT infrastructure needed. To meet the sustainable development goals, different parties (international organizations, local governments, philanthropic institutions, large ICT companies) will therefore need to work in partnership.

2.6.7 THE NEED FOR DIGITAL ETHICS

Digital ethics is not a new concept; in fact, it was first touched upon in the mid-1940s by Norbert Wiener, who coined the term cybernetics in his book “Cybernetics: or control and communication in the animal and the machine” (1948). At the time, it was not taken very seriously by the scientific community. The last decade has, however, witnessed a sharp increase in the interest in digital ethics.

Much of this originally started with discussions about the decisions made by self-driving cars, and whether algorithms should protect the passengers in the car or the people in the street in the event of an accident [151]. However, even philosophers do not agree on the best decision in such cases. Some argue that the car should try to minimize the overall harm; others argue that the people in the street should be protected at all costs because passengers in the vehicle decided to use the car, so they have more responsibility than the people in the street; while still others believe that it depends on the pedestrian’s intentions. If someone deliberately jumps in front of the car (to test the car, for example, or to commit suicide), this argument goes, it is OK to protect the passengers in the car; unfortunately the car cannot know the intentions of the person in the street. Given the unlikelihood of situations where a car will have to choose between two lives, such dilemmas are not very helpful in practice and self-driving cars will have to avoid accidents in the first place.

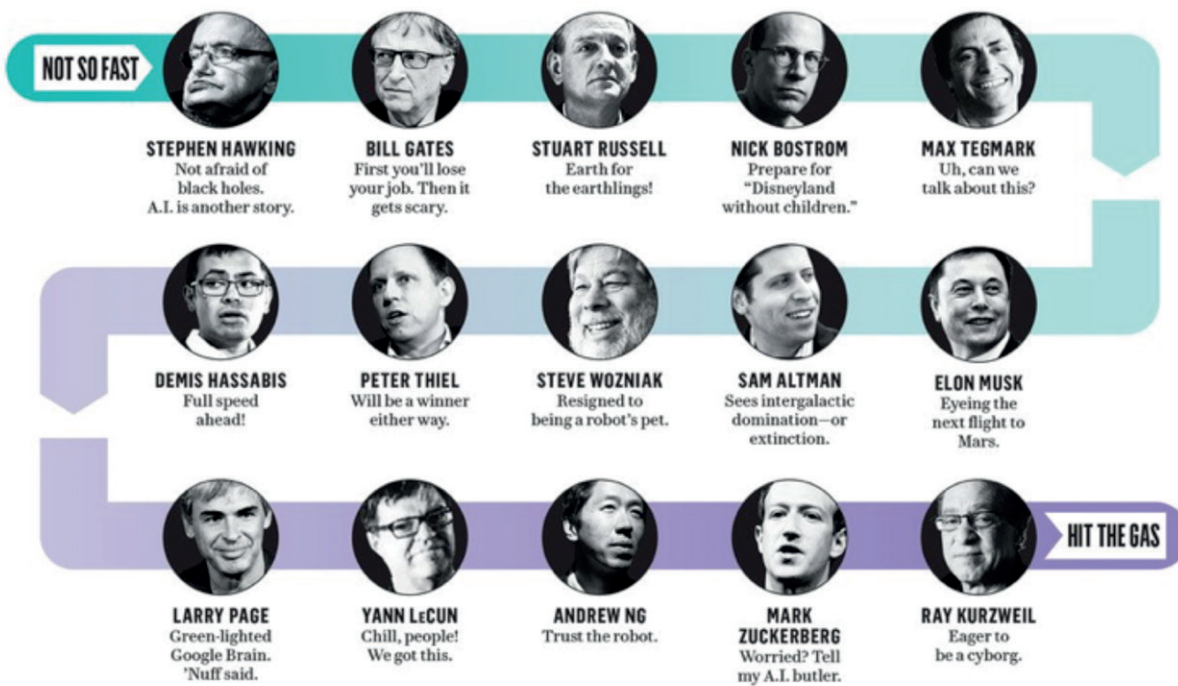


Figure 165: People who warned about the unwanted side effects of AI
Source: [133].

Following recent successes in artificial intelligence development, influential commentators such as Elon Musk, Bill Gates, and Stephen Hawking warned about the unwanted side effects of artificial intelligence. Others, on the other hand, are eager to see it deployed on a large scale.

The key problem is that in the past, computers did the calculations and humans made the decisions based on the calculations. Decision makers are assumed to have an ethical framework to guide them in the decision making process. With artificial intelligence, computers not only do the calculations, they also make the decisions – small decisions at the moment, but the expectation is that they might be asked to make important, even life-changing decisions in the near future too, as in the example of the self-driving car. This means that decision-making algorithms have to include an ethical framework to guide the decision-making process. If not, whatever the poor computer scientist instructs the program to do will be done for the foreseeable future by the system. Not all computer scientists have developed an ethical framework comparable to that of people involved in building public policy, for example.

That said, recently 3,100 Google employees wrote a letter to Google's CEO asking for project Maven to be cancelled that wants to analyse images produced by US military drones, because they do not want to be involved in the business of war [208]. Thousands of AI researchers in several countries have expressed similar concerns on the use of artificial intelligence for military applications [230]. This movement has spurred a number of organizations to come up with set of ethical guidelines.

Notable examples are:

- The Asilomar AI Principles by the Future of Life Institute. This list of 23 guidelines is very comprehensive [258] and focuses a lot on values and share ethical ideals. They require interpretation to apply them in a particular context.
- ACM updated its Code of Ethics [434]. The previous version dated from 1992. These guidelines are logical and straightforward, but not so easy to implement in practice. For example, what is the meaning of harm in "avoid harm"? Is a robot causing harm to the workers it replaces? Or is not installing a robot causing harm to the company because it is less competitive and might go bankrupt (and hence cause harm to all the employees)? Did the computer scientists who tweaked the emissions software in Volkswagen cause harm? There is also the tension between the code of ethics and loyalty towards an employer.

With respect to the issue of bias in AI, they make a separate set of seven principles for algorithmic transparency and accountability. These are clearly targeted at computer scientists.

- 1 Awareness: Owners, designers, builders, users, and other stakeholders of analytic systems should be aware of the possible biases involved in their design, implementation, and use and the potential harm that biases can cause to individuals and society.
- 2 Access and redress: Regulators should encourage the adoption of mechanisms that enable questioning and

redress for individuals and groups that are adversely affected by algorithmically informed decisions.

- 3 Accountability: Institutions should be held responsible for decisions made by the algorithms that they use, even if it is not feasible to explain in detail how the algorithms produce their results.
 - 4 Explanation: Systems and institutions that use algorithmic decisionmaking are encouraged to produce explanations regarding both the procedures followed by the algorithm and the specific decisions that are made. This is particularly important in public policy contexts.
 - 5 Data Provenance: A description of the way in which the training data was collected should be maintained by the builders of the algorithms, accompanied by an exploration of the potential biases induced by the human or algorithmic data gathering process. Public scrutiny of the data provides maximum opportunity for corrections. However, concerns over privacy, protecting trade secrets, or revelation of analytics that might allow malicious actors to game the system can justify restricting access to qualified and authorized individuals.
 - 6 Auditability: Models, algorithms, data, and decisions should be recorded so that they can be audited in cases where harm is suspected.
 - 7 Validation and Testing: Institutions should use rigorous methods to validate their models and document those methods and results. In particular, they should routinely perform tests to assess and determine whether the model generates discriminatory harm. Institutions are encouraged to make the results of such tests public.
- Nicolas Economou from H5 proposes a framework of six principles [243] targeted at lawyers and law firms when applying AI in electronic discovery.

• Another interesting set of principles for AI Ethics has been proposed by Oren Etzioni of the Allen Institute for Artificial Intelligence [387]. The second principle, that AI systems must clearly disclose that they are not human, is particularly interesting.

- 1 An A.I. system must be subject to the full gamut of laws that apply to its human operator.
- 2 An A.I. system must clearly disclose that it is not human.
- 3 An A.I. system cannot retain or disclose confidential information without explicit approval from the source of that information.

• Recently, the Atomium European Institute has created the AI4People forum to develop an ethical framework for the use of artificial intelligence. This framework consists of the four ethics principles used in bioethics, along with explicability.

• The European Commission recently created a High-Level Expert Group on Artificial Intelligence [283]. Among other responsibilities, this will be tasked with proposing draft AI ethics guidelines to the commission.

• Some universities have established centres for digital ethics (for example, the Digital Ethics Lab of the Oxford Internet Institute [389] founded in 2017, and Center for Digital Ethics and Policy of The Loyola University of Chicago [345]). Courses on digital ethics are being introduced into several computer science courses in order to ensure that graduates have a basic understanding of ethical aspects of their profession.

Topics that are generally discussed are:

- Privacy: what happens to people's data, including images and video.
- Equal access to information for everybody, censorship, ethical issues relating to the algorithms that populate timelines in social media.

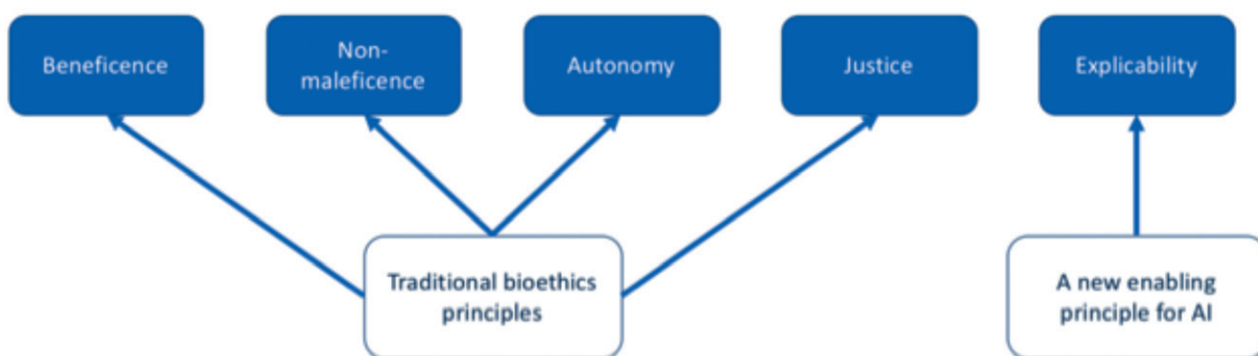


Figure 166: An ethical framework for AI, formed of four traditional and one new principle
Source: [123]

- Sharing and related aspects including author rights, intellectual property and plagiarism.
- Robotics and their application in industry and the military with implications on employment, safety and so on.
- Digital behaviour with aspects of netiquette, cyberbullying etc.
- Accurate information: the ethical consequences of spreading fake information (hoaxes, conspiracy theories, fake news, deep fake videos) in order to mislead people.
- Professional conduct: the importance of delivering high-quality products that work as specified, do not have undisclosed functions, are reliable and safe to use

It seems that digital ethics is starting to evolve into a separate domain of ethics, on a par with bioethics, medical ethics, war ethics and the like. This means that practitioners agree that computing technology is so powerful that its applications should somehow be controlled, or in other words that not everything that can be done, should be done. This is a significant change in comparison with the general vision of computing during the 20th century that advances in computing were almost by definition good for humankind, by creating productivity gains and economic growth, making life more enjoyable, healing diseases and so on.

Today, an increasing number of people are afraid that computing might get out of control, especially artificial intelligence, particularly general artificial intelligence. In the latter case, there is even no agreement on how to react [452]. On the other hand, other people are afraid that fears about artificial general intelligence could result in regulations that control the further development of artificial intelligence in Europe, comparable to the strong European regulations on genetically modified organisms. This could weaken Europe in the global competition for artificial intelligence, and potentially lead to a brain drain of AI researchers to other continents with more liberal regulations. They believe that ethical regulations should be global or they will be useless.

One particular proposition launched by Cédric Villani is to set up an international scientific panel independent from governments. It could take a form similar to that of the Intergovernmental Panel on Climate Change (IPCC) and would provide the world with a clear scientific view on the current state of knowledge in ICT and AI. Europe should play a major role in such an endeavour.

2.7 THE POSITION OF EUROPE IN THE WORLD

2.7.1 EUROPEAN POSITION (SWOT)

The Lamy report [446] was created by a group of experts with the aim of formulating a vision for future EU research and innovation and making strategic recommendations on how the impact of EU research and innovation programmes can be maximized.

The observations at the start were that (i) in the last twenty years, two-thirds of economic growth in industrialized countries is attributed to research and innovation, (ii) Europe has just 7% of the world's population but produces 24% of global GDP and around 30% of the world's scientific publications, and (iii) compared to other major economies, Europe suffers from a growth deficit caused by an innovation deficit because Europe does not capitalize enough on the knowledge it has and produces. The Lamy report proposes 11 recommendations to address Europe's innovation deficit through maximizing the impact of future EU research and innovation programmes.

- 1 Prioritize research and innovation in EU and national budgets by doubling the budget of the post 2020 EU research and innovation programme.
- 2 Build a true EU innovation policy that creates future markets by fostering ecosystems for researchers, innovators, industries and governments and by promoting and investing in innovative ideas with rapid scale-up potential through a European Innovation Council.
- 3 Educate for the future and invest in people who will make the change by modernizing, rewarding and resourcing the education and training of people for a creative and innovative Europe.
- 4 Design the EU R&I programme for greater impact by making the future programme's pillars driven by purpose and impact, fine-tune the proposal evaluation system and increase flexibility.
- 5 Adopt a mission-oriented, impact-focused approach to address global challenges by setting research and innovation missions that address global challenges and by mobilizing researchers, innovators and other stakeholders to realize them.
- 6 Rationalize the EU funding landscape and achieve synergy with structural funds by cutting the number of R&I funding schemes and instruments, and by making those remaining reinforce each other and make synergy with other programmes work.
- 7 Simplify further to become the most attractive R&I funder in the world, privileging impact over process.
- 8 Mobilize and involve citizens by stimulating co-design and co-creation through citizen involvement.
- 9 Better align EU and national R&I investment by ensuring EU and national alignment where it adds value to the EU's R&I ambitions and missions.
- 10 Make international R&I cooperation a trademark of EU research and innovation by opening up the R&I programme to association by the best and participation by all, based on reciprocal co-funding or access to co-funding in the partner country.
- 11 Capture and better communicate impact by branding EU research and innovation and by ensuring wide communication of its results and impacts.

	Strengths	Weaknesses
Science and Technology	<ul style="list-style-type: none"> • High-quality education • Large number of PhDs • Largest publication and citation count of the world • World leader in lithography 	<ul style="list-style-type: none"> • Weak academia-industry link • Strong in research, but not in commercialization
Market and Industry	<ul style="list-style-type: none"> • Second largest market in the world • Large embedded market 	<ul style="list-style-type: none"> • EU ICT contributes less to GDP than in other advanced countries • Europe lacks advanced foundries
Policy and Measurements	<ul style="list-style-type: none"> • Common market • Variety of research funding instruments • Decent public funding level of R&D 	<ul style="list-style-type: none"> • Europe lacks VC culture • Lack of ICT-workers • Fragmentation of funding
	Opportunities	Threats
Science and Technology	<ul style="list-style-type: none"> • The end of Moore's law 	<ul style="list-style-type: none"> • Economic stagnation • Brain drain
Market and Industry	<ul style="list-style-type: none"> • Embedded systems, IoT, CPS • Cybersecurity 	<ul style="list-style-type: none"> • Saturating markets • Computing initiatives in countries such as China, Russia and Japan
Policy and Measurements	<ul style="list-style-type: none"> • Solutions for societal challenges 	<ul style="list-style-type: none"> • Political instability

In this section, we present a SWOT (strengths, weaknesses, opportunities, threats) analysis of the European computing systems community. We make a distinction between three stakeholders: (i) publicly funded universities and research institutions (“Science and Technology”), (ii) the computing industry and its market (“Market and Industry”), and (iii) the local and European governments responsible for creating an environment in which research, innovation and commercialization can take place (“Policy and Measurement”).

Most of the data in this section are taken from [445, 451].

2.7.1.1 STRENGTHS

2.7.1.1.1 High-quality education

Europe has a good educational system. Higher education is more affordable than in the USA, and of the top one hundred best universities worldwide in the 2018 Times “Higher Education Ranking”, Europe has 38 institutions (North America has 45, and Asia 17) [418]. Unfortunately, since 2016, Europe has lost 4 universities in the top one hundred. Two went to the USA and two went to Asia Pacific. This shows that Europe is losing its leading position. Another remarkable observation is that all 38 European universities are located in the north-western part of Europe. Given the position of the United Kingdom in this ranking, its leaving the European Union will be a great loss.

Country	#
United Kingdom	12
Germany	10
The Netherlands	7
Sweden	3
Switzerland	3
Finland	1
Belgium	1
France	1
Total	38

2.7.1.1.2 Large number of PhDs

European universities produce on average significantly more PhD degrees per 1,000 of the population than American universities, South Korea or Japan. Even better, the majority of individual European countries produce considerably more PhD degrees than the USA, even in science and technology. This is therefore a clear strength.

2.7.1.1.3 Largest publication and citation count of the world

With respect to scientific output, Europe is among the strongest regions in the world. More than one quarter of all scientific publications in 2016 originated in Europe. The USA was second with 19.5% of the global publications, followed by China. This

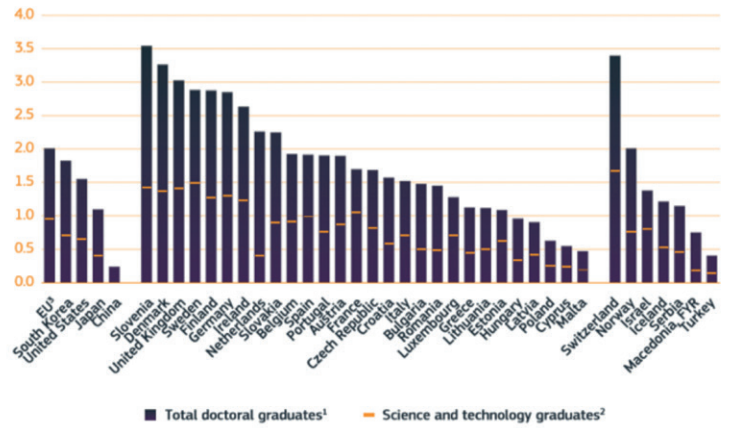


Figure 167: New doctoral graduated per thousand populations aged 25-34, 2015 – Source: DG Research and Innovation

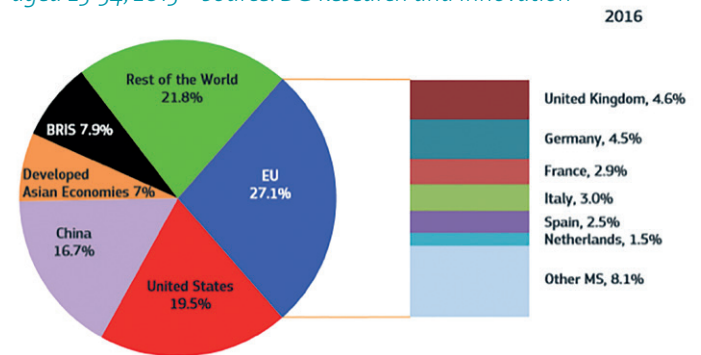


Figure 168: World share of scientific publications, 2000 and 2016 – Source: DG Research and Innovation

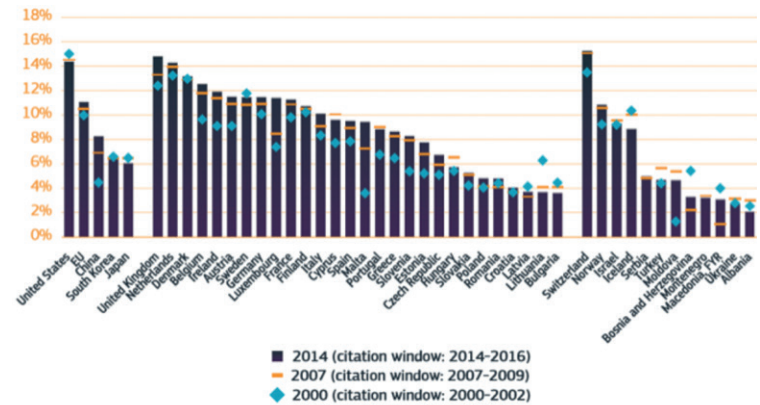


Figure 170: Top 10% highly cited scientific publications, 2000, 2007 and 2014 – Source: DG Research and Innovation

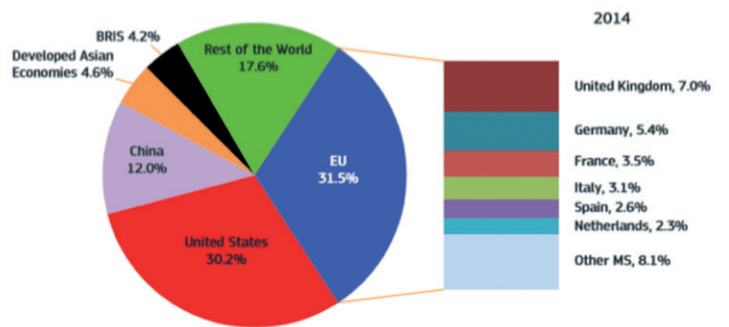


Figure 169: World share of top 10% highly cited scientific publications, 2000 and 2014 – Source: DG Research and Innovation

means that research in Europe is of excellent quality and can compete globally.

When looking at the top 10% highly cited papers, Europe is still leading (in absolute numbers). Most striking in this diagram is that China only produced 1.2% of the 10% highly cited papers in 2000 which means that they have realized a tenfold increase in their high-impact papers over a period of 15 years. This is a spectacular result, and if they keep growing at this pace, China might one day overtake the United States and Europe in number of high-impact papers published.

However, if we relate the number of highly cited papers to the total number of papers published, the view changes. 14.2% of all published US papers are high impact, while only 10.6% of the published EU papers are high impact. Only two European countries match the performance of the USA: The United Kingdom and The Netherlands. Europe is gradually improving its performance, while China is improving spectacularly, and is on a clear path to become a world leader in research.

Compared by sector, the USA has more highly cited publications in most domains. In ICT, Europe is second after the USA. Surprisingly, China is leading security research with regards to the number of cited papers.

2.7.1.1.4 World leader in lithography

Europe has several research institutes and companies that are key players in technology development (including CEA, Imec and ASML). They are Europe's biggest asset when it comes to the further development of CMOS-technology, and their expertise might also be crucial to the development of post-CMOS technology. With the recently approved quantum computing flagship, Europe has demonstrated its intention to take the lead in quantum computing too.

