# Scheduling Variable Capacity Resources for Sustainability Workshop

Anne Benoit, Andrew A Chien, Yves Robert

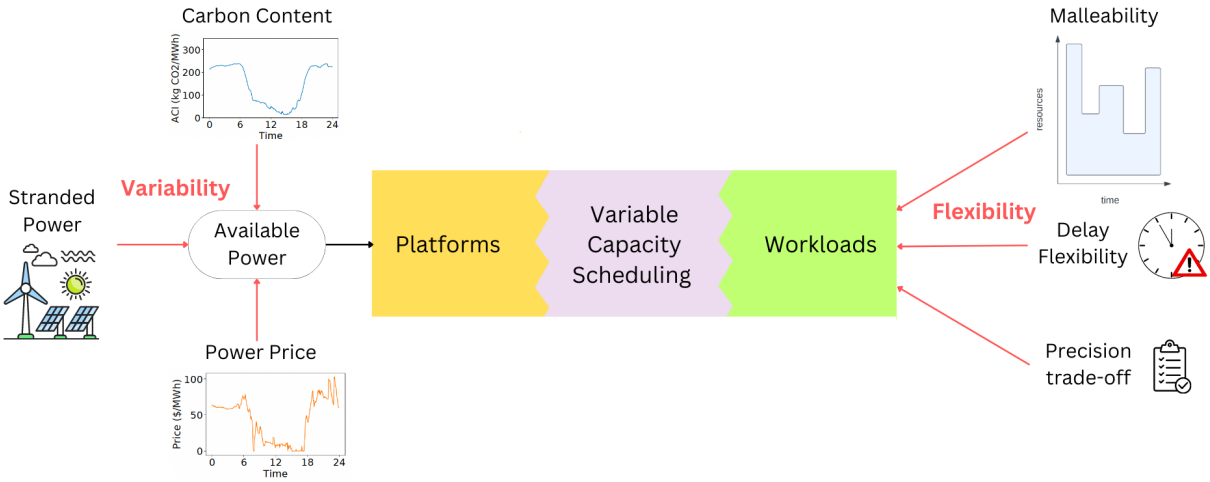
# Scheduling Variable Capacity Resources for Sustainability Workshop

Organizers: Anne Benoit (ENS Lyon), Andrew A Chien (UChicago and Argonne), Yves Robert (ENS Lyon)

Student Recorders: Svetlana Kulagina (Univ. Berlin), Lucas Perotin (ENS Lyon), Rajini Wijayawardana (UChicago)

Additional contributors to the report: Florina Ciorba (Univ. Basel), Denis Trystram (Univ. Grenoble Alpes)

March 29-31, 2023
UChicago Paris Center, Paris, France



# Executive Summary

We gathered the research community focused on compute scheduling in the new age of renewable power generation -- where variation in weather and solar radiation drives renewable generation, and thereby the carbon-intensity of power, and efforts to reduce environmental damage then create variation in compute capacity (opportunity). In this age, effective utilization of that opportunity is the key to sustainable computing. We report important findings on the attributes of future platforms, characterizing the power grid variation, reliability challenges, and two-way relationships between datacenters and grid that are emerging. Workloads are changing as well, with growing understanding of how computing can be malleable, delay flexible, or even acceptably approximate. Between these, we focus on how scheduling can meet these challenges, managing between flexible workloads and dynamically varying platforms – performing to new metrics that reflect both performance and non-performance attributes (e.g. carbon emissions). The concerns that underlie a shift to variable capacity derive from a broader societal concern about the sustainability of computing, and our last section addresses critical challenges in awareness, responsibilities, and action.

The talks and position papers from attendees are available at https://people.cs.uchicago.edu/~aachien/workshops/varsched23/.

**Table of Contents**

# 1. Objective

The aim of the “Scheduling Variable Capacity Resources for Sustainability" workshop was to nucleate a research community focused on compute scheduling in the new age of renewable power generation. The growing fraction of renewable energy in grids worldwide has led to a situation where the variation in weather and solar radiation drives variation in compute capacity. This creates new challenges in job scheduling and datacenter resource management, as well as opportunities for harvesting renewable energy to drive the future growth of computing. The effective utilization of platform resources while adapting to this dynamic resource environment is key to the progression of sustainable computing. This workshop was designed to foster collaboration and bridge the traditional scheduling research community, corporate leaders in cloud resource management research, funding agencies, and leaders in government policy, whose innovation in adaptive datacenters loads has accelerated clean renewable energy usage and thereby power grid decarbonization.

# 2. Organization

The “Scheduling Variable Capacity Resources for Sustainability" workshop was held at the Paris Center of the University of Chicago from 29th to 31st of March 2023, with the support of the University of Chicago under the “France And Chicago Collaborating in The Sciences” (FACCTS) program.[1]

The workshop was by invitation only and gathered European, US, and Asian researchers from universities, national laboratories, and industry. Specifically, attendees traveled from France, Switzerland, Poland, Germany, Luxembourg, the US and Japan and had a diverse range of expertise in areas of sustainability, carbon emissions analysis, cloud and HPC resource management, scheduling, stochastic processes and power management. They contributed towards the successful discussion and formulation of the variable capacity research domain. To facilitate the interaction between different communities, the workshop began with presentations by attendees, interleaved with discussions and panels. The workshop concluded with working groups on the identified topics of interest: scheduling and metrics, workloads and scheduling, policy and societal factors, and future platforms and grids. We extend our sincere gratitude to the FACCTS program at UChicago without whose generous support, and the staff of the UChicago Paris center without which the workshop would not have been possible. We expect to organize a follow up workshop in 2024, to support continued collaboration and discussion.

