Job scheduling with jobs' energy profiles

Danilo Carastan-Santos

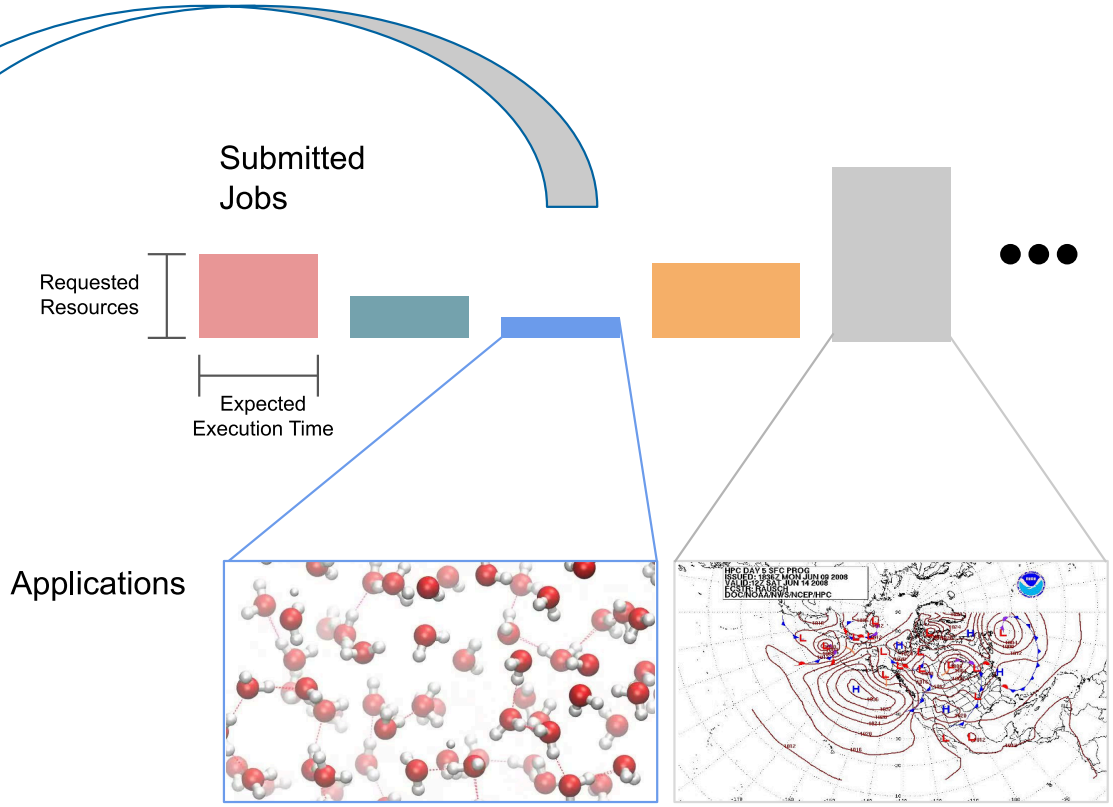
[1]Université Grenoble Alpes, Grenoble INP, Inria, LIG, France
email:danilo.carastan-dos-santos@inria.fr

