Placing leverages on Clouds for footprint reduction

Thomas Stavis[1, 2], Laurent Lefèvre[1], Anne-Cécile Orgerie[2]

[1]Univ Lyon1, Inria, LIP, ENS Lyon, CNRS. Lyon, France.
[2]Univ Rennes, Inria, IRISA, CNRS. Rennes, France.

CONTEXT

- In 2018, data centers consumed about 1% of global energy *[1]* and this consumption could increase up to 4% in 2030 *[2]*.
- Resources consumption (water, rare metals and earths) and greenhouse gas (GHG) emissions follow this evolution, especially in France *[3]*.
- Techniques, called "leverages", can help reduce these footprints.
- Even if some leverages are well-studied *[4, 5]*, there is actually no methodology for using a large set of leverages at the same time.

CHALLENGE

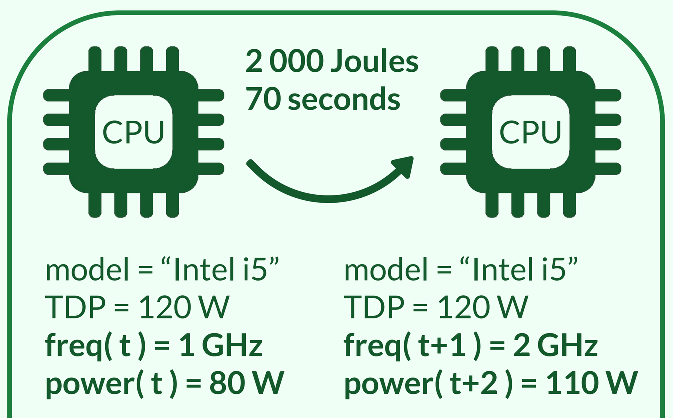
Propose a methodology for:
- placing numerous and heterogeneous leverages
- on large-scale data centers
- to satisfy objectives of footprint reduction

This work follows Vladimir Ostapenco's thesis *[6]*.

PROBLEM MODELING

Leverage

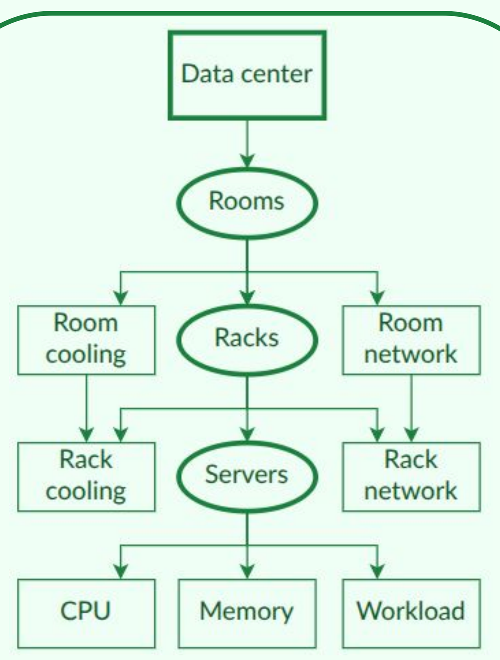
- Space of parameters
  - time
  - component
  - intensity
- Changes
  - in component states
  - in architecture
- Costs
  - additional footprints
  - time before / after application



Example of the application of a leverage on a CPU

Architecture (DAG)

- Components
  - states: time dependant
  - features: time **in**dependant
- Links of the Directed Acyclic Graph
  - notion of ownership
    U → *V* ⇔ *U* can modify *V*
    but *V* can't modify *U*



Example of a Cloud architecture

Objectives

- Optimization functions
  (ex: minimize rack's power)
- Constraints on costs, footprints, components or leverages
  (ex: exclusion leverages A & B, max power = X Watts)
- Strategies for resolution
  (ex: choice of algorithm, heuristics given by experts, optional constraints, time granularity)

The tuple (Leverages, Architecture, Objectives) is called a Scenario.

Placement problem

For an large architecture and a large set of leverages, find a set of leverage placements such that, when applied, it satisfies constraints and optimality, according to specific strategies.

METHODOLOGY VALIDATION

For this scenario of instantaneous power capping, we solve the placement problem and compare different strategies to find out which one is better.