[1] For more information on the program, see https://fcc.uchicago.edu/faccts/

## 2.1 Workshop Agenda

| | |
|---|---|
| **Day 0 ( Tuesday, 28th March)** | |
| 19h | Reception and Dinner |
| **Day 1 ( Wednesday, 29th March)** | |
| 8h30-9h | Coffee |
| 9h-9h30 | Keynote: Pierre Segone, Electricity Maps |
| 9h30-10h10 | Morning Session 1 - **Power and renewable energy** |
| | Power-aware resource management for the Supercomputer Fugaku, Keiji Yamamoto, Riken, Japan |
| | Network-aware energy-efficient virtual machine management in distributed Cloud infrastructures with on-site photovoltaic production, Anne-Cécile Orgerie, CNRS & U. Rennes, France |
| 10h10-10h40 | Discussion |
| 10h40-11h | Coffee |
| 11h-12h | Morning Session 2 - **Variable capacity** |
| | Managing Variable Capacity in the Chameleon Testbed, Kate Keahey, Argonne and UChicago, USA |
| | Scheduling on variable capacity resources, Anne Benoit, ENS Lyon and CNRS/LIP |
| | Time series models for variable capacity resources, Varun Gupta, UChicago Booth, USA |
| 12h-12h30 | Discussion |
| 12h45-14h | Lunch |
| 14h-15h20 | Afternoon Session - **Batch schedulers** |
| | Why current Batch Schedulers are not sustainable, Denis Trystram, U. Grenoble, France |
| | Scalable and Efficient Big Data Management in Distributed Systems: Addressing performance variability for Data processing in the Cloud, Shadi Ibrahim, Inria Rennes, France |
| | ElastiSim: A Batch-System Simulator for Malleable Workloads, Felix Wolf, U. Darmstadt, Germany |
| | Understanding your optimization criteria, Guillaume Pallez, Inria & U. Bordeaux, France |
| 15h20-15h50 | Discussion |
| 15h50-16h15 | Coffee |
| 16h15-17h30 | **Panel: Radical Futures** (led by Andrew Chien, with Maciej Drozdowski, Pierre Segonne and Denis Trystram) |
| 19h | Dinner at Anco |
| **Day 2 (Thursday, 30th March)** | |
| 8h30-9h | Coffee |
| 9h-9h30 | Keynote: Ana Radovanovic, Google, Switzerland |
| 9h30-10h10 | Morning Session 1 - **Workloads** |
| | Workload Limits on Cloud tolerance of Capacity Variation, Chaojie Zhang and Andrew A Chien, UChicago, USA |

| | |
|---|---|
| 10h10-10h40 | Discussion |
| 10h40-11h | Coffee |
| 11h-12h | Morning Session 2 - **DVFS and extensions** |
| | Combining Uncore Frequency and Dynamic Power Capping to Improve Power Savings, Amina Guermouche, U. Bordeaux, France |
| | Online regulation of power draw for energy savings in HPC kernels, Piotr Luszczek, UT Knoxville, USA |
| | Automated Scheduling Algorithm Selection in OpenMP, Florina Ciorba, U. Basel, Switzerland |
| 12h-12h30 | Discussion |
| 12h45-14h | Lunch |
| 14h-15h20 | Afternoon Session - **Time-energy trade-offs** |
| | Time-energy trade-offs in processing divisible loads on heterogeneous hierarchical memory systems, Maciej Drozdowski, Poznan University, Poland |
| | On scheduling algorithms for minimizing the energy consumption and execution time of Federated Learning, Laércio Lima Pilla, CNRS and U. Bordeaux, France |
| | Bridging the Cloud & HPC, Johnatan E. Pecero, University of Luxembourg, Luxembourg |
| | Scheduling with Two Unbounded Resources with Communication Costs, Alix Munier Kordon, U. Sorbonne, France |
| 15h20-15h50 | Discussion |
| 15h50-16h15 | Coffee |
| 16h15-17h30 | **Panel: Unspoken Challenges** (led by Yves Robert, with Florina Ciorba, Guillaume Pallez, and Ana Radovanovic) |
| 19h | Dinner at Le Train Bleu |
| **Day 3 (Friday, 31st March)** | |
| 8h30-8h45 | Coffee |
| 8h45-10h15 | Working groups by topics |
| 10h15-10h30 | Coffee |
| 10h30-11h30 | Report out to group |
| 11h30-12h | Discussion |
| 12h-13h | Lunch |

## 2.2 Workshop Attendees and Affiliation

**Participants/Speakers**

- Anne Benoit, ENS Lyon, France
- Andrew A Chien, University of Chicago and Argonne National Laboratory, USA
- Florina Ciorba, University of Basel, Switzerland
- Maciej Drozdowski, Poznań University of Technology, Poland
- Amina Guermouche, University of Bordeaux, France
- Varun Gupta, University of Chicago, Booth School of Business, USA
- Shadi Ibrahim, Inria Rennes, France
- Kate Keahey, Argonne and University of Chicago, USA
- Piotr Luszczek, University of Tennessee, Knoxville, USA
- Alix Munier-Kordon, Sorbonne University, France
- Anne-Cécile Orgerie, CNRS & University of Rennes, France
- Guillaume Pallez, Inria & University of Bordeaux, France
- Johnatan E. Pecero, University of Luxembourg, Luxembourg
- Laercio Lima Pilla, CNRS and University of Bordeaux, France
- Ana Radovanovic, Google, Switzerland
- Yves Robert, ENS Lyon, France
- Pierre Segonne, Electricity Maps
- Denis Trystram, University of Grenoble Alpes, France
- Felix Wolf, TU Darmstadt, Germany
- Keiji Yamamoto, Riken, Japan

**Student Recorders**

- Svetlana Kulagina, University of Berlin, Germany
- Lucas Perotin, ENS Lyon, France
- Rajini Wijayawardana, University of Chicago, USA

# 3. Platforms and Resources

Andrew A Chien and Rajini Wijayawardana, University of Chicago

## 3.1 Findings

The rising prominence of sustainability of computing concerns creates new dimensions of variability in the properties and quantities of computing resources. It adds the complexity of time-varying resource properties to the already daunting explosion in heterogeneity both in hardware (servers, GPUs, embedded, and even mobile/IoT devices) and deployment (datacenters, edge, etc.) [13]. As background, we highlight several findings that frame the research areas and questions:

1. Power costs already vary 5-20 fold over periods of days/weeks (2022), and this variation is projected to increase as global and local initiatives drive an increasing fraction of power generation to use variable generation sources (eg. wind, solar) [7].
2. The carbon-intensity of power used by datacenters already varies 2-20x over periods of days/weeks and this variation is projected to increase as power generation is shifted increasingly to variable renewable generation sources (eg. wind, solar). In some advanced grids, electric-power's carbon-intensity can fluctuate down to zero [25], and this variation is greater in more isolated power settings such as edge datacenters [15].
3. Power grids face increasing challenges to maintain their reliability cost-effectively in the face of increasingly variable generation, and thus likely to constrain power use of large loads, producing variable capacity.
4. Driven by overall carbon-emissions accounting (and the desire for reduction), many platforms will operate as variable capacity, adapting based on power/carbon/price circumstance [24,15].