April 11, 2023

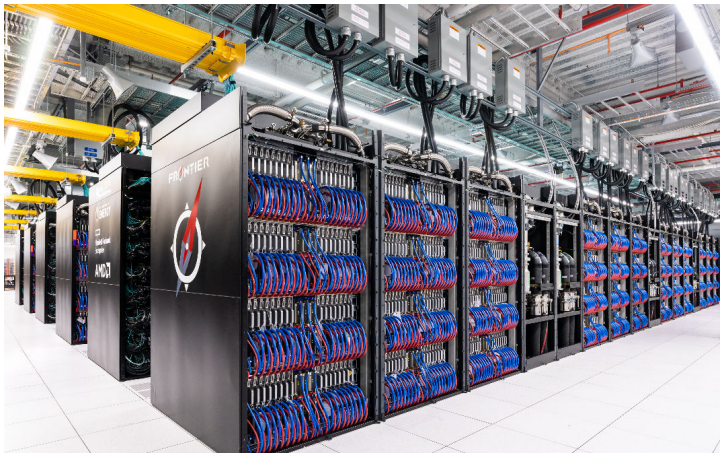
Outline

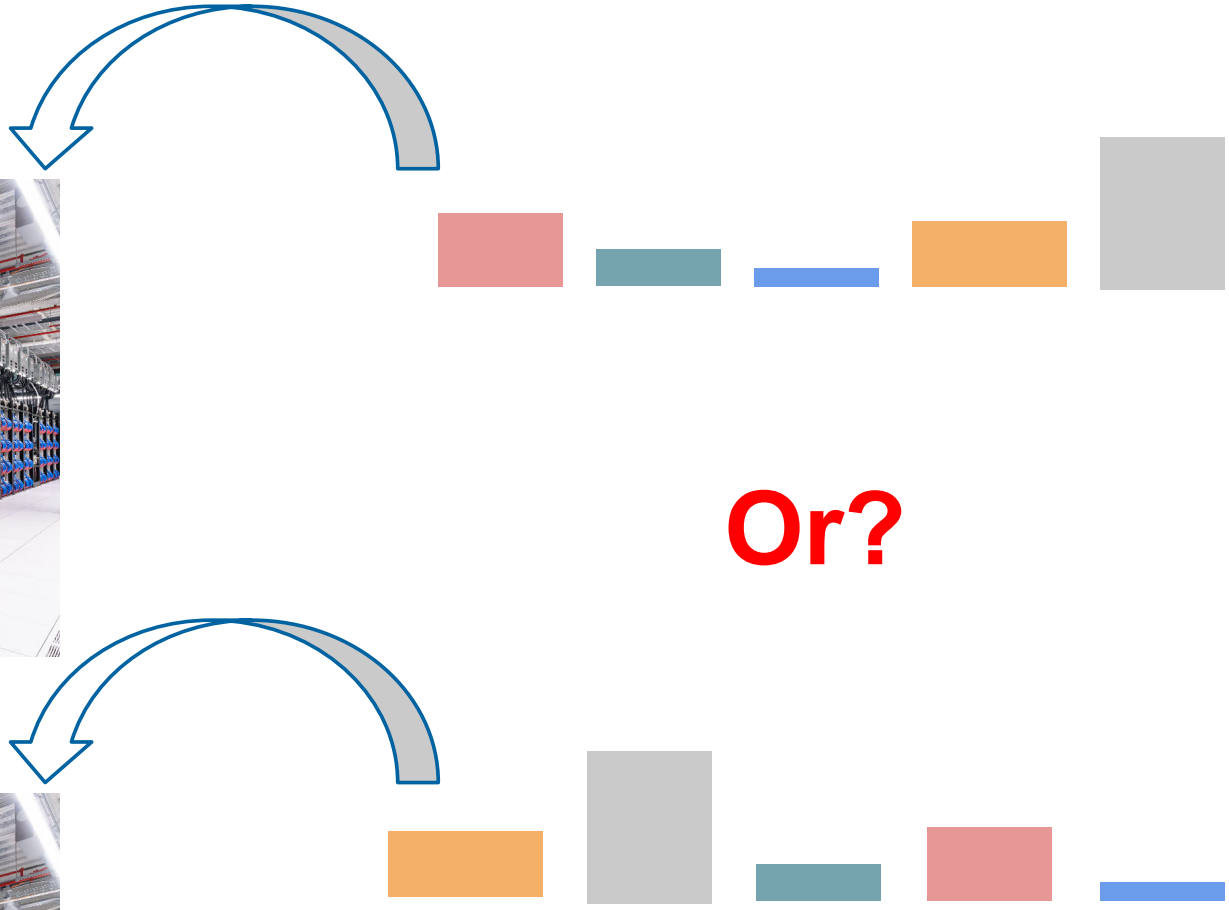
1. Overview of the problem
   - High-Performance Computing (HPC) resource management
   - HPC job scheduling
2. Monitoring/gathering jobs' energy consumption
   - Monitoring tools (wattmeters, RAPL)
   - A use case
3. Job scheduling with energy information
   - Research challenges/perspectives

Overview of the problem



Overview of the problem





Overview of the problem

- Add a new jobs' data: **Jobs' power consumption**
- Add new objectives:
  - **Respect a platform power cap**
  - **As low power as possible**

