Adaptive Resource and Job Management for limited power consumption

23/09/14

Yiannis Georgiou
David Glesser
Matthieu Hautreux
Denis Trystram

- Introduction
- DVFS & Switch-off

- The model
- Algorithm and implementation
- Experimentations

- Conclusion and future works

- **Introduction**
- **DVFS & Switch-off**

- **The model**
- **Algorithm and implementation**
- **Experimentations**

- **Conclusion and future works**

Powercap: limit the power consumption during a certain amount of time

- Why control?
  - Power peak = O(power of a city)
  - Power installations lifetime
  - Electricity providers limitations
  - Controling energy consumption = Controling cost

- How control?
  - DVFS
  - Switch-off
    - (or shutdown, or sleep mode, or hibernation...)

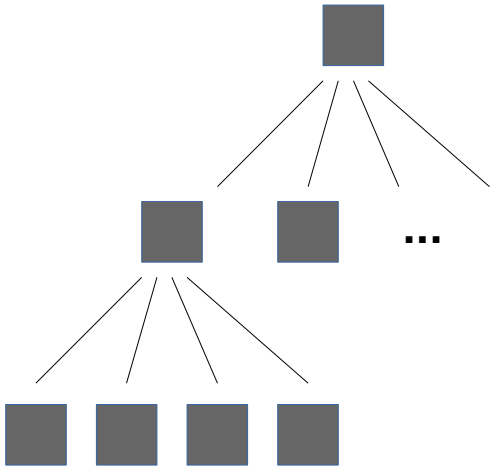
- Introduction
- **DVFS & Switch-off**

- The model
- Algorithm and implementation
- Experimentations

- Conclusion and future works

- Switch-off
  - Switch-off some resources
  - switched-off has a cost
  - Not possible on all clusters
  - Jobs can not run on switched-off nodes!

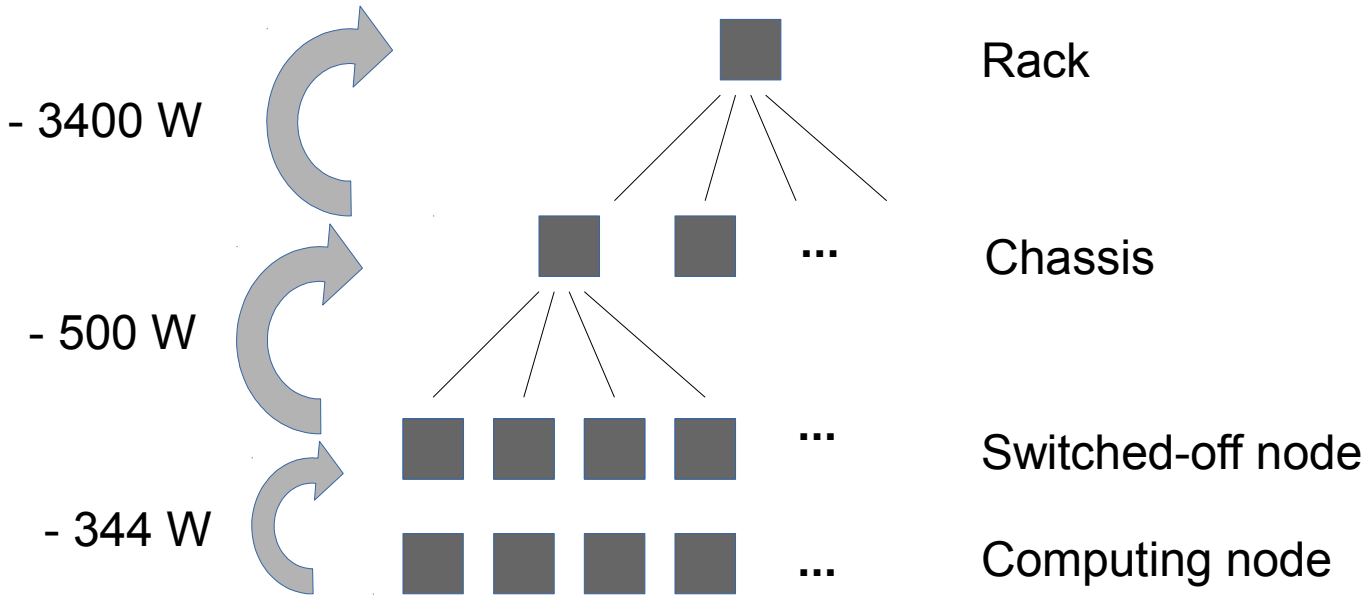
- « Power Bonuses »
  - If all components of a level are switched-off, the component of the upper level can be switched-off and provide an additional gain
  - Exemples :
  - Nodes are made of processors
  - Chassis are made of nodes
  - Rack are made of Chassis