2.7.1.1.5 Second largest market in the world

According to the International Monetary Fund, Europe (EU-28) has the second largest GDP in the world:

Country	GDP in billion USD (2017)
USA	19 390
EU	17 308
China	12 014
Japan	4 872
Germany	3 684
United Kingdom	2 624
India	2 611
France	2 583
Brazil	2 054

ICT market by region (€ bn, source: Digiworld)

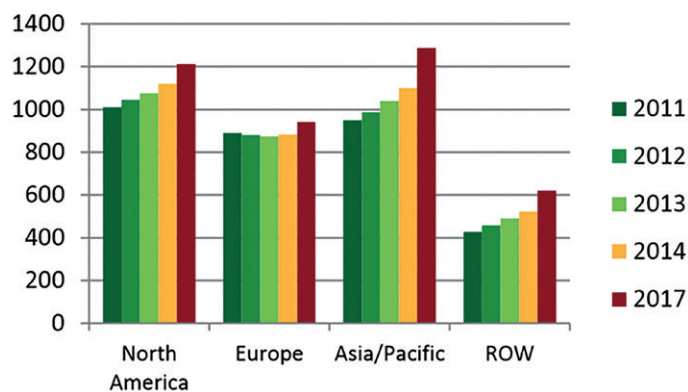


Figure 171: ICT market by region

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 (like China and Europe) might be an important advantage in times of troubled international trade relations.

Unfortunately, the ICT market in Europe has been growing more slowly than in other parts of the world [454].

2.7.1.1.6 Large embedded market

According to Global Markets Insight [362], the embedded systems market will reach a total size of US\$258 billion in 2023 at an average annual growth rate of 5.6%.

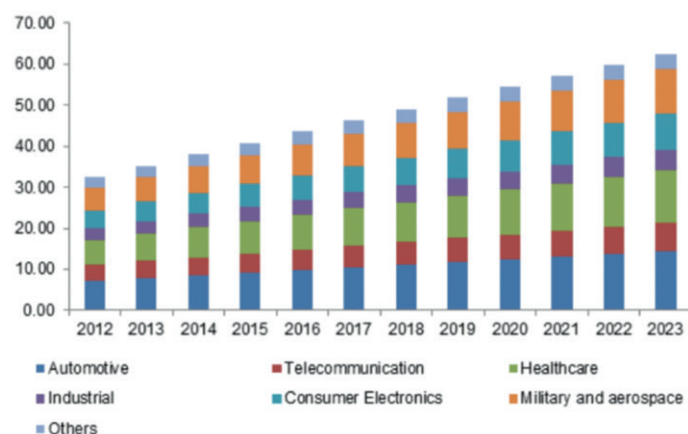


Figure 172: Europe embedded system market size
Source: Global Markets Insights

The European embedded systems market is the third largest in the world after North America and Asia, and will have an estimated size of US\$62 billion in 2023 (North America will attain US\$84 billion, and Asia US\$81 billion in the same year). The biggest embedded systems sectors in Europe are automotive, followed by healthcare and military and aerospace. The automotive market is spread out over a large geographical area in Europa, with the centre of gravity in Germany.

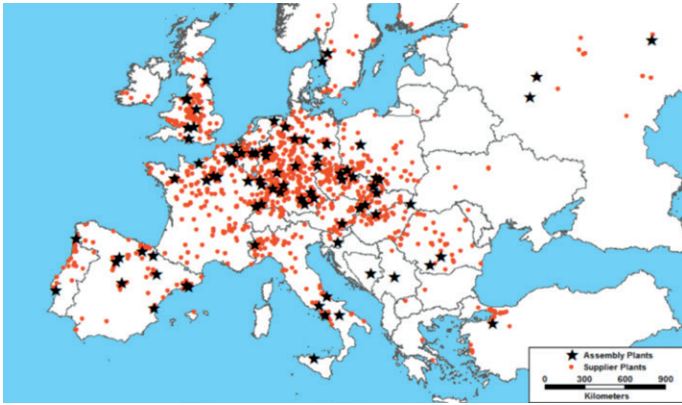


Figure 173: Automotive assembly plants and supplier plants
Source: REVEL, revues électronique de l'UNS

With annual growth of 5,3% per year, the potential of the embedded systems industry to create added value and employment cannot be underestimated. On the other hand, the fact that the embedded systems market in Europe is not the largest might suggest that the embedded market is weaker in Europe than in the USA. According to [362], the embedded hardware market will grow to US\$144 billion, while the embedded software market will only grow to US\$18 billion. In order to grow, Europe's focus should be on embedded hardware, not software. The good news is that Europe has some important key players in this area: Infineon Technologies, STMicroelectronics, NXP Semiconductor. Non-European players are Renesas Electronics, Texas Instruments, and Microchip.

2.7.1.1.7 Common market

At the policy level, one of the strengths is the common market, and the fact that Europe can act as one economic block 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:

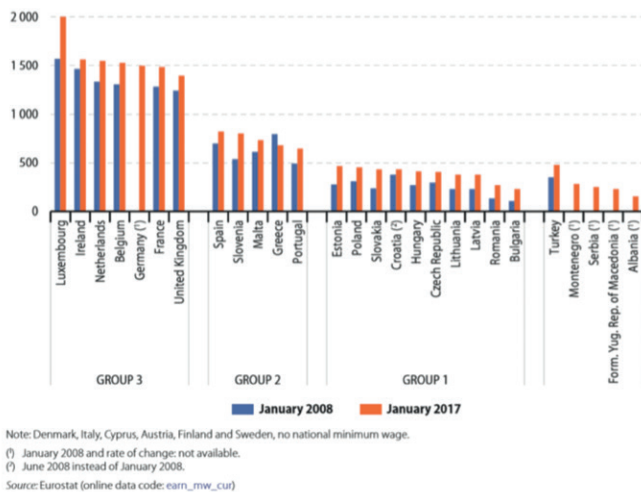


Figure 174: Minimum wages, January 2008 and 2017 (EUR per month) – Source: Eurostat

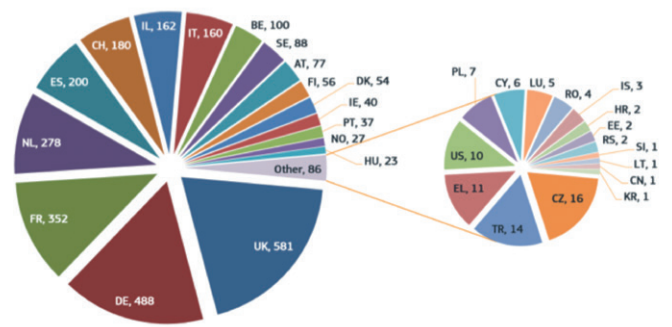


Figure 175: Number of European Research Council (ERC) grants by country 2017 – Source: DG Research and Innovation

The fact that the United Kingdom voted to leave the common market is a sign that building a common market will remain a challenge in the future.

2.7.1.1.8 Variety of research funding instruments

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. ERC 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 like ECSEL aim to pool local and European funding to encourage research and innovation.

2.7.1.1.9 Decent public funding level of R&D

The total amount of public funding available make Europe a good place to carry out R&D (at 0.7% of GDP). Worldwide, Europe is in second place after South Korea.

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 the other geographies. The aim for Europe is to spend 3% of GDP, but it is still far away from that target.

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 continent, this does not lead to more researchers in employment.

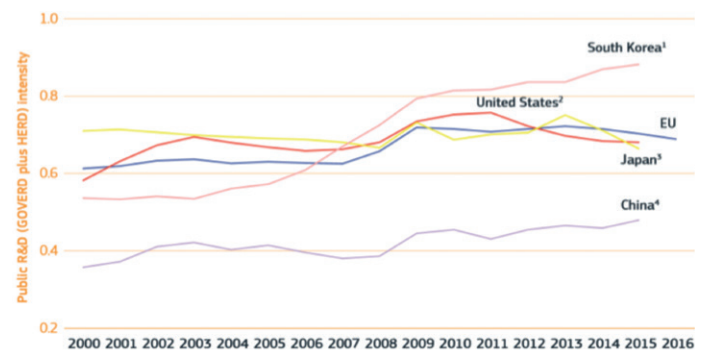


Figure 176: Evolution of public R&D intensity 2000-2016
Source: DG Research and Innovation

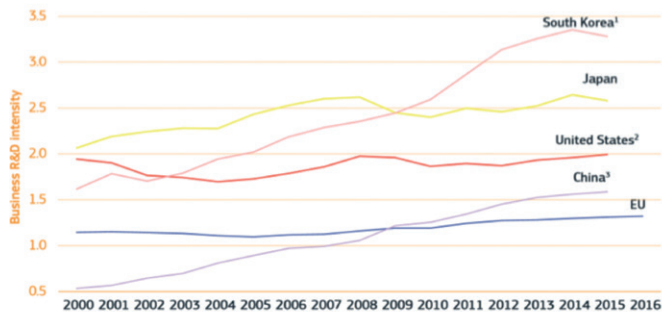


Figure 177: Evolution of business R&D intensity 2000-2016
Source: DG Research and Innovation



Figure 178: Total researchers (FTE) as % of total employment 2007 and 2015 – Source: DG Research and Innovation

The total picture of R&D intensity is depicted below. Asian countries are apparently preparing for the future. Their R&D intensity is higher than the European average, and (apart from Japan) also growing faster than the average growth in Europe.

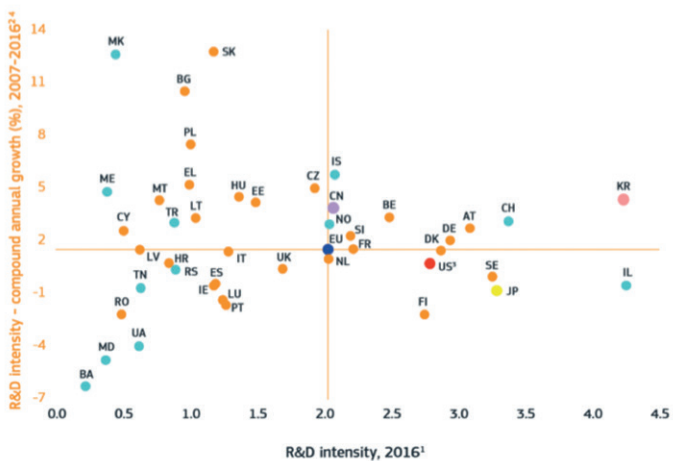


Figure 179: R&D intensity 2016 and compound annual growth 2007-2016 – Source: DG Research and Innovation

2.7.1.2 WEAKNESSES

2.7.1.2.1 Weak academia-industry link

The collaboration between academia and industry (quantified as the number of joint scientific publications) is weak in Europe (about 50% of those in the USA), and decreased between 2008 and 2015. Europe has the highest share of publications and citations, but these are not the result of collaboration between academia and industry.

2.7.1.2.2 Strong in research, but not in commercialization

Europe is lagging behind the USA and Japan with respect 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). The USA and Japan have improved a bit, while Europe is stagnating. There are large differences in innovation performance between member states.

However, Europe outperforms the USA in start-up creation in the knowledge-intensive sector, and this rate is growing. In fact, most European countries perform better than the average in the USA.

In recent years, five European cities have emerged as start-up ecosystems in the global top-20.



Figure 180: Public-private co-authored scientific publications per million population 2008 and 2015
Source: DG Research and Innovation

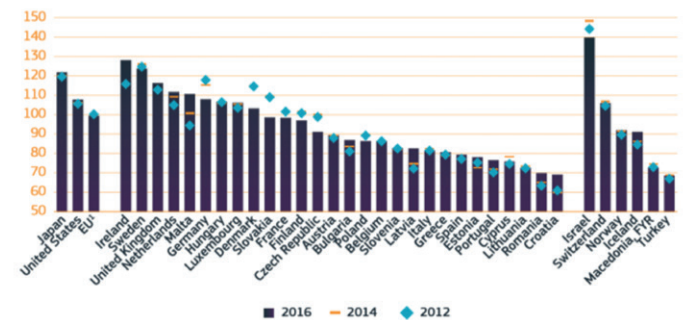
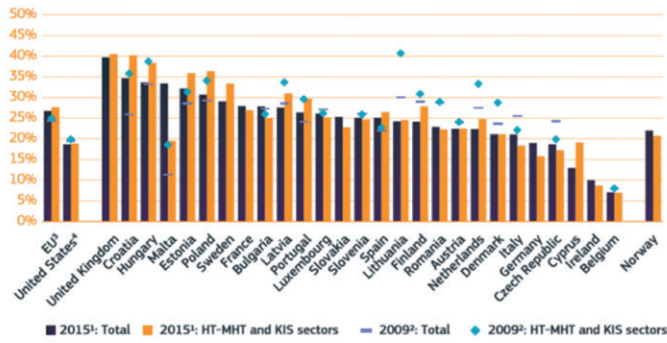


Figure 181: Innovation output indicator (EU2011=100), 2012, 2014 and 2016 – Source: DG Research and Innovation



Science, Research and Innovation performance of the EU 2018
 Source: DG Research and Innovation - Unit for the Analysis and Monitoring of National Research and Innovation Policies
 Data: Eurostat, OECD

Figure 182: Start-ups (0 to 2 years old) as % of employer enterprises, 2009 and 2015 – Source: DG Research and Innovation

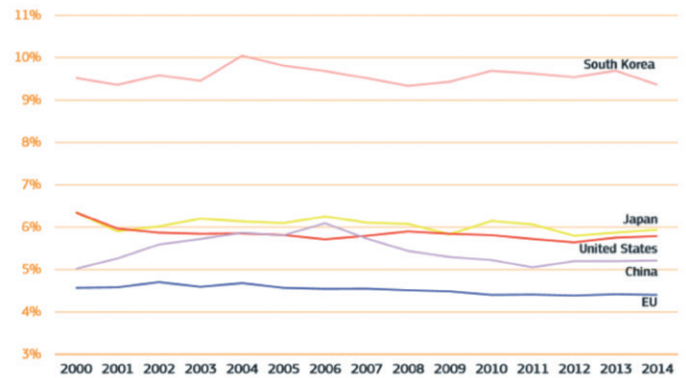
RANKING 2017	Performance ¹	Funding ²	Market reach ³	Talent ⁴	Start-up experience ⁵
1	Silicon Valley				
2	New York				
3	London				
4	Beijing				
5	Boston				
6	Tel Aviv				
7	Berlin				
8	Shanghai				
9	Los Angeles				
10	Seattle				
11	Paris				
12	Singapore				
13	Austin				
14	Stockholm				
15	Vancouver				
16	Toronto-Waterloo				
17	Sydney				
18	Chicago				
19	Amsterdam				
20	Bangalore				

Science, Research and Innovation performance of the EU 2018
 Source: DG Research and Innovation - Unit for the Analysis and Monitoring of National Research and Innovation Policies
 Data: Global Startup Ecosystem Report 2017, Startup Genome
 Notes: ¹Performance includes start-up output, exits, valuations, early-stage success, growth-stage success, and overall ecosystem value. ²Funding concerns growth in early-stage investments, and funding quality through the presence of experienced VC firms. ³Market reach is linked to global connectedness and global and local reach, based on the start-ups' proportion of foreign customers and the national GDP. ⁴Talent-access, cost, and quality of talent. ⁵Start-up experience: team experience and ecosystem experience in terms of knowledge and networks available from which start-ups can develop.
 Stat. link: https://ec.europa.eu/info/sites/info/files/sriopart01_6-a_figuresif_1.6-a_17.xlsx

Figure 183: World Top 20 start-up ecosystems 2017
 Source: DG Research and Innovation

2.7.1.2.3 EU ICT contributes less to GDP than in other advanced countries

The European ICT-industry contributes less than 5% to GDP, as compared to more than 5% in competing countries. One explanation is that Europe lacks GAFAM (Google, Apple, Facebook, Amazon, Microsoft), and other major ICT-companies like HP, Dell, IBM, and the ecosystem supporting them. This is a structural weakness which also limits the innovation potential for the ICT sector (the smaller the sector, the fewer the resources available to invest in research and development). The lack of such large corporations can be explained by the lack of venture capital culture in Europe [99, 193]. In order for a company to grow to US\$50-100 million, they have to enter non-European markets like the USA or China. The US market is very competitive and



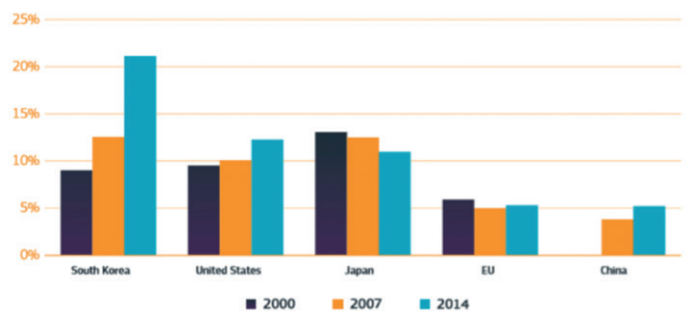
Science, Research and Innovation performance of the EU 2018
 Source: DG Research and Innovation - Unit for the Analysis and Monitoring of National Research and Innovation Policies
 Data: PREDICT Project (DG JRC)
 Note: ¹The operational definition of ICT, as defined in the PREDICT project, was used. The operational definition of ICT allows for international comparison with non-EU countries.
 Stat. link: https://ec.europa.eu/info/sites/info/files/sriopart01_3-b_figuresif_1.3-b_3.xlsx

Figure 184: Value added in ICT as % of GDP 2000-2014
 Source: DG Research and Innovation

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, cultures and so on. Therefore, it is difficult for a company to address the whole of Europe without extra work 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. That is one of the explanations why European VCs are more cautious; they doubt whether many companies have the potential to successfully break into markets outside Europe.

There seems to be a correlation between the contribution to the GDP and the intensity of R&D.

The value added primarily stems from software services. Apart from Ireland, the differences between the European countries are



Science, Research and Innovation performance of the EU 2018
 Source: DG Research and Innovation - Unit for the Analysis and Monitoring of National Research and Innovation Policies
 Data: PREDICT Project (DG JRC)
 Note: ¹Business enterprise expenditure on R&D as % of value added. The operational definition of ICT, as defined in the PREDICT project, was used. The operational definition of ICT allows for international comparison with non-EU countries.
 Stat. link: https://ec.europa.eu/info/sites/info/files/sriopart01_3-b_figuresif_1.3-b_5.xlsx

Figure 185: R&D intensity of ICT, 2000, 2007 and 2014
 Source: DG Research and Innovation

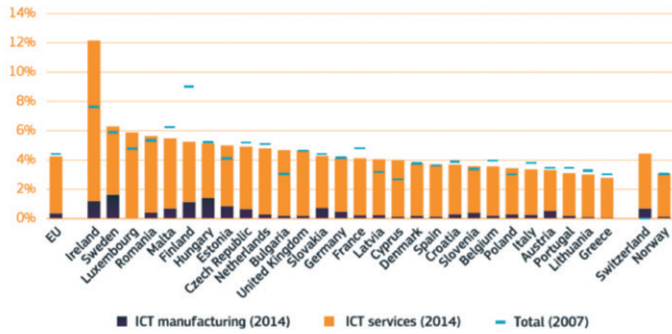


Figure 186: Value added in ICT as % of GDP broken down by manufacturing and services, 2014 (and for 2007 without breakdown) – Source: DG Research and Innovation

not that great. The situation in Ireland can be explained by the presence of a number of large USA-based ICT-companies (Apple, Dell, IBM and so on).

Employment in the manufacturing sector is very low in the USA and in the EU. China, Japan and South Korea are the ICT-factories of the world. The USA and the EU are strong in services, and on a par with South Korea and Japan.

Labour productivity in the EU is lower than the USA, but similar to or better than other advanced economies.

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, ARM by Softbank, Movidius by Intel, for example. There are few counterexamples like Sysgo, which was acquired by Thales.

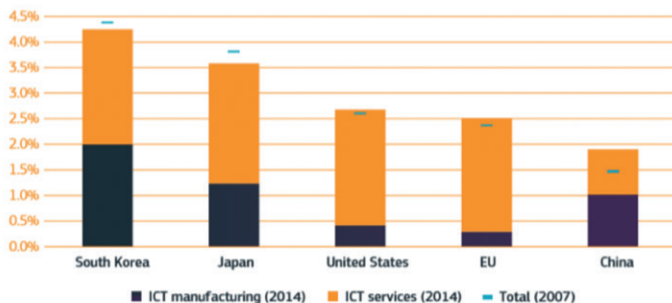


Figure 187: Employment in ICT as % of total employment broken down by manufacturing and services 2014 (and for 2007 without breakdown) – Source: DG Research and Innovation

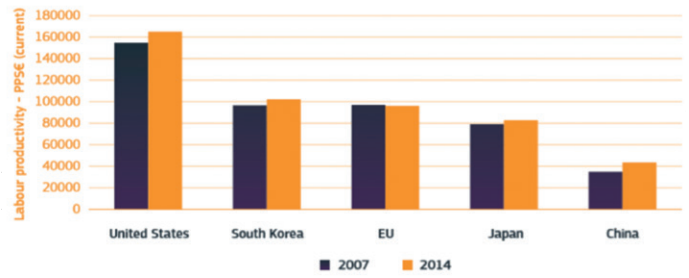


Figure 188: Labour Productivity (GDP per person employed) in ICT 2007 and 2014 – Source: DG Research and Innovation

2.7.1.2.4 Europe lacks 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. Since the embedded hardware market is many times bigger than the embedded software market, this is a lost market opportunity. 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, ASML), it is surprising that no large foundries are left in Europe and that Global Foundries recently 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 USA).

1Q18 Top 15 Semiconductor Sales Leaders (\$M, Including Foundries)

1Q18 Rank	1Q17 Rank	Company	Headquarters	1Q17 Tot IC	1Q17 Tot O-S-D	1Q17 Tot Semi	1Q18 Tot IC	1Q18 Tot O-S-D	1Q18 Tot Semi	1Q18/1Q17 % Change
1	2	Samsung	South Korea	12,811	770	13,581	18,581	820	19,401	43%
2	1	Intel	U.S.	14,220	0	14,220	15,832	0	15,832	11%
3	3	TSMC (1)	Taiwan	7,524	0	7,524	8,473	0	8,473	13%
4	4	SK Hynix	South Korea	5,346	109	5,455	8,016	125	8,141	49%
5	5	Micron	U.S.	4,931	0	4,931	7,360	0	7,360	49%
6	6	Broadcom Ltd. (2)	U.S.	3,740	368	4,108	4,160	430	4,590	12%
7	7	Qualcomm (2)	U.S.	3,676	0	3,676	3,897	0	3,897	6%
8	9	Toshiba	Japan	2,747	265	3,012	3,517	310	3,827	27%
9	8	TI	U.S.	2,960	204	3,164	3,339	227	3,566	13%
10	11	Nvidia (2)	U.S.	1,965	0	1,965	3,110	0	3,110	58%
11	15	WD/SanDisk	U.S.	1,795	0	1,795	2,350	0	2,350	31%
12	10	NXP	Europe	1,965	246	2,211	2,017	252	2,269	3%
13	12	Infineon	Europe	1,130	754	1,884	1,360	907	2,267	20%
14	13	ST	Europe	1,378	440	1,818	1,696	518	2,214	22%
15	17	Apple* (2)	U.S.	1,600	0	1,600	1,830	0	1,830	14%
Top 10 Total				59,920	1,716	61,636	76,285	1,912	78,197	26.9%
Top 15 Total				67,788	3,156	70,944	85,538	3,589	89,127	25.6%

(1) Foundry (2) Fabless *Custom devices for internal use.
Source: Company reports, IC Insights' Strategic Reviews database.

