Power and Energy aware job scheduling techniques

Yiannis Georgiou
R&D Software Architect

02-07-2015

Your business technologists. Powering progress

Bull
atos technologies

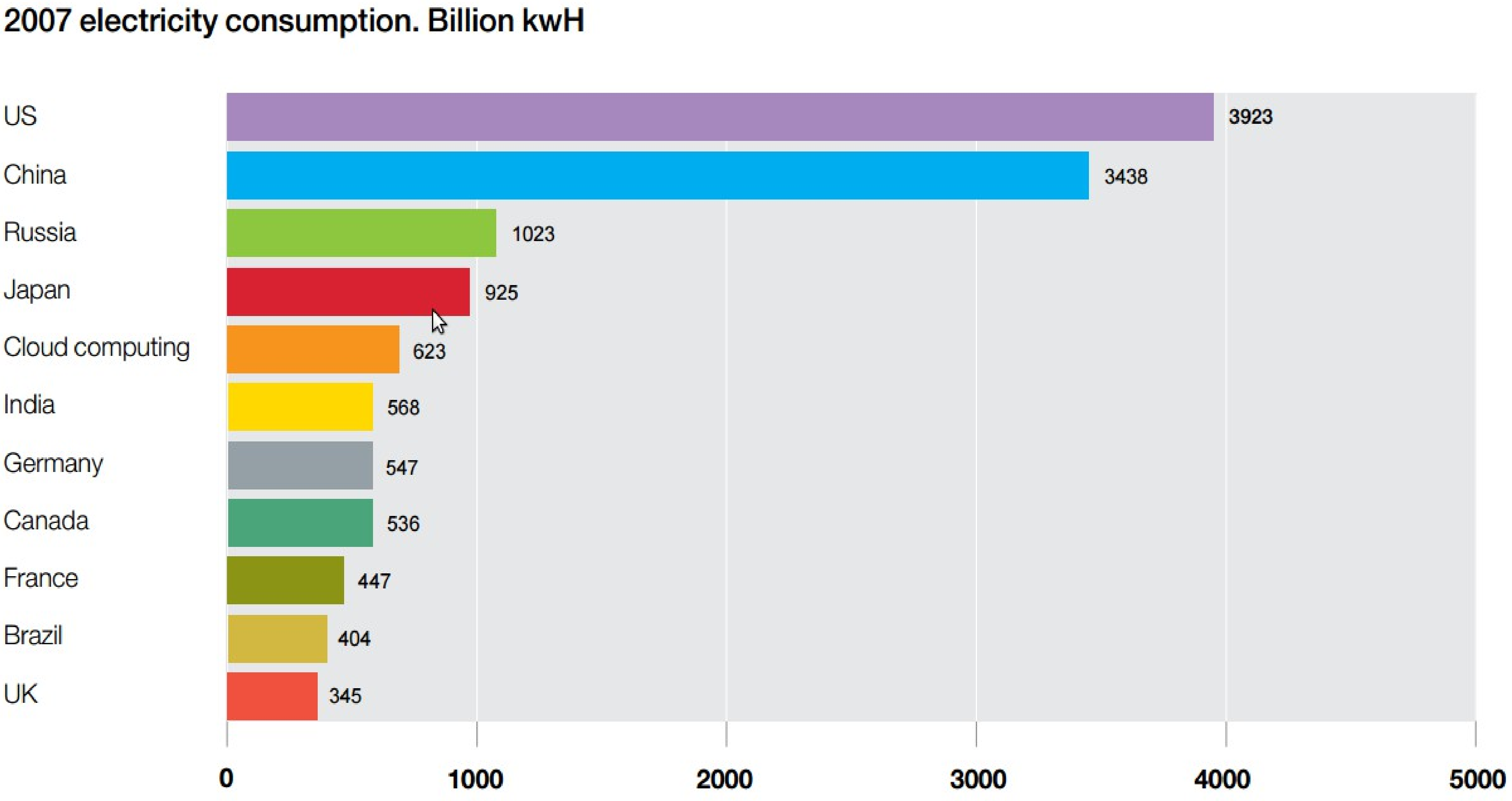
Top500 HPC supercomputers

Rank	Site	System	Cores	Rmax (Tflop/s)	Rpeak (Tflop/s)	Power (kW)
1	NUDT, China	Tianhe-2	3,120,000	33,862.7	54,902.4	17,808
2	ORNL, USA	Titan	560,640	17,590.0	27,112.5	8,209
3	LLNL, USA	Sequoia	1,572,864	17,173.2	20,132.7	7,890
4	Riken, Japan	K computer	705,024	10,510.0	11,280.4	12,660
5	ANL, USA	Mira	786,432	8,586.6	10,066.3	3,945
6	CSCS, Switzerland	Piz Daint	115,984	6,271.0	7,788.9	2,325
7	Texas, USA	Stampede	462,462	5,168.1	8,520.1	4,510
8	Juelich, Germany	JUQUEEN	458,752	5,008.9	5,872.0	2,301
9	LLNL, USA	Vulcan	393,216	4,293.3	5,033.2	1,972
10	Government,USA	Cray XC30	72,800	3,577	6,131.8	1,499