- « Power Bonuses » on CURIE cluster:
  - 18 nodes per chassis, 5 chassis per rack
  - Power gained by switching off a **Chassis**

    ~= Power(**computing node**)
  - Power gained by switching off a **Rack**

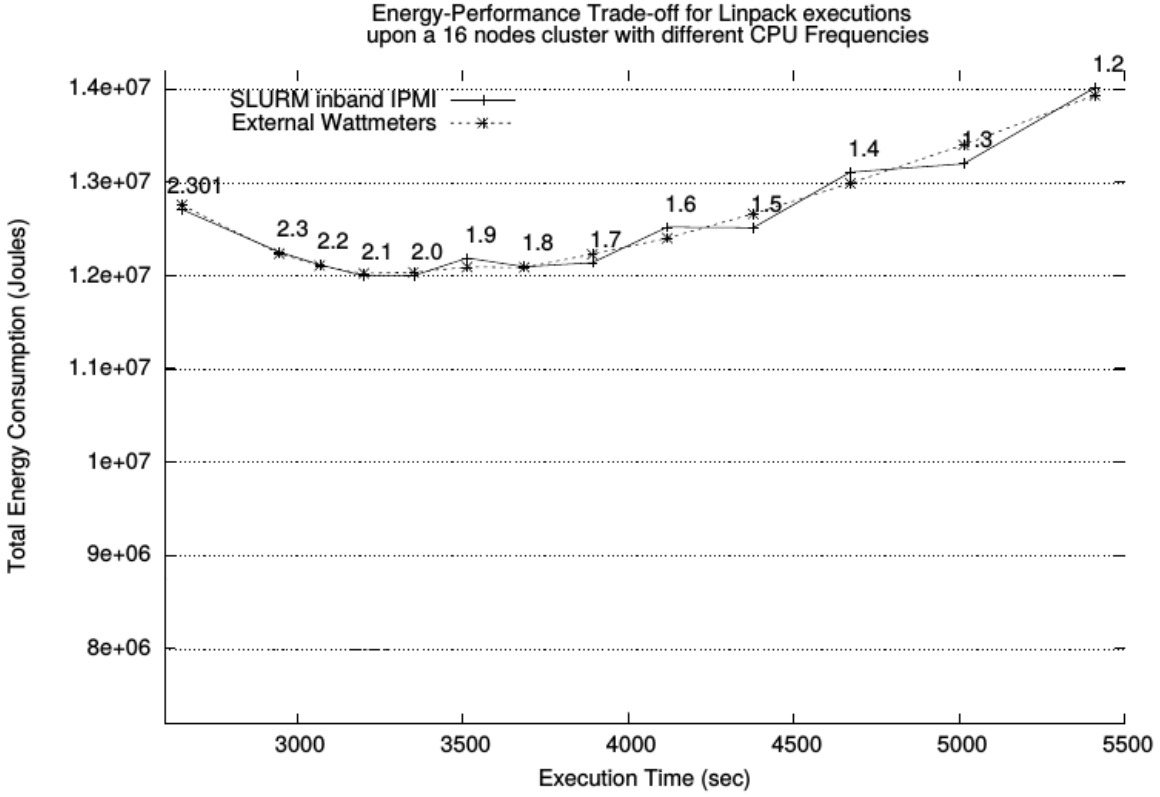
    ~= **10** * Power(**computing node**)



