Adaptive Resource and Job Management for limited power consumption

02/07/14

Yiannis Georgiou
David Glesser
Matthieu Hautreux
Denis Trystram

- Target: Big clusters
  - >10k cores
  - Biggest has 3M cores
- Lot of resources, managed by the RJMS
  - Resource and Job Management System
  - Famous ones: Slurm, PBS, OAR
  - Resources: CPU, GPU, networks, energy…
- How this works?
  - Users submit jobs
  - The RJMS chooses when and where to launch them

- This work targets the RJMS level

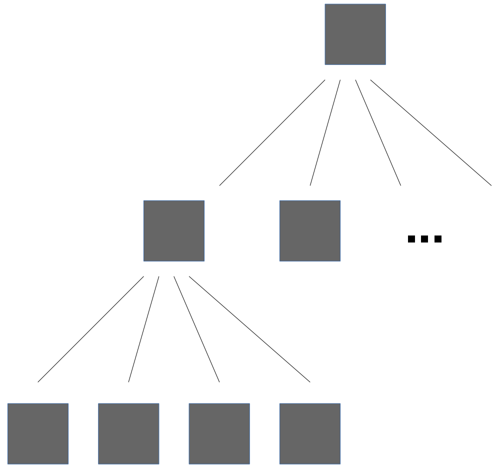
- What we know on each app at this level?
  - Max(runtime)
  - Resources needed (cores and other specific resources)
  - User
  - History of submissions

Energy is a driven constraint, going to the exascale requires to be able to gain 2 orders of magnitude in Power

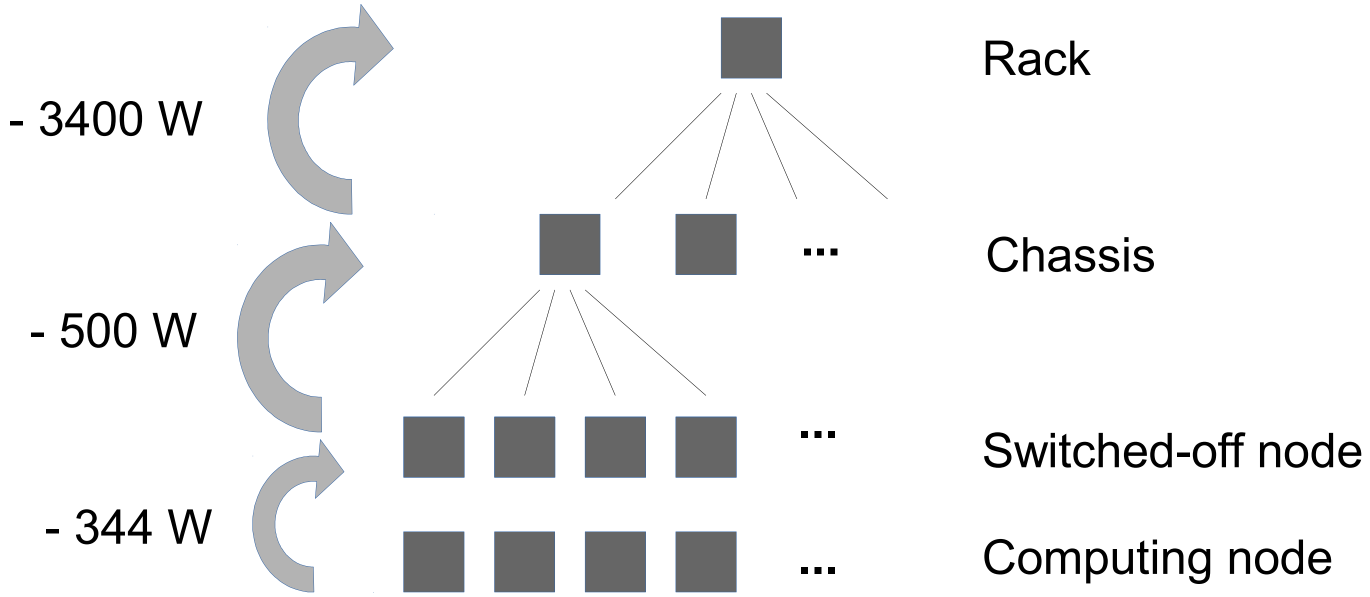
- What can we do to manage energy?
  - Architecture design
  - Applications optimizations
  - DVFS (dynamic frequency and voltage scaling)
  - Switch-off