Our use-case

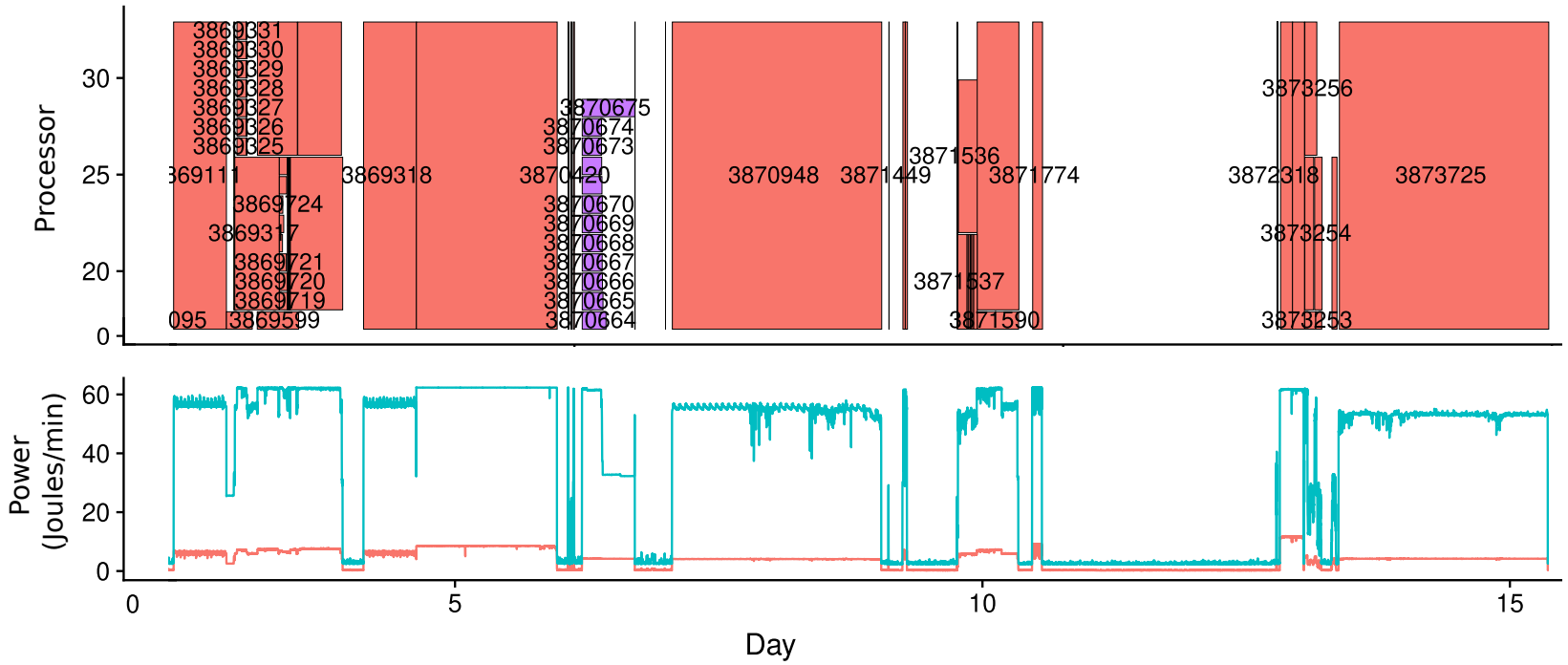
- Gricad[a] large-scale computing platform
- **Dahu** cluster (Grenoble site)
  - Each node: dual-socket Intel Xeon Gold 6130 (16 physical cores, 32 virtual)
  - Nodes' energy data collected with **Colmet**[b](Oar-team in Grenoble)

[a] https://gricad.univ-grenoble-alpes.fr/

[b] https://github.com/oar-team/colmet

Monitoring the energy consumption of the Dahu Cluster[2]

- Two sources of data:
  - **Jobs** (OAR, upper graph): processing time and number of processors
  - **Energy consumption** (Colmet, lower graph)