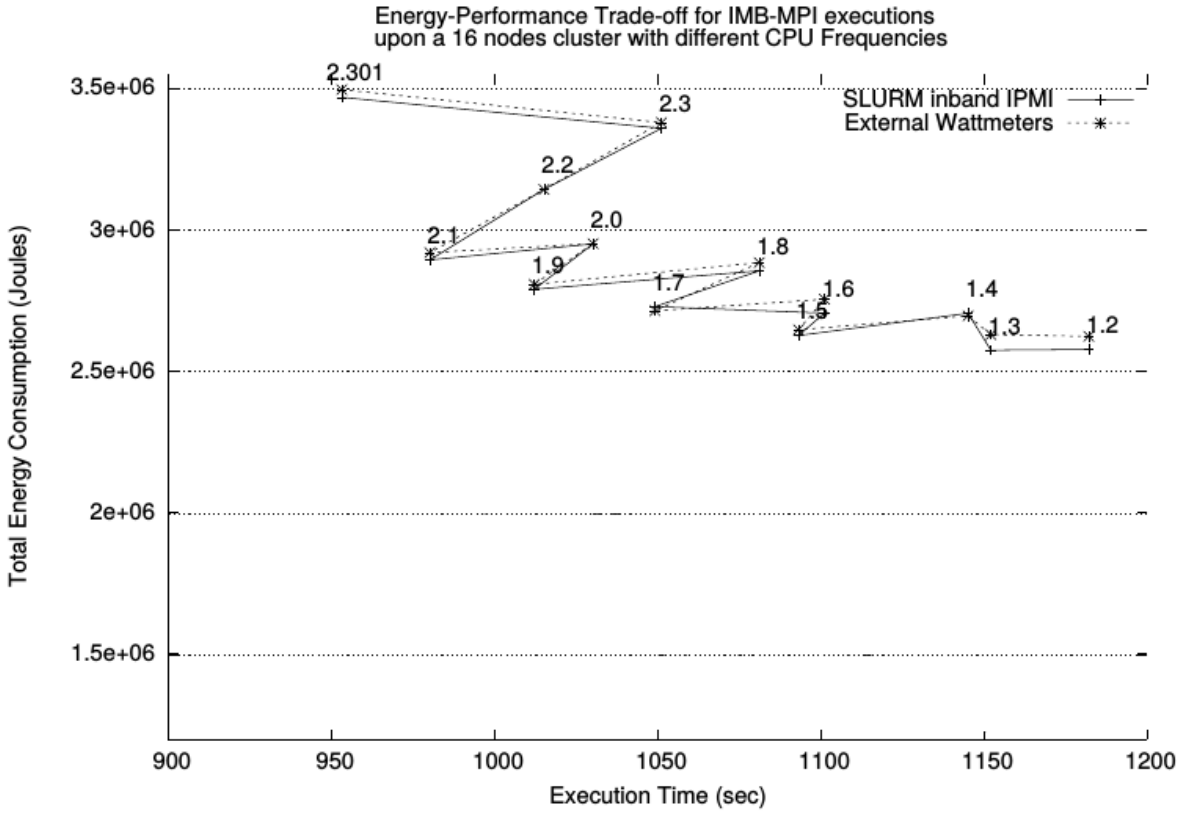
- DVFS
  - It's a trade-off between **performance** and **power consumption**
  - What about **performance** / **energy** trade-off ?

$$\int POWER \,.\, dt = Energy$$

- DVFS
  - It's a trade-off between **performance** and **power consumption**
  - What about **performance** / **energy** trade-off ?



- DVFS
  - It's a trade-off between **performance** and **power consumption**
  - What about **performance** / **energy** trade-off?



- DVFS is a trade-off between **completion time** and **power**

- No obvious **performance** / **energy** trade-off
  - Minimizing power != minimizing energy
  - The impact of DVFS is highly dependant on the job

- Introduction
- DVFS & Switch-off

- **The model**
- Algorithm and implementation
- Experimentations

- Conclusion and future works

- We work with maximum power consumptions
- W is the maximal computational work possible

$$W = T.(\frac{N - N_{off} - N_{dvfs}}{\sigma_{Max}} + \frac{N_{dvfs}}{\sigma_{Min}})$$

- Powercap limitation

$$N_{off}.P_{off} + N_{dvfs}.P_{Min} + (N - N_{off} - N_{dvfs}).P_{Max} \leq P$$

$N_X = number\ of\ node\ in\ state\ X$

$\Sigma_Z = speed\ degradation\ at\ state\ Z$

$P_Y = power\ consumption\ at\ Y$

$P = powercap$

- In the space 3D (N_{dvfs}, N_{off}, W)

$$W = T.(\frac{N - N_{off} - N_{dvfs}}{\sigma_{Max}} + \frac{N_{dvfs}}{\sigma_{Min}})$$ is a plane

$$N_{off}.P_{off} + N_{dvfs}.P_{Min} + (N - N_{off} - N_{dvfs}).P_{Max} \leq P$$ is an half space