From Top500 November 2014 list

Bull atos technologies

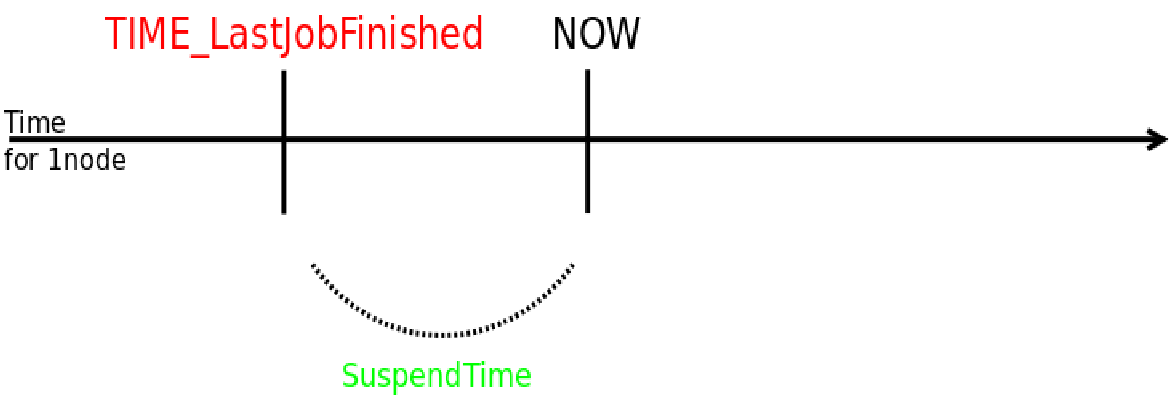
IT Energy Consumption



http://www.greenpeace.org/international/Global/international/publications/climate/2012/iCoal/HowCleanisYourCloud.pdf

Energy Reduction Techniques

- Framework for energy reductions through unutilized nodes
  - Administrator configurable actions (hibernate, DVFS, power off, etc)
  - Automatic 'wake up' when jobs arrive



Algorithm for SLURM Energy Reduction Techniques

Nodes Sleep Actions
    if SuspendTime > A_PreDefined_Idle_TIME
        exec SuspendProgram upon SuspendRate nodes per minute

Nodes WakeUp Actions
    if SleepingNode_isNeeded then
        exec ResumeProgram upon ResumeRate nodes per minute