These findings support the creation of a new two-way, perhaps many-way relationship between computing workloads (and resource managers) and underlying platforms. The critical differences implied include opportunities as follows:

1. Workloads and resource managers can shape the capacity (and even heterogeneity) properties for a given point in time (present, future), based on willingness to bear costs (price, carbon emissions) and available workload demand.
2. External entities (power grid, and circumstances (weather, time of day, competitive load) can shape the capacity (and even heterogeneity) properties; for example based on availability of power, carbon intensity, or reliability needs.
3. These interactions are not confined to a single site or facility, but with widespread high speed networking and geo-replicated data can involve multi-site planning and optimization

4. These interactions involve no complex federated and integrated resource and reward structures (e.g. different power grids, nations, international networks of datacenters operated by a single corporation, portable workloads, etc.)

These exciting and challenging changes form the basis for a new set of research areas and questions about platforms/resources, and of course the implications of these questions for resource management and scheduling.

## 3.2 Research Areas and Questions

**New and more Complex definitions of Capacity.** Traditional notions of capacity include compute (MIPS[2] and BIPS[3]), memory (gigabytes) and network rate (Gbps). Sustainability intrinsically adds time-variation to these properties as the carbon intensity or cost of power varies with time (5 minutes, hours, days, weeks, seasons). Further, in some cases the total power available (at any price) may be less than computer capacity (eg. shortage). Also, limits may be soft - a target power level to reduce average carbon emissions - or hard - hardware damage if exceeding a power cap. These problems compound current challenges already imposed by growing hardware heterogeneity (CPUs, older CPUs, GPUs, older GPUs, media accelerators, smaller CPUs and GPUs – edge, mobile, etc.).

Describing Resource Capacity as a Function of Time [28]

- How to describe instantaneous and future resource capacity? (compute, networking, and storage)
- How accurately can it be forecast? How to deal with uncertainty and change?
- What is the role of soft limits (goals) versus hard limits (strict)?
- Within this, how to manage growing heterogeneity?
- Is capacity malleable at some cost? How do malleable workloads affect effective ways to describe capacity?
- How can these more complex definitions be made simple enough to be understandable to applications? Users? Schedulers?
- Can we separate these concerns into distinct dimensions or subproblems, enabling modular or composable solutions?

Negotiation of Capacity between Compute Management and External Factors

- How should a computing resource's capacity be best determined? Ex. using a scope (local, regional, global optimization) and optimization procedure over some objective? (cost, carbon emissions, lifetime,...)
    - How does changing resource capacity affect the resource's utility for computing?

[2] Million instructions per second
[3] Billion instructions per second

    - How do constraints over data and computation movement (eg. govt regulation, network limits, other?) factor in?
- How can capacity planning be coordinated across different autonomous systems (with different incentives and objectives) for better overall outcomes? What are the limits of such coordination? (technical competition, govt regulatory)
    - How should multiple networks of datacenters coordinate with each other? What positive and negative effects?
    - How should other adaptive load pools (eg. energy storage, building/home heating/cooling, EV charging) play into this?
- What role can regional power grids, as localized players have in this coordination?
    - How should carbon-emissions effects on the power grid affect the negotiation?
    - How does power grid reliability limit or incentivize collaboration?
    - With power grid effects spanning time scales of hours, days, weeks, months, for longer than the typical computation resource management, how do we effectively bridge these time scales?
    - How do we shape incentive structures to support/drive coordination?

Impact of Structure/Nature of Variable Capacity on External Utility (for the Grid or other coupled systems, e.g. natural gas, water, environment)

- Does the structure/nature of variable capacity enable optimization of external metrics such as carbon emissions or cost?
- How does it fit into a larger class of pilotable (controllable) loads in the power grid? (eg advanced demand response) [15]
- How should these controllable loads be coordinated?
- How do we shape incentive structures to support/drive coordination?

Reporting non-Performance Attributes

- How are carbon-emissions captured in resource description and tied to use (both operational and embodied [10]), for compute, storage, and networking
- How are other sustainability attributes such as e-waste and water use captured in resource description and tied to use?
- How are impacts on others through competition (eg. elevated power prices for others) reflected?
- Some attributes may be conditioned on dynamic actions (eg. load shifting) or attributes (eg. flexible load), how do we report/incentivize credit for dynamic actions?
- How do we make reporting of these attributes auditable? When they are collected by corporations that do not make sufficient information available? (eg. cloud or datacenter operators)
- How do we define new non-performance attributes that track and incentivize sustainable use? How to support the needs of the future dynamic grid for stability?

Cloud, Edge, Mobile, and Internet of Things resources

- The issues above apply to cloud datacenters, but also collections of cloud, edge, and mobile devices as well as each of those platforms independently. Edge and mobile devices have unique characteristics outlined below.
- Edge datacenters have remarkably higher diversity (heterogeneity) – hardware resources, load and service requirements, and power environment (with and without renewables, siting, etc.). How do we solve all of these problems, but with 1000x more sites and increased heterogeneity?
- Mobile devices from smartphones to electric vehicles to ships, drones, and airplanes represent new opportunities and challenges for sustainability. Similar problems involve the acquisition of low-carbon power, but new opportunities include logistics for both the transport of power, and additional grid/microgrid flexibility by exploiting the internal energy storage. Can we solve all of these problems? What new opportunities arise? And with 1,000,000x more devices?