⇒ The intersection is a straight line

- Within the bound of the total number of nodes, W is maximized when:

$$\begin{cases} N_{off} = \frac{P - N.P_{Max}}{P_{off} - P_{Max}} \\ N_{dvfs} = 0 \end{cases} \quad \text{or} \quad \begin{cases} N_{off} = 0 \\ N_{dvfs} = \frac{P - N.P_{Max}}{P_{Min} - P_{Max}} \end{cases}$$

- 3 cases:
  - DVFS is better ⇒ we only use DVFS
  - Switch-off is better ⇒ we only use Switch-off
  - The powercap is so low that we should use both

$$\begin{cases} N_{off} = \frac{P - N.P_{Max}}{P_{off} - P_{Max}} \\ N_{dvfs} = 0 \end{cases} \quad \text{or} \quad \begin{cases} N_{off} = 0 \\ N_{dvfs} = \frac{P - N.P_{Max}}{P_{Min} - P_{Max}} \end{cases}$$

How to choose ?

$$\rho = 1 - \frac{\sigma_{Max}}{\sigma_{Min}} - \frac{P_{Max} - P_{dvfs}}{P_{max} - P_{off}}$$

When ρ <0, switch-off is prefered

- On CURIE cluster:

Benchmark	Degradation	ρ	Best mechanism
NA	2.27	0	-
linpack	2.14	-0.027	Switch-off
IMB	2.13	-0.029	Switch-off
SPEC Float [11]	1.89	-0.088	Switch-off
SPEC Integer [11]	1.74	-0.134	Switch-off
Common value [22]	1.63	-0.174	Switch-off
NAS suite [11]	1.5	-0.225	Switch-off
STREAM	1.26	-0.350	Switch-off
GROMACS	1.16	-0.422	Switch-off

Fig. 5: Comparison between DVFS and switch-off in Curie for various benchmarks.