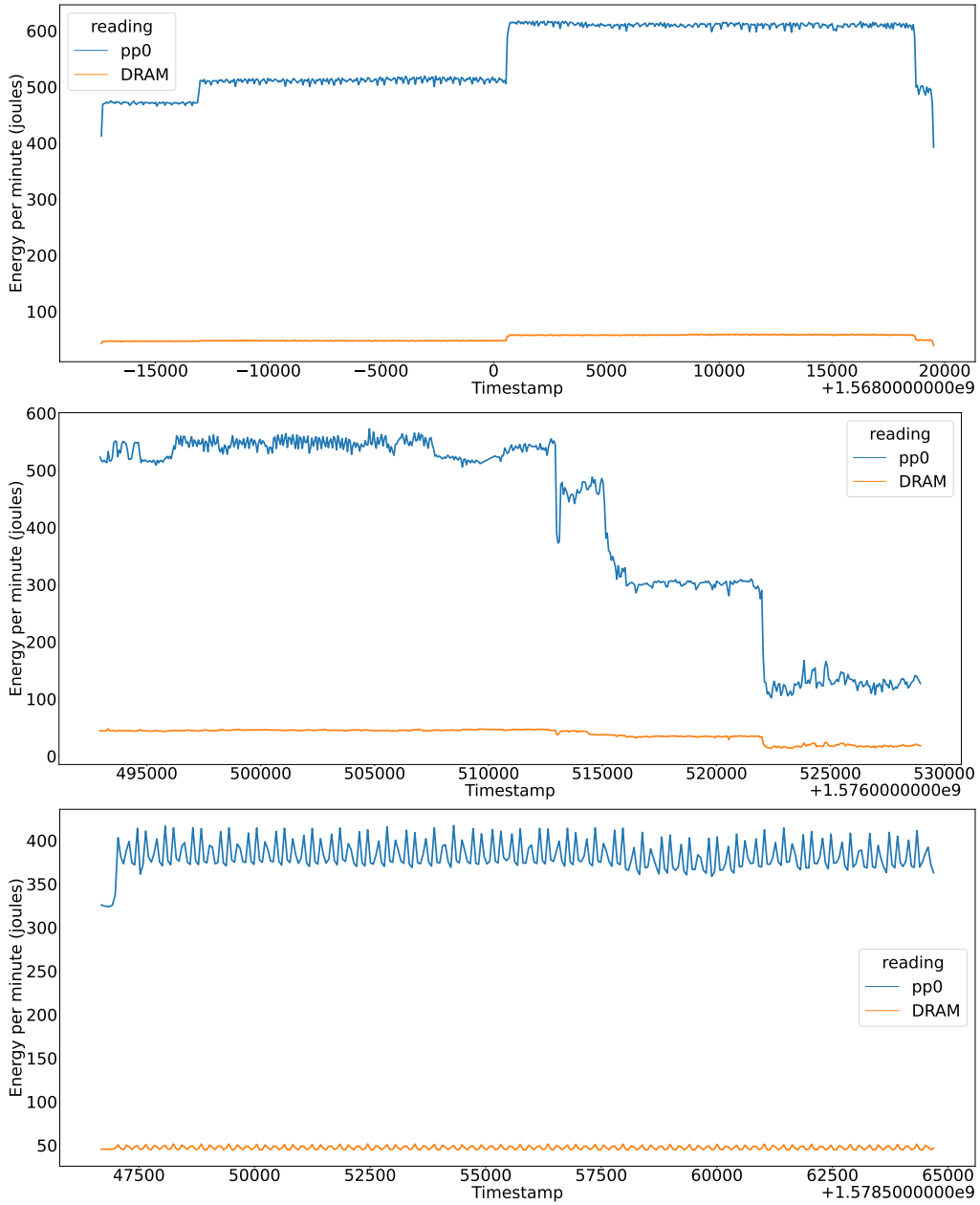
Problems: Jobs that share a node, jobs that run in multiple nodes, incomplete energy traces, container jobs[1]

[1] Jobs that host other jobs inside. This is a standard OAR feature

[2] Example illustrating a single socket of a Dahu node, with hyperthreading enabled.