Figure 189: Top 15 semiconductors sales leaders Source: IC Insights

2.7.1.2.5 Europe lacks VC culture

More generally, Europe lacks a VC culture, and in this metric, the gap between the USA and Europe could not be bigger. This observation in combination with the large number of young start-up companies is problematic. It means that they have to fight hard to get the funding to become a scale-up company.

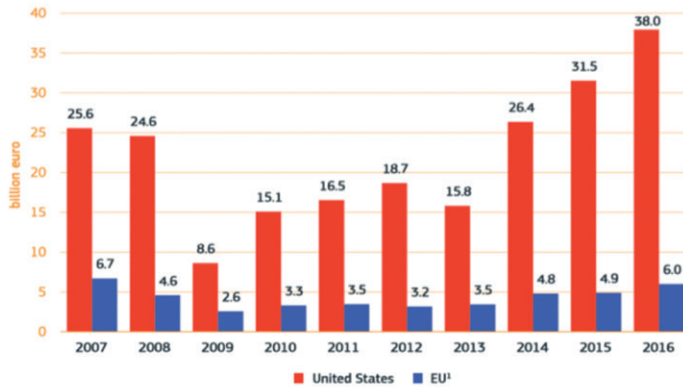


Figure 190: Venture capitalist funds raised (billion euro) in the EU and in the United States 2007-2019

Source: DG Research and Innovation

2.7.1.2.6 Lack of ICT-workers

Europe lacks hundreds of thousands of ICT-workers. Most European countries are witnessing positive growth in the number of graduates overall, but a significant number are reporting declining numbers of ICT graduates. Apparently, Europe is not succeeding in convincing high-school students to start a career in the ICT-sector. This is unfortunate because the competitiveness of the European ICT-sector will depend on the size of its workforce in order to innovate in big data analytics, artificial intelligence, robotics and so on.

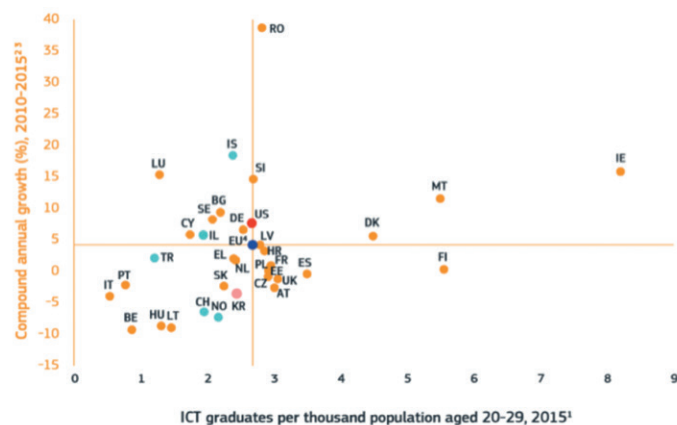


Figure 191: Graduated in the field of ICT per thousand population aged 20-29, 2015 and compound annual growth, 2010-2015

Source: DG Research and Innovation

Massively importing well-trained foreign workers to Europe to help mitigate the shortage is not an effective solution. First of all, 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 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.

2.7.1.2.7 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 part of the proposal. On top of this, the success rate for research proposals is sometimes lower than 10%.

Within a funding agency, different committees deal with particular topics, which makes multidisciplinary project proposals very hard to get funded because committees tend to give priority to the proposals that belong to the core of a domain, leading to lower acceptance rates for interdisciplinary projects. The organizational structure of the funding agency thus ends up constraining the research work that can be proposed in one single project. The design of a novel, secure, cloud-based IoT solution will cut across the topics of at least three units of DG CONNECT. The fact that European Regional Development Funds have also started to be used to fund research only adds to the complexity.

2.7.1.3 OPPORTUNITIES

2.7.1.3.1 The end of Moore's law

With respect to opportunities, the end of Moore's law is a clear opportunity for research. The increase of sequential performance at the pace of Moore's law already ended a decade ago; parallelism kicked in to keep performance increasing in lockstep with number of transistors and cores, but now power consumption has started limiting the number of active cores.

This means that the computing systems community has to start thinking outside the box, and come up with clever solutions to make the best use of the computing resources offered by the computing substrate and available power envelope. Today, specialized accelerators seem to be the preferred solution. There is however room (and also a need) for more disruptive solutions, possibly replacing the (rather inefficient) von Neumann architecture by another computing paradigm.

2.7.1.3.2 Embedded systems, IoT, CPS

The number one market opportunity in computing systems is the strongly growing market of embedded systems (including the IoT, CPS, and the digitization of European industry). Europe has

the second largest economy in the world, it has a number of world-class players producing the key enabling technology for advanced embedded systems, and it has strong automotive, health and aerospace industries. Furthermore, there are no dominant companies like Google, Apple, Facebook, Amazon or Microsoft (GAFAM) in this space yet. The stars of the IoT era will probably not be the same as the ones of the internet era (which are different from the ones in the mainframe era). Could the company dominating computing in 2030 be European? The only way to win this race is to create as many innovative IoT-start-ups as possible, support them to scale up, and hope that they will become world leaders.

2.7.1.3.3 Cybersecurity

Cybersecurity is a growing challenge, and it will become even bigger in the coming years. According to Cybersecurity Ventures [231], the cybersecurity market grew from US\$4 billion in 2004, to US\$75 billion by 2015, and it is forecasted to grow to US\$170 billion by 2020. This is comparable to the size of the global embedded systems market of a couple of years ago. The annual growth rate will be twice the growth rate of the embedded systems market, which makes it one of the fastest growing markets in computing.

On 20 June 2016, the European industry created ECSO (European Cyber Security Organisation) with the objective of supporting all types of initiatives to develop, promote and encourage European Cybersecurity [244]. According to ECSO [245], the European cybersecurity market is about 25% of the global market while the North American market is 43%. The share of the global market secured by companies originating in Europe is only 8.5% (or 35% of the European market) and representing around 100,000 jobs. Given the importance of cybersecurity for the future, Europe needs to catch up. In July 2015, the European Commission signed a public private partnership with ECSO and will invest € 450 million in research and innovation via Horizon 2020. The objective is to raise three times more investments from industry, leading to a total investment of € 1.8 billion by 2020.

In order to increase European digital autonomy, in 2018 the European Commission adopted a proposal to create a European Cybersecurity Competence Centre and a Network of National Cybersecurity Coordination Centres. The Competence Centre will be responsible for managing European financial resources for cybersecurity.

2.7.1.3.4 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 [256]. Only Japan has an older population. That means that Europe and Japan 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 older people first and to sell them to the rest of the world.

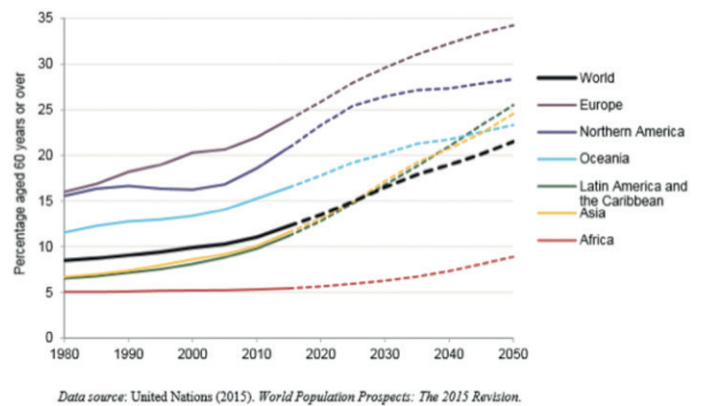


Figure 192: Percentage of the population aged 60 or over

The same reasoning holds for the environment. The European population (together with the USA) has one of the largest ecological footprints of the world. Solutions for reducing our footprint may also work on other continents, and thus may create opportunities for European businesses.

2.7.1.4 THREATS

2.7.1.4.1 Economic stagnation

So far, Europe has seemed to be unable to find effective solutions to end economic stagnation in the region.

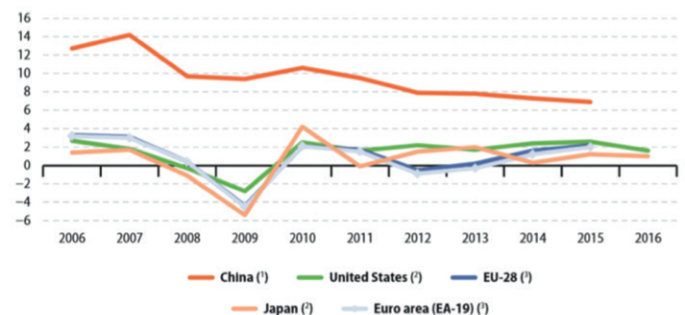


Figure 193: Real GDP growth 2006-2016 (% change compares with previous years)

The lack of economic growth, decline of the middle class, and growth of income inequality [253] have put stress on both businesses and governments. Current approaches need to be reassessed and replaced by more adequate solutions. If this stagnation keeps affecting Europe more than other regions, Europe could quickly lose its leading position in the global market.

For the EU-28, the cost of pensions was already more than 12.5% of GDP in 2014. The cost of the pensions will continue to grow until 2040 (when “baby boomers” will have reached their life expectancy).

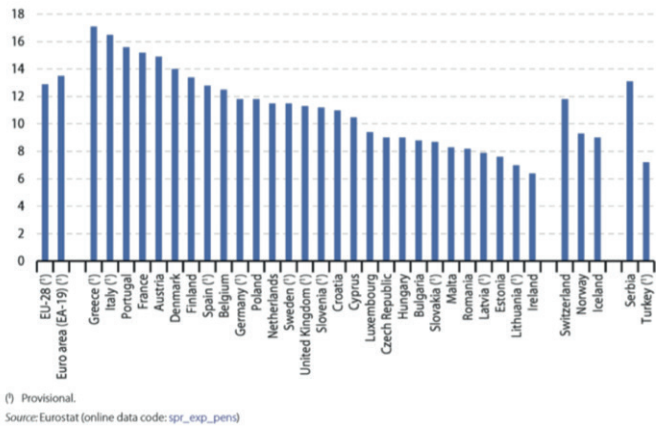


Figure 194: Expenditure on pensions 2014 (% of GDP)

The weakness of research is its dependence on investments by industry or governments. Low or no economic growth easily leads to cuts in R&D budgets, especially when these budgets are requested to fund long-term research that might not lead to short-term results and new market opportunities.

2.7.1.4.2 Brain drain

There is a lot of public attention on the topic of immigration in Europe, and it is indeed the case that immigration has increased since the fall of the Berlin wall in 1989 and is now the major source of population growth in Europe.

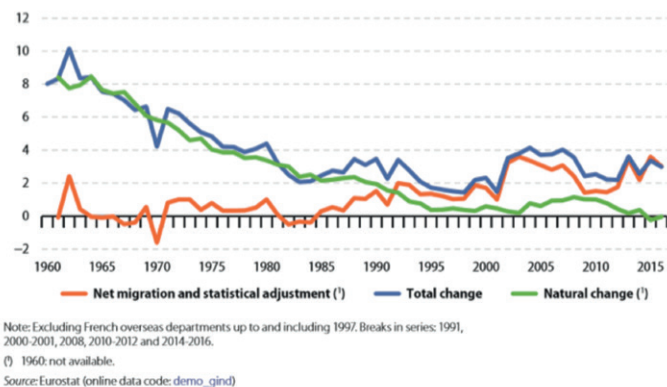


Figure 195: Population change by component (annual crude rates), EU-28, 1960-2016 (per 1.000 persons) – Source: Eurostat

This graph, however, masks the fact that the net immigration is the difference between immigration and emigration. Most migrants are young people: the average age of immigrants in Europe is 27.9, as compared to an average age of 42.9 for the European population as a whole. There are more males (55%) than females among immigrants [447]. Migration 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, and from North-Western Europe to the USA and other rich countries in the world.

In computing, there seems to be a brain drain from Europe to the USA. Top researchers and ambitious entrepreneurs are attracted

by the merit-based American society and top salaries for high potential in both academia and in industry. Large multinational ICT-companies are attractive employers for young European talent eager to travel the world and make a fast career. If they don't want to move, USA-based companies acquire European companies in order to have access to their talent. Particularly in machine learning, there has been a very strong pull on the top talent in Europe by companies like Facebook and Google.

Europe should 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 [271] and cybersecurity [286] launched recently will hopefully help fulfil this need.

2.7.1.4.3 Saturating markets

The market of desktop computers and laptops is shrinking, and the market of smartphones is likely to shrink too (having cannibalized the markets of other devices like navigation systems, cameras, music and video players). This puts pressure on the companies to cut costs and jobs, and to focus on short-term results instead of mid-term innovations or long-term research.

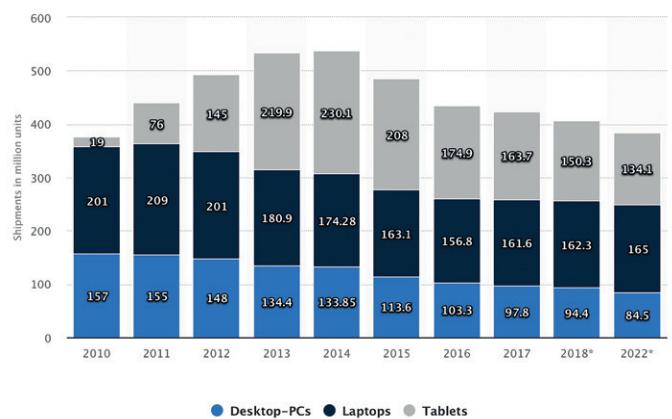


Figure 196: Shipment forecast of tablets, laptops, desktop PCs

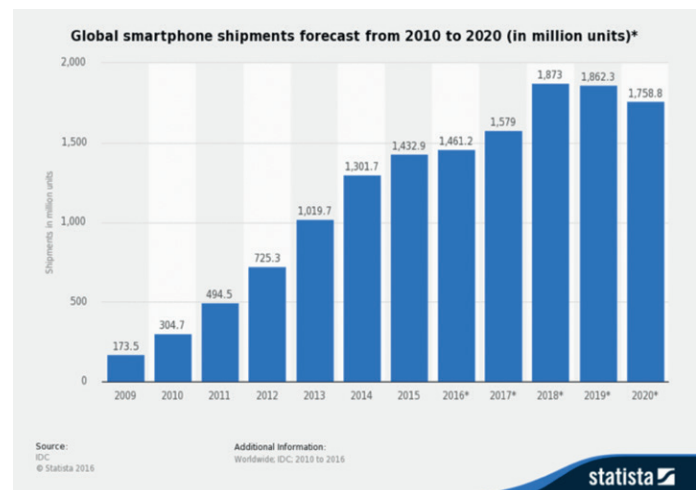


Figure 197: Global smartphone shipment forecasts 2010-2020

2.7.1.4.4 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 China, Russia and Japan. Many countries understand that computing is a key enabling technology of strategic importance, and are investing in their own research, products and companies (see 2.6.5, “Computing technology and the future of Europe”). If Europe fails to do the same, it might eventually become dependent on technology which 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 few sectors where it does not have the ambition to become a world leader (artificial intelligence and renewable energy being just two examples). This is evident from the quickly growing number of patent applications by Chinese companies.

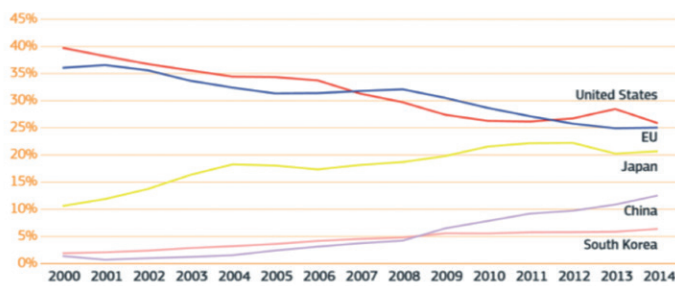


Figure 198: World share (%) of PCT patent applications 2000-2014
Source: DG Research and Innovation

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 [270]. 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 program, 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 (BAT) is an asset in developing advanced AI applications [88].

2.7.1.4.5 Political instability

Another threat is the political instability that Europe and the rest of the world are currently experiencing. Terrorist attacks, Brexit, financial problems and the refugee crisis are influencing business and consumer confidence. The current uncertainty regarding how the UK will leave the European Union is having an impact on the international relations of certain UK companies and universities. The refugee crisis is also contributing to political changes in many European countries.

2.7.1.5 CONCLUSION

Europe’s position in the world is weakening. Not because Europe is doing worse, but because the rest of the world is getting better faster. There are a number of hard-to-change facts that make it more difficult to compete with the rest of the world. To name a few:

- Europe has the oldest average age of all continents. It has double the number of people aged 60 or older than the average of the world. In 2040 one third of the European population will be 60 years or older. The more active professionals there are in a country, the more that country can innovate.
- The European population is predicted to grow by 3.9% by 2040 compared to 2015, while the population of the world as a whole will grow by more than 15%. Europe’s share in the world population will drop from 6.6% to 5.8%. Europe’s impact and influence in the world will decrease. The power of demographics is absolute.
- Despite efforts by the European Union to create a digital single market, Europe will stay a fragmented market with respect to languages, currencies and culture, which makes scaling up companies in their home market more challenging than in the USA or China, for example.
- Viewed on a global scale, Europe is relatively small, densely populated, and does not have many natural resources of its own (oil, gas, minerals, ...). Its economy heavily depends on global trade.

Given the above, Europe will have a hard time to compete and stay ahead of countries with a very young and dynamic population, eager to build a life, and to work hard. This does not mean that Europe won’t be a good place to live in the future, but the solutions that work well in other major countries might not work equally well in Europe.

The biggest challenge seems to be how to sustain economic growth with a shrinking active population and a growing retired population that depends on the government for healthcare, social care and retirement benefits. This can only be done by improving the productivity per person, or by importing young qualified workers. The latter is difficult for two reasons: (i) many of the countries of origin of these workers also face a shrinking workforce, and (ii) there are limits to the number of migrant workers countries want to admit. Fully compensating for the retirement of the “baby boom” generation won’t be possible.

There are measures that can help increase productivity, including:

- Further automating routine and non-routine tasks in the manufacturing and service sectors.
- Stimulating innovation and start-up creation in the whole population. A large number of start-ups is a prerequisite to ensure scale-ups.
- Adapting education even more to the needs of the job market and make sure that there are enough graduates in disciplines where there is a lack of workers (including technology and healthcare).
- In any event, such measures will need to be implemented sooner rather than later, in order to try and ensure that Europe retains its place on the global stage.

3

GLOSSARY

III-V	Chemical compounds with at least one group III element and at least one group V element.
ACAS	Airborne Collision Avoidance Systems
AES	Advanced Encryption Standard
AGI	Artificial General Intelligence
AI	Artificial Intelligence
AlphaZero	A computer program developed by DeepMind that can master Go, chess and shogi. This is in contrast with DeepMind's better known AlphaGo, which could only master Go.
ALU	Arithmetic Logic Unit
ANT	Multicast wireless sensor network technology, designed by ANT Wireless
API	Application Programming Interface
ASIC	Application-Specific Integrated Circuits are integrated circuits designed for a particular purpose, as opposed to being applicable for general use in many different situations.
Auto-ML	Techniques to design the meta-parameters associated with deep learning networks
AWS	Amazon Web Services
B2B	business-to-business
B2C	business-to-consumer
BAITX	Baidu, Alibaba, Tencent, Xiaomi
Bayesian computing	Bayesian computing refers to computational methods that are based on Bayesian (probabilistic) statistics.
BDVA	Big Data Value Association
Big data	Complex and exceedingly large data sets
BLE	Bluetooth Low Energy
C2PS	Cognitive Cyber-Physical Systems
C3PS	Connected Cognitive Cyber-Physical Systems
CAD	Computer-Aided Design
CAGR	Compound annual growth rate is a specific business and investing term for the smoothed annualised gain of an investment over a given time period.
CBRAM	Conductive-Bridging RAM
CGRA	Coarse-Grained Reconfigurable Architecture
CIS	CMOS Image Sensor
Cloud computing	Cloud computing is a paradigm whereby computing power is abstracted as a virtual service over a network. Executed tasks are transparently distributed.
CMOS	Complementary Metal–Oxide–Semiconductor is a common technology for constructing integrated circuits. CMOS technology is used in microprocessors, microcontrollers, static RAM, and other digital logic circuits.

CNTK	Microsoft Cognitive Toolkit
CPS	Cyber-Physical Systems combine computing resources and sensors/actuators that directly interact with and influence the real world. Robotics is one of the primary fields that works on such systems.
CPU	Central Processing Unit
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
D ₂ NN	Diffraction Deep Neural Network
DAB	Digital Audio Broadcasting
Data analytics	Data analytics examines large amounts of data to uncover hidden patterns, correlations and other insights.
DCC	Digital Compact Cassette
Declarative programming	Declarative programming is a programming paradigm that expresses the logic of a computation without describing its control flow. Many languages applying this style attempt to minimize or eliminate side effects by describing what the program should accomplish, rather than describing how to go about accomplishing it (the how is left up to the language's implementation). The opposite concept is imperative programming.
Deep learning	A class of machine learning techniques characterised by having deeply stacked layers that process the input.
DIY	Do-it-yourself
DLX	An ISA developed at Berkeley
DMD	Digital Micro-Mirror Devices
DNA	Deoxyribonucleic Acid
DRAM	Dynamic RAM
DSL	Domain Specific Language
DSSTNE	Deep Scalable Sparse Tensor Network Engine
DVR	Digital Video Recorder
ECG	Electrocardiography
ECSO	European Cyber Security Organisation
Edge computing	Edge Computing is pushing the frontier of computing applications, data, and services away from centralized nodes to the logical extremes of a network. It enables analytics and knowledge generation to occur at the source of the data.
EMIB	Embedded Multi-Die Interconnect Bridge
EPI	European Processor Initiative
ERI	Electronics Resurgence Initiative
EUV	Extreme ultraviolet lithography is a next-generation lithography technology using an extreme ultraviolet (EUV) wavelength, currently expected to be 13.5 nm.
FDSOI	Fully Depleted Silicon On Insulator (MOSFETs). For a FDSOI MOSFET the sandwiched p-type film between the gate oxide (GOX) and buried oxide (BOX) is very thin so that the depletion region covers the whole film. In FDSOI the front gate (GOX) supports less depletion charges than the bulk transistors so an increase in inversion charges occurs resulting in higher switching speeds. Other drawbacks in bulk MOSFETs, like threshold voltage roll off, higher sub-threshold slop body effect, etc. are reduced in FDSOI since the source and drain electric fields cannot interfere, due to the BOX (adapted from Wikipedia).