- 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:
  - Node is the smallest switched-off level
  - 18 nodes per chassis, 5 chassis per rack
  - Power(**switch-off node**) ~= **5** * Power(**computing node**)
  - Power(**Chassis only**) ~= Power(**computing node**)
  - Power(**Rack**) ~= **10** * Power(**computing node**)

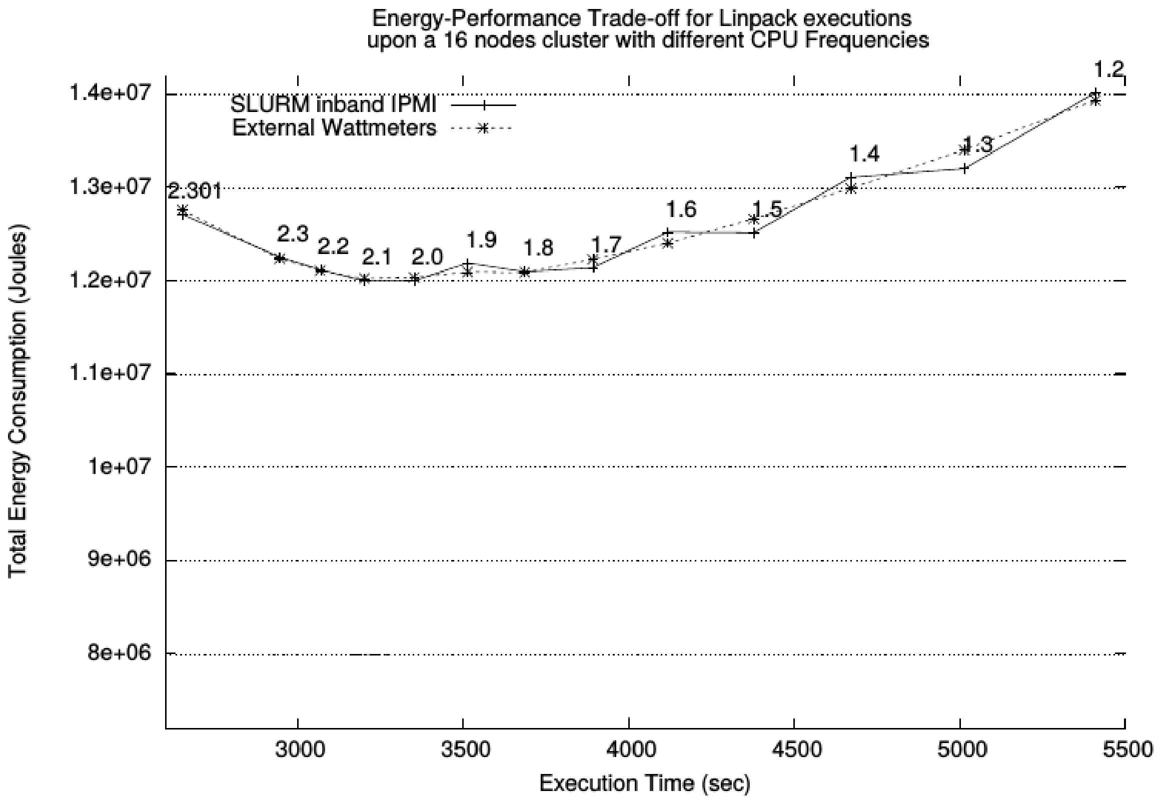


- DVFS
  - It's a trade-off between **performance** and **power consumption**
  - What about **performance** / **energy** trade-off ?