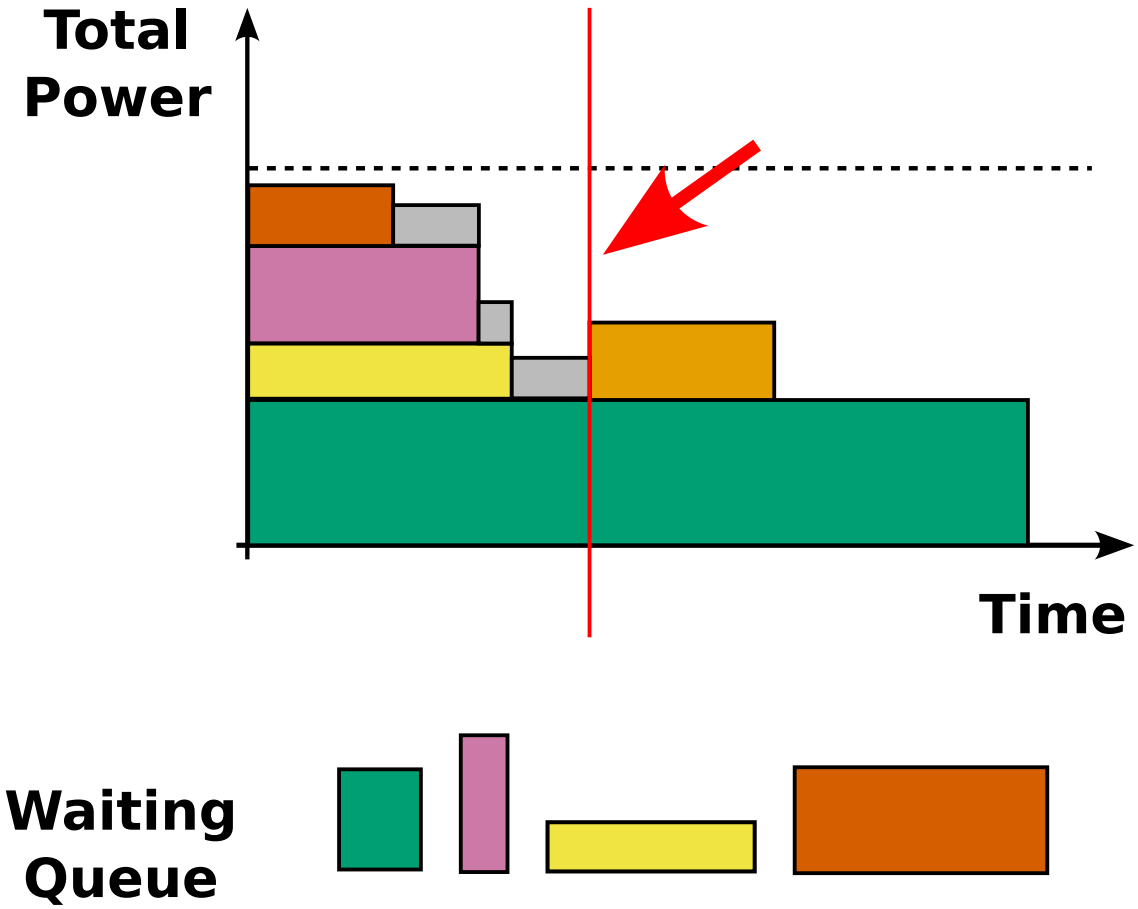
Some job energy profiles

- **We can see what's inside the "job box"** regarding the energy consumption
- **We want to integrate** this in the RJMS to do online **scheduling decisions**
- This requires **predicting** the jobs' energy profile
  - Mean/Median/Max: "easy"
  - Full profile: "complicated" (this is why it is interesting)



Job scheduling with energy information

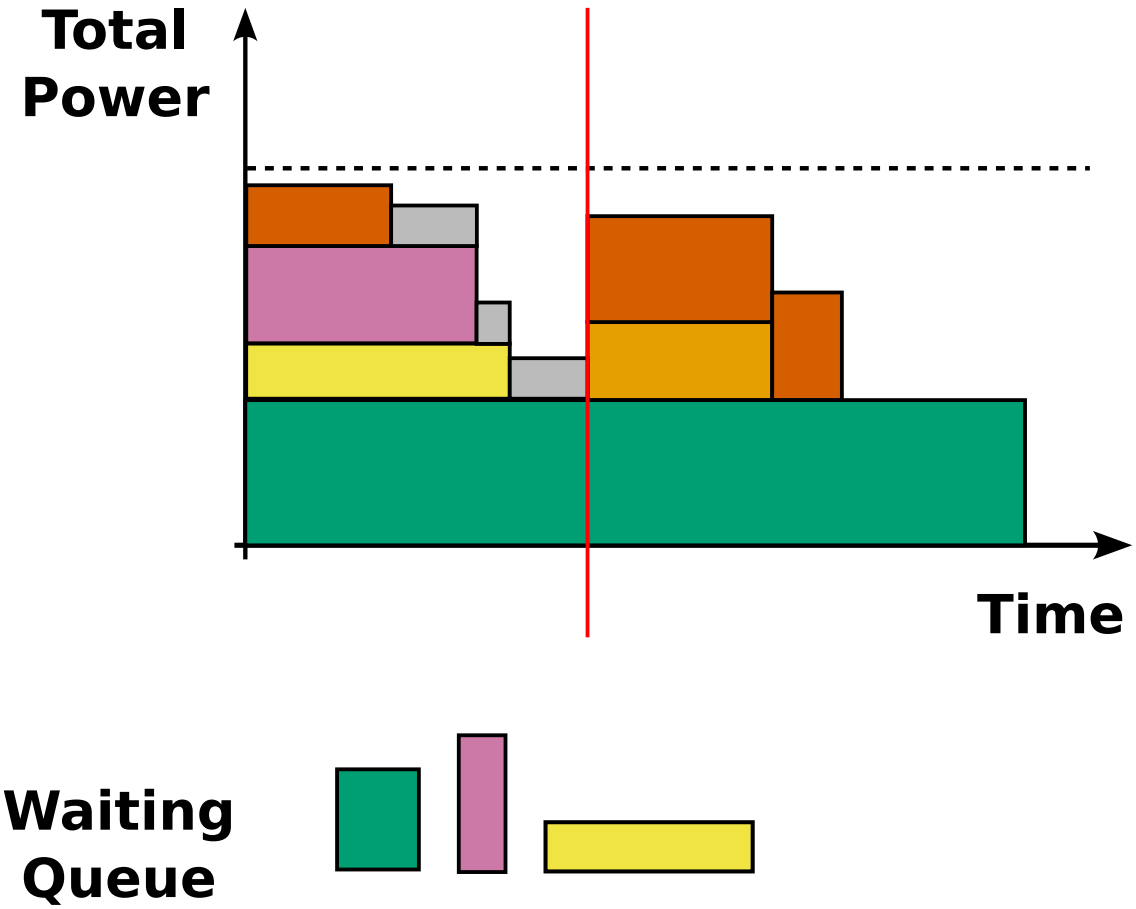
An example with maximum job power consumption