### 3.3 Contributors / Participants

Andrew A Chien, University of Chicago and Argonne
Varun Gupta, University of Chicago Booth School of Business
Shadi Ibrahim, Inria Rennes, France
Kate Keahey, Argonne National Laboratory and UChicago
Anne-Cécile Orgerie, CNRS & U. Rennes, France
Pierre Segonne, Electricity Maps
Rajini Wijayawardana, University of Chicago
Keiji Yamamoto, Riken, Japan

## 4. Flexible Workloads

Anne Benoit and Lucas Perotin, ENS Lyon

### 4.1 Findings

Current workloads appear to be more and more flexible, and it seems very important that future workloads become, if possible, even more flexible, in order to adapt to the capacity of the platform and effectively utilize available resources [17,11,21,24]. There are several facets to flexibility, which we detail below.

1. **There should be flexible start dates and flexible resources used for any given workload.** Rather than having a rigid requirement to start immediately, workloads should be designed to be executed anytime before a flexible deadline. Unless it is a time-sensitive task (like a weather forecast), we should run any workload whenever

enough resources become available, rather than discarding it if it cannot run immediately.
Furthermore, when requesting resources for a job, the number of cores needed is often fixed. However, a more dynamic approach involves placing the job in a queue, and as resources become available, the system chooses the allocation based on defined priorities and could further adapt. This allows for better resource management and efficient execution. We should be able to give incentives to users so that they propose flexible workloads.

2. **Flexibility can also be achieved in terms of accuracy.** This involves the concept of approximate computing, where less precise results are acceptable within certain bounds. By defining a function that improves over iterations, the number of iterations can be determined to achieve an acceptable level of precision. This approach allows for trade-offs between computation time and precision.

3. **The primary goal should be to reduce the carbon footprint generated by computing activities.** Currently, we focus solely on the ability to run applications, but we should also look at the environmental impact. By shifting the focus to carbon reduction, greater attention can be placed on mitigating the ecological consequences of computing. To do so, we should evaluate the cost implications of different approaches and resource allocation strategies. This involves assessing how expensive certain operations or computing tasks can be, and incorporating a cost function to make informed decisions based on cost factors.

4. **Service level agreements (SLAs).** Cloud platforms are commonly using SLAs, and maybe HPC platforms should adapt such ideas and solutions from the Cloud, as there is no SLA in most current systems.

5. **Locality of data.** Current (and future) workloads are handling large amounts of data. Because of data governance and privacy, moving data might be difficult. Furthermore, it comes at a cost. Flexible workloads should be able to be moved, together with their data, if this allows for a more efficient execution, reducing the carbon footprint as highlighted above.

6. **Workflows.** Workflows are a good way to represent flexibility; if the application is made of different parts, the parts can be moved more easily to different locations so that they can take advantage of variable capacity at the platform level. There are different ways of defining workflows, in many languages.

## 4.2 Research Areas and Questions

**What will future workloads look like, and what should they look like?**

Should future workloads possess a full awareness of their execution requirements? How should uncertainties be accounted for in workload design? Which types of jobs are best suited for flexibility? How should workload specifications be defined to incorporate flexibility? Is it advantageous to advocate for non-urgentness whenever possible? Is there a need for flexibility in terms of precision, allowing for approximate computing?

In addition to these questions, it is crucial to consider different categories of workload urgency and timing, such as mandatory and instant, mandatory and non-instant, and non-mandatory and (non-)instant. How should workload representation be approached to capture these nuances effectively? Should we consider malleable jobs and suspendable jobs as much as possible?

**How to promote the (necessary change) to flexible HPC applications?**

How can we motivate users to design applications with as much flexibility as possible? As incentives, one can consider pricing policies, but also making the user aware of the environmental impact and how flexibility may lower carbon footprint. For the sake of transparency, it should be clearly stated what happens to jobs with lower priority. Also, users' SLAs should be respected. An important role is to be played by educating the developers (users) that flexible HPC applications will also be finished faster. Hence, there is a natural incentive to frame HPC applications as flexible as possible.

How to motivate flexibility and guarantee that there will not be too much loss in performance? Can we identify stragglers and migrate jobs as needed, in order to reach trade-offs between performance and energy consumption? If a task takes too long to complete, we would like to be able to interrupt it, if it is a mandatory task, and eventually migrate it to a processor where it could complete in a more timely way.

**Which kind of flexibilities are interesting for workloads?**

When considering valuable flexibilities for workloads, important aspects include workload grouping for cost-effective migration, running at lower precision with the ability to do additional computations later for improving accuracy, and incorporating affordable interruptions through checkpoints and non-volatile memory. These flexibilities optimize resource utilization, minimize costs, and enhance the efficiency and reliability of workloads.

**Which kind of flexibilities are interesting for infrastructures/hardware?**

Evaluating the effectiveness of flexible workloads involves various approaches. Conducting experiments on real systems is one method, although it can be costly and time-consuming. Simulations using tools like SimGrid offer a more efficient alternative, providing a simulated

environment for evaluating workload performance. Additionally, in-house simulators can be developed to model the behavior of real systems using specific assumptions for a given work.

To approximate the execution time of flexible workloads, the ElastiSim simulator [21] can be used, as it is the first simulator specifically designed to support malleability, allowing for the evaluation of workload adaptability and performance under different scenarios.

By employing a combination of real system experiments, simulations using tools like SimGrid, and the use of specialized simulators like ElastiSim, comprehensive evaluations of flexible workloads can be conducted efficiently and effectively.