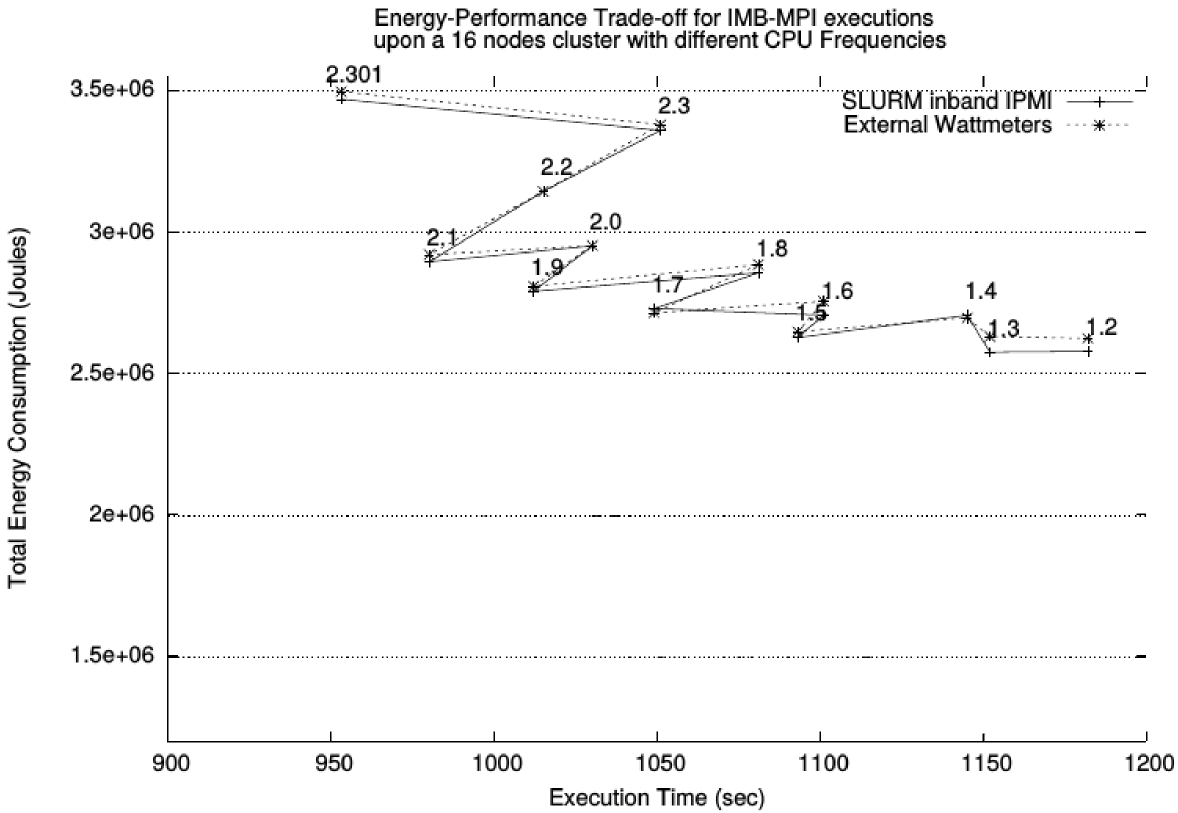
- DVFS
  - It's a trade-off between **performance** and **power consumption**
  - What about **performance** / **energy** trade-off ?

$$\int POWER . dt = Energy$$

- 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 energy != minimizing power
  - The impact of DVFS is highly dependant on the job

⇒ let's concentrate on power control

Intro - conclusion

Let's powercap!

- Why reduce?
  - Reduce cost
  - 50% of the annual cost
  - Reduce CO_2

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

- We work with maximum power consumptions
- 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:

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

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

How to choose ?

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

When RHO <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.