- Introduction
- DVFS & Switch-off

- The model
- **Algorithm and implementation**
- Experimentations

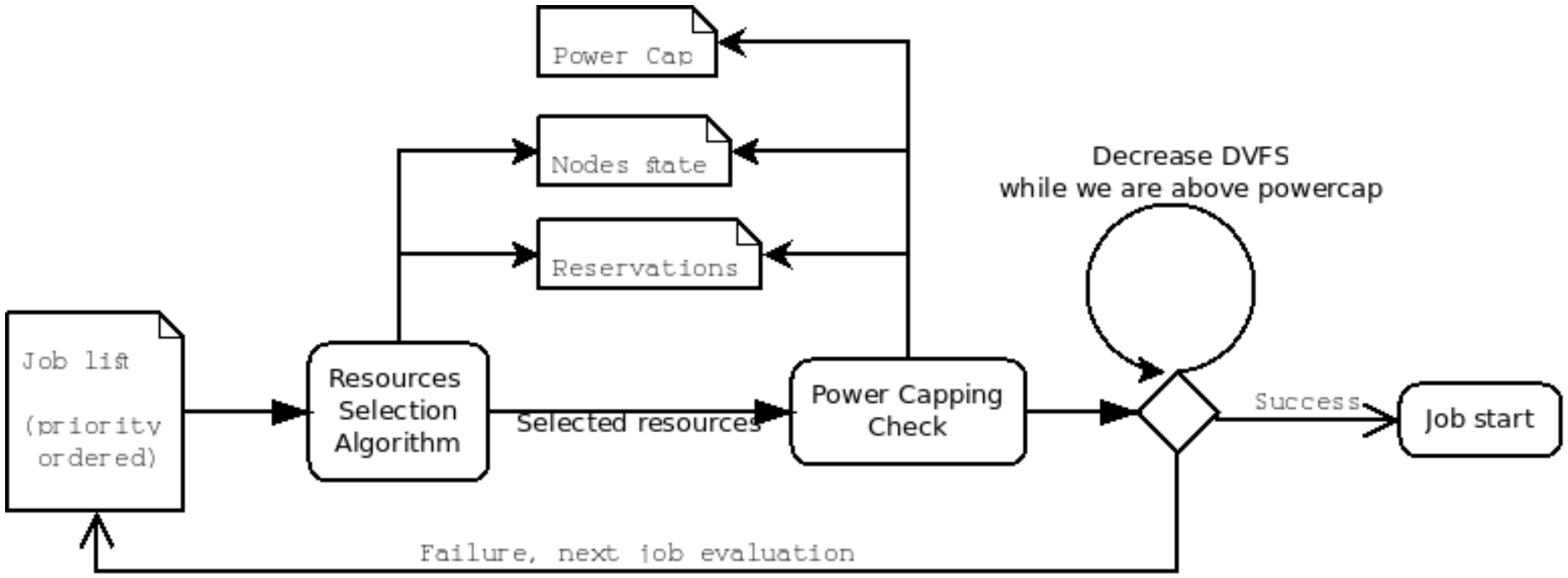
- Conclusion and future works

- When a powercap limit is set
- Choose between DVFS and switch-off

- If DVFS
  - When a job is being launched,
  - Try to schedule it at the highest frequency

- If switch-off
  - switch-off nodes at runtime,
  - mark these nodes as « reserved » for the scheduler

$ scontrol create res Watts=123151 ...



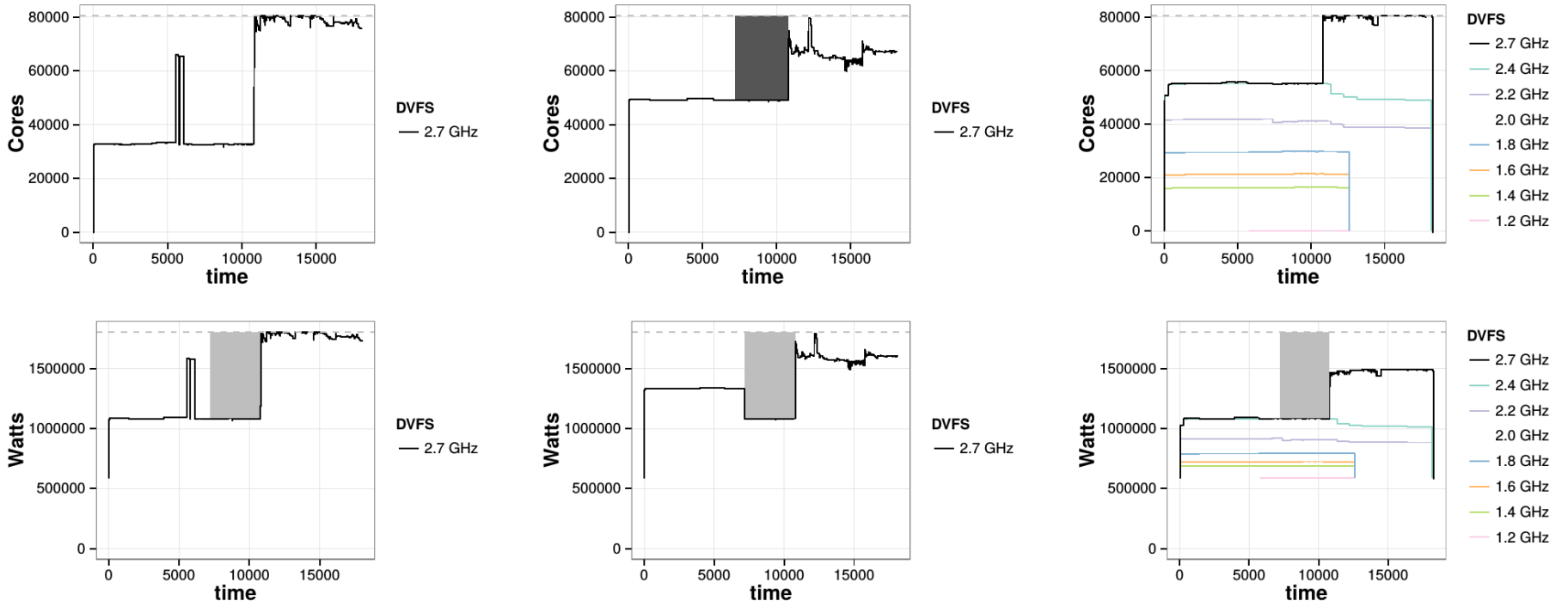
- Introduction
- DVFS & Switch-off

- The model
- Algorithm and implementation
- **Experimentations**

- Conclusion and future works

- Replay interesting parts of the CURIE workload
  - 5 hours, high utilization, jobs representative of the whole workload
- Slurm can emulate his environement
  - 336 Slurm nodes on 1 physical node
  - *Sleep* instead of real computational jobs
- Add a powercap
  - Case study: 1 hour, in the middle of the trace, at different powers