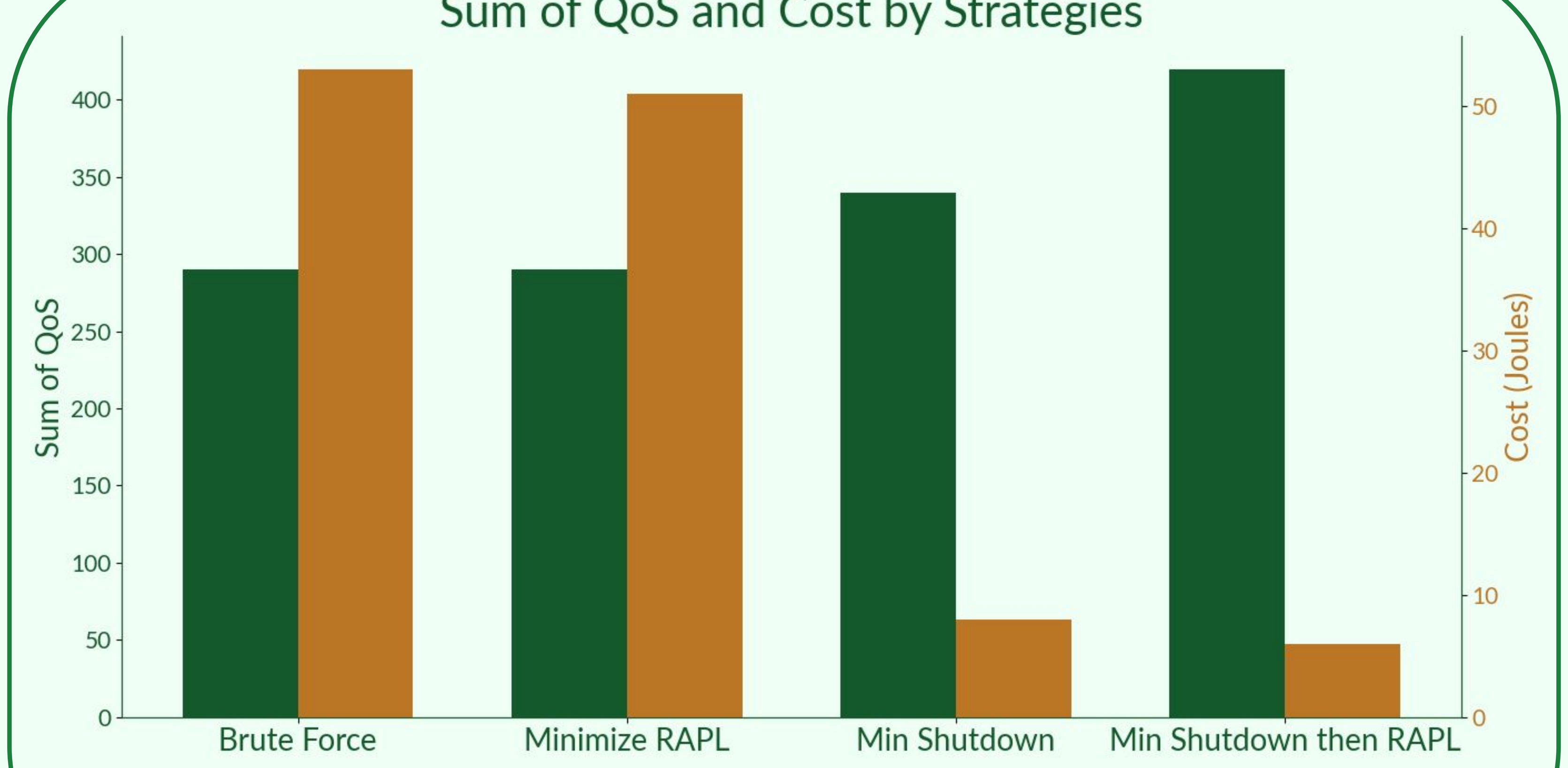
Leverages

- RAPL(*p*) on *CPU [7]*
  - limits power consumption to *p* Watts
  - affects linearly QoS and Power
  - costs 1 Joule
- Shutdown on *Server [8]*
  - turns off server
  - highly decreases QoS and sets Power to 0
  - costs 50 Joules

Objectives

- Optimization function
  - maximizing the Sum of QoS
- Constraints
  - Shutdown & RAPL are exclusive
  - architecture power must be below 45% of its maximum
- The Quality of Service (QoS) is a scalar between -100 and 100
- $p \in [20, 120]$ with a step of 10 W

Architecture One rack with two servers. First includes four "Intel X" CPUs and second includes four "Intel Y" CPUs. Both servers and the rack consume little inherent power.