The algorithm

- A usable algorithm
  - Implemented in Slurm
  - We keep the original algorithm (ordered list + backfilling)

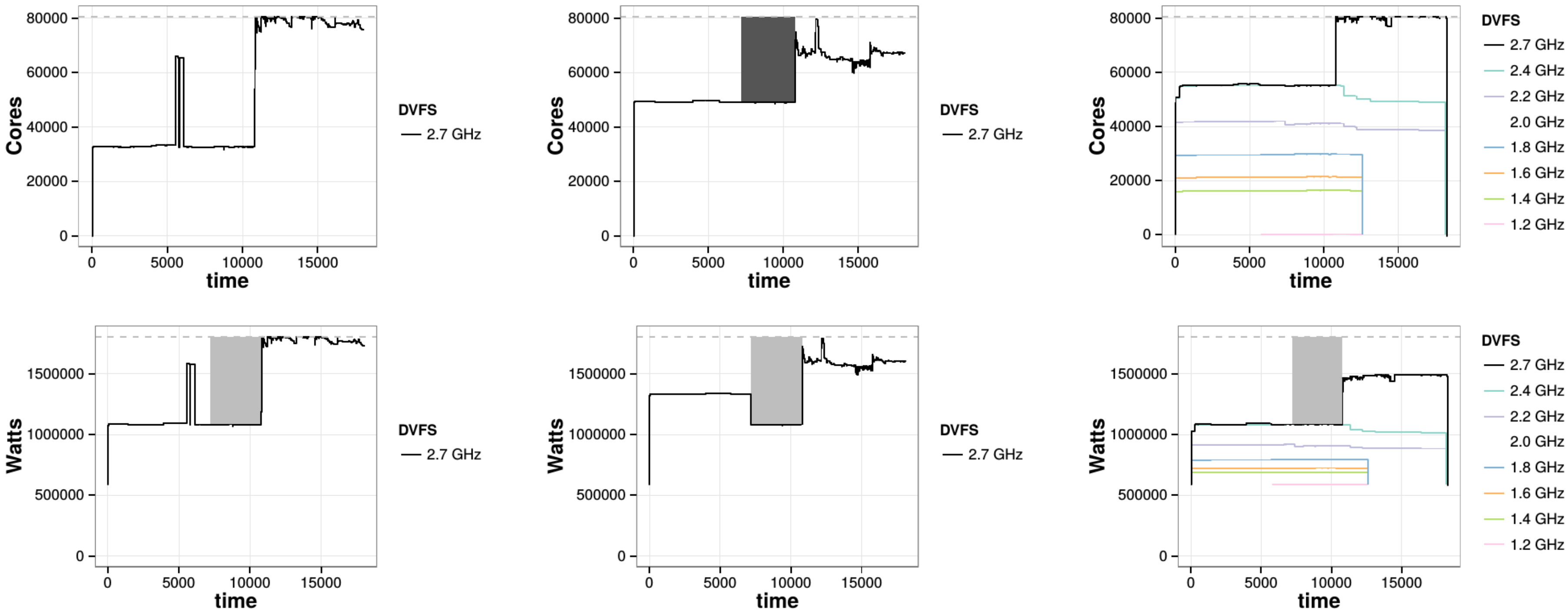
- Compute less thing at runtime

- 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

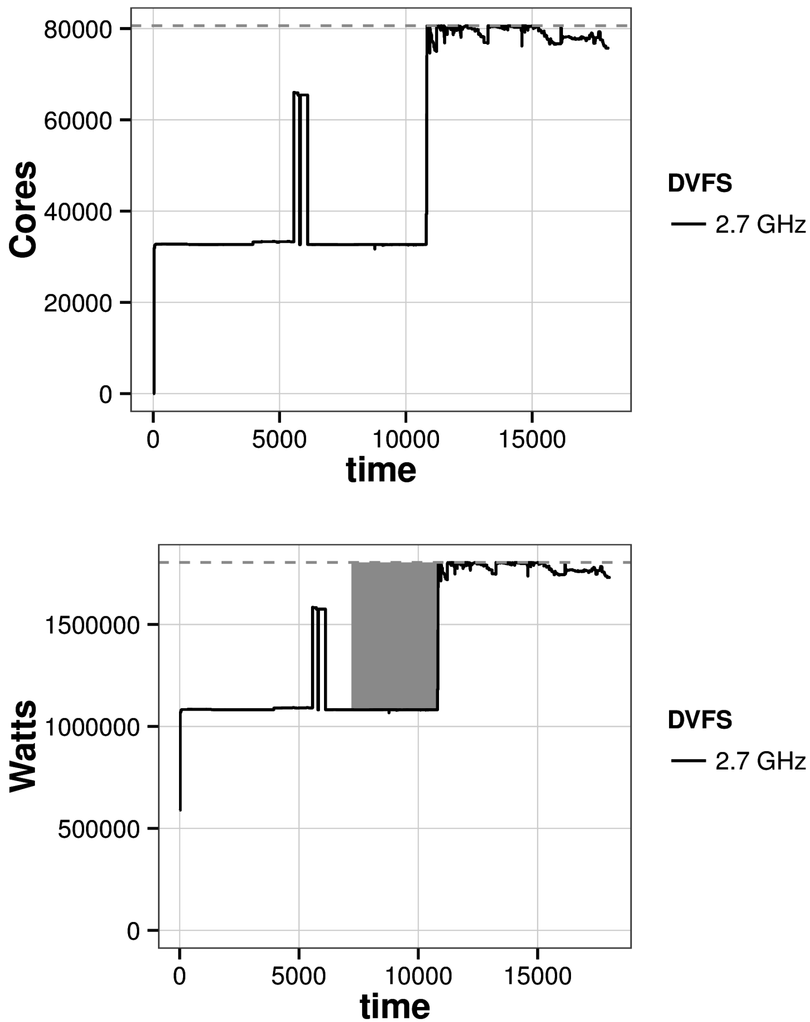
- Slurm can emulate his environement
  - 336 Slurm nodes on 1 physical node
  - *Sleep* instead of real job
- Replay interesting part of the original log
  - 5 hours, high throughput, jobs representative of the whole log
- Add a powercap
  - Case study: 1 hour, in the middle of the trace, at different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