### 4.3 Contributors / Participants

Anne Benoit, ENS Lyon, France
Amina Guermouche, University of Bordeaux, France
Shadi Ibrahim, Inria Rennes, France
Svetlana Kulagina, University of Berlin
Piotr Luszczek, University of Tennessee, Knoxville, USA
Lucas Perotin, ENS Lyon
Felix Wolf, TU Darmstadt, Germany

## 5. Scheduling Models and Metrics

Florina Ciorba, University of Basel, and Yves Robert, ENS Lyon

### 5.1 Findings

In the past, parallel and distributed computing environments consisted of stable capacity resources, with homogeneous architectures, executing jobs with equal priority [9,23,26,27,20,18,2]. Approaches used to model and evaluate performance were agnostic to environmental issues, including platform-oriented metrics such as utilization or goodput, as well as user-oriented metrics like response time and stretch. Note that the goodput is a variant of platform utilization where only work executed for successful jobs is accounted for.

However, with the advent of **more complex and diverse parallel and distributed computing environments** (see Platforms and Resources), these traditional scheduling models and metric approaches are no longer sufficient [16]. We highlight below the **findings** that guide the research areas and questions discussed at the workshop:

a. Modern parallel and distributed computing environments require the development of **new models for resource variability.** These models must consider both homogeneous and heterogeneous resources, as well as the availability of machines,

which can vary over time depending on factors such as user demand or energy availability in a geographic location (e.g., “follow the sun”) [5,3,28,24]. Additionally, the concept of *system malleability* has emerged, which refers both to (i) the ability to adapt to changing resource availability by dynamically adjusting resource allocation and scheduling and (ii) the ability to increase or decrease the number of required resources in response to the requirements of the flexible workloads.

b. Modern parallel and distributed computing environments also require **new models for job classification** (see Workloads). These models must consider “*not-instant*” computing, allowing for delays in starting and completing jobs, as well as the option for *approximate computing* [19] by specifying mandatory or non-mandatory parts of a job to trade off result accuracy or resolution with job completion. These models must also provide *flexibility* in iteratively refining these non-/mandatory parts of a job and account for data and work locality and estimation of the data volumes involved.

c. **New metrics** are needed to evaluate the **performance** of modern parallel and distributed computing systems. In addition to traditional metrics of efficiency and responsiveness, these metrics must now take into account environmental issues (also known as non-performance attributes) such as e-waste, datacenter water use, power, energy consumption, and CO2 emissions (see Platforms and Resources) [28,24], as well as network congestion and system heterogeneity. Capping and reducing brown energy consumption, which involves lowering the power level bought on the fixed annual contract and increasing intermittently the power level using green energy and daily contracts, are becoming increasingly important. Furthermore, there is a growing need to shift from "fast" to "green" computing, which emphasizes energy efficiency and sustainability in addition to performance.

These requirements for **new models and metrics** reflect the changing demands of modern parallel and distributed computing environments and the need for more sophisticated scheduling approaches to optimize **both performance and sustainability** (see figure on the front page of the report).

When designing a new metric, particularly when considering environmental issues, one should always care about the risk of the rebound effect or the Jevons paradox (see https://en.wikipedia.org/wiki/Rebound_effect_(conservation)): by improving energy efficiency, users change the way they use a technology, and the energy gain disappears.

In addition, with the advent of machine learning-based algorithms, resource management strategies may lose in explainability–why does the algorithm take such and such decisions? A clear understanding of the limits of the target optimization objective(s) becomes particularly important [1].

## 5.2 Research Areas and Questions

1. With new models and metrics for the emerging parallel and distributed computing environments with variable capacities, novel job scheduling algorithms need to be developed. From the point of view of batch job scheduler (e.g. SLURM [27]), the traditional rigid jobs will be complemented by moldable, malleable, and evolvable jobs, for which the allocated resources will vary over time, either triggered by the system at submission time (as it is the case for moldable jobs), by the system during execution (as is the case for malleable jobs), or by the application during execution (as is the case for evolvable jobs). Here is a more detailed description of these job categories [9,23] (see also Flexible Workloads):
- **rigid**: classical parallel job whose execution uses a fixed, given number of processors
- **moldable**: parallel job whose execution uses a fixed number of processors, but this number is chosen at submission and can vary in range or shape. A typical example is a matrix product for which the user specifies a list of possible grid configurations (say 10x10, 8x12 or 12x8) and the batch scheduler selects one of them. In this example MPI-I/O would be used to distribute input data according to selected grid size
- **malleable**: parallel job whose execution uses a number of processors that can vary over time. A typical example is a parallel job that can survive losing one processor due to a fail-stop error. Another example is a master-worker job whose number of workers can be increased or decreased by the batch scheduler, depending upon resource availability;
- **evolvable**: malleable job whose number of resources can be varied during execution at the request of the user. A typical example is a parallel job which during its execution will temporarily require more resources (e.g., due to adaptive mesh refinement or adaptive time sub-stepping) to spread out/speed up parts of the simulation, but then would revert to the initial number of allocated resources.

For **rigid jobs**, there are significant opportunities for scheduling, but allocated machines no longer sustain equal capacity computing. The scheduler must thus be risk-aware and consider machines with high risks when backfilling with jobs of lower priority or higher tolerance to risks.

For all jobs, placement is important for maintaining locality, which may significantly reduce energy costs and communication delays. Based on power grid constraints, the scheduler will have to decide which nodes will be on or off. The scheduler may also overdrive or use dynamic voltage and frequency scaling (DVFS) on allocated resources to finish jobs sooner. Oversubscription is an increasingly common approach, initially used in smaller compute centers, but recently present in top-tier supercomputers (e.g., LUMI).
In constant resource capacity scenarios, job preemption and suspension remain sensible approaches, with checkpointing as the mechanism, which requires non-volatile memory.

In the case of **moldable jobs**, the system decides the number of resources to allocate at job submission time. Therefore, the scheduler must also decide the shape of jobs (degree of parallelism, job length) to allocate the resources, and the shape can vary across different executions. With queues that support preemption and suspension of running jobs, it is important to support resumption of a previously suspended job under a different shape than the initial shape.

**Malleable jobs**, for which the system decides at runtime how many and which resources to allocate, require the same considerations as moldable jobs (above).
However, reshaping of jobs may even take place during execution to avoid suspending them (via checkpointing) and to allow for on-the-fly redistribution of the memory footprint.
The goal is to *shave the load* now and *shift it later* in time. Non-volatile memory will be required to support reliable suspension and checkpointing.