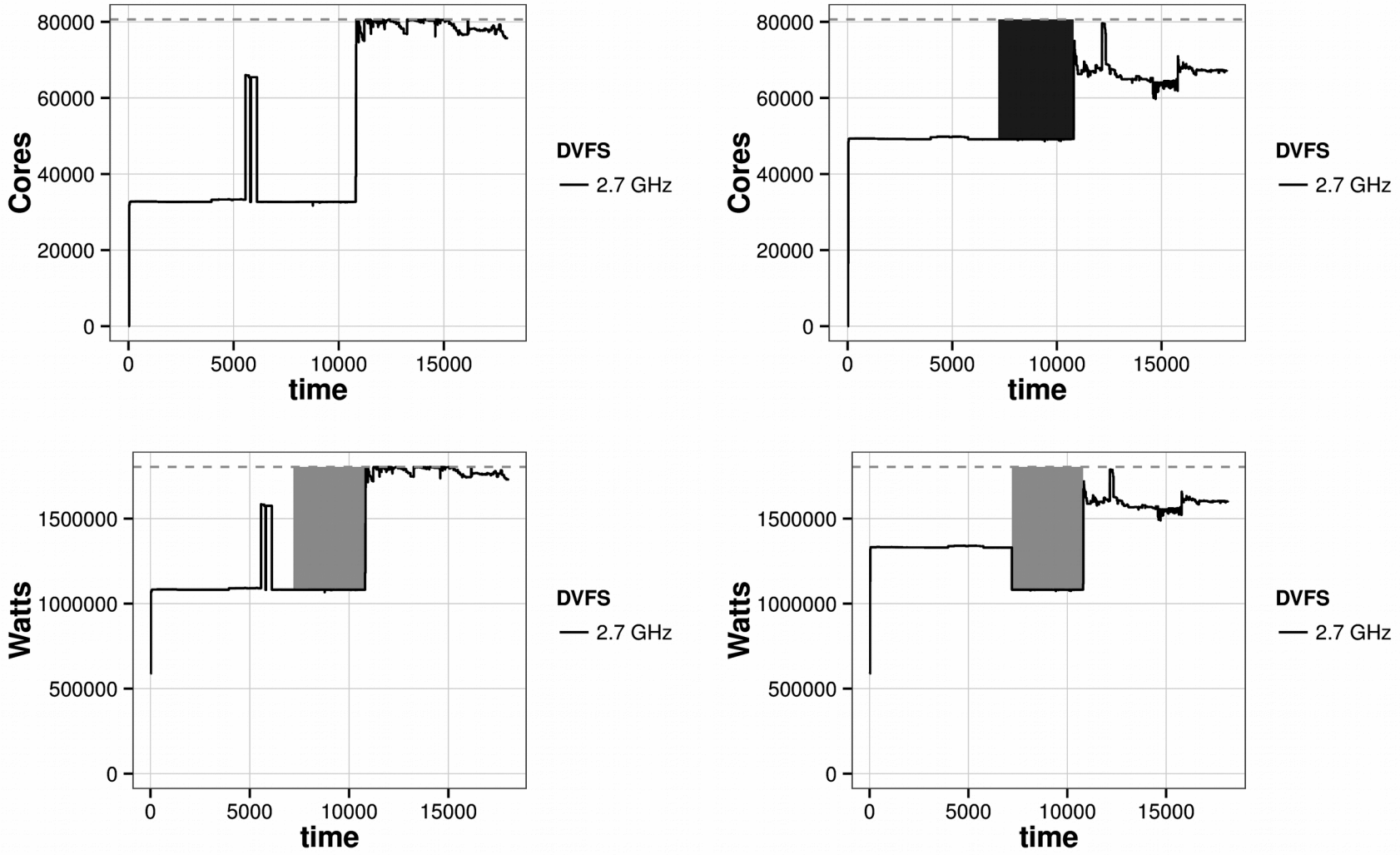
Experimental validation



(a) IDLE (b) SHUT (c) DVFS

Fig. 7: System utilization for the IDLE, DVFS and SHUT policies in terms of cores (up) and power (bottom) during the 5 hours workload with a reservation of 60% of total powercap