- Gantt Chart **Power** *versus* **time**
- Rectangles: Predictions of the **maximum** job power consumption (height)
- **Which job to choose without passing the power cap?**

Job scheduling with energy information

An example with maximum job power consumption

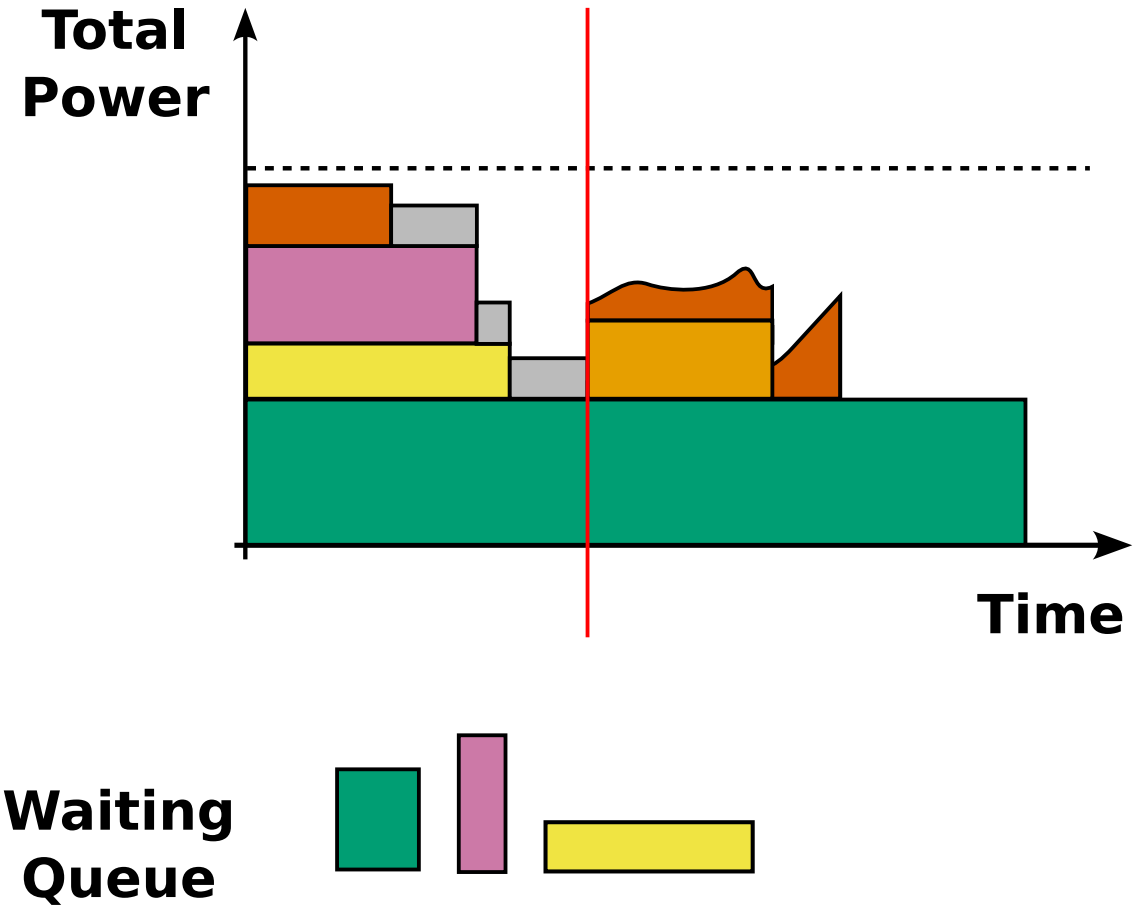


- Gantt Chart **Power** *versus* **time**
- Rectangles: Predictions of the **maximum** job power consumption (height)
- SoA: Best-fit[a]
- Better ways to choose jobs? e.g., Knapsack

[a] Ryuichi Sakamoto et al. "Analyzing resource trade-offs in hardware overprovisioned supercomputers". In: *2018 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. IEEE. 2018, pp. 526–535.