Finally, **evolvable jobs**, which allow the user to decide their size during execution, are supported by batch schedulers offering more or fewer resources for adaptive resizing under the constraints discussed above.

2. There are several **approaches** to consider for designing new job scheduling algorithms on parallel and distributed computing systems with variable capacity.

Firstly, it is crucial to have an *exchange of information between jobs and the job scheduler* as a prerequisite. Recent efforts show both the challenges and promise of this approach [8,6]. This exchange will provide transparency and visibility about what the job scheduler will do with *non-instant jobs* (which accept execution delays) and whether it respects the users' Service-Level Objectives (SLOs). Establishing effective communication and coordination between jobs and job schedulers requires defining compelling incentives and rewards. Pricing models that incentivize later deadlines could additionally be explored.

Another important aspect is supporting the batch scheduler with implementation of *job checkpointing and reshaping*.
The batch scheduler should also be *risk-aware* regarding available resources for allocation and their variable capacity.
It is necessary to incorporate a *global perspective* of jobs' new parameters (as described in point 1 above), as well as their impact on resource usage. Lastly, being aware of the *current state of the platforms' capacity* is crucial for effective scheduling on variable capacity resources.

**3. Questions related to Scheduling Models and Metrics**
The group discussion led to the following open questions related to workloads and their models, computing platforms, scheduling objectives and approaches to support scheduling on variable capacity resources.

**Models**: How do we expose, express, **exchange**, and **exploit** all this (new) information about the workloads (shape flexibility and SLO) and the platforms (variable capacity) to a scheduler?

**Workloads**: How do we express and support job suspendability, evolvability, moldability, and malleability? How to combine these considerations into combinations of non-mandatory and non-instant computing job requirements?

**Platforms**: Under the premise that capacity resources exhibit high variability, how can arriving jobs be classified into rigid, moldable, malleable, evolvable? How to predict the flexibility of a workload in terms of job shaping? How to model and predict system malleability?

**Uncertainty**: both workloads and platforms encompass high degrees of uncertainty. A workload with uncertain requirements may lead to overestimating or underestimating resource requirements. Platforms with high capacity variations may lead to interrupting or reshaping applications. Uncertainty must be accounted for globally by scheduling and resource management algorithms.

**Objectives**: How do we expose and exchange information between jobs and schedulers to enable cooperation? What is the interplay between the user-facing incentives and the job scheduling objectives (which include delays, carbon-saving, migration, etc.)? Do we need pricing models and negotiations to model and study this interplay?
Do we still want a single-objective optimization function? Or should we consider multiple objectives at a time, but optimize for one at a time: performance capping and energy minimization (not always *-capping and performance maximization).

**Algorithms**: What are the scheduling techniques that allow us to move from single platforms to distributed platforms, to allow us to exploit workload flexibility and variable resource capacity? Are online algorithm selection approaches a viable solution for adaptive job scheduling on variable capacity resources?

**Approaches**: How can the new job types be supported by the new job scheduler in practice (rigid+suspendable, moldable+suspendable, malleable+suspendable, evolvable+suspendable)? Do we need a time-quantum approach to support preemption in job scheduling (à la Operating System scheduling)?

## 5.3 Contributors / Participants

Florina Ciorba, University of Basel
Laércio Lima Pilla, CNRS Bordeaux
Guillaume Pallez, INRIA Bordeaux

Lucas Perotin, ENS Lyon
Yves Robert, ENS Lyon

# 6. Policy and Societal Factors

Denis Trystram, University of Grenoble Alpes

## 6.1 Findings

**Positioning**
The IT community, as a whole, witnesses the massive growth of digital technologies and its diffusion in all industrial/commercial sectors and in services offered to citizens. HPC has a significant impact on the climate crisis through CO2 emissions [4,22]. On the positive side, it can provide solutions to mitigate the crisis; on the negative side, it is part of the problem through the energy expended in large-scale computing platforms.

**Digital is everywhere**
The carbon emission impact of the IT sector is hard to estimate since it is diluted in all the other sectors from agriculture to health or transportation. Any modern car contains more than a hundred computing devices and integrated circuits. Depending on the studies and the perimeter involved, it represents about 3 to 4 percent of the greenhouse gas emissions per year, which are only part of a larger phenomenon of the fast degradation of the planet as a consequence of the development of human activities since the 1950s. The growth is estimated with an acceleration of 6 to 9 percent per year. That is not sustainable in a time when we should reduce the GHG emissions (see the 2019 report https://theshiftproject.org/wp-content/uploads/2019/03/Lean-ICT-Report_The-Shift-Project_2019.pdf and its 2021 update https://theshiftproject.org/wp-content/uploads/2023/04/Environmental-impacts-of-digital-technology-5-year-trends-and-5G-governance_March2021.pdf).
.
The challenge for the community is therefore to study the means to get out (or to mitigate) of the climate crisis and to behave in a constrained world where the resources are finite and not always available. The general consensus in the HPC community is to develop IT to solve complex environmental problems and help in determining solutions. Such solutions may be to better understand earth sciences or simulations of extreme events. It also includes the pathways to promote better practices. Another way is being heard which advocates the renunciation of any existing and future IT solutions if it is not strictly demonstrated that they do not have negative impacts in their direct, indirect and rebound effects. At the very least, the total carbon footprint taking all the life cycle assessment into account should not be negative. The right attitude will rely on solid scientific analyses that measure such impacts and propose better practices for using less computing power.
We may help on two levels.

1. First, within our academic computing community, on the one hand, to study the negative effects, which requires pluri-disciplinary collaborations with other hard sciences like physicists or climatology, and to imagine mechanisms to circumvent them. On the other hand, to focus our research on the optimization of complex problems to help imagine better solutions to be deployed efficiently in the field.

2. Second, outside our community, at the society level. The main challenge here is the acceptability of recommendations towards good practices, respectful of the environment, which are necessarily binding. Individuals should be educated in particular by systematically providing the carbon cost of any usage of digital equipment, but that will take time. We can also influence decision makers (local, national or corporate) by convincing studies and well-founded alternatives.