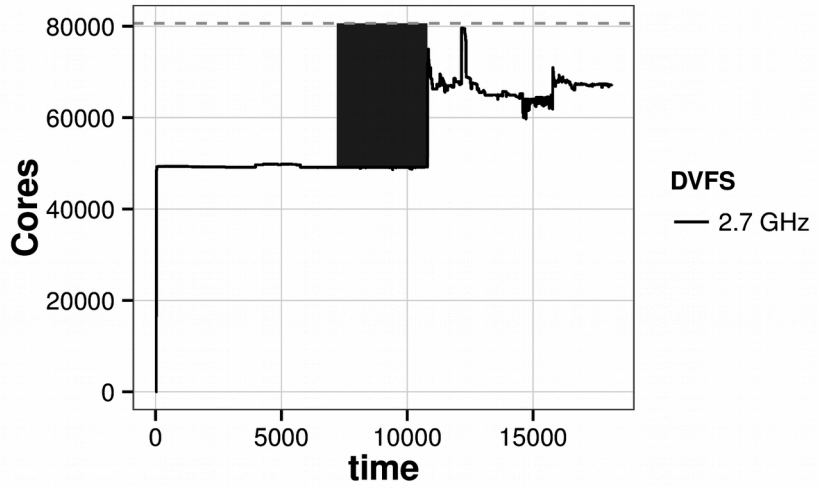
Experimental validation

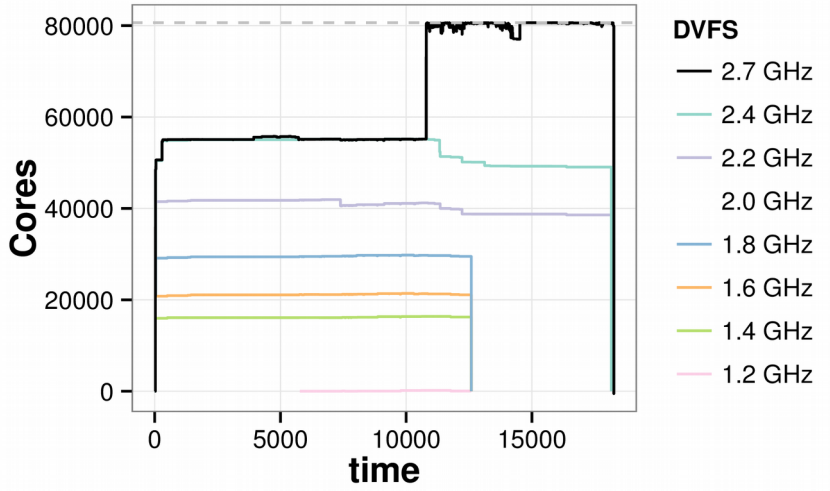


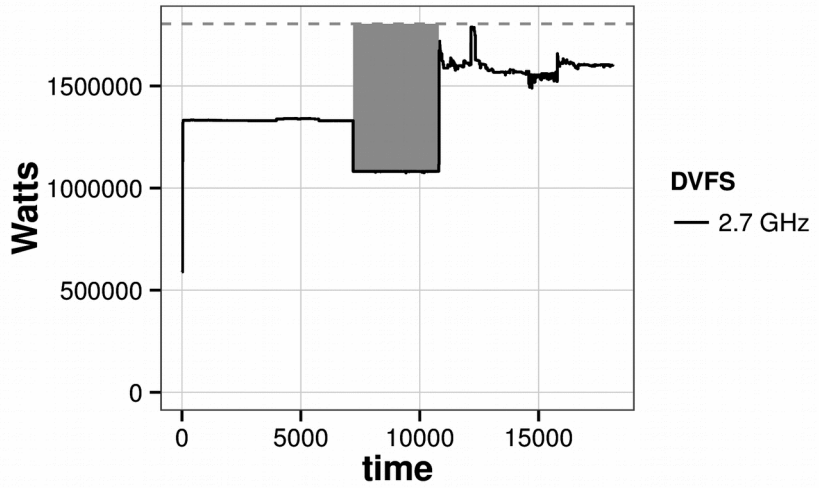
Idle

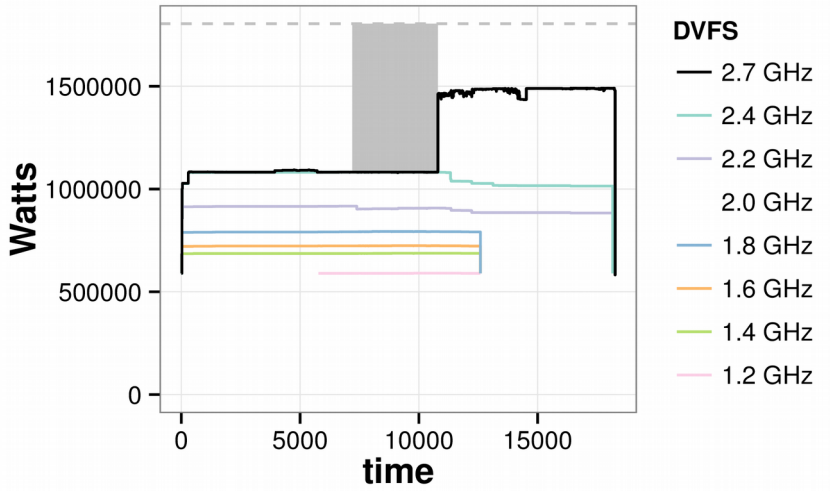
Switch-off

Experimental validation









Switch-off

DVFS

- Introduction
- DVFS & Switch-off

- The model
- Algorithm and implementation
- Experimentations

- **Conclusion and future works**

Current and future works

- Powercap on live power values
  - Implemented using Layouts
- Powercap on nodes
- DVFS
  - What about reproducibility of jobs runs?
  - To do DVFS right, we need to know the job
- Switch-off
  - New scheduling algorithms
  - Switch-off (with bonuses) whithout powercaps
  - Switch-off particular components (cpus, gpus, network...)

Architect of an Open World™