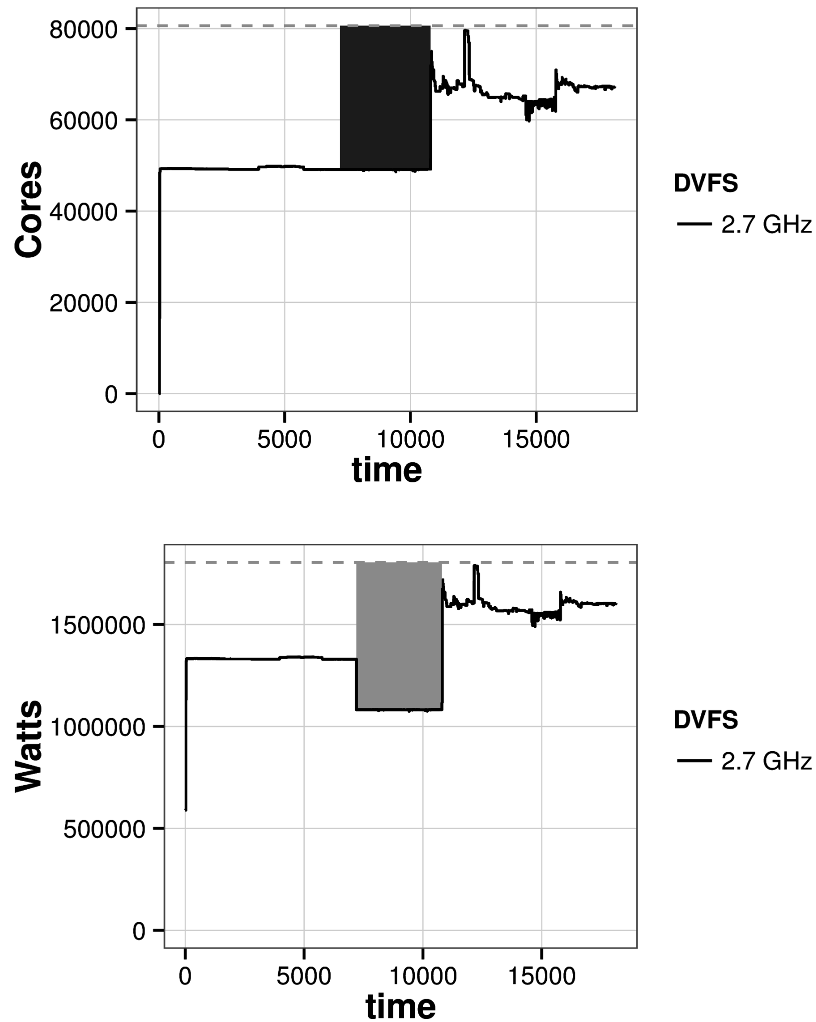
(a) IDLE (b) SHUT (c) DVFS

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

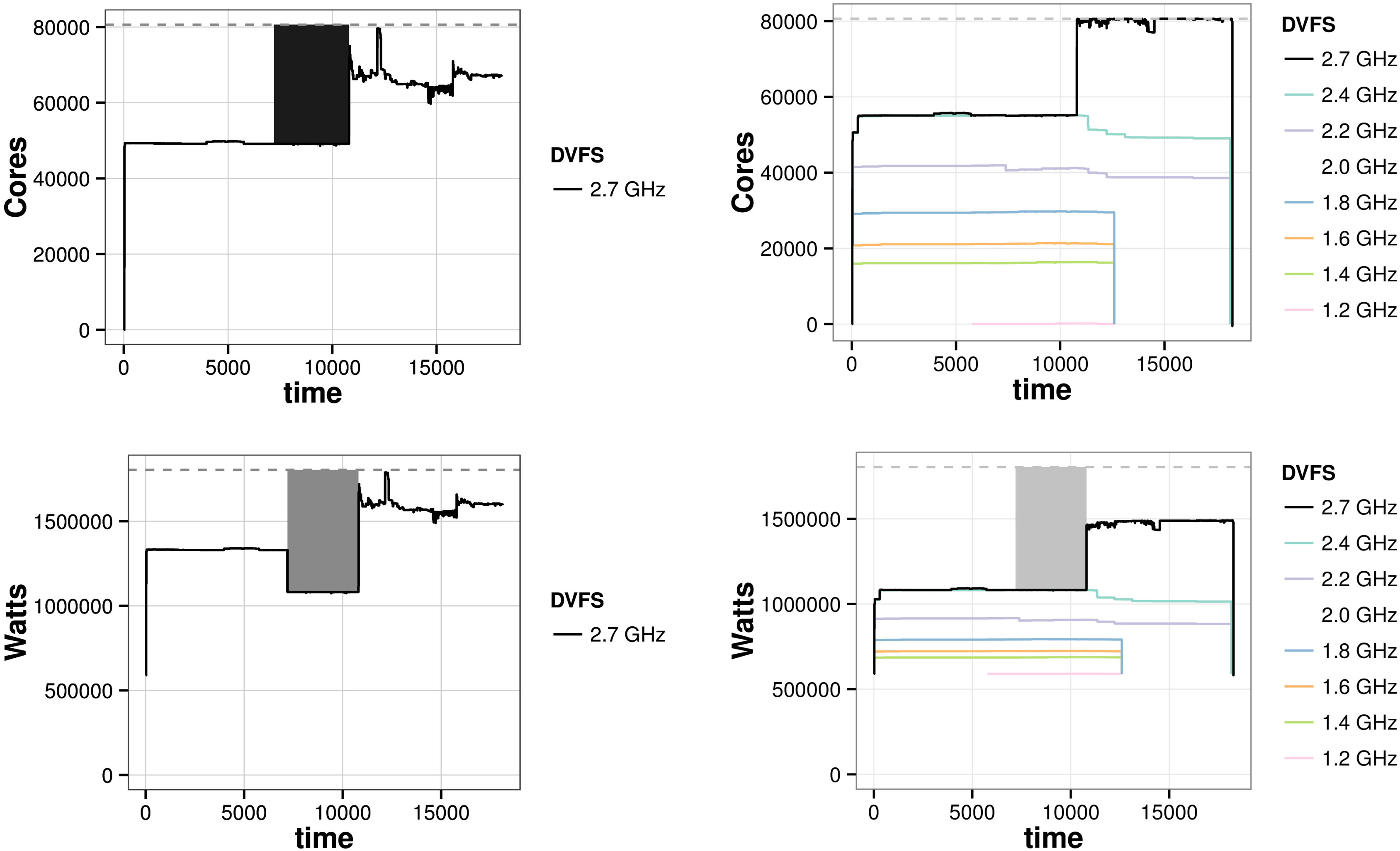


Idle



Switch-off

Experimental validation



Switch-off DVFS

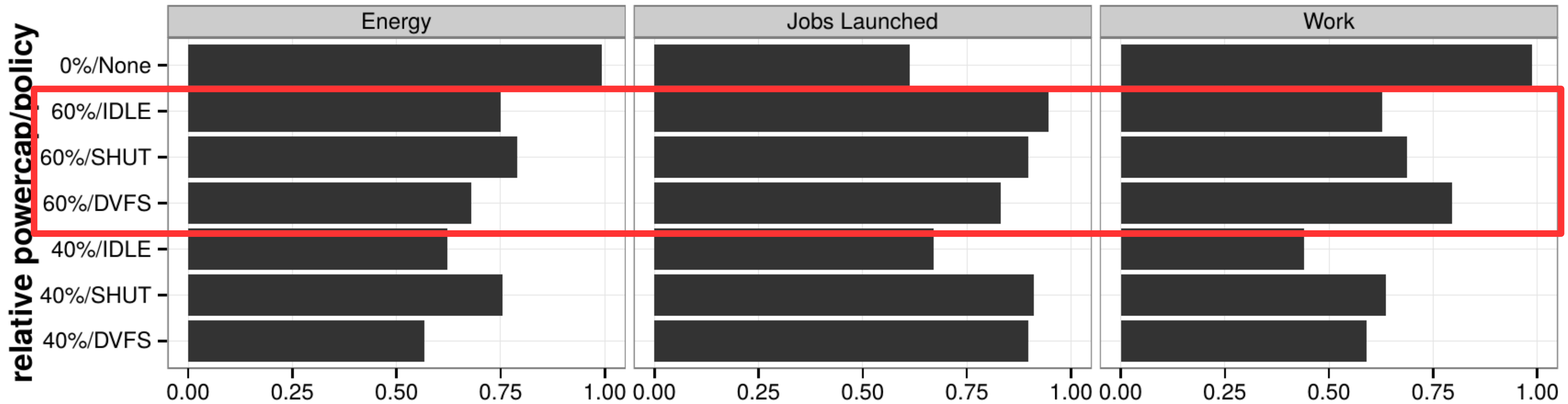


Fig. 8: Comparison of different scenarios of policies and powercaps based on normalized values of launched jobs, accumulated cpu time and total consumed energy during the 5 hours workload interval

- Powercap on real power values ?

- More switch-off
  - New scheduling algorithms
  - Switch-off (with bonuses) whithout powercaps

- Less DVFS
  - At least not at our level
  - What about reproducibility of jobs runs?
  - To do DVFS right, we need to know the job

Architect of an Open World™