## 6.2 Directions for Research and Action

**Analysis of constraints and potential solutions**
**Society and Environment informs our research**
We, as computer engineers and researchers, are aware of the urgency of the climate warming situation. We have to do something in order to limit the uncontrollable growth of IT without necessarily knowing what to do yet. Based on well-established facts, we are aware that material resources of the Earth are limited and become a strong constraint for building future computers and digital devices. Most resources will be lacking very soon, such as the several rare metals that are used for building digital components or such as the energy to run them [14]. We have the choice between the continuation of the race of performance as we did in the past decades or to adapt our research subjects and the way we are conducting them to more sustainable purposes. For instance, the classical objective functions that are targeted in batch schedulers reflect a fair sharing of computing resources among users who are competing for achieving their jobs as soon as possible. Introducing flexibility reflecting the diversity of the users may relax the performance constraints and save energy. No matter if I get my results in the next 2 hours if I submit my job before going home in the evening!  It is very hard to guess what new proposals may be actionable and acceptable by the users of an HPC platform. This holds for every usage of computer devices that should be socially acceptable (like video streaming). Thus, we can only recommend some directions of actions with the hope that some of them may become effective in the future.

Some directions may be deployed immediately but most of them will take time (in terms of years).

1. To meet these challenges would induce big changes that will highly impact the users. Thus, an important step is to identify and classify existing mechanisms that could

help people to accept the constraints linked to environmental rules. Such changes will be long-term although we need to urgently move on and the effort is collective: we will not be able to succeed without the help of colleagues coming from other disciplines, including sociologists, economists, philosophers.

2. Another important direction is to get out of our academic ivory tower and propose new/efficient mechanisms with the target of practically implementing them, respectful of the planet's limits. This must be done with the issue of environmental preservation in mind, the reparability of software tools, their flexibility, at the lowest possible cost. Preserving fairness among all users, this is the only way common people will accept the constraints.
3. The changes must be societally prepared e.g. via public and democratic discussions: the challenges, the goals to attain and the tools to achieve them should be clearly identified. There must be transparency in these actions otherwise imposing necessary changes on unaccepting society will not be actionable and acceptable. Education is the main pillar here.
4. Computer science should be taught to children as other sciences. Consequently, the required mentality changes in society and companies will take time.

**Recommendations**

**Our responsibilities to Society**

The main condition for convincing people (citizens) to move on and integrate new constraints that impact their usage is to get verified and trustable data and information about the environmental and climate crisis and its human origin. The interest of computing scientists to study and develop such measurement tools is growing [12].

Don’t forget that Science needs time, it is impossible to provide quick information or recommendations. The process will take some time and should start as soon as possible.
Each group making politics, namely, decision makers, lobbyists (often led by industrial groups), “grass root” (ordinary) people should be addressed with specific arguments convincing them on the need for change. For lobbyists, the “business as usual” is not a long term solution since it will certainly collapse if no adequate measure is taken. The ordinary people should be aware that resigning from the massive production of cheap goods is in their interest. Decision makers must question the utopia of growth that has led to the depletion of nature's resources. All these goals can be actioned by education, public communication in order to have a chance to be adopted.

Also in scientific circles, the “superficial feeling of abundance” does exist in the form of abuse of computational resources. Thus, the recommendation for scientific circles would be just think twice how to solve your problem before jumping onto computationally expensive methods (heaps of unneeded data, training speculative/unneeded AI models, or other tools of high energetic costs).

Ethical issues (in IT and more generally, in Science) should be promoted. We, the academic IT community, should not accept to make reports/promote not ethical thesis or companies. Scientific world should pay attention to “green washing” which may be unintentional. So, the scientists should evaluate the usefulness of the proposed approaches. The scientists should also be independent from their employer. One of the most promising leverage mechanisms is to force people (citizens and companies) to pay the right price for computing resources, because we must take care of the common resources. Again, we are not living in a world with infinite resources!

### 6.3 Contributors / Participants

Denis Trystram, University of Grenoble Alpes
Maciej Drozdowski, Poznań University of Technology
Alix Munier-Kordon, Sorbonne University
Johnatan E. Pecero, University of Luxembourg
Andrew A Chien, University of Chicago and Argonne National Laboratory
Anne-Cécile Orgerie, CNRS & Univ of Rennes

## 7. References

[1] BOEZENNEC R., DUFOSSE F., PALLEZ G., 2023. Optimization Metrics for the Evaluation of Batch Schedulers in HPC. Job Scheduling Strategies for Parallel Processing: 26th International Workshop, JSSPP 2023.

[2] BUYYA, R., ABRAMSON, D. & GIDDY, J. Nimrod/G: An architecture for a resource management and scheduling system in a global computational grid. Proceedings Fourth International Conference/Exhibition on High Performance Computing in the Asia-Pacific Region, 2000. IEEE, 283-289.

[3] CHASAPIS, D., MORETÓ, M., SCHULZ, M., ROUNTREE, B., VALERO, M. & CASAS, M. Power efficient job scheduling by predicting the impact of processor manufacturing variability. Proceedings of the ACM International Conference on Supercomputing, 2019. 296-307.

[4] COROAMĂ, V. C., BERGMARK, P., HÖJER, M. & MALMODIN, J. 2020. A Methodology for Assessing the Environmental Effects Induced by ICT Services: Part I: Single Services. Proceedings of the 7th International Conference on ICT for Sustainability. Bristol, United Kingdom: Association for Computing Machinery.

[5] D'AMICO, M. & GONZALEZ, J. C. 2021. Energy hardware and workload aware job scheduling towards interconnected HPC environments. IEEE Transactions on Parallel and Distributed Systems.

[6] DROZDOWSKI M. Scheduling for Parallel Processing, Springer-Verlag, London, 2009.