References

- [11] V. W. Freeh, D. K. Lowenthal, F. Pan, N. Kappiah, R. Springer, B. L. Rountree, and M. E. Femal, “Analyzing the energy-time trade-off in high-performance computing applications,” IEEE Transactions on Parallel and Distributed Systems, 2007.

- [22] M. Etinski, J. Corbalan, J. Labarta, and M. Valero, “BSLD threshold driven power management policy for HPC centers,” in 2010 IEEE International Symposium on Parallel Distributed Processing, Workshops and Phd Forum (IPDPSW), 2010.

Experimental validation

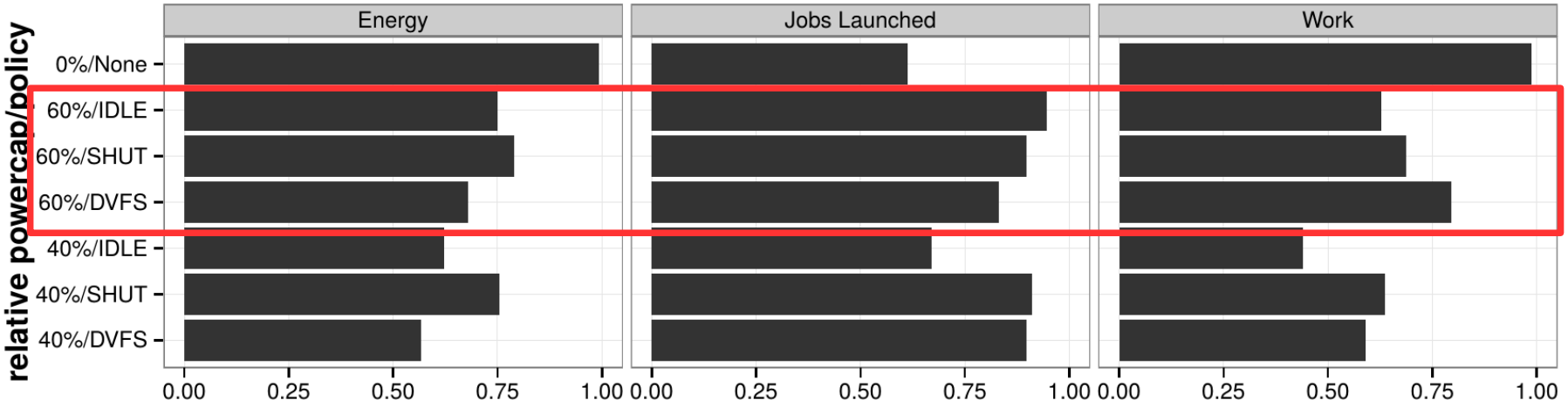


Fig. 8: Comparison of different scenarios of policies and powercaps based on normalized values of launched jobs, accumulated cpu time and total consumed energy during the 5 hours workload interval