FEOL	Front-End-Of-Line, is the first step in fabricating an IC, in which devices on the wafer (such as transistors, resistors, etc.) are formed.
FET	Field-effect transistor
FHE	Fully Homomorphic Encryption, a form of encryption that allows computations being performed on data without having the key to decrypt that data
FinFET	The term FinFET was coined by University of California, Berkeley researchers (Profs. Chenming Hu, Tsu-Jae King-Liu and Jeffrey Bokor) to describe a nonplanar, double-gate transistor built on an SOI substrate.... The distinguishing characteristic of the FinFET is that the conducting channel is wrapped by a thin silicon 'fin', which forms the body of the device. In the technical literature, FinFET is used somewhat generically to describe any fin-based, multigate transistor architecture regardless of number of gates (from Wikipedia).
FMCG	Fast-moving consumer goods
Fog computing	Fog computing is an architecture that uses one or more end-user clients or near-user edge devices to carry out a substantial amount of storage (rather than stored primarily in cloud data centres), communication (rather than routed over the internet backbone), control, configuration, measurement and management.
FPGA	Field-Programmable Gate Array
GaAs	Gallium arsenide
GaN	Gallium nitride
Generative Design	Generative design is a technology that starts with your design goals and then explores all of the possible permutations of a solution to find the best option. Using cloud computing, generative design software quickly cycles through thousands—or even millions—of design choices, testing configurations and learning from each iteration what works and what doesn't. The process lets designers generate brand new options, beyond what a human alone could create, to arrive at the most effective design
GAFAM	Google, Apple, Facebook, Amazon, Microsoft
GCC	The GNU Compiler Collection
GDP	Gross Domestic Product
GDPR	General Data Protection Regulation (EU) 2016/679
GPL	General Purpose Language
GPU	A Graphics Processing Unit refers to the processing units on video cards. In recent years, these have evolved into massively parallel execution engines for floating point vector operations, reaching performance peaks of several gigaflops.
HBM	High-Bandwidth Memory
HDD	Hard Disk Drive
HHS	Department of Health and Human Services
HiPEAC	The European Network of Excellence on High Performance and Embedded Architecture and Compilation coordinates research, facilitates collaboration and networking, and stimulates commercialization in the areas of computer hardware and software research.
HMC	Hybrid Memory Cube
Homomorphic encryption	Homomorphic systems send encrypted data to an application (generally executed on a remote server) and let application perform its operations without ever decrypting the data. As a result the application never knows the actual data, nor the results it computes.
HPC	High Performance Computing

HPDA	High-Performance Data Analytics
HTML	HyperText Markup Language
HTTP	HyperText Transfer Protocol
IARPA	Intelligence Advanced Research Projects Activity
ICT	Information & Communication Technology is a generic term used to refer to all areas of technology related to computing and telecommunications.
IDM	Integrated Device Manufacturer
IGZO	Indium-Gallium-Zinc-Oxide
Imperative programming	Imperative programming is a programming paradigm that describes computation in terms of statements that change a program state. In much the same way that the imperative tense in natural languages expresses commands to take action, imperative programs define sequences of commands for the computer to perform. The opposite concept is declarative programming.
InFO	Integrated Fan-Out (InFO) advanced packaging technology from Taiwan Semiconductor Manufacturing Company (TSMC).
Internet of Things	The Internet of Things (IoT) is a computing concept that describes a future where everyday physical objects will be connected to the Internet and will be able to identify themselves to other devices.
IP	Internet Protocol
IP block	Intellectual property block, is a reusable unit of logic, cell, or chip layout design that is the intellectual property of one party. IP cores may be licensed to another party or can be owned and used by a single party alone. IP blocks can be used as building blocks within ASIC chip designs or FPGA logic designs.
IPCC	Intergovernmental Panel on Climate Change
ISA	An Instruction Set Architecture is the definition of the machine instructions that can be executed by a particular family of processors.
ISS	Integrated Smart Systems
ITAR	International Traffic in Arms Regulations
JVM	Java Virtual Machine
LCD	Liquid Crystal Display
LED	Light Emitting Diode
LIDAR	Light Detection And Ranging is a technology that measures distance by illuminating a target with a laser.
LLVM	The LLVM Project is a collection of modular and reusable compiler and toolchain technologies.
LTE	Long Term Evolution, a standard for mobile internet communications
LTPS	Low Temperature Polycrystalline Silicon
MCU	Micro Controller Unit
MEMS	Micro-Electrical-Mechanical Systems
MFM	Magnetic Force Micrograph
MIPS	Microprocessor without Interlocked Pipeline Stages, a RISC ISA
MNIST	A large database of handwritten digits
MOS	Metal-Oxide-Semiconductor

MPU	Micro Processor Unit
MRAM	Magnetic RAM
MRI	Magnetic Resonance Imaging
MTJ	Magnetic Tunnel Junction
NAND	NOT-AND, a type of logic gate
NAS	Network attached storage
NES	Nintendo Entertainment System
Neural networks	Neural networks are computational entities that operate in a way that is inspired by how neurons and synapses in an organic brain are believed to function. They need to be trained for a particular application, during which their internal structure is modified until they provide adequately accurate responses for given inputs.
Neuromorphic	Analog, digital, or mixed-mode analogue/digital VLSI and software systems that implement models of neural systems.
NFC	Near Field Communication
NIST	National Institute of Standards and Technology
NML	Nanomagnet Logic Quantum Cellular Automata
NoC	Network-on-Chip
NOR	NOT-OR, a type of logic gate
NRE	Non-Recurring Engineering costs refer to one-time costs incurred for the design of a new chip, computer program or other creation, as opposed to marginal costs that are incurred per produced unit.
NSA	National Security Agency
NVM	Non-Volatile Memory
OCP	Open Compute Project
OECD	Organisation for Economic Co-operation and Development
OLED	Organic Light Emitting Diode
OS	Operating system
Open source	Projects (software, schematics, etc.) in which the relevant source files are distributed to end users. Depending on the type of license, users can also be allowed to modify and redistribute these projects.
Operational research	Mathematical study of making decisions.
OPU	Optical Processing Unit, produced by Lighton
OPV	Organic Photovoltaics
ORNL	Oak Ridge National Laboratory
OSAT	Outsourced Semiconductor Assembly & Test, companies performing IC packaging and testing
Ox RAM	Oxide based RAM
PCB	Printed Circuit Board
PCM / PCRAM	Phase Change Memories
PCR	Polymerase Chain Reaction

PDMS	PolyDimethylSiloxane
PHP	A programming language.
PII	Personally Identifiable Information
Post-quantum cryptography	Field of study in which cryptography is made secure in the presence of quantum computers
Programming model	A programming model is a collection of technologies and semantic rules that enable the expression of algorithms in an efficient way. Often, such programming models are geared towards a particular application domain, such as parallel programming, real-time systems, image processing ...
Pseudo-quantum computing	Pseudo-quantum computing is a term used to refer to machines that allegedly are quantum computers, but that in practice have not been proven to be actually faster than regular computers executing very optimized algorithms.
Python	A programming language
QoS	Quality of Service.
RAM	Random-Access Memory
Reservoir computing	Reservoir Computing is similar to neural networks, but rather than modifying the internal structure during the training phase, the way to interpret the output is adjusted until the desired accuracy has been obtained.
REST	Representational State Transfer. A paradigm for transferring, accessing, and manipulating textual data in a stateless manner.
RFID	Radio-Frequency Identification is the use of a wireless non-contact system that uses radio-frequency electromagnetic fields to transfer data from a tag attached to an object, for the purposes of automatic identification and tracking.
RISC	Reduced Instruction Set Computing, a type of Instruction Set Architecture generally characterised by a simple and general design rather than having a large set of instructions, many of which are complex or specialised
RISC-V	An open RISC Instruction Set Architecture, developed at UC Berkeley
RNA	Ribonucleic Acid
ROM	Read-Only Memory
RSA	A cryptographic algorithm, named after its inventors Ron Rivest, Adi Shamir and Len Adleman
RTL	Register-Transfer Level
SAN	Storage area network, a dedicated network that connects a set of storage devices that are able to share low-level data with each other.
Secure multi-party computation	A computation in which several parties compute the result of a function on different inputs together, while at the same time keeping these different inputs secret from each other.
SEM	Scanning Electron Microscope
SGX	SGX Software Guard Extensions, an extension to Intel's x86 ISA
Si	Silicon
SIMD	Single Instruction, Multiple Data
SME	Small and Medium-sized Enterprise, a company of up to 250 employees.
SoC	A System on Chip refers to integrating all components required for the operation of an entire system, such as processors, memory, and radio, on a single chip.

SPARC	Scalable Processor Architecture, a RISC ISA developed by Sun Microsystems.
SPARK	A formally defined computer programming language based on the Ada programming language.
Spike computations	A programming model where large collections of devices, modelled after neurons, interact through the transmission of spike signals
SRAM	Static RAM
STDP	Spike-Timing-Dependent Plasticity is a biological process that adjusts the strength of connections between neurons in the brain. The process adjusts the connection strengths based on the relative timing of a particular neuron's input and output action potentials (or spikes).
STEM	Science, Technology, Engineering and Mathematics
Streaming analytics	Streaming analytics, also called event stream processing, is the analysis of large, in-motion data called event streams. The growing number of connected devices—the Internet of Things—will exponentially increase the volume of events that surround business activity. The more data is generated, the greater the potential benefits from streaming analytics.
SVM	Support Vector Machine
SWOT	Strengths, Weaknesses, Opportunities, Threats
TCP	Transmission Control Protocol
TFET	Tunnel FET
TFLOPS	TeraFLOPs, 10 ¹² floating-point operations per second
TFT	Thin-Film Transistor
TLS	Transport Layer Security
TOF	Time-of-Flight
TPU	Tensor Processing Unit
TRL	Technology Readiness Level
TSV	Through Silicon Via, a (vertical) electrical interconnect that goes through a silicon die or wafer (“via” = vertical interconnect access)
TSX	Transactional Synchronization Extensions, an extension to Intel's x86 ISA
UML	Unified Modelling Language is a general-purpose, developmental, modelling language in the field of software engineering, that is intended to provide a standard way to visualize the design of a system.
URL	Uniform Resource Locator
USB	Universal Serial Bus
UTP	Unshielded Twisted Pair
VHDL	VHSIC (Very High Speed Integrated Circuit) Hardware Description Language
VLSI	Very-large-scale integration is the process of creating integrated circuits by combining thousands of transistors into a single chip.
VUCA	Volatile, Uncertain, Complex and Ambiguous
WSN	Wireless Sensor Network
XAI	Explainable Artificial Intelligence

4

REFERENCES

- [1] Ari Sorsaniemi, "5G and Energy Efficiency", Future Connectivity Systems https://docbox.etsi.org/Workshop/2017/20171123_ITU_ETSI_ENV_REQ_5G/KEYNOTE/KEYNOTE_5G_EE_SORSANIEMI_EC.pdf
- [2] EC Digital Single Market, "A wearable device to detect cardiac arrhythmias". <https://ec.europa.eu/digital-single-market/en/news/wearable-device-detect-cardiac-arrhythmias>
- [3] A. Arbesman, "It's complicated". Aeon Newsletter. Jan 2014 <https://aeon.co/essays/is-technology-making-the-world-indecipherable>
- [4] A. D. McKinnon, S. R. Thompson, R. A. Doroshchuk, G. A. Fink and E. W. Fulp, "Bio-Inspired Cyber Security for Smart Grid Deployments", Proc. IEEE Innovative Smart Grid Technologies, 2013
- [5] A. Dascalescu. "10 Alarming Cyber Security Facts that Threaten Your Data". Heimdal Security, May 2016. <https://heimdalsecurity.com/blog/10-surprising-cyber-security-facts-that-may-affect-your-online-safety/>
- [6] 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., <https://ieeexplore.ieee.org/document/1159651/>
- [7] A. Tang, S. Simha, and S. Stolfo. "CLKSCREW: exposing the perils of security-oblivious energy management." 26th USENIX Security Symposium. 2017.
- [8] AMASS: Architecture-driven, Multi-concern and Seamless Assurance and Certification of Cyber-Physical Systems" https://cordis.europa.eu/project/rcn/202642_en.html
- [9] Akopyan, F., J. Sawada, A. Cassidy, R. Alvarez-Icaza, J. Arthur, P. Merolla, N. Imam, et al. 2015. "TrueNorth: Design and Tool Flow of a 65 mW 1 Million Neuron Programmable Neurosynaptic Chip". IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems 34 (10): 1537-57. <https://doi.org/10.1109/TCAD.2015.2474396>
- [10] Alex Gray, "The 10 skills you need to thrive in the Fourth Industrial Revolution", 2016, <https://www.weforum.org/agenda/2016/01/the-10-skills-you-need-to-thrive-in-the-fourth-industrial-revolution/>
- [11] Amirali Boroumand et al, "Google Workloads for Consumer Devices: Mitigating Data Movement Bottlenecks" Proceedings of the 23rd International Conference on Architectural Support for Programming Languages and Operating Systems, 2018.
- [12] Anna Swanson, "The job market just recovered from the recession. Men and white people haven't", August 2017, <https://www.washingtonpost.com/news/wonk/wp/2017/08/04/the-job-market-just-recovered-from-the-recession-men-and-white-people-havent>
- [13] Annie Lowrey, "The Great Recession Is Still With Us", 2017, <https://www.theatlantic.com/business/archive/2017/12/great-recession-still-with-us/547268/>
- [14] Anthony Rose, Ben Ramsey, "Picking Bluetooth Low Energy Locks from a Quarter Mile Away". DEF CON 24, 2016
- [15] Anzt H, Quintana-Ortí. "Improving the energy efficiency of sparse linear system solvers on multicore and manycore systems". Phil.Trans.R.Soc.A 372:20130279. 2014. <http://dx.doi.org/10.1098/rsta.2013.0279> <http://rsta.royalsocietypublishing.org/content/roypta/372/2018/20130279.full.pdf>
- [16] Araci et al, "An Implantable Microfluidic Device for Self-monitoring of Intraocular Pressure", Nature Medicine, 2014.
- [17] Arman Shehabi, Sarah Smith, Dale Sartor, Richard Brown, Magnus Herrlin, Jonathan Koomey, Eric Masanet, Nathaniel Horner, Inês Azevedo, William Lintner, "United States Data Center Energy Usage Report (Lawrence Berkeley National Laboratory)", http://eta-publications.lbl.gov/sites/default/files/lbnl-1005775_v2.pdf
- [18] Ashley Carman, "A security developer wrote such a scathing Amazon review that the product disappeared", The Verge, Jul 5, 2016 <https://www.theverge.com/circuitbreaker/2016/7/5/12096520/bad-amazon-review-product-pulled-auyou>
- [19] Augur, <https://www.augur.net>
- [20] Ben Pallant, "A 180-Day CGM: Senseonics' Eversense XL Approved in Europe", 2017, <https://diatribe.org/180-day-cgm-senseonics-eversense-xl-approved-europe>
- [21] Dmitri E. Nikonov, Ian A. Young, "Benchmarking of Beyond-CMOS Exploratory Devices for Logic Integrated Circuits", 2015
- [22] Benjamin, B. V., P. Gao, E. McQuinn, S. Choudhary, A. R. Chandrasekaran, J. Bussat, R. Alvarez-Icaza, J. V. Arthur, P. A. Merolla, et K. Boahen. 2014. "Neurogrid: A Mixed-Analog-Digital Multichip System for Large-Scale Neural Simulations", Proceedings of the IEEE 102 (5): 699-716. <https://doi.org/10.1109/JPROC.2014.2313565>

- [23] Bernadette Houghton, "Preservation Challenges in the Digital Age", Jul 2016, <http://www.dlib.org/dlib/july16/houghton/07houghton.html>
- [24] Brandon Lucia et al., "Intermittent Computing: Challenges and Opportunities", 2nd Summit on Advances in Programming Languages (SNAPL 2017), 2017
- [25] Nasdaq, Building on the Blockchain, Nasdaq's Vision of Innovation", https://business.nasdaq.com/Docs/Blockchain%20Report%20March%202016_tcm5044-26461.pdf
- [26] C. Richardson et al., "New Development Platforms Emerge For Customer-Facing Applications", June 2014, <https://www.forrester.com/report/New+Development+Platforms+Emerge+For+CustomerFacing+Applications/-/E-RES113411#>
- [27] COMBEST: COMponent-Based Embedded Systems design Techniques" https://cordis.europa.eu/project/rcn/85417_en.html
- [28] CONCERTO: Guaranteed Component Assembly with Round Trip Analysis for Energy Efficient High-integrity Multi-core Systems" https://cordis.europa.eu/project/rcn/110387_en.html
- [29] CONTREX: Design of embedded mixed-criticality CONTRol systems under consideration of EXtra-functional properties" https://cordis.europa.eu/project/rcn/109948_en.html
- [30] Jina Moore, "Cambridge Analytica Had a Role in Kenya Election", Too, 20 Mar 2018, New York Times <https://www.nytimes.com/2018/03/20/world/africa/kenya-cambridge-analytica-election.html>
- [31] Stephen Cheng, "Can China build a US\$145 million superconducting computer that will change the world? South China Morning Post, 27 Aug 2018 <https://amp.scmp.com/news/china/society/article/2161390/can-china-build-us145-million-superconducting-computer-will>
- [32] Carbon NanoTechnology, nantero.com/technology/
- [33] Cathy O'Neil, "Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy", Crown, 2016
- [34] Chris Hughes, "What are the negative long term effects if the middle class is not rebuilt?", 2018, <https://www.quora.com/What-are-the-negative-long-term-effects-if-the-middle-class-is-not-rebuilt>
- [35] Chris Merriman, "California just rushed through its own GDPR-style privacy law", <https://www.theinquirer.net/inquirer/news/3035171/california-just-rushed-through-its-own-gdpr>
- [36] Phui San Cheong, Johan Bergs, Chris Hawinkel, Jeroen Famaey, "Comparison of LoRaWAN classes and their power consumption". In 2017 IEEE Symposium on Communications and Vehicular Technology (SCVT) <https://ieeexplore.ieee.org/document/8240313>
- [37] IARPA - Office of the Director of National Intelligence, "Cryogenic Computing Complexity (C3)", <https://www.iarpa.gov/index.php/research-programs/c3>
- [38] Swamit S. Tannu, Douglas M. Carmean, Moinuddin K. Qureshi, "Cryogenic-DRAM based Memory System for Scalable Quantum Computers: A Feasibility Study", MEMSYS 2017, memlab.ece.gatech.edu/papers/MEMSYS_2017_2.pdf
- [39] Mattathias Schwartz, "Cyberwar for Sale, 4 Jan 2017, New York Times <https://www.nytimes.com/2017/01/04/magazine/cyberwar-for-sale.html>
- [40] D. Castelvecchi, "The Black Box", Nature, Vol 538, Oct. 2016
- [41] DARPA, "DARPA Announces \$2 Billion Campaign to Develop Next Wave of AI Technologies", 2018, <https://www.darpa.mil/news-events/2018-09-07>
- [42] DZone Refcard #129, <https://dzone.com/refcardz/rest-foundations-restful?chapter=1>
- [43] Dan Nosowitz, "Your Set-Top Box Is Murdering Your Electric Bill. Here's What You Can Do", Popular Science, 2011. <https://www.popsci.com/gadgets/article/2011-06/least-electric-bill-murdering-dvrs-every-provider>
- [44] David McKay, "Sustainable Energy – without the hot air", 2011, <https://www.withouthotair.com/>
- [45] David Roberts, "Sucking carbon out of the air won't solve climate change", 2016, <https://www.vox.com/energy-and-environment/2018/6/14/17445622/direct-air-capture-air-to-fuels-carbon-dioxide-engineering>
- [46] David Robson, "The reasons why exhaustion and burnout are so common", <http://www.bbc.com/future/story/20160721-the-reasons-why-exhaustion-and-burnout-are-so-common>
- [47] Davies, M., N. Srinivasa, T. Lin, G. Chinya, Y. Cao, S. H. Choday, G. Dimou, et al. Loihi: A Neuromorphic Manycore Processor with On-Chip Learning". IEEE Micro 38 (1): 82-99, 2018. <https://doi.org/10.1109/MM.2018.112130359>
- [48] Suyog Gupta, Ankur Agrawal et al., "Deep Learning with Limited Numerical Precision", arXiv:1502.02551 [cs.LG], <https://arxiv.org/abs/1502.02551>
- [49] Diane Whitmore Schanzenbach, Ryan Nunn, Lauren Bauer, and Audrey Breitwieser, "The Closing of the Jobs Gap A Decade of Recession and Recovery," Aug 2017, http://www.hamiltonproject.org/assets/files/closing_jobs_gap_recession_recovery.pdf
- [50] Directive 95/46/EC, <https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:en:HTML>
- [51] Dmitri E. Nikonov , and Ian A. Young, "Benchmarking of Beyond-CMOS Exploratory Devices for Logic Integrated Circuits", 2015
- [52] Dong Up Lee et al. IEEE Journal of Solid-State Circuits, VOL. 50, NO. 1, JANUARY 2015