[7] ELECTRICITY MAPS. 2022. Electricity Maps | Reduce carbon emissions with actionable electricity data [Online]. Copenhagen, Denmark: Electricity Maps. Available: https://www.electricitymaps.com/ [Accessed June 9 2023].

[8] ELELIEMY, A. & CIORBA, F. M. A Resourceful Coordination Approach for Multilevel Scheduling. 2021. Proceedings of the International Conference on High Performance Computing & Simulation (HPCS 2020), Barcelona, Spain, virtual event, March 2021.

[9] FEITELSON, D. G. & RUDOLPH, L. Toward convergence in job schedulers for parallel supercomputers. Job Scheduling Strategies for Parallel Processing: IPPS'96 Workshop Honolulu, Hawaii, April 16, 1996 Proceedings 2, 1996. Springer, 1-26.

[10] GUPTA, U., ELGAMAL, M., HILLS, G., WEI, G., LEE, H. S., BROOKS, D. & WU, C. 2022. ACT: Designing Sustainable Computer Systems with an Architectural Carbon Modeling Tool. Proceedings of the 49th Annual International Symposium on Computer Architecture. Association for Computing Machinery.

[11] HILBRICH, M., MÜLLER, S., KULAGINA, S., LAZIK, C., DE MECQUENEM, N. & GRUNSKE, L. A Consolidated View on Specification Languages for Data Analysis Workflows. 2022 Cham. Springer Nature Switzerland, 201-215.

[12] JAY, M., OSTAPENCO , V., LEFEVRE, L., TRYSTRAM, D., ORGERIE, A.-C. & FICHEL, B. 2023. An experimental comparison of software-based power meters: focus on CPU and GPU. 23rd IEEE/ACM international symposium on cluster, cloud and internet computing. Bangalore, India.

[13] KEAHEY, K., ANDERSON, J., ZHEN, Z., RITEAU, P., RUTH, P., STANZIONE, D., CEVIK, M., COLLERAN, J., GUNAWI, H. S., HAMMOCK, C., MAMBRETTI, J., BARNES, A., HALBACH, F. C., ROCHA, A. & STUBBS, J. 2020. Lessons Learned from the Chameleon Testbed. Proceedings of the 2020 USENIX Conference on Usenix Annual Technical Conference. USENIX Association.

[14] LEE, H., CALVIN, K., DASGUPTA, D., KRINNER, G., MUKHERJI, A., THORNE, P., TRISOS, C., ROMERO, J., ALDUNCE, P. & BARRETT, K. 2023. AR6 Synthesis Report: Climate Change 2023. Summary for Policymakers.

[15] LIN, L. & CHIEN, A. A. 2023. Adapting Datacenter Capacity for Greener Datacenters and Grid. Proceedings of the Fourteenth ACM International Conference on Future Energy Systems.

[16] LUDEMA, J. & NOSSOKOFF, M. 2023. Sustainability and Energy Efficiency Found to be of Strategic Importance for HPC Datacenters.

[17] MATSUOKA, S., DOMKE, J., WAHIB, M., DROZD, A., CHIEN, A. A., BAIR, R., VETTER, J. S. & SHALF, J. 2022. Preparing for the Future—Rethinking Proxy Applications. Computing in Science & Engineering, 24, 85-90.

[18] MEGOW, N., UETZ, M. & VREDEVELD, T. 2006. Models and algorithms for stochastic online scheduling. Mathematics of Operations Research, 31, 513-525.

[19] MENON, H., DIFFENDERFER, J., GEORGAKOUDIS, G., & LAGUNA, I. 2023. Approximate High-Performance Computing: A Fast and Energy-Efficient Computing Paradigm in the Post-Moore Era. IT Professional, Volume: 25, Issue: 2.

[20] MÖHRING, R. H. Scheduling under uncertainty: Optimizing against a randomizing adversary. Approximation Algorithms for Combinatorial Optimization: Third International Workshop, APPROX 2000 Saarbrücken, Germany, September 5–8, 2000 Proceedings, 2003. Springer, 15-26.

[21] ÖZDEN, T., BERINGER, T., MAZAHERI, A., FARD, H. M. & WOLF, F. 2023. ElastiSim: A Batch-System Simulator for Malleable Workloads. Proceedings of the 51st International Conference on Parallel Processing. Bordeaux, France, 2022. Association for Computing Machinery.

[22] PIRSON, T., GOLARD, L. & BOL, D. 2023. Evaluating the (ir)relevance of IoT solutions with respect to environmental limits based on LCA and backcasting studies. LIMITS'23: Workshop on Computing within Limits.

[23] PRABHAKARAN, S. 2016. Dynamic resource management and job scheduling for high performance computing. PhD Thesis, T.U. Damstadt.

[24] RADOVANOVIC, A., KONINGSTEIN, R., SCHNEIDER, I., CHEN, B., DUARTE, A., ROY, B., XIAO, D., HARIDASAN, M., HUNG, P., CARE, N., TALUKDAR, S., MULLEN, E., SMITH, K., COTTMAN, M. & CIRNE, W. 2022. Carbon-Aware Computing for Datacenters. IEEE Transactions on Power Systems, 38, 1270--1280.

[25] SOMMER, L. 2022. California just ran on 100% renewable energy, but fossil fuels aren't fading away yet. Oregon Public Broadcasting. https://www.opb.org/article/2022/05/13/california-renewable-energy-fossil-fuels.

[26] TIRMAZI, M., BARKER, A., DENG, N., HAQUE, M. E., QIN, Z. G., HAND, S., HARCHOL-BALTER, M. & WILKES, J. 2020. Borg: the next generation. EuroSys '20. Heraklion, Crete.

[27] YOO, A. B., JETTE, M. A. & GRONDONA, M. SLURM: Simple Linux Utility for Resource Management. Job Scheduling Strategies for Parallel Processing, 2003.

[28] ZHANG, C. & CHIEN, A. A. 2021. Scheduling Challenges for Variable Capacity Resources. Job Scheduling Strategies for Parallel Processing: 24th International Workshop, JSSPP 2021, Virtual Event, May 21, 2021, Revised Selected Papers. Springer-Verlag.