Job scheduling with energy information

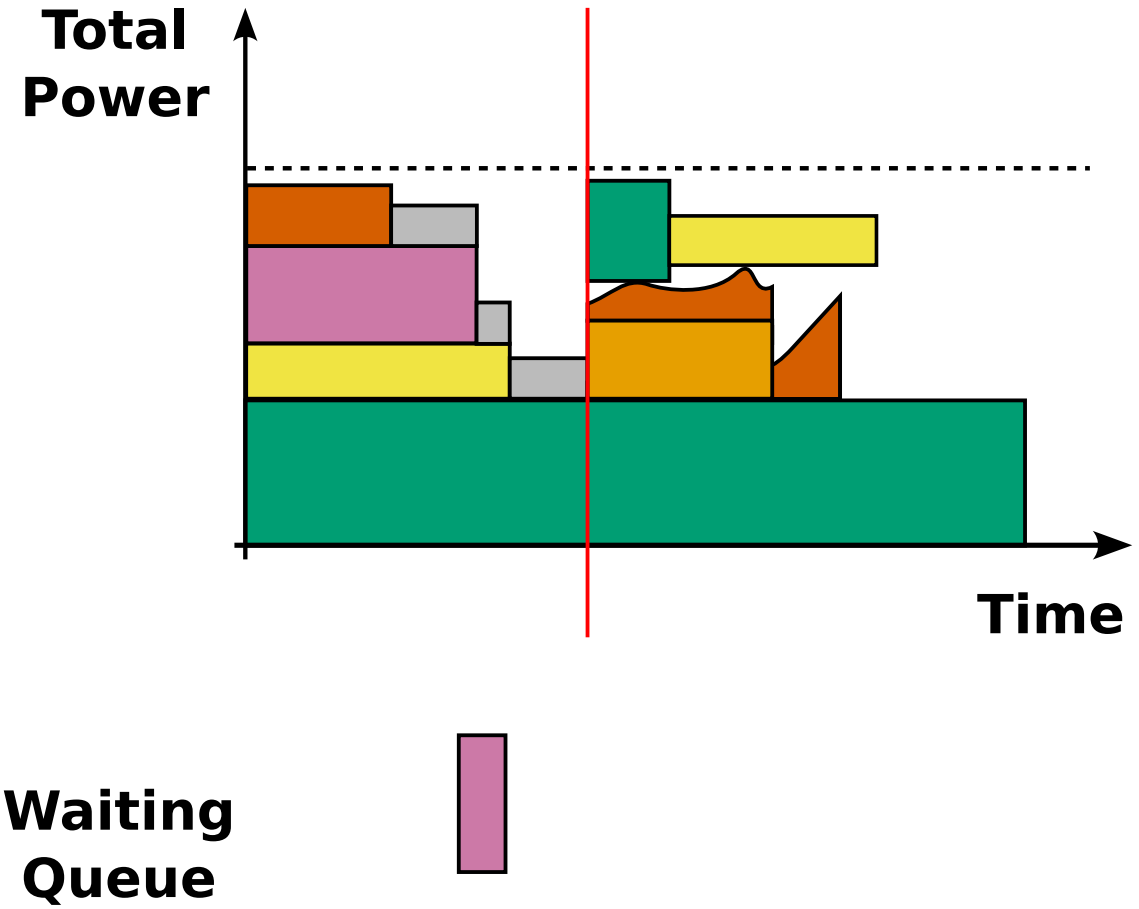
A more sophisticated approach



- Maximum $\rightarrow$ **"bounding box"** of the jobs' energy consumption
- We could still **do more** with the given cap