- [53] E. Kjeang et al., "Microfluidic Fuel Cells: A Review", *Journal of Power Sources*, 2009.
- [54] E. Ozer, "Toward a Self-Learning and Energy-Neutral IoT", *IEEE Micro*, Nov-Dec. 2016
- [55] E. W. Fulp, H. Donald Gage, David J. John, Matthew R. McNiece, William H. Turkett, and Xin Zhou, "An Evolutionary Strategy for Resilient Cyber Defense," *Proc. IEEE Globecom*, 2015
- [56] Kocabas, Ovunc, et al. "Assessment of cloud-based health monitoring using homomorphic encryption." 2013 IEEE 31st International Conference on Computer Design (ICCD). IEEE, 2013
- [57] Eleanor Krause, Isabel Sawhill, "Seven reasons to worry about the American middle class", 2018, <https://www.brookings.edu/blog/social-mobility-memos/2018/06/05/seven-reasons-to-worry-about-the-american-middle-class/>
- [58] Tim Simonite, "Even Artificial Neural Networks Can Have Exploitable 'Backdoors'", *Wired*, 25 August 2017 <https://www.wired.com/story/machine-learning-backdoors/>
- [59] Timothy Revell, "Eyes of Things Horizon 2020 project: Smart Doll with Emotion Recognition" <https://www.newscientist.com/article/2137835-smart-doll-fitted-with-ai-chip-can-read-your-childs-emotions/>
- [60] F.J. Cazorla et al. "PROARTIS: Probabilistically Analyzable Real-Time Systems". *ACM TECS*. 12:2s (2013)
- [61] F.J. Cazorla et al. "Probabilistic Worst-Case Timing Analysis: Taxonomy and Comprehensive Survey", *ACM CSUR* (2018), to appear
- [62] "Fiat Chrysler recalls 1.4 million cars after Jeep hack", *BBC News*, 24 July 2015 <https://www.bbc.com/news/technology-33650491>
- [63] Fiona McKenzie, "The fourth industrial revolution and international migration", *Lowy Institute for international policy*, 2017, <https://www.lowyinstitute.org/publications/fourth-industrial-revolution-and-international-migration>
- [64] Abbott Newsroom, "Freestyle Libre", www.abbott.com/corpnewsroom/product-and-innovation/revolutionizing-cgm-with-freestyle-libre.html
- [65] Furber, S. B., D. R. Lester, L. A. Plana, J. D. Garside, E. Painkras, S. Temple, et A. D. Brown. 2013. "Overview of the SpiNNaker System Architecture. *IEEE Transactions on Computers* 62 (12): 2454-67. <https://doi.org/10.1109/TC.2012.142>
- [66] G. De Micheli, R. Ernest, W. Wolf, "Readings in Hardware/Software Co-Design", *Morgan Kaufmann* (697 pp.), doi: <https://doi.org/10.1016/B978-1-55860-702-6.X5000-5>
- [67] G. M. Church, M. B. Elowitz, C. D. Smolke, C. A. Voigt and R. Weiss, "Realizing the Potential of Synthetic Biology," *Molecular Cell Biology*, 2014.
- [68] G. M. Whitesides, "The origins and the future of microfluidics", *Nature*, July 2006
- [69] Marco Brenner, "GDPR and protecting data privacy with cryptographic pseudonyms", *IBM Services Blog*, 9 Apr April 2018 <https://www.ibm.com/blogs/insights-on-business/gbs-strategy/gdpr-protecting-data-privacy-cryptographic-pseudonyms>
- [70] C. Gamrat, O. Bichler, et al. "Memristive based device arrays combined with Spike based coding can enable efficient implementations of embedded neuromorphic circuits". In 2015 IEEE International Electron Devices Meeting (IEDM), 4.5.1-4.5.7. <https://doi.org/10.1109/IEDM.2015.7409626>
- [71] Gene Amdahl, "Validity of the Single Processor Approach to Achieving Large Scale Computing Capabilities". In *Proceedings of the April 18-20, 1967, Spring Joint Computer Conference*, 483-485. *AFIPS '67 (Spring)*. New York, NY, USA: ACM. <https://doi.org/10.1145/1465482.1465560>.
- [72] Geoff V. Merrett, "Energy Harvesting and Transient Computing: A Paradigm Shift for Embedded Systems?", *DAC*, June 2016
- [73] George Papadopoulos et al., "Statistics on small and medium-sized enterprises", 2018, https://ec.europa.eu/eurostat/statistics-explained/index.php/Statistics_on_small_and_medium-sized_enterprises
- [74] Georgios Goumas, "Delivering the data: How HIPEAC is revolutionizing data centres, storage and networking", *HIPEACinfo* 55 pp.17-19, October 2018 https://www.hipeac.net/assets/public/publications/newsletter/hipeacinfo55_final_web.pdf
- [75] Gilad Lotan, "Israel, Gaza, War & Data: Social networks and the art of personalizing propaganda", 2014
- [76] H. Ye and M. Fussenegger, "Synthetic Therapeutic Gene Circuits in Mammalian Cells," *FEBS Letters*, August 2014
- [77] Haydn Thompson et al, "Platforms4HPCS: Key Outcomes and Recommendations", 2018, https://www.platforms4cps.eu/fileadmin/user_upload/E-Book_-_Platforms4CPS_Key_Outcomes_and_Recommendations.pdf
- [78] Heather Cleland Woods, Holly Scott, "#Sleepyteens: Social media use in adolescence is associated with poor sleep quality, anxiety, depression and low self-esteem", *Journal of Adolescence*, 51, August 2016, Pages 41-49
- [79] Bernard Marr, "How Blockchain Will Transform The Supply Chain And Logistics Industry". *Forbes Blog*, 23 Mar 2018, <https://www.forbes.com/sites/bernardmarr/2018/03/23/how-blockchain-will-transform-the-supply-chain-and-logistics-industry/#699060115fec>
- [80] I. Agirre et. al., "Fitting Software Execution-Time Exceedance into a Residual Random Fault in ISO-26262". *IEEE TREL* 67(3): 1314-1327 (2018)

- [81] ISO/IEC/JTC1/SC22/WG23, "Guidance to avoiding programming language vulnerabilities", http://www.open-std.org/jtc1/sc22/wg23/docs/ISO-IEC/JTC1-SC22-WG23_No823-tr24772-1-language-independent-guidance-after-meeting-27-20180827.docx
- [82] Joshua Althaus, "Illinois Government Pilots Blockchain Technology in Medical Licenses Issuance". Joshua Althaus, 12 Aug 2017, Cointelegraph <https://cointelegraph.com/news/illinois-government-pilots-blockchain-technology-in-medical-licenses-issuance>
- [83] Indiveri, Giacomo, Bernabe Linares-Barranco, Tara Julia Hamilton, André van Schaik, Ralph Etienne-Cummings, Tobi Delbruck, Shih-Chii Liu, et al. 2011. "Neuromorphic silicon neuron circuits". *Neuromorphic Engineering* 5: 73. <https://doi.org/10.3389/fnins.2011.00073>
- [84] Intel Development Forum 2009, <https://pdfs.semanticscholar.org/presentation/4dac/aaf7ca5fffc74a3e417fd69757be40823760.pdf>
- [85] Matthew H., "Intel® SGX for Dummies (Intel® SGX Design Objectives)", Intel Developer Zone, September 26, 2016. <https://software.intel.com/en-us/blogs/2013/09/26/protecting-application-secrets-with-intel-sgx>
- [86] Timothy Prickett Morgan, "Intel's Exascale Dataflow Engine Drops X86 And Von Neumann". [nextplatform.com](https://www.nextplatform.com/2018/08/30/intels-exascale-dataflow-engine-drops-x86-and-von-neuman/) August 30, 2018. <https://www.nextplatform.com/2018/08/30/intels-exascale-dataflow-engine-drops-x86-and-von-neuman/>
- [87] "International Magnetic Tape Storage Roadmap", 2011, <http://www.insic.org/news/A&S%20Roadmap.pdf>
- [88] Iris Deng, "China's AI industry gets the most funding, but lags the USA in key talent, says Tsinghua", 2018, <https://www.scmp.com/tech/china-tech/article/2155600/chinas-ai-industry-gets-most-funding-lags-us-key-talent-says>
- [89] J. A. N. Brophy and C. A. Voigt, "Principles of Genetic Circuit Design," *Nature Methods*, 2014
- [90] J. Bornholt, R. Lopez, D. M. Carmean, L. Ceze, G. Seelig and K. Strauss, "A DNA-Based Archival Storage System," in *Proceedings of the Twenty-First International Conference on Architectural Support for Programming Languages and Operating Systems*, 2016
- [91] J. Thönes, "Microservices," *IEEE Software*, 32(1):116, Jan 2015.
- [92] Janna Anderson, Lee Rainie, "The negatives of digital life", <http://www.pewinternet.org/2018/07/03/the-negatives-of-digital-life/>
- [93] Janna Anderson, Lee Rainie, "The positives of digital life", <http://www.pewinternet.org/2018/07/03/the-positives-of-digital-life/>
- [94] Jaron Lanier, "Ten Arguments for Deleting Your Social Media Accounts Right Now", Henry Holt and Co, 2018
- [95] Jean Twenge, Gabrielle Martin, and Brian Spitzberg, "Trends in U.S. Adolescents' Media Use, 1976-2016: The Rise of Digital Media, the Decline of TV, and the (Near) Demise of Print", 2018. *Psychology of Popular Media Culture*, <https://www.apa.org/pubs/journals/releases/ppm-ppm0000203.pdf>
- [96] Jean Twenge, Zlatan Krizan, Garrett Hisler, "Decreases in self-reported sleep duration among US adolescents 2009–2015 and association with new media screen time", *Sleep Medicine*, 39, Nov 2017, pp. 47-53
- [97] Jean Twenge, "Analysis: Teens are sleeping less. Why? Smartphones", <https://www.pbs.org/newshour/science/analysis-teens-are-sleeping-less-why-smartphones>
- [98] Jeremie S. Kim, Damla Senol Cali, Hongyi Xin, Donghyuk Lee, Saugata Ghose, Mohammed Alser, Hasan Hassan, Oguz Ergin, Can Alkan, and Onur Mutlu, "GRIM-Filter: Fast Seed Location Filtering in DNA Read Mapping Using Processing-in-Memory Technologies", *BMC Genomics*, 2018
- [99] John Detrixhe, "Why can't Europe create tech giants like the US and China?", 2018, <https://qz.com/1320983/why-arent-europes-technology-companies-as-big-as-in-the-us-and-china/>
- [100] Alex de Vries, "Bitcoin's Growing Energy Problem", *Joule*, Volume 2, ISSUE 5, P801-805, 16 May 2018, [https://www.cell.com/joule/fulltext/S2542-4351\(18\)30177-6](https://www.cell.com/joule/fulltext/S2542-4351(18)30177-6)
- [101] Jouppi, Norman P., et al. "In-datacentre performance analysis of a tensor processing unit." *Computer Architecture (ISCA)*, 2017 ACM/IEEE 44th Annual International Symposium on. IEEE, 2017.
- [102] Junwhan Ahn, Sungjoo Yoo, Onur Mutlu, and Kiyong Choi, "PIM-Enabled Instructions: A Low-Overhead, Locality-Aware Processing-in-Memory Architecture", *Proceedings of the 42nd International Symposium on Computer Architecture (ISCA)*, 2015
- [103] Junwhan Ahn, Sungpack Hong, Sungjoo Yoo, Onur Mutlu, and Kiyong Choi, "A Scalable Processing-in-Memory Accelerator for Parallel Graph Processing" *Proceedings of the 42nd International Symposium on Computer Architecture (ISCA)*, 2015
- [104] K. Clancy and C. A. Voigt, "Programming Cells: Towards an Automated Genetic Compiler," *Current Opinion in Biotechnology*, 2010.
- [105] K. Moammer. "US Government Bans Intel, NVIDIA and AMD From Selling High End Chips To The Chinese Government". *Wccf Tech*, April 2015. <http://wccftech.com/us-government-bans-intel-nvidia-amd-chips-china/#ixzz4HfzB874b>

- [106] K. Myny, "The development of flexible integrated circuits based on thin-film transistors", *Nature Electronics*, Jan 2018. <https://www.nature.com/articles/s41928-017-0008-6>
- [107] K. Yoongu, et al. "Flipping bits in memory without accessing them: An experimental study of DRAM disturbance errors." *ACM SIGARCH Computer Architecture News*. 42(3). IEEE Press, 2014.
- [108] Katarina Brooker, "I Was Devastated: Tim Berners-Lee, the Man Who Created the World Wide Web, Has Some Regrets", 2018
- [109] Kate Raworth, "Doughnut Economics", 2017, Chelsea Green Pub Co, <https://www.amazon.de/DOUGHNUT-ECONOMICS-Kate-Raworth/dp/1603586741>
- [110] Kea: A Computation Offloading System for Smartphone Sensor Data", 2017 IEEE 9th International Conference on Cloud Computing Technology and Science
- [111] Kevin Delaney, "The robot that takes your job should pay taxes, says Bill Gates", 2017, <https://qz.com/911968/bill-gates-the-robot-that-takes-your-job-should-pay-taxes/>
- [112] Khan et al, "Technologies for Printing Sensors and Electronics over Large Flexible Substrates: A Review", *IEEE Sensors Journal*, 2013. <https://ieeexplore.ieee.org/document/6974982>
- [113] Kimberley Holland, "How to Identify and Manage Phubbing", 2018, <https://www.healthline.com/health/phubbing>
- [114] Klaus Schwab, Richard Samans, "The Future of Jobs Employment, Skills and Workforce Strategy for the Fourth Industrial Revolution", 2016, http://www3.weforum.org/docs/WEF_Future_of_Jobs.pdf
- [115] Klaus Schwab, "The Fourth Industrial Revolution: what it means, how to respond", 2016, <https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond>
- [116] L. Cong, F. A. Ran, D. Cox, S. Lin, R. Barretto, N. Habib, P. D. Hsu, X. Wu, W. Jiang, L. A. Marraffini and F. Zhang, "Multiplex Genome Engineering using CRISPR/Cas Systems," *Science*, Feb 2013
- [117] L. Millet et al., "550ofps 85GOPS/W 3D stacked BSI vision chip based on parallel in-focal-plane acquisition and processing", *VLSI* 2018
- [118] Laurent Larger, Antonio Baylón-Fuentes, Romain Martinenghi, Vladimir S. Udaltsov, Yanne K. Chembo, and Maxime Jacquot, "High-Speed Photonic Reservoir Computing Using a Time-Delay-Based Architecture: Million Words per Second Classification", *Phys. Rev. X* 7, 011015, 6 Feb 2017
- [119] Lauren Sherman et al, "The Power of the Like in Adolescence", <http://journals.sagepub.com/doi/abs/10.1177/09567976166645673>
- [120] Laurence C. Smith, "The World in 2050: Four Forces Shaping Civilization's Northern Future", 2011, Plume, ISBN 978-0452297470
- [121] Lee Jong-Wha, "How will Asia's growing middle class change the world?", 2015, <https://www.weforum.org/agenda/2015/03/how-will-asias-growing-middle-class-change-the-world/>
- [122] Lee et al, "Human-level Concept Learning through Probabilistic Program Induction", *Science*, Vol. 350 Issue 6266, Dec. 2015
- [123] Luciano Floridi, Raja Chatila, Patrice Chazerand, and Christoph Luetge, "AI4People -An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations", 2018
- [124] M. AbdelBaky, et al., "Computing in the Continuum: Combining Pervasive Devices and Services to Support Data-Driven Applications," *ICDCS*, Atlanta, GA, 2017, pp. 1815-1824. doi: 10.1109/ICDCS.2017.323, URL: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7980120&isnumber=7979941>
- [125] M. Duranton et al., "Rapid Technology-Aware Design Space Exploration for Embedded Heterogeneous Multiprocessors" in *Processor and System-on-Chip Simulation*, Ed. R. Leupers, 2010
- [126] M. Karczynski, A. Hamieh, J. H. Huh, H. Holm, S. R. Rajagopalan, and N. H. Fefferman, "Hive Oversight for Network Intrusion Early-Warning Using DIAMoND (HONIED): A Bee-Inspired Method for Fully Distributed Cyber Defense", *IEEE Communications Magazine*, June 2016
- [127] M. Rouse. "Edge Computing". *SearchDataCenter*, 2016.
- [128] M.L. Abbot and M.T. Fisher, "The Art of Scalability", 2nd Ed., Addison Wesley, ISBN-10: 0-13-403280-2
- [129] Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin, "Why Should I Trust You? Explaining the Predictions of Any Classifier", 2016
- [130] Marjan Mernik, University of Maribor, Slovenia, "Formal and Practical Aspects of Domain-Specific Languages: Recent Developments": <http://reed.cs.depaul.edu/peterh/class/csc458/Examples/Formal%20and%20Practical%20Aspects%20of%20Domain-Specific%20Languages%20-%20Recent%20Developments.pdf>
- [131] Mark Lantz, "Why the future of Data Storage is (still) Magnetic Tape, 2018, <https://spectrum.ieee.org/computing/hardware/why-the-future-of-data-storage-is-still-magnetic-tape>
- [132] Martin Fowler, *Microservices*, <https://martinfowler.com/articles/microservices.html>, Mar 2014
- [133] Maureen Dowd, "Elon Musk's Billion-Dollar Crusade to Stop the A.I. Apocalypse", 2017, <https://www.vanityfair.com/news/2017/03/elon-musk-billion-dollar-crusade-to-stop-ai-space-x>
- [134] Meltdown and Spectre, <https://meltdownattack.com>

- [135] Michael Ashby, "Materials and Sustainable Development", 2015, Elsevier Science & Technology
- [136] Michael J. Miller, "Google I/O: 11 Big Trends", 2016
- [137] "Intel Microcode Revision Guidance", 2 Apr 2018, <https://newsroom.intel.com/wp-content/uploads/sites/11/2018/04/microcode-update-guidance.pdf>
- [138] Tom Warren, "Microsoft reveals how Spectre updates can slow your PC down", The Verge, 9 Jan 2018. <https://www.theverge.com/2018/1/9/16868290/microsoft-meltdown-spectre-firmware-updates-pc-slowdown>
- [139] Mike Tom, "Tech giants paying big money for AI talent", 2016, <https://pitchbook.com/news/articles/tech-giants-paying-big-money-for-ai-talent>
- [140] Miran Kim, Kristin Lauter, "Private genome analysis through homomorphic encryption", BMC Med Inform Decis Mak. 2015; 15(Suppl 5): S3.
- [141] Paulius Micikevicius, "Mixed-Precision Training of Deep Neural Networks", October 11, 2017, <https://devblogs.nvidia.com/mixed-precision-training-deep-neural-networks/>
- [142] Mos Zhang, "Tencent ML Team Trains ImageNet In Record Four Minutes", 2018
- [143] Myny, K. et al. "An 8-bit, 40-instructions-per-second organic microprocessor on plastic foil". IEEE J. Solid-St. Circ. 47, 284–291 (2012) <https://ieeexplore.ieee.org/document/6069822>
- [144] NEXOF-RA: NESSI Open Framework - Reference Architecture https://cordis.europa.eu/project/rcn/86242_en.html
- [145] 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.
- [146] Nicholas Carr. "The Shallows: What the Internet Is Doing to Our Brains", 2010, Norton & Company
- [147] Nick Routley, "From vinyl to streaming: 40 years of the global music industry visualized", 2018, <https://www.weforum.org/agenda/2018/10/visualizing-40-years-of-music-industry-sales/>
- [148] Nicole Kobie, "What is the gig economy and why is it so controversial?", WIRED, 2018, <https://www.wired.co.uk/article/what-is-the-gig-economy-meaning-definition-why-is-it-called-gig-economy>
- [149] Noah Kulwin, "The Internet Apologizes", The New York magazine, 2018, <http://nymag.com/selectall/2018/04/an-apology-for-the-internet-from-the-people-who-built-it.html>
- [150] Olga Khazan, "How Smartphones Hurt Sleep", 2015
- [151] Olivia, Goldhill, "Should driverless cars kill their own passengers to save a pedestrian?", <https://qz.com/536738/should-driverless-cars-kill-their-own-passengers-to-save-a-pedestrian/>
- [152] Organick et al. "Random access in large-scale DNA data storage", Nature Biotechnology. 36 (3), 2018
- [153] "Overview of the national laws on electronic health records in the EU Member States and their interaction with the provision of cross-border eHealth services, Final report and recommendations", https://ec.europa.eu/health/sites/health/files/ehealth/docs/laws_report_recommendations_en.pdf
- [154] P-Yu Chen, C-Chao Lin, S-Ming Cheng, H-Chun Hsiao and C-Ying Huang, "Decapitation via Digital Epidemics: A Bio-Inspired Transmissive Attack", IEEE Communication Magazine, June 2016
- [155] P. Caserta, O. Zendra. "Visualization of the Static aspects of Software: a survey". IEEE Transactions on Visualization and Computer Graphics, Institute of Electrical and Electronics Engineers, 2011, 17 (7), pp.913-933
- [156] PROWSSS: Property-based testing of Web services https://cordis.europa.eu/project/rcn/105389_en.html
- [157] Bryce Kellogg, Vamsi Talla, Shyamnath Gollakota, and Joshua R. Smith, "Passive Wi-Fi: Bringing Low Power to Wi-Fi Transmissions". 13th USENIX Symposium on Networked Systems Design and Implementation (NSDI '16), 2016. <https://www.usenix.org/system/files/conference/nsdi16/nsdi16-paper-kellogg.pdf>
- [158] Patrick Watson, "The Middle Class Might Nearly Disappear In The Next Decade", 2018, <https://www.forbes.com/sites/patrickwatson/2018/03/08/the-middle-class-might-nearly-disappear-in-the-next-decade/#630cf492251e>
- [159] Peter Temin, "The Vanishing Middle Class", 2017, MIT Press
- [160] Philipp Blom, "Was auf dem Spiel steht", 2017, Carl Hanser Verlag GmbH
- [161] Artem Dementyev, Steve Hodges, Stuart Taylor, Joshua Smith, Power consumption analysis of Bluetooth Low Energy, ZigBee and ANT sensor nodes in a cyclic sleep scenario". In 2013 IEEE International Wireless Symposium (IWS) <https://ieeexplore.ieee.org/document/6616827>
- [162] Douglas M. Carmean, "Quantum and Cryo and DNA, oh my! Sights Along the New Yellow Brick Road, ISCA 2016 Keynote, June 2016 dcarmean.azurewebsites.net/ISCA2016.pdf
- [163] Max Tegmark, 7 Dec 2017, <https://twitter.com/tegmark/status/938820518680993792?lang=en>
- [164] R. Iyer, "Visual IoT: Architectural Challenges and Opportunities", IEEE Micro, Nov-Dec. 2016