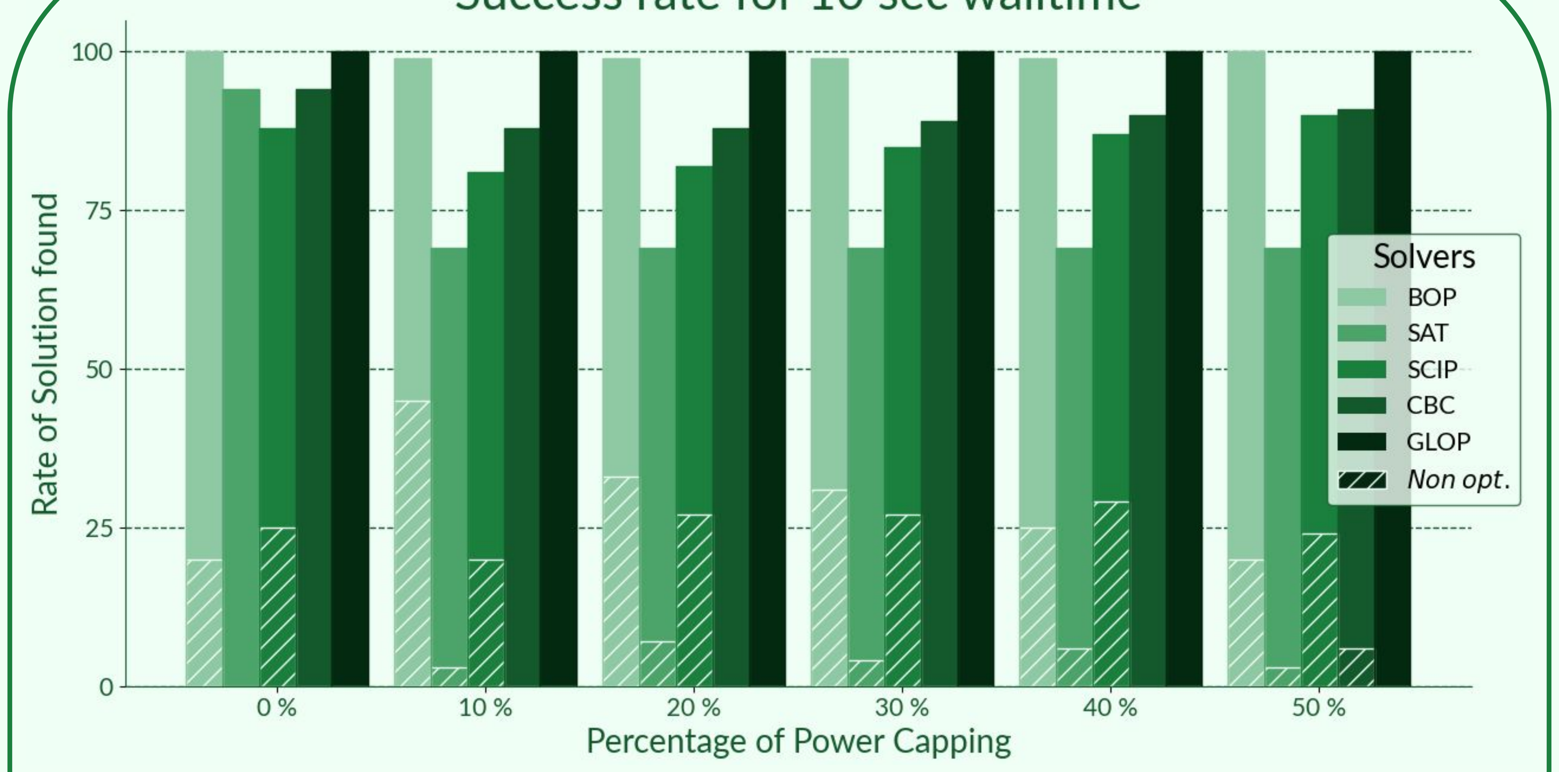


Comparison of four different strategies: the two left cost a lot and have reasonable Quality of Service, whereas the two right cost less and have better QoS

LARGE-SCALE EXPERIMENTATION

- The space of solutions is **exponential**. Without simplifying assumptions, browsing all solutions to find a good one is very ineffective at large scale.
- To tackle this, we delegate the browsing to **constraint solvers** from Google's OR-Tools library *[9]*.
- We made an experiment to determine the efficiency of solvers on different scales of instantaneous power capping scenarios.

We explore scales by varying the number of components, the number of leverage parameters, the power limit, and the walltime.



Experiment made with 10 repetitions of each case, on Montcalm's cluster from Grid'5000 [10], simulating a power capping on a tree with 10 to 10 000 CPU nodes, and RAPL(p) leverage with p ∈ [20, 120] with a step of 1,25 to 10 W

- Solvers SAT, SCIP, and CBC have non-negligible failure rates.
- BOP achieves nearly 100% success but produces a high percentage of non-optimal solutions.
- **GLOP consistently finds 100% optimal solutions.**

References

[1] Masanet et al., "Recalibrating global data center energy-use estimates", 2020
[2] IEA, "Data Centre Energy Use: Critical Review of Models and Results", 2025
[3] ARCEP, "Pour un numérique soutenable", 2025
[4] Raïs, "Discover, model and combine energy leverages for large scale energy efficient infrastructures", 2018
[5] Beloglazov et al., "OpenStack Neat: a framework for dynamic and energy-efficient consolidation of virtual machines in OpenStack clouds", 2015.
[6] Ostapenco et al., "Modeling, evaluating, and orchestrating heterogeneous environmental leverages for large-scale data center management", 2023.
[7] Ostapenco et al., "Exploring RAPL as a Power Capping Leverage for Power-Constrained Infrastructures," 2024.
[8] Benoit et al., "Reducing the energy consumption of large scale computing systems through combined shutdown policies with multiple constraints", 2018
[9] P. Laurent, F. Vincent, "Or-tools framework v9.8", google [Online]
[10] Grid'5000, "Toulouse: Hardware – Grid'5000", 2024 [Online accessed December-2024]

Contact

- Avalon Team, LIP, ENSL, Lyon, France
- Magellan Team, IRISA, Rennes, France
- Mails: thomas.stavis@ens-lyon.fr, laurent.lefevre@ens-lyon.fr, anne-cecile.orgerie@irisa.fr