Job scheduling with energy information

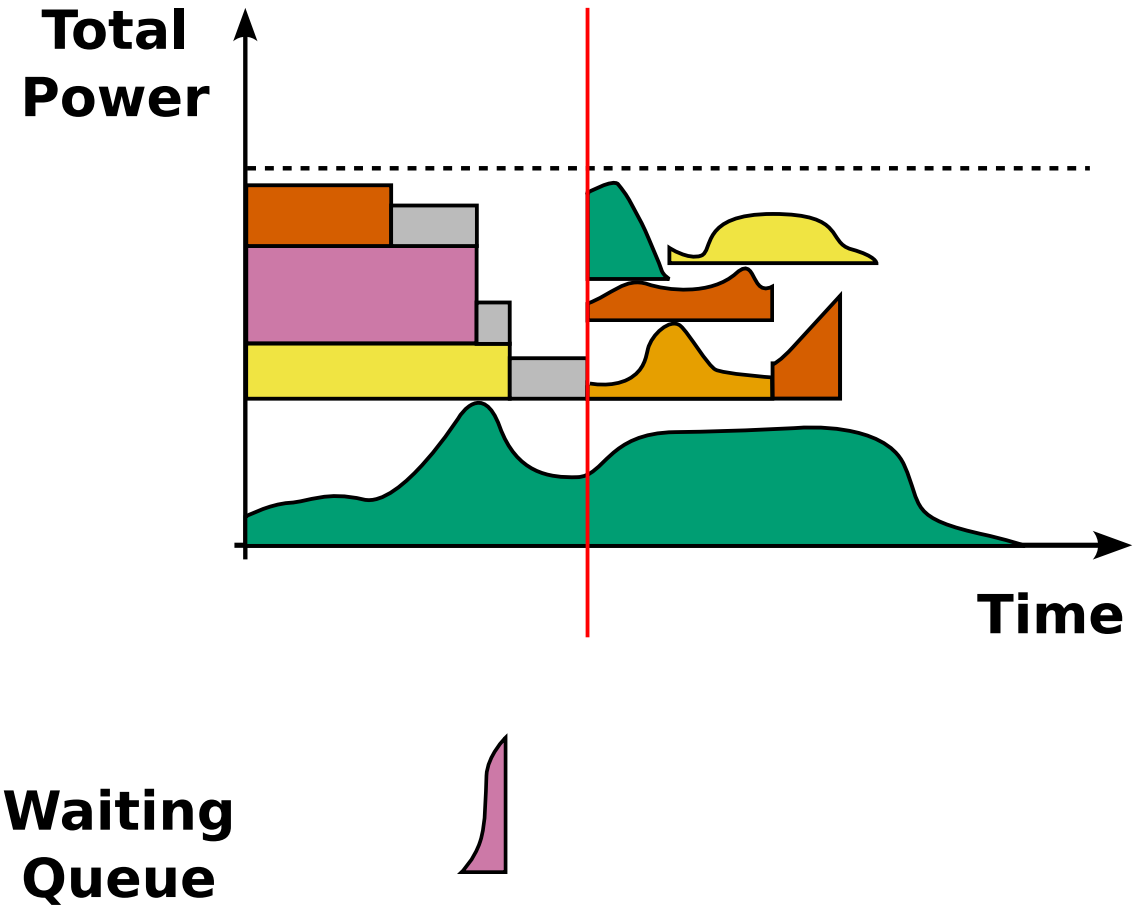
A more sophisticated approach



- Maximum is just a **"bounding box"** of the jobs' energy consumption
- We could still **do more** with the given cap

Job scheduling with energy information

A more sophisticated approach



- Each job has its own energy profile
- We can potentially **do more** with the given power cap
- **Questions**
- Predict the profile before job execution?
- Forecast the energy profile on the fly?
- Choose the appropriate jobs?

Job scheduling with energy information

To conclude

- There is room for improving power-capped HPC platforms if we have information about the **jobs' energy profile**

Job scheduling with energy information

To conclude

- There is room for improving power-capped HPC platforms if we have information about the **jobs' energy profile**
- Many **challenges** need to be addressed

Job scheduling with energy information

To conclude

- There is room for improving power-capped HPC platforms if we have information about the **jobs' energy profile**
- Many **challenges** need to be addressed
  - HPC workload data with energy information
  - Jobs energy profile predictions
  - Efficient scheduling methods

Job scheduling with energy information

To conclude

- There is room for improving power-capped HPC platforms if we have information about the **jobs' energy profile**
- Many **challenges** need to be addressed
  - HPC workload data with energy information
  - Jobs energy profile predictions
  - Efficient scheduling methods
  - **All of the above remaining frugal (lightweight)**

Contact

- Email: danilo.carastan-dos-santos@inria.fr
- Website: https://danilo-carastan-santos.github.io/ **(QR code on the right)**