- [165] RJ Reinhart, "Most US Workers Unafraid of Losing Their Jobs to Robots", 2018, <https://news.gallup.com/poll/226841/workers-unafraid-losing-jobs-robots.aspx>
- [166] Richard Chastelein, BlockchainNews, Jam 2017, <https://www.the-blockchain.com/2017/01/05/move-bitcoin-mit-cryptographer-silvio-micali-public-ledger-algorand-future-blockchain/>
- [167] Richard Freed, "The Tech Industry's War on Kids: How psychology is being used as a weapon against children", 2018, <https://medium.com/@richardnfreed/the-tech-industrys-psychological-war-on-kids-c452870464ce>
- [168] Richard Gray, "Are you a sleep procrastinator?", 2017, <http://www.bbc.com/capital/story/20170911-are-you-a-sleep-procrastinator>
- [169] Richard Heinberg, "Systemic Change Driven by Moral Awakening is Our Only Hope", <https://www.ecowatch.com/climate-change-heinberg-2471869927.html>
- [170] S. Coughlan. Digital dependence 'eroding human memory'. BBC News, October 2015. <http://www.bbc.com/news/education-34454264>
- [171] S. Forrest, A. S. Perelson, L. Allen and R. Cherukuri, "Self-nonsel Self Discrimination in a Computer", Proceedings of the IEEE Symp. on Research in Security and Privacy, 1994
- [172] S. Forrest, A. Somayaji and D. H. Ackley, "Building Diverse Computer Systems", Proceedings of the 6th Workshop on Hot Topics in Operating Systems, 1997
- [173] S. L. Shipman, J. Nivala, J. D. Macklis and G. M. Church, "CRISPR-Cas Encoding of a Digital Movie into the Genomes of a Population of Living Bacteria," Nature, July 2017.
- [174] S. Wiesner, Conjugate Coding, SIGACT News 15:1, pp. 78–88, 1983.
- [175] S.-M. Cheng, W. C. Ao, P-Yu Chen and K-Cheng Chen, "On Modeling Malware Propagation in Generalized Social Networks," IEEE Commun. Lett., Jan. 2011
- [176] SAFURE: SAFety and secURity by design for interconnected mixed-critical cyber-physical systems https://cordis.europa.eu/project/rcn/194149_en.html
- [177] SATURN: SysML bAsed modeling, architecTUre exploRation, simulation and syNthesis for complex embedded systems https://cordis.europa.eu/project/rcn/85357_en.html
- [178] SESAMO: Security and Safety Modelling https://cordis.europa.eu/project/rcn/104955_en.html
- [179] Samidh Chakrabarti, "Hard Questions: What Effect Does Social Media Have on Democracy?", <https://newsroom.fb.com/news/2018/01/effect-social-media-democracy/>
- [180] Samuel K. Mooren "Two Startups Use Processing in Flash Memory for AI at the Edge", 2018, <https://spectrum.ieee.org/tech-talk/computing/embedded-systems/two-startups-use-processing-in-flash-memory-for-ai-at-the-edge>
- [181] Saugata Ghose, Kevin Hsieh, Amirali Boroumand, Rachata Ausavarungnirun, and Onur Mutlu, "Enabling the Adoption of Processing-in-Memory: Challenges, Mechanisms, Future Research Directions", 2018, <https://arxiv.org/pdf/1802.00320.pdf>
- [182] Schemmel, J., D. Brüderle, A. Grübl, M. Hock, K. Meier, et S. Millner. 2010. "A wafer-scale neuromorphic hardware system for large-scale neural modeling". In Proceedings of 2010 IEEE International Symposium on Circuits and Systems (ISCAS), 1947-50 <https://doi.org/10.1109/ISCAS.2010.5536970>
- [183] Schien D., Coroama V.C., Hilty L.M., Preist C. (2015) The Energy Intensity of the Internet: Edge and Core Networks. In: Hilty L., Aebischer B. (eds) ICT Innovations for Sustainability". Advances in Intelligent Systems and Computing, vol 310. Springer, Cham
- [184] "Seagate's Enterprise Storage: Helium-Filled Drives", New Technology To Drive Future Growth. Forbes, September 2015. <http://www.forbes.com/sites/greatspeculations/2015/09/10/seagates-enterprise-storage-helium-filled-drives-new-technology-to-drive-future-growth/2/#6025c43f6dfd>
- [185] "ServFace: Service Annotations for User Interface Composition" https://cordis.europa.eu/project/rcn/85557_en.html
- [186] "Servitization in Industry", Gunter Lay (editor), 2014, Springer, ISBN 978-3-319-06934-0, doi: 10.1007/978-3-319-06935-7.
- [187] "Social Media as a Tool of Hybrid Warfare", Nato Strategic Communications Centre Of Excellence, ISBN 978-9934-8582-6-0, <https://www.stratcomcoe.org/social-media-tool-hybrid-warfare>
- [188] Sourced from <https://insights.stackoverflow.com/survey/2018/>
- [189] Sourced from: <https://insights.stackoverflow.com/survey/2018/>
- [190] Sourced from: <https://spectrum.ieee.org/at-work/innovation/the-2018-top-programming-languages>
- [191] Stacey Higginbotham, "The Internet of Trash: IoT Has a Looming E-Waste Problem", 2018, <https://spectrum.ieee.org/telecom/internet/the-internet-of-trash-iot-has-a-looming-ewaste-problem>
- [192] Stijn Baert et al., "Smartphone Use and Academic Performance: Correlation or Causal Relationship?" <http://ftp.iza.org/dp11455.pdf>
- [193] Stuart Summer, "Why the UK can't produce billion dollar companies", 2018, <https://www.computing.co.uk/ctg/news/3062574/why-the-uk-cant-produce-billion-dollar-companies>

- [194] "Students, Computers and Learning: Making the connection", OECD Publishing, 2015, https://read.oecd-ilibrary.org/education/students-computers-and-learning_9789264239555-en
- [195] Sunil, Archana Bindu, Zekeriya Erkin, and Thijs Veugen. "Secure matching of Dutch car license plates." Signal Processing Conference (EUSIPCO), 2016 24th European. IEEE, 2016
- [196] Susan Weinschenk, "The True Cost Of Multi-Tasking", 2018, <https://www.psychologytoday.com/us/blog/brain-wise/201209/the-true-cost-multi-tasking>
- [197] TOP500.org. "TOP 10 Sites for June 2016". June 2016. <https://www.top500.org/lists/2016/06>
- [198] Tom Bawden, "Global warming: Data centres to consume three times as much energy in next decade, experts warn, The Independent, 23 Jan 2016, <https://www.independent.co.uk/environment/global-warming-data-centres-to-consume-three-times-as-much-energy-in-next-decade-experts-warn-a6830086.html>
- [199] Andy Greenberg, "The Jeep Hackers Are Back to Prove Car Hacking Can Get Much Worse", Wired, 01 August 2016 <https://www.wired.com/2016/08/jeep-hackers-return-high-speed-steering-acceleration-hacks/>
- [200] The New York Magazine, April 16, 2018
- [201] The Royal Academy of Engineering, "Synthetic Biology: Scope, Applications and Implications," 2009.
- [202] Timothy Prickett Morgan, "The Skinny On Future Cascade Lake Xeons". nextplatform.com, August 22, 2018 <https://www.nextplatform.com/2018/08/22/the-skinny-on-future-cascade-lake-xeons/>
- [203] Tianyu Gu, Brendan Dolan-Gavitt, Siddharth Garg, "BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain", arXiv:1708.06733v1 [cs.CR], 2017
- [204] Toffoli, "Programmable Matter Methods", <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.72.226&rep=rep1∓type=pdf>
- [205] Tom Wijman, "Mobile Revenues Account for More Than 50% of the Global Games Market as It Reaches \$137.9 billion in 2018", 2018
- [206] Tony Hey, Stewart Tansley, Kristin Tolle, "The Fourth Paradigm: Data-Intensive Scientific Discovery", 2009, Microsoft Press
- [207] 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
- [208] Tucker Davey, "Lethal Autonomous Weapons: An Update from the United Nations", Future of Life Institute, 2018, <https://futureoflife.org/ai-open-letter/?cn-reloaded=1>
- [209] Ujjwal Kumar, "Which country has the most effective cyberarmy?", Quora, 2016, <https://www.quora.com/Which-country-has-the-most-effective-cyberarmy>
- [210] Alexandru Oprunenco, Chami Akmeemana, "Using blockchain to make land registry more reliable in India", LSE Business Review blog, 13 Apr 2018 blogs.lse.ac.uk/businessreview/2018/04/13/using-blockchain-to-make-land-registry-more-reliable-in-india/
- [211] Victor Tangermann, "I Tried Apple's Screen Time Tool. The Results Were Not Heartening." Futurism, 13 Aug 2018, <https://futurism.com/screen-time-smartphone-obsession/>
- [212] Virginia Eubanks, "Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor", St Martin's Press, 2018
- [213] Vivek Seshadri, Donghyuk Lee, Thomas Mullins, Hasan Hassan, Amirali Boroumand, Jeremie Kim, Michael A. Kozuch, Onur Mutlu, Phillip B. Gibbons, and Todd C. Mowry, "Ambit: In-Memory Accelerator for Bulk Bitwise Operations Using Commodity DRAM Technology", 2017, Proceedings of the 50th International Symposium on Microarchitecture (MICRO)
- [214] W. Mazurczyk and E. Rzeszutko, "Security – A Perpetual War: Lessons from Nature", IT Professional, 2015
- [215] W. S. Harrison, N. Hanebutte, P.W. Oman, and J. Alves-Foss. "The MILS Architecture for a Secure Global Information Grid" Crosstalk: The Journal of Defense Software Engineering, Oct-2005
- [216] Angela Chen, "What the Apple Watch's FDA clearance actually mean's", The Verge, 13 Sept 2018, <https://www.theverge.com/2018/9/13/17855006/apple-watch-series-4-ekg-fda-approved-vs-cleared-meaning-safe>
- [217] Wikipedia, "Metamaterial", <https://en.wikipedia.org/wiki/Metamaterial>
- [218] Wikipedia, "Fog Computing", https://en.wikipedia.org/w/index.php?title=Fog_computing&oldid=751138994
- [219] Arvind Narayanan, "Written Testimony of Arvind Narayanan, Associate Professor of Computer Science, Princeton University", United States Senate, Committee on Energy and Natural Resources, Hearing on Energy Efficiency of Blockchain and Similar Technologies, August 21, 2018 https://www.energy.senate.gov/public/index.cfm/files/serve?File_id=8A1CECD1-157C-45D4-A1AB-B894E913737D
- [220] Xing Lin, Yair Rivenson, Nezh T. Yardimci, Muhammed Veli, Yi Luo, Mona Jarrahi, Aydogan Ozcan, "All-optical machine learning using diffractive deep neural networks", 2018
- [221] Xuan Qi, Chen Liu, Stephanie Schuckers, "IoT Edge Device Based Key Frame Extraction for Face in Video Recognition", Cluster

Cloud and Grid Computing (CCGRID) 2018 18th IEEE/ACM International Symposium on, pp. 641-644, 2018., <https://ieeexplore.ieee.org/document/8091072/>

- [222] Yogesh Malik, "Internet of Things Bringing Fog, Edge & Mist Computing", 2017
- [223] Shoshana Zuboff, "Big Other: Surveillance Capitalism and the Prospects of an Information Civilization" (April 4, 2015). *Journal of Information Technology* (2015) 30, 75–89. doi:10.1057/jit.2015.5. Available at SSRN: <https://ssrn.com/abstract=2594754>
- [224] Charles H. Bennett, Gilles Brassard, "Quantum cryptography: Public key distribution and coin tossing", *Theoretical Computer Science*, Volume 560, Part 1, 4 December 2014, Pages 7-11. doi:10.1016/j.tcs.2011.08.039
- [225] John Preskill, "Quantum computing and the entanglement frontier", arXiv:1203.5813v3 [quant-ph], 10 Nov 2012, <http://arxiv.org/abs/1203.5813>
- [226] Alaa Saade, Francesco Caltagirone et al., "Random Projections through multiple optical scattering: Approximating kernels at the speed of light", arXiv:1510.06664v2 [cs.ET], 25 Oct 2015, <http://arxiv.org/abs/1510.06664>
- [227] Ryan LaRose, "Overview and Comparison of Gate Level Quantum Software Platforms", arXiv:1807.02500 [quant-ph], 6 Jul 2018, <http://arxiv.org/abs/1807.02500>
- [228] Edward Farhi, Jeffrey Goldstone et al., "Quantum Computation by Adiabatic Evolution", arXiv:quant-ph/0001106v1, 28 Jan 2000, <http://arxiv.org/abs/quant-ph/0001106>
- [229] Peter W. Shor, "Polynomial-Time Algorithms for Prime Factorization and Discrete Logarithms on a Quantum Computer", arXiv:quant-ph/9508027v2, 30 Aug 1995, <http://arxiv.org/abs/quant-ph/9508027>
- [230] Ban Lethal Autonomous Weapons, "Compilation of open letters against autonomous weapons", accessed december 2018, <http://autonomousweapons.org/compilation-of-open-letters-against-autonomous-weapons/>
- [231] Steve Morgan, "2018 Cybersecurity Market Report", 31 May 2018, <http://cybersecurityventures.com/cybersecurity-market-report/>
- [232] Benoît Valiron, Neil J. Ross et al., "Programming the quantum future", *Communications of the ACM*, Volume 58 Issue 8, August 2015, <http://doi.acm.org/10.1145/2699415>
- [233] X. Fu, L. Rieseboos et al., "A heterogeneous quantum computer architecture", *ACM International Conference on Computing Frontiers*, 16 - 19 May 2016, <http://doi.acm.org/10.1145/2903150.2906827>
- [234] Paul Milgram, Fumio Kishino, "A Taxonomy of Mixed Reality Visual Displays", *IEICE Transactions on Information Systems*, Vol E77-D, No.12 December 1994, http://etclab.mie.utoronto.ca/people/paul_dir/IEICE94/ieice.html
- [235] David Thomson, Aaron Zilkie et al., "Roadmap on silicon photonics", *Journal of Optics*, Volume 18, Number 7, 24 Jun 2016, <http://iopscience.iop.org/article/10.1088/2040-8978/18/7/073003/meta>
- [236] <http://map.norsecorp.com/#/>
- [237] Stephen Jordan, "Algebraic and Number Theoretic Algorithms", 22 Apr 2011 - 13 Jun 2018, <http://math.nist.gov/quantum/zoo>
- [238] Jarvis, accessed december 2018, <http://openjarvis.com/>
- [239] Herbert Jaeger, Harald Haas, "Harnessing Nonlinearity: Predicting Chaotic Systems and Saving Energy in Wireless Communication", *Science*, Vol. 304, Issue 5667, 02 Apr 2004, <http://science.sciencemag.org/content/304/5667/78>
- [240] N. Samkharadze, G. Zheng et al, "Strong spin-photon coupling in silicon", *Science*, 25 Jan 2018, <http://science.sciencemag.org/content/early/2018/01/24/science.aar4054>
- [241] Xing Lin, Yair Rivenson et al, "All-optical machine learning using diffractive deep neural networks", *Science*, 26 Jul 2018, <http://science.sciencemag.org/content/early/2018/07/25/science.aat8084>
- [242] Zoë Bernard, "Vladimir Putin, and fluffy buzzwords: 6 clues the blockchain project you're thinking about investing in is a scam", *Business Insider UK*, 6 Jul 2018, <http://uk.businessinsider.com/how-to-tell-ico-scam-blockchain-2018-7?r=US&IR=T>
- [243] Nicolas Economou, "A 'principled' artificial intelligence could improve justice", *ABA Journal - Legal Rebels*, 3 Oct 2017, http://www.abajournal.com/legalrebels/article/a_principled_artificial_intelligence_could_improve_justice
- [244] European Cyber Security Organisation, accessed december 2018, <http://www.ecs-org.eu/>
- [245] European Cyber Security Organisation, "European Cybersecurity Industry Proposal for a contractual Public-Private-Partnership", Jun 2016, <http://www.ecs-org.eu/documents/ecs-cppp-industry-proposal.pdf>
- [246] Peter Clarke, "Alibaba forms chip subsidiary Pingtougé", *EETE Analog*, 27 Sept 2018, <http://www.eenewsanalog.com/news/alibaba-forms-chip-subsiary-pingtougé>
- [247] Fujitsu, "Fujitsu Completes Post-K Supercomputer CPU Prototype", Begins Functionality Trials, 21 Jun 2018, <http://www.fujitsu.com/global/about/resources/news/press-releases/2018/0621-01.html>
- [248] Fujitsu, "Digital Annealer", accessed december 2018, <http://www.fujitsu.com/global/digitalannealer/>
- [249] IDTechEx, "Printed and Flexible Sensors 2017-2027: Technologies, Players, Forecasts", accessed december 2018, <http://www.idtechex.com/research/reports/printed-and-flexible-sensors-2015-2025-technologies-players-forecasts-000428.asp>

- [250] IDTechEx, “Flexible, Printed and Organic Electronics 2019-2029: Forecasts, Players & Opportunities”, Oct 2018, <http://www.idtechex.com/research/reports/printed-organic-and-flexible-electronics-forecasts-players-and-opportunities-2016-2026-000457.asp>
- [251] David Verstraeten, Benjamin Schrauwen et al., “Oger: Modular Learning Architectures For Large-Scale Sequential Processing”, *JMLR* 13(Oct):2995–2998, 2012, <http://www.jmlr.org/papers/v13/verstraeten12a.html>
- [252] oe-a, “OE-A Roadmap for Organic and Printed Electronics”, 2017, <http://www.oe-a.org/roadmap>
- [253] OECD, “Inequality”, accessed december 2018, <http://www.oecd.org/inequality.htm#income>
- [254] Pipeline WorkSpaces, “Unhealthy Smartphone? You’re undateable!”, accessed december 2018, <http://www.pipelineworkspaces.com/unhealthy-smartphone-youre-undateable/>
- [255] D. Verstraeten, B. Schrauwen et al., “An experimental unification of reservoir computing methods”, *Neural Networks*, Volume 20, Issue 3, April 2007, <http://www.sciencedirect.com/science/article/pii/S089360800700038X>
- [256] UN Department of Economic and Social Affairs: Population Division, “World Population Ageing”, 2015, http://www.un.org/en/development/desa/population/publications/pdf/ageing/WPA2015_Report.pdf
- [257] Sergey Brin, “2017 Founders’ Letter”, 2017, <https://abc.xyz/investor/founders-letters/2017/index.html>
- [258] AI Ethics, “Draft Principles”, accessed <https://ai-ethics.com/past-principles/>
- [259] Julian Kelly, “A Preview of Bristlecone, Google’s New Quantum Processor”, *Google AI Blog*, 5 Mar 2018, <https://ai.googleblog.com/2018/03/a-preview-of-bristlecone-googles-new.html>
- [260] AI Superpowers, accessed december 2018, <https://aisuperpowers.com>
- [261] AIY Projects, “Do-it-yourself artificial intelligence”, accessed december 2018, <https://aiyprojects.withgoogle.com/>
- [262] Marco Tulio Ribeiro, Sameer Singh et al., ““Why Should I Trust You?”: Explaining the Predictions of Any Classifier”, *arXiv:1602.04938v3 [cs.LG]*, 9 Aug 2016, <https://arxiv.org/pdf/1602.04938.pdf>
- [263] Felipe Petroski Such, Vashisht Madhavan, et al., “Deep Neuroevolution: Genetic Algorithms Are a Competitive Alternative for Training Deep Neural Networks for Reinforcement Learning”, *arXiv:1712.06567v3 [cs.NE]*, 20 Apr 2018, <https://arxiv.org/pdf/1712.06567.pdf>
- [264] Philip Colangelo, Nasibeh Nasiri et al., “Exploration of Low Numeric Precision Deep Learning Inference Using Intel FPGAs”, *arXiv:1806.11547v1 [cs.DC]*, 12 Jun 2018, <https://arxiv.org/pdf/1806.11547.pdf>
- [265] Atos, “Quantum Machine Learning”, accessed december 2018, <https://atos.net/en/products/quantum-learning-machine>
- [266] Amazon, “Deploy an Elastic HPC Cluster”, accessed december 2018, https://aws.amazon.com/getting-started/projects/deploy-elastic-hpc-cluster/?nc1=h_ls
- [267] Catherine Hsiao, “Pushing the limits of streaming technology”, *Google Blog*, 1 Oct 2018, <https://blog.google/technology/developers/pushing-limits-streaming-technology/>
- [268] Francois Chollet, “The limitations of deep learning”, *The Keras Blog*, 17 July 2017, <https://blog.keras.io/the-limitations-of-deep-learning.html>
- [269] Steve Rosenbush, “Microsoft Says UBS Moves Key Platform to Azure Cloud”, *The Wall Street Journal blogs*, 26 Apr 2017, <https://blogs.wsj.com/cio/2017/04/26/microsoft-says-ubs-moves-key-platform-to-azure-cloud/>
- [270] Rogier 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/>
- [271] Claire, “Confederation of Laboratories for Artificial Intelligence Research in Europe”, accessed december 2018, <https://claire-ai.org/>
- [272] CLASS Project, “about”, accessed december 2018, <https://class-project.eu/about>
- [273] Google, “Cloud AutoML”, accessed december 2018, <https://cloud.google.com/automl/>
- [274] Norm Jouppi, “Quantifying the performance of the TPU, our first machine learning chip”, *Google Blog*, 5 Apr 2017, <https://cloudplatform.googleblog.com/2017/04/quantifying-the-performance-of-the-TPU-our-first-machine-learning-chip.html>
- [275] CORDIS, “Cloud Computing via Homomorphic Encryption and Multilinear Maps”, accessed december 2018, https://cordis.europa.eu/project/rcn/214880_en.html
- [276] Peter Mell, Tim Grance et al., “The NIST Definition of Cloud Computing”, Sept 2011, <https://csrc.nist.gov/publications/detail/sp/800-145/final>
- [277] DappRadar, “Ranked list of blockchain dapps”, accessed december 2018, <https://dappradar.com>
- [278] Apple, “ARKit 2”, accessed december 2018, <https://developer.apple.com/arkit/>
- [279] Google, “ARCore”, accessed december 2018, <https://developers.google.com/ar/>
- [280] Google, “Google VR for everyone”, accessed december 2018, <https://developers.google.com/vr/>
- [281] Google, “Progressive Web Apps”, accessed december 2018, <https://developers.google.com/web/progressive-web-apps>

- [282] Brandon Bray, Nick Schonning et al., “What is mixed reality?”, Microsoft Docs, accessed december 2018, <https://docs.microsoft.com/en-us/windows/mixed-reality/mixed-reality>
- [283] EC Digital Single Market, “High-Level Expert Group on Artificial Intelligence”, accessed december 2018, <https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence>
- [284] EC Digital Single Market Digibyte, “European Processor Initiative: consortium to develop Europe’s microprocessors for future supercomputers”, 23 Mar 2018, <https://ec.europa.eu/digital-single-market/en/news/european-processor-initiative-consortium-develop-europes-microprocessors-future-supercomputers>
- [285] Council Recommendation of 22 May 2018 on key competences for lifelong learning.
- [286] European Cyber Security Organisation”, accessed december 2018, <https://ecs-org.eu/>
- [287] Wikipedia, “Airborne collision avoidance system”, accessed december 2018, https://en.wikipedia.org/wiki/Airborne_collision_avoidance_system
- [288] Wikipedia, “Autopilot”, accessed december 2018, <https://en.wikipedia.org/wiki/Autopilot>
- [289] Wikipedia, “Cloud gaming”, accessed december 2018, https://en.wikipedia.org/wiki/Cloud_gaming
- [290] Wikipedia, “Machine learning”, accessed december 2018, https://en.wikipedia.org/wiki/Machine_learning
- [291] Wikipedia, “Neuromorphic engineering”, accessed december 2018, https://en.wikipedia.org/wiki/Neuromorphic_engineering
- [292] Wikipedia, “Open Compute Project”, accessed december 2018, https://en.wikipedia.org/wiki/Open_Compute_Project
- [293] Wikipedia, “Quantum supremacy”, accessed december 2018, https://en.wikipedia.org/wiki/Quantum_supremacy
- [294] Wikipedia, “WannaCry ransomware attack”, accessed december 2018, https://en.wikipedia.org/wiki/WannaCry_ransomware_attack
- [295] Daniel J. Bernstein, Tanja Lange, “Post-quantum cryptography – dealing with the fallout of physics success”, 17 Arr 217, <https://eprint.iacr.org/2017/314.pdf> Post-quantum cryptography –dealing with the fallout of physics success
- [296] Jen Weedon, William Nuland et al., “Information Operations and Facebook”, 27 Apr 2017, <https://fbnewsroomus.files.wordpress.com/2017/04/facebook-and-information-operations-v1.pdf>
- [297] Future of Life Institute, “AI Policy”, accessed december 2018, <https://futureoflife.org/ai-policy/>
- [298] Future of Life Institute, “National and International AI Strategies”, accessed december 2018, <https://futureoflife.org/national-international-ai-strategies/>
- [299] GladysProject, “Gladys”, accessed december 2018, <https://github.com/GladysProject/Gladys>
- [300] Mark Seaborn and Thomas Dullien, “Exploiting the DRAM rowhammer bug to gain kernel privileges”, Google Project Zero blog, 9 Mar 2015, <https://googleprojectzero.blogspot.com/2015/03/exploiting-dram-rowhammer-bug-to-gain.html>
- [301] Dickey Singh, “Self-supervised learning gets us closer to autonomous learning”, 6 Aug 2018, <https://hackernoon.com/self-supervised-learning-gets-us-closer-to-autonomous-learning-be77e6c86b5a>
- [302] HEAT Project, “Homomorphic Encryption Applications and Technology H2o2o-ICT-644209”, accessed december 2018, <https://heat-project.eu/index.html>
- [303] Dick Pelletier, “Shape-shifting claytronics: wild future here by 2020, experts say”, 24 Mar 2017, <https://ieet.org/index.php/IEET2/more/pelletier20140324>
- [304] IFTTT, “IFTTT helps your apps and devices work together”, accessed december 2018, <https://ifttt.com/>
- [305] IPFS, “IPFS is the Distributed Web”, accessed december 2018, <https://ipfs.io>
- [306] Laurent Larger, Antonio Baylón-Fuentes et al., “High-Speed Photonic Reservoir Computing Using a Time-Delay-Based Architecture: Million Words per Second Classification”, Physical Review X, 6 Feb 2017, <https://journals.aps.org/prx/abstract/10.1103/PhysRevX.7.011015>
- [307] Alain Aspect, Jean Dalibard, et al. “Experimental Test of Bell’s Inequalities Using Time-Varying Analyzers”, Physical Review Letters, 20 Dec 1982, <https://link.aps.org/doi/10.1103/PhysRevLett.49.1804>
- [308] Laurent Larger, Antonio Baylón-Fuentes et al., “High-Speed Photonic Reservoir Computing Using a Time-Delay-Based Architecture: Million Words per Second Classification”, Physical Review X, 6 Feb 2017, <https://link.aps.org/doi/10.1103/PhysRevX.7.011015>
- [309] Nicolai Friis, Oliver Marty et al., “Observation of Entangled States of a Fully Controlled 20-Qubit System”, Physical Review X, 10 Apr 2018, <https://link.aps.org/doi/10.1103/PhysRevX.8.021012>
- [310] Martin Fowler, “BeckDesignRules, 2 Mar 2015, <https://martinfowler.com/bliki/BeckDesignRules.html>
- [311] Tim Dutton, “An Overview of National AI Strategies”, Medium, 28 Jun 2018, <https://medium.com/politics-ai/an-overview-of-national-ai-strategies-2a70ec6edfd>
- [312] Mos Zhang, “Tencent ML Team Trains ImageNet” In Record Four Minutes, 31 Jul 2018, <https://medium.com/syncedreview/tencent-ml-team-trains-imagenet-in-record-four-minutes-d3d85eff2062>

- [313] MetaMask, “MetaMask”, accessed december 2018, <https://metamask.io>
- [314] Mycroft, “Get Mycroft”, accessed december 2018, <https://mycroft.ai/get-mycroft/>
- [315] Namecoin, “Decentralized secure names”, accessed december 2018, <https://namecoin.org>
- [316] MIT News, “Material could bring optical communication onto silicon chips”, 23 Oct 2017, <https://news.mit.edu/2017/ultrathin-films-semiconductor-optical-communication-silicon-chips-1023>
- [317] Joshua Aslan, Kieren Mayers et al., “Electricity Intensity of Internet Data Transmission: Untangling the Estimates”, *Journal of Industrial Ecology*, 1 Aug 2017, <https://onlinelibrary.wiley.com/doi/abs/10.1111/jiec.12630>
- [318] OpenAI, accessed december 2018, <https://openai.com/>
- [319] Google, “Poly”, accessed december 2018, <https://poly.google.com>
- [320] Peter H. Diamandis, “China Is Quickly Becoming an AI Superpower”, *SingularityHub*, 29 Aug 2018, https://singularityhub.com/2018/08/29/china-ai-superpower/?utm_source=Singularity+Hub+Newsletter&utm_campaign=0080bde973-Hub_Weekly_Newsletter&utm_medium=email&utm_term=o_focf6ocdae-0080bde973-57449729#sm.00018nlmu9b8geogxgn2fgpd08xpb
- [321] Snips, “Using Voice to Make Technology Disappear”, accessed december 2018, <https://snips.ai/>
- [322] Intel, “Introducing The Intel® Neural Compute Stick 2 (Intel® Ncs 2)”, accessed december 2018, <https://software.intel.com/en-us/neural-compute-stick/>
- [323] Samuel K. Moore, “Intel Starts R&D Effort in Probabilistic Computing for AI”, *IEEE Spectrum*, 10 May 2018, <https://spectrum.ieee.org/automaton/robotics/artificial-intelligence/intel-starts-rd-effort-in-probabilistic-computing-for-ai>
- [324] Evan Ackerman, “Mayfield Robotics Cancels Kuri Social Home Robot”, *IEEE Spectrum*, 25 Jul 2018, <https://spectrum.ieee.org/automaton/robotics/home-robots/mayfield-robotics-cancels-kuri-social-home-robot>
- [325] Tim Enwall, “Why the Pursuit of a “Killer App” for Home Robots Is Fraught With Peril”, *IEEE Spectrum*, 9 Aug 2018, <https://spectrum.ieee.org/automaton/robotics/home-robots/why-the-pursuit-of-a-killer-app-for-home-robots-is-fraught-with-peril>
- [326] Samuel K. Moore, “DARPA Plans a Major Remake of U.S. Electronics”, *IEEE Spectrum*, 16 Jul 2018, <https://spectrum.ieee.org/tech-talk/computing/hardware/darpas-planning-a-major-remake-of-us-electronics-pay-attention>
- [327] Samuel K. Moore, “IBM Edges Closer to Quantum Supremacy with 50-Qubit Processor”, *IEEE Spectrum*, 15 Nov 2017, <https://spectrum.ieee.org/tech-talk/computing/hardware/ibm-edges-closer-to-quantum-supremacy-with-50qubit-processor>
- [328] Samuel K. Moore, “DARPA Picks Its First Set of Winners in Electronics Resurgence Initiative”, *IEEE Spectrum*, 24 Jul 2018, <https://spectrum.ieee.org/tech-talk/semiconductors/design/darpa-picks-its-first-set-of-winners-in-electronics-resurgence-initiative>
- [329] Stack Overflow, accessed december 2018, <https://stackoverflow.com/>
- [330] Steam, “Welcome to Steam”, accessed december 2018, <https://store.steampowered.com>
- [331] Mitch Pronschinske, “9 code and framework trends to watch in 2018”, accessed december 2018, <https://techbeacon.com/9-code-framework-trends-watch-2018>
- [332] Scott Wasson, “Errata prompts Intel to disable TSX in Haswell, early Broadwell CPUs”, *The Tech Report*, 12 Aug 2014 <https://techreport.com/news/26911/errata-prompts-intel-to-disable-tsx-in-haswell-early-broadwell-cpus>
- [333] Raphaël Couturier, Michel Salomon, “Parallelization and optimization of the neuromorphic simulation code. Application on the MNIST problem”, Nov 2015, https://trimestres-lmb.univ-fcomte.fr/IMG/pdf/dynamical_systems_salomon.pdf
- [334] Alexis Chemblette, “How China Is Trying to Become the World’s Leader in Artificial Intelligence”, *Adweek*, 8 May 2018, <https://www.adweek.com/digital/how-china-is-trying-to-become-the-worlds-leader-in-artificial-intelligence/>
- [335] Lucia Maffei, “More Layoffs Hit Jibo—This Time, They’re “Significant””, *Bostinno*, 11 Jun 2018, <https://www.americaninno.com/boston/inno-news-bosthttps://www.bbc.com/news/business-43090226on/more-layoffs-hit-jibo-this-time-theyre-significant/>
- [336] Tim Bowler, “The low-cost mini satellites bringing mobile to the world”, *BBC News*, 23 Feb 2018, <https://www.bbc.com/news/business-43090226>
- [337] BSC, “OmpSs@FPGA on track to become embedded systems programming standard, thanks to BSC’s work in the AXIOM project”, 23 Jul 2018, <https://www.bsc.es/news/bsc-news/ompsffpga-track-become-embedded-systems-programming-standard-thanks-bsc%E2%80%99s-work-the-axiom-project>
- [338] Kate Taylor, “Only a single Blockbuster remains open in all of America. Here’s what it’s like to visit.”, *Business Insider España*, 9 Aug 2018, <https://www.businessinsider.es/blockbuster-survives-in-bend-oregon-2018-8?page=3>
- [339] BusinessWire, “Global Artificial Intelligence Chipsets Market to Reach \$59.26 Billion by 2025 - ResearchAndMarkets.com”, 8 May 2018, <https://www.businesswire.com/news/home/20180508005878/en/Global-Artificial-Intelligence-Chipsets-Market-Reach-59.26>
- [340] Aaron Mehta, “Google’s Schmidt: US losing edge in AI to China”, *C4ISRNET*, 2 Nov 2017, <https://www.c4isrnet.com/it-networks/2017/11/02/china-on-path-to-eclipse-us-with-ai-warns-google-head/>

- [341] RECORD-IT, "Reservoir Computing with Real-time Data for future IT (RECORD-IT)", accessed december 2017, <https://www.chalmers.se/en/projects/Pages/RECORD-IT.aspx>
- [342] Jordan Novet, "Why tech companies are racing each other to make their own custom A.I. chips", CNBC, 24 Apr 2018, <https://www.cnbc.com/2018/04/21/alibaba-joins-google-others-in-making-custom-ai-chips.html>
- [343] Ben Fox Rubin, "Amazon's Alexa assistant now works with over 20k devices", CNET, 1 Sept 2018, <https://www.cnet.com/news/amazon-alexa-assistant-is-now-in-20k-devices/#ftag=CAD590a51e>
- [344] David Gunning, "Explainable Artificial Intelligence (XAI)", DARPA, accessed december 2018, <https://www.darpa.mil/program/explainable-artificial-intelligence>
- [345] Center for Digital Ethics, "Welcome to the Center for Digital Ethics & Policy", accessed december 2018, <https://www.digitaletics.org>
- [346] D-WAVE, "Home", accessed december 2018, <https://www.dwavesys.com/home>
- [347] Rick Merritt, "TSMC Goes Photon to Cloud", EETimes, 4 Oct 2018, https://www.eetimes.com/document.asp?_mc=RSS%5FEET%5FEDT&utm_source=newsletter&utm_campaign=link&utm_medium=EETimesDaily%2D20181004&doc_id=1333827&page_number=1
- [348] Rick Merrott, "LinkedIn Group Preps Server Specs", EETimes, 23 May 2017, https://www.eetimes.com/document.asp?doc_id=1331772
- [349] Dylan McGrath, "Intel Delays 10-nm Volume Production Until 2019", EETimes, 27 Apr 2018, https://www.eetimes.com/document.asp?doc_id=1333230
- [350] Ann R. Thryft, "Why the IIoT is So Vulnerable to Cyberattacks", EETimes, 14 Sept 2018, https://www.eetimes.com/document.asp?doc_id=1333693&page_number=1
- [351] Ann R. Thryft, "Real-Life Industrial IoT Cyberattack Scenarios", EETimes, 14 Sept 2018, https://www.eetimes.com/document.asp?doc_id=1333709&page_number=1
- [352] Doug Black, "GlobalFoundries Drops 7nm Development Program", EnterpriseTech, 28 Aug 2018, <https://www.enterprisetech.com/2018/08/28/globalfoundries-drops-7nm-fenfet-development-program/>
- [353] Elise Gould, "Looking at the latest wage data by education level", Economic Policy Institute Blog, 1 Sept 2016, <https://www.epi.org/blog/looking-at-the-latest-wage-data-by-education-level/>
- [354] ESA, "The Use of Reprogrammable FPGAs in Space", accessed december 2018, https://www.esa.int/Our_Activities/Space_Engineering_Technology/Microelectronics/The_use_of_reprogrammable_FPGAs_in_space
- [355] Ethereum, "Ethereum Project", <https://www.ethereum.org>
- [356] FBI, "Cyber's Most Wanted", accessed december 2018, <https://www.fbi.gov/wanted/cyber>
- [357] Ashraf Eassa, "How Apple Dethroned Intel As the World's Most Innovative Chipmaker", The Motley Fool, 28 May 2018, <https://www.fool.com/investing/2018/05/28/how-apple-dethroned-intel-as-the-worlds-most-innov.aspx>
- [358] Janet Burns, "There's Now Evidence That Online Dating Causes Stronger, More Diverse Marriages", Forbes blog, 25 Oct 2017, <https://www.forbes.com/sites/janetwburns/2017/10/25/theres-now-evidence-that-online-dating-causes-stronger-more-diverse-relationships/#3edecac558bd>
- [359] Gartner, "Gartner Hype Cycle", accessed december 2018, <https://www.gartner.com/en/research/methodologies/gartner-hype-cycle>
- [360] Ginkgo Bioworks, "Foundries", accessed december 2018, <https://www.ginkgobioworks.com/foundries/>
- [361] GlobalFoundries, "GLOBALFOUNDRIES Reshapes Technology Portfolio to Intensify Focus on Growing Demand for Differentiated Offerings", 27 Aug 2018, <https://www.globalfoundries.com/news-events/press-releases/globalfoundries-reshapes-technology-portfolio-to-intensify-focus-on-growing-demand-for-differentiated-offerings>
- [362] Global Market Insights, "Embedded system market", accessed december 2018, <https://www.gminsights.com/industry-analysis/embedded-system-market>
- [363] TETRACOM delivers four-fold return on EU tech transfer investment, HiPEAC, 5 Jul 2016, <https://www.hipeac.net/press/6787/tetacom-delivers-four-fold-return-on-eu-tech-transfer-investment/>
- [364] Tiffany Trader, "Nvidia Debuts Turing Architecture, Focusing on Real-Time Ray Tracing", HPCWire, 16 Aug 2018, <https://www.hpcwire.com/2018/08/16/nvidia-debuts-turing-architecture-focusing-on-real-time-ray-tracing/>
- [365] Human Brain Project, "Hardware", accessed december 2018, <https://www.humanbrainproject.eu/en/silicon-brains/how-we-work/hardware/>
- [366] Human Brain Project, "The Neuromorphic Computing platform: Getting started", accessed december 2018, <https://www.humanbrainproject.eu/en/silicon-brains/neuromorphic-computing-platform/>
- [367] NSA, "Commercial National Security Algorithm Suite", 19 Aug 2015, <https://apps.nsa.gov/iaarchive/programs/iad-initiatives/cnsa-suite.cfm>

- [368] Tom Bawden, “Global warming: Data centres to consume three times as much energy in next decade, experts warn”, *The Independent*, 23 Jan 2016, <https://www.independent.co.uk/environment/global-warming-data-centres-to-consume-three-times-as-much-energy-in-next-decade-experts-warn-a6830086.html>
- [369] Andrew Ross, “IoT adoption perceived as risky, as failures plague 64% of users worldwide”, *InformationAge*, 21 Aug 2018, <https://www.information-age.com/iot-adoption-123474305/>
- [370] Peter Wayner, “21 hot programming trends—and 21 going cold, InfoWorld”, 16 Apr 2018, <https://www.infoworld.com/article/3188464/application-development/21-hot-programming-trends-and-21-going-cold.html>
- [371] Dan Clark, “Top 16 Open Source Deep Learning Libraries and Platforms”, *KDnuggets*, Apr 2018, <https://www.kdnuggets.com/2018/04/top-16-open-source-deep-learning-libraries.html>
- [372] Michèle Finck, “Law and Autonomous Systems Series: Blockchains and the Right to be Forgotten”, *Oxford Business Law Blog*, 20 Apr 2018, <https://www.law.ox.ac.uk/business-law-blog/blog/2018/04/law-and-autonomous-systems-series-blockchains-and-right-be-forgotten>
- [373] Bernard Marr, “Big Data: The 5 Vs Everyone Must Know”, *LinkedIn pulse*, 6 Mar 2014, <https://www.linkedin.com/pulse/20140306073407-64875646-big-data-the-5-vs-everyone-must-know/>
- [374] Wolfgang Maass, “Thomas Natschläger et al., Real-Time Computing Without Stable States: A New Framework for Neural Computation Based on Perturbations,” *Neural Computation*, Volume 14, Issue 11, November 2002, <https://www.mitpressjournals.org/doi/10.1162/089976602760407955>
- [375] M. Veldhorst, C. H. Yang et al., “A two-qubit logic gate in silicon”, *Nature* volume 526, pages 410–414, 15 October 2015, <https://www.nature.com/articles/nature15263>
- [376] Jens Jakob W. H. Sørensen, Mads Kock Pedersen, “Exploring the quantum speed limit with computer games”, *Nature* volume 532, pages 210–213, 14 April 2016, <https://www.nature.com/articles/nature17620>
- [377] Jacob Biamonte, Peter Wittek et al., “Quantum machine learning”, *Nature* volume 549, pages 195–202, 14 September 2017, <https://www.nature.com/articles/nature23474>
- [378] X. Mi, M. Benito et al., “A coherent spin–photon interface in silicon”, *Nature* volume 555, pages 599–603, 29 March 2018, <https://www.nature.com/articles/nature25769>
- [379] R. Maurand, X. Jehl et al., “A CMOS silicon spin qubit”, *Nature Communications* volume 7, Article number: 13575, 2016, <https://www.nature.com/articles/ncomms13575>
- [380] Chao Du, Fuxi Cai et al., “Reservoir computing using dynamic memristors for temporal information processing”, *Nature Communications* volume 8, Article number: 2204, 2017, <https://www.nature.com/articles/s41467-017-02337-y>
- [381] Xiaogang Qiang, Xiaoqi Zhou et al., “Large-scale silicon quantum photonics implementing arbitrary two-qubit processing”, *Nature Photonics* volume 12, pages 534–539, 2018, <https://www.nature.com/articles/s41566-018-0236-y>
- [382] François Duport, Anteo Smerieri et al., “Fully analogue photonic reservoir computer”, *Scientific Reports* volume 6, Article number: 22381, 2016, <https://www.nature.com/articles/srep22381>
- [383] Abhinav Kandala, Antonio Mezzacapo, “Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets”, *Nature* volume 549, pages 242–246, 14 September 2017, <https://www.nature.com/nature/journal/v549/n7671/full/nature23879.html>
- [384] Timothy Prickett Morgan, “Intel’s Exascale Dataflow Engine Drops X86 And Von Neumann”, *TheNextPlatform*, 30 Aug 2018, <https://www.nextplatform.com/2018/08/30/intels-exascale-dataflow-engine-drops-x86-and-von-neuman/>
- [385] Nicole Hemsoth, “Full Qubit Access a Game-Changer for Quantum Development”, *TheNextPlatform*, 4 Oct 2018, <https://www.nextplatform.com/2018/10/04/full-qubit-and-tooling-access-a-game-changer-for-quantum-development/>
- [386] NVIDIA, “Cloud Gaming - Gaming as a Service (GaaS)”, accessed december 2018, <https://www.nvidia.com/object/cloud-gaming.html>
- [387] Oren Etzioni, “How to Regulate Artificial Intelligence”, *The New York Times Opinion*, 1 Sept 2017, <https://www.nytimes.com/2017/09/01/opinion/artificial-intelligence-regulations-rules.html>
- [388] Olivier Ezratty, “Comprendre l’informatique quantique – outils de développement”, 31 Jul 2018, <https://www.oezratty.net/wordpress/2018/comprendre-informatique-quantique-outils-de-developpement/>
- [389] Oxford Internet Institute, “Digital Ethics Lab”, accessed december 2018, <https://www.oii.ox.ac.uk/research/digital-ethics-lab/>
- [390] Kenneth O. Stanley, “Neuroevolution: A different kind of deep learning”, *O’Reilly ideas*, 13 Jul 2018, <https://www.oreilly.com/ideas/neuroevolution-a-different-kind-of-deep-learning>
- [391] Jean Twenge, “Analysis: Teens are sleeping less. Why? Smartphones”, *PBS*, 19 Oct 2017, <https://www.pbs.org/newshour/science/analysis-teens-are-sleeping-less-why-smartphones>
- [392] Jarred Walton, “Nvidia Turing architecture deep dive”, *PC Gamer*, 21 Sept 2018, <https://www.pcgamer.com/nvidia-turing-architecture-deep-dive/>

- [393] Private Biometrics, accessed december 2018, <https://www.privatebiometrics.com/>
- [394] PULP Platform, accessed december 2018, <https://www.pulp-platform.org>
- [395] Kevin Hartnett, "Major Quantum Computing Advance Made Obsolete by Teenager", Quanta Magazine, 31 Jul 2018, <https://www.quantamagazine.org/teenager-finds-classical-alternative-to-quantum-recommendation-algorithm-20180731/>
- [396] Kevin Hartnett, "To Build Truly Intelligent Machines, Teach Them Cause and Effect", Quanta Magazine, 15 May 2018, <https://www.quantamagazine.org/to-build-truly-intelligent-machines-teach-them-cause-and-effect-20180515/>
- [397] Quantiki, "List of QC simulators", accessed december 2018, <https://www.quantiki.org/wiki/list-qc-simulators>
- [398] Reuters, "Global Automotive IoT Market Share, Size, Estimates, Trends and Forecast 2023", 22 Jan 2018, <https://www.reuters.com/brandfeatures/venture-capital/article?id=25649>
- [399] John Markoff, "Silicon Valley investors to bankroll artificial-intelligence center", The Seattle Times, 13 Dec 2015, <https://www.seattletimes.com/business/technology/silicon-valley-investors-to-bankroll-artificial-intelligence-center/>
- [400] Scotten Jones, "IEDM 2017 - Leti Gate-All-Around Stacked-Nanowires", SemiWiki, 12 Feb 2018, <https://www.semiwiki.com/forum/content/7282-iedm-2017-leti-gate-all-around-stacked-nanowires.html>
- [401] Scotten Jones, "7nm, 5nm and 3nm Logic, current and projected processes", SemiWiki, 25 Jun 2018, <https://www.semiwiki.com/forum/content/7544-7nm-5nm-3nm-logic-current-projected-processes.html>
- [402] SenseTime, "SenseTime Raises US\$600 Million in Series C Funding", 9 Apr 2018, <https://www.sensetime.com/news/669.html>
- [403] Mateo Valero, "European Processor Initiative & RISC-V", 9 May 2018, <https://www.slideshare.net/insideHPC/european-processor-initiative-riscv>
- [404] Software Heritage, accessed december 2018, <https://www.softwareheritage.org/>
- [405] MIT Technology Review, "Evolutionary algorithm outperforms deep-learning machines at video games", 18 Jul 2018, <https://www.technologyreview.com/s/611568/evolutionary-algorithm-outperforms-deep-learning-machines-at-video-games/>
- [406] Mike Moore, "Worldwide AI investment to top \$200bn by 2025", 31 Jul 2018, <https://www.techradar.com/news/worldwide-ai-investment-to-top-dollar200bn-by-2025>
- [407] Climate Home News - Guardian Environment Network, "'Tsunami of data' could consume one fifth of global electricity by 2025", The Guardian, 11 Dec 2017, <https://www.theguardian.com/environment/2017/dec/11/tsunami-of-data-could-consume-fifth-global-electricity-by-2025>
- [408] Jonathan Watts, "We have 12 years to limit climate change catastrophe, warns UN, The Guardian", 8 Oct 2018, <https://www.theguardian.com/environment/2018/oct/08/global-warming-must-not-exceed-15c-warns-landmark-un-report>
- [409] Nick Evershed, "Carbon countdown clock: how much of the world's carbon budget have we spent?", The Guardian, 19 Jan 2017, <https://www.theguardian.com/environment/datablog/2017/jan/19/carbon-countdown-clock-how-much-of-the-worlds-carbon-budget-have-we-spent>
- [410] Larry Elliott, "World's eight richest people have same wealth as poorest 50%", The Guardian, 19 Jan 2017, <https://www.theguardian.com/global-development/2017/jan/16/worlds-eight-richest-people-have-same-wealth-as-poorest-50>
- [411] Rupert Neate, "Richest 1% own half the world's wealth, study finds", The Guardian, 14 Nov 2017, <https://www.theguardian.com/inequality/2017/nov/14/worlds-richest-wealth-credit-suisse>
- [412] Julian Borra, "Digital obesity: our high-tech lives may be bad for our health", The Guardian, 24 Apr 2013, <https://www.theguardian.com/sustainable-business/digital-obesity-high-tech-health>
- [413] Priya Nanad, "The Reality Behind Voice Shopping Hype", The Information, 6 Aug 2018, <https://www.theinformation.com/articles/the-reality-behind-voice-shopping-hype>
- [414] Kieren McCarthy, "Google bod wants cookies to crumble and be remade into something more secure", The Register, 15 Aug 2018, https://www.theregister.co.uk/2018/08/15/killing_off_cookies
- [415] James Vincent, "China and the US are battling to become the world's first AI superpower", The Verge, 3 Aug 2017, <https://www.theverge.com/2017/8/3/16007736/china-us-ai-artificial-intelligence>
- [416] James Vincent, "Putin says the nation that leads in AI 'will be the ruler of the world'", The Verge, 4 Sept 2017, <https://www.theverge.com/2017/9/4/16251226/russia-ai-putin-rule-the-world>
- [417] James Vincent, "Watch Jordan Peele use AI to make Barack Obama deliver a PSA about fake news", The Verge, 17 Apr 2018, <https://www.theverge.com/tldr/2018/4/17/17247334/ai-fake-news-video-barack-obama-jordan-peeel-buzzfeed>
- [418] Times Higher Education, "World University Rankings", accessed december 2018, <https://www.timeshighereducation.com/world-university-rankings>
- [419] Monica Chin, "Amazon's New Echo Plus Controls Gadgets Without Internet", Tom's Guide, 20 Sept 2018, <https://www.tomsguide.com/us/new-echo-plus-price-release-date,news-28132.html>
- [420] TOP500 Supercomputer Sites, "The Green500 list", June 2018, <https://www.top500.org/green500/list/2018/06/>

- [421] Michael Feldman, "IBM Enters Quantum Computing Business with First Paying Customers", TOP500 Supercomputer Sites, 15 Dec 2017, <https://www.top500.org/news/ibm-enters-quantum-computing-business-with-first-paying-customers/>
- [422] EVA by Voicera, "Artificial Intelligence Assistant," accessed december 2018, <https://www.voicera.com/>
- [423] Lily Hay Newman, "An Elaborate Hack Shows How Much Damage IoT Bugs Can Do", Wired, 16 Apr 2018, <https://www.wired.com/story/elaborate-hack-shows-damage-iot-bugs-can-do/>
- [424] Jason Pontin, "Greedy, Brittle, Opaque, and Shallow: The Downsides to Deep Learning", Wired, 2 Feb 2018, <https://www.wired.com/story/greedy-brittle-opaque-and-shallow-the-downsides-to-deep-learning/>
- [425] Jack Stewart, "Tesla Says Its New Self-Driving Chip Is Finally Baked", Wired, 4 Aug 2018, <https://www.wired.com/story/tesla-self-driving-car-computer-chip-nvidia/>
- [426] Cristian S. Calude, "De-Quantizing the Solution of Deutsch's Problem", International Journal of Quantum Information Vol. 05, No. 03, pp. 409-415 (2007), <https://www.worldscientific.com/doi/abs/10.1142/S021974990700292X>
- [427] Brent Leary, "Alexa's land-and-expand strategy is racking up the numbers", ZDNet, 11 Sept 2018, <https://www.zdnet.com/article/alexas-land-and-expand-strategy-is-racking-up-the-numbers/>
- [428] Liam Tung, "Developers, despair: Half your time is wasted on bad code", ZDNet, 11 Sept 2018, <https://www.zdnet.com/article/developers-despair-half-your-time-is-wasted-on-bad-code>
- [429] Amy Talbott, "Infographic: Why companies are switching to Everything as a Service", ZDNet, 30 Oct 2017, <https://www.zdnet.com/article/infographic-why-companies-are-switching-to-everything-as-a-service/>
- [430] Xenproject.org Security Team, "Xen Security Advisory CVE-2013-0153 / XSA-36, interrupt remap entries shared and old ones not cleared on AMD IOMMUs", 5 Feb 2013, <https://xenbits.xen.org/xsa/advisory-36.html>
- [431] nSafeCer: "nSafety Certification of Software-Intensive Systems with Reusable Components", accessed december 2018, https://cordis.europa.eu/project/rcn/105610_en.html
- [432] NeuRAM3, "NeuRAM Cube", accessed december 2018, www.neuram3.eu
- [433] NeuRAM3, "DynapSEL chip results", accessed december 2018, www.neuram3.eu/achievements/neuram3-chips/dynapsel-results
- [434] ACM Code 2018 Task Force, "ACM Code of Ethics and Professional Conduct", 22 Jun 2018, <https://www.acm.org/code-of-ethics>
- [435] Julien Happich, "ARM's next bet for plastic chips: Neural networks", eeNews Europe, 25 May 2018, <http://www.eenewseurope.com/news/arms-next-bet-plastic-chips-neural-networks>
- [436] Adrienne Jeffries, "'Blockchain' is meaningless", 7 Mar 2018, The Verge, <https://www.theverge.com/2018/3/7/17091766/blockchain-bitcoin-ethereum-cryptocurrency-meaning>
- [437] Cleanaway Waste Management, "E-waste: An inconvenient consequence of the digital age", 30 Jun 2018, <https://www.cleanaway.com.au/about-us/sustainable-future/e-waste-problem/>
- [438] ENCOS consortium, "European Nanoelectronics consortium on sustainability (ENCOS)" https://cdn.uclouvain.be/groups/cms-editors-icteam/iot/ENCOS_white%20paper.pdf
- [439] European Environment Agency, "Ecological footprint of European countries", 31 Mar 2015, <https://www.eea.europa.eu/data-and-maps/indicators/ecological-footprint-of-european-countries/ecological-footprint-of-european-countries-2>
- [440] European Commission - Entrepreneurship and SMEs, "Entrepreneurship and Small and medium-sized enterprises (SMEs)", accessed december 2018, https://ec.europa.eu/growth/smes_en
- [441] Tegegnetwork Gettu, Gilbert Hougbo et al., 'Fast-forward progress Leveraging tech to achieve the global goals', 2017, https://www.itu.int/en/sustainable-world/Documents/Fast-forward_progress_report_414709%20FINAL.pdf
- [442] diaTribe Learn, "Google Secures Patent for Glucose-Sensing Contact Lens", 16 Apr 2015, <https://diatribe.org/google-secures-patent-glucose-sensing-contact-lens>
- [443] Farnam Street Blog, "Half Life: The Decay of Knowledge and What to Do About It", Mar 2018, <https://fs.blog/2018/03/half-life/>
- [444] Brenden M. Lake, Ruslan Salakhutdinov et al., Science, Vol. 350 Issue 6266, Dec. 2015, "Human-level Concept Learning through Probabilistic Program Induction"
- [445] eurostat, "Key figures on Europe", 2017, <https://ec.europa.eu/eurostat/documents/3217494/8309812/KS-EI-17-001-EN-N.pdf/b7df53f5-4faf-48a6-aca1-c650d40c9239>
- [446] Directorate-General for Research and Innovation, European Union, "LAB – FAB – APP: Investing in the European future we want", 2017, http://ec.europa.eu/research/evaluations/pdf/archive/other_reports_studies_and_documents/hlg_2017_report.pdf
- [447] eurostat, "Migration and migrant population statistics", 12 Mar 2018, <https://ec.europa.eu/eurostat/statistics-explained/pdfscache/1275.pdf>
- [448] Univ. Grenoble Alpes, "Need for IOT Project", accessed december 2018, <https://need.univ-grenoble-alpes.fr/need-for-iot-home-742767.htm>

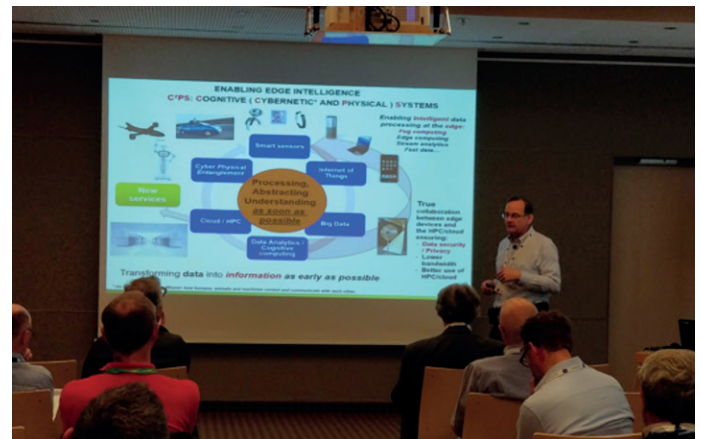
- [449] Enveil, Encrypted Veil, accessed december 2018, <https://www.enveil.com/products/>
- [450] Zoubin Ghahramani, "Probabilistic machine learning and artificial intelligence", *Nature*, Vol 521, May 2015
- [451] Directorate A — Policy Development and Coordination, Unit A4 — Analysis and monitoring of national research and innovation policies, "SCIENCE, RESEARCH AND INNOVATION PERFORMANCE OF THE EU 2018", 2018 https://ec.europa.eu/info/sites/info/files/rec-17-015-srip-report2018_mep-web-20180228.pdf
- [452] Dan Robitzski, "Should Evil AI Research Be Published? Five Experts Weigh In", 27 Aug 2018, <https://futurism.com/should-evil-ai-research-be-published-five-experts-weigh-in/>
- [453] Julien Happich, "Starting all over again on plastic: ARM", *eeNews Europe*, 24 Nov 2015, <http://www.eenewseurope.com/news/starting-all-over-again-plastic-arm>, Nov. 2015.
- [454] Wilke, Maack und Partner, "Strategic study on the anticipation of changes in the European ICT sector", Sept 2016, http://www.industriall-europe.eu/proj/ictstrat/industriAllEurope_2016-09_ICT_StrategicStudy_EN.pdf
- [455] Pew Research Center, "The American Middle Class Is Losing Ground", 9 Dec 2015, <http://www.pewsocialtrends.org/2015/12/09/the-american-middle-class-is-losing-ground/>
- [456] Rob Davies, "Uber loses appeal in UK employment rights case", *The Guardian*, 10 Nov 2017, <https://www.theguardian.com/technology/2017/nov/10/uber-loses-appeal-employment-rights-workers>
- [457] "What They Don't Tell You about Climate Change", *The Economist*, 16 Nov 2017, <https://www.economist.com/leaders/2017/11/16/what-they-dont-tell-you-about-climate-change>
- [458] "'It's the memory, stupid": A conversation with Onur Mutlu', *HiPEACinfo* 55, October 2018
- [459] Catherine D. Schuman, Thomas E. Potok et al., 'A Survey of Neuromorphic Computing and Neural Networks in Hardware', [arXiv:1705.06963 \[cs\]](http://arxiv.org/abs/1705.06963), 19 May 2017, <http://arxiv.org/abs/1705.06963>
- [460] Randi Klett, Erico Guizzo, "Sony's Aibo Robot Dog Is Coming to America", *IEEE Spectrum*, 23 Aug 2018, <https://spectrum.ieee.org/automaton/robotics/home-robots/sony-aibo-robot-dog-is-coming-to-america>
- [461] Verily, "Update on our Smart Lens program with Alcon", 16 Nov 2016, <https://blog.verily.com/2018/11/update-on-our-smart-lens-program-with.html>
- [462] Jared Newman, "Amazon Settles Kindle "1984" Lawsuit", *PCWorld*, 1 Oct 2009, https://www.pcworld.com/article/172953/amazon_kindle_1984_lawsuit.html

5 PROCESS

The HiPEAC Vision is a bi-annual document that presents the trends that have an impact on the community of High Performance and Embedded Architecture and Compilation. The document is based on information collected through different channels.

- Meetings with teachers and industrial partners during the ACACES 2017 and ACACES 2018 Summer Schools;
- A survey circulated to all HiPEAC members, and which received 35 responses;
- A co-organised a vision workshop between ETP4HPC and HiPEAC just before the start of the ISC High Performance Conference in Frankfurt, Germany on Sunday 24 June 2018;
- Two vision meeting with the HiPEAC community in Brussels (on invitation): 27 November 2017 and 27 April 2018;
- Presentations at Road4CPS meeting on 15 May 2018, Brussels;
- Presentations at Road4CPS meeting on 12 September 2018, Paris;
- Presentation at IWES-2018 on 14 September 2018, Siena;
- A dedicated feedback workshop during the HiPEAC Computing Systems Week on 29 October 2018 in Heraklion;
- A presentation at EFCS 2018 on 21 November 2018, Lisbon;
- A presentation at ICT 2018 on 5 December 2018, Vienna.

The 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.



Marc Duranton presenting at ETP4HPC/HiPEAC Vision Workshop, Frankfurt

6

ACKNOWLEDGEMENTS

This document is based on the valuable inputs from the HiPEAC members. The editorial board, composed of Marc Duranton (CEA), Koen De Bosschere (Ghent University), Bart Coppens (Ghent University), Christian Gamrat (CEA), Madeleine Gray (BSC), Harm Munk (TNO), Emre Özer (ARM), Tullio Vardanega (University of Padua), Olivier Zendra (INRIA), and would like to thank particularly: Vicky Wandels (Ghent University), Eneko Illarramendi (Ghent University), Carlo Reita (CEA-Leti), Séverine Cheraamy (CEA-Leti), Thomas Ernst (CEA), Jean-Marc Denis (EPI) for their active contribution to the document and Magnus Sjalander (NTNU), Jennifer Sartor (Ghent University), Katrien Van Impe (Ghent University), Dimitrios Soudris (NTUA), Marisa Gil (Barcelona Supercomputing Center), Paul Carpenter (Barcelona Supercomputing Center) for their useful comments and all the numerous people that provided support and information.

7 HIPEAC VISION 2019 RECOMMENDATIONS

ACCELERATE, ACCELERATE, SPECIALIZE AND AUTOMATE

The only way to continue performance scaling in the short to medium term is to specialize hardware for important application domains, such as artificial intelligence and processing near memory.

This specialization will require significant investment, and will only be economically viable if the specialisation is automated. Open-source hardware could boost innovative solutions. To facilitate the integration of accelerators into a system and to manage the increasing complexity, new automation intelligence and frameworks will be needed for both hardware and software.

DEVELOP ALTERNATIVE ARCHITECTURES

Not only hardware technology but also architectures need to be revisited in light of the end of exponential scaling in computing power, and to improve energy and efficiency. Alternatives to the von Neumann architecture should be investigated, responding to the needs of modern computing especially the need to process vast amounts of data. Taking inspiration from the example of neural networks, Europe should revisit innovative concepts from the past, which may have been made viable by new technology and production techniques. The new computing models can be applied to specific application areas for efficiency benefits.

BUILD COMPUTERS WE CAN TRUST

With computers forming part of every aspect of our lives, any solutions developed must lead to trustable computing systems. They need to be secure – with watertight protections against malicious attacks – and safe, not harming people when they interact with their environment. This is particularly important for connected and cyber-physical systems. They also need to be reliable despite being increasingly complex, and here artificial intelligence could help by writing software and developing systems.

As ICT systems increasingly make decisions based on machine learning, the algorithms and the decisions they provide should be explainable enough to build trust.

GET LOOKING FOR CMOS ALTERNATIVES

The end of complementary-metal-oxide semiconductor – or CMOS – scaling means that all bets are on as to what technology will look like in 2030. With no one technology emerging as a clear frontrunner, Europe should continue investing in research and help get results to market so that it will be at the heart of new technology developments. These technologies are unlikely to supplant CMOS, but instead will complement it.

Post-CMOS technologies might throw up good solutions for the innovative sensor / actuator / interface technologies, which will play a crucial role in cyber-physical systems and wireless sensor networks.

TREAT THE COMPUTING INFRASTRUCTURE AS A CONTINUUM, FROM THE EDGE TO THE CLOUD

From microcontrollers with sensors and actuators, to concentrators, to micro-servers, to cloud and high-performance computing, computing is on a continuum, and self-contained systems are now themselves components of a large system. Interoperability is key; systems need to collaborate to give the best service to users. There is a need for dynamic devices, which can adapt intelligently.

Europe should encourage collaboration between different communities – such as software versus hardware, cloud and high-performance computing (HPC) versus the edge, to help break down silos – thereby making better use of resources, reducing energy consumption or latency as needed.

SHIFT VALUE TOWARDS THE EDGE

Europe needs to play to its strengths. That means building on strong industries like automotive, aerospace and trains, and electronic components and systems for embedded computing. Bringing intelligence at the edge should be a major priority, aiming for a wide range of cognitive cyber-physical devices, and not necessarily always chasing the most advanced CMOS technology.

Investing in mature technologies (above 10nm) doesn't mean giving up on ambition. Interposer and chiplet technology will lower costs and allow different technologies to slot together, such as analogue, power converters, memories, digital and photonics.

LEAD ON THE USE OF COLLECTIVE DATA

Europe should develop the ethical use of state-owned, collective or domain data. This will allow the continent to develop its strength in AI-based solutions based on large amounts of data, without relying on the big B2C technology companies. Solutions to ensure the privacy and security of data should be developed and enforced.

BECOME A LEADER IN ENERGY-EFFICIENT, SUSTAINABLE ELECTRONICS

Europe should become a leader in the design of sustainable electronics, the recycling of computing devices and modularity, prolonging the life of ICT systems. Innovative approaches should be developed to increase the longevity of electronic systems, through certification and virtualization, modularity, specific supervision, etc.

Conversely, computing should also be used to find solutions to the sustainability crisis facing the planet and bring the ecological footprint of Europe to within the continent's biocapacity.

INVEST IN THE FUTURE WORKFORCE

The impact of computing on employment cannot be underestimated. Disappearing medium-skilled jobs will increase income inequality and may lead to social unrest. Europe should continue to invest in training programs for workers at risk of losing their jobs, and try to reintegrate them in the job market at the highest level.

On the other hand, automation in AI may deliver the productivity the continent needs. In addition, Europe must invest in digital skills to maintain its innovation potential and remain competitive. Areas to be prioritized include machine learning, security, blockchain, architecture, system design and tools.

DEVELOP A ROBUST DIGITAL ETHICS FRAMEWORK

Computing has become such a powerful commodity that we should start thinking about whether everything that *can* be done *should*, in fact, be done. It is time to invest in digital ethics as a discipline to guide us to the future and to make sure that all computing professionals receive basic training in it. Digital ethics should also support policy makers to make decisions.

HIGHLIGHTS OF THE HIPEAC VISION 2019

The current surge of attention around artificial intelligence will hopefully lead to seeking solutions that help increase efficiency, quality and designer productivity for both hardware and software, despite the massively increasing complexity of modern ICT systems. This very much correlates with the **“keep it simple for humans, and let the computer do the hard work”** element of the 2009 HiPEAC Vision. Intelligent or more “cognitive” solutions need to be developed at the edge (in what will give rise to *connected cognitive cyber-physical systems*) for safety, privacy and efficiency reasons.

Efficiency, particularly energy efficiency, is still a key challenge that will lead to increasing heterogeneity in hardware, with the “classical” silicon processor acting more as an orchestrator of various accelerators, using various technologies (GPU, neuromorphic, quantum computing, and so on). However, there is still no credible “successor” of silicon on the horizon.

Software, applications and infrastructures will increasingly be aggregates of heterogeneous artefacts, including legacy ones, with a variety of deployment requirements. Software will be distributed, becoming a **“continuum of computing”** across platforms and devices. Programming has to be reinvented for this, 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.

To successfully be accepted, our ICT systems should inspire trust in their users, so they should ensure security, privacy and safety, hardening against cyber-attacks.

Therefore, the HiPEAC 2019 recommendations for Europe are:

- Accelerate, accelerate, specialize and automate
- Develop alternative architectures
- Build computers we can trust
- Get looking for CMOS alternatives
- Treat the computing infrastructure as a continuum
- Shift value towards the edge
- Lead on the use of collective data
- Become a leader in energy-efficient sustainable electronics
- Invest in the future workforce
- Develop a robust digital ethics framework.

ISBN 978-90-90-31364-1



9 789090 313641