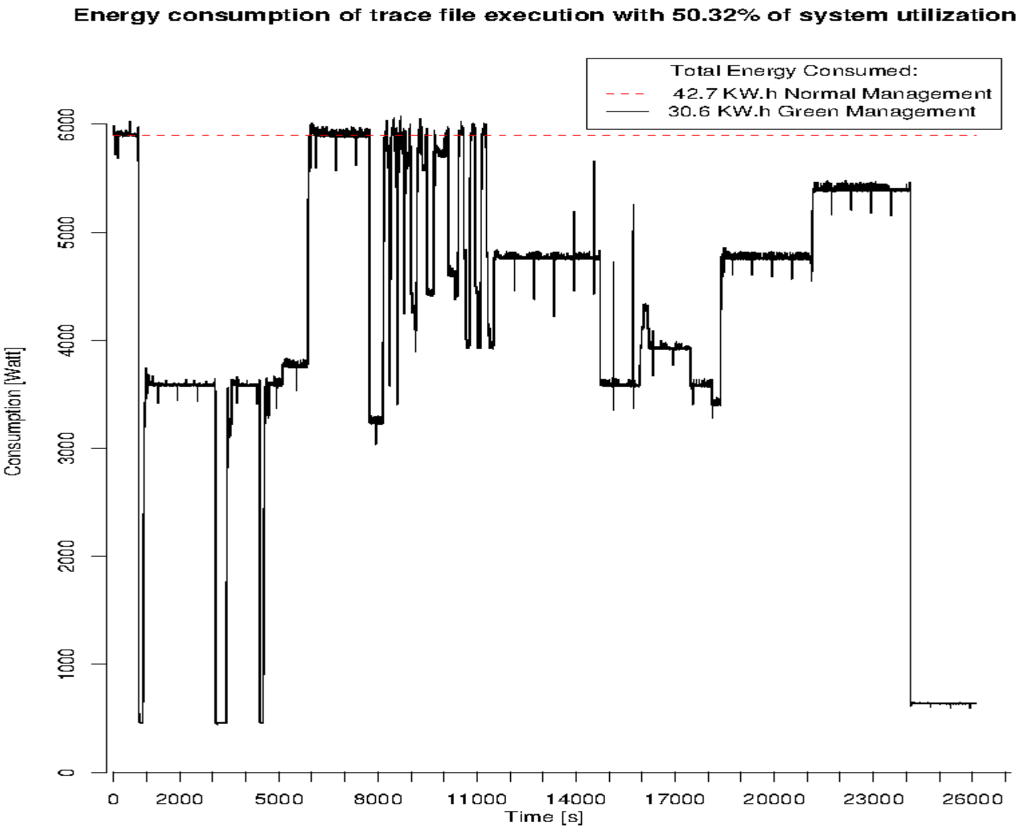
LIG

inria
INVENTEURS DU MONDE NUMÉRIQUE

Green-Net

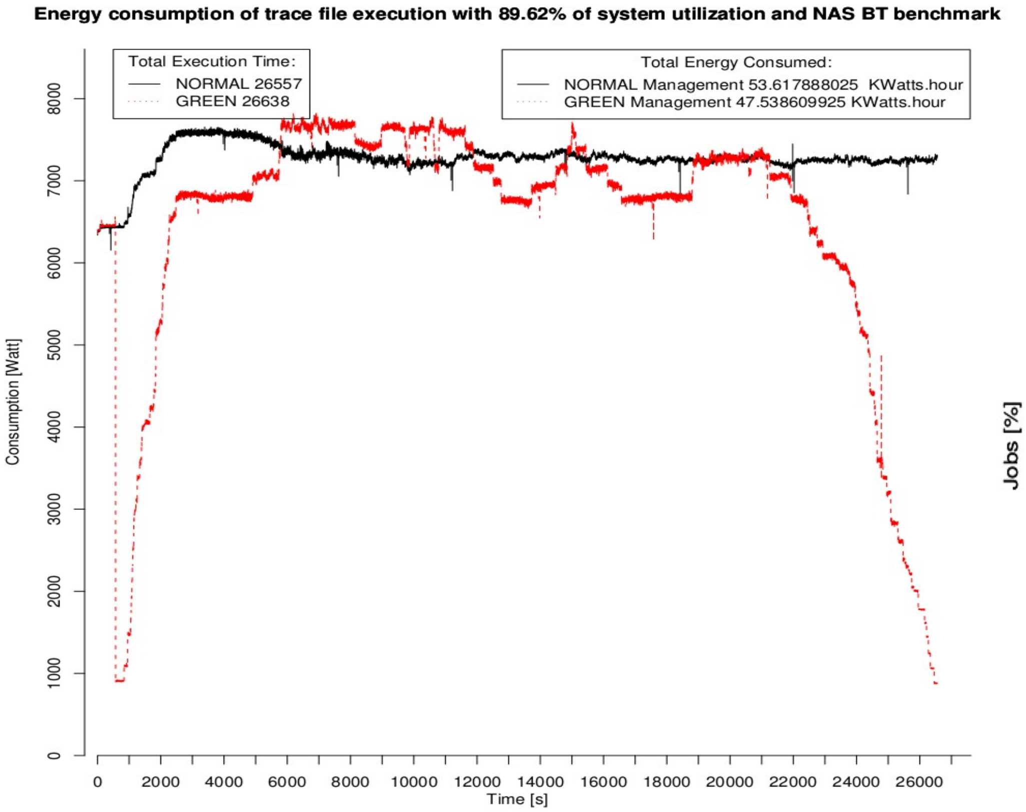
Bull
atos technologies

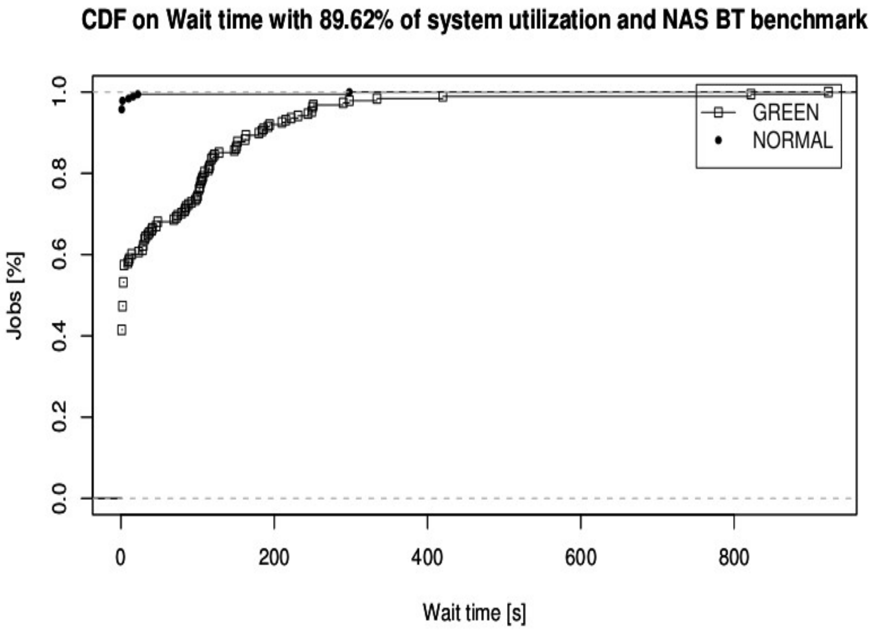
Energy Reduction Techniques



Georges Da Costa, Marcos Dias de Assuncao, Jean-Patrick Gelas, Yiannis Georgiou, Laurent Lefevre, Anne-Cecile Orgerie, Jean-Marc Pierson, Olivier Richard and Amal Sayah
Multi-facet approach to reduce energy consumption in clouds and grids: The green-net framework.
(In proceedings of e-Energy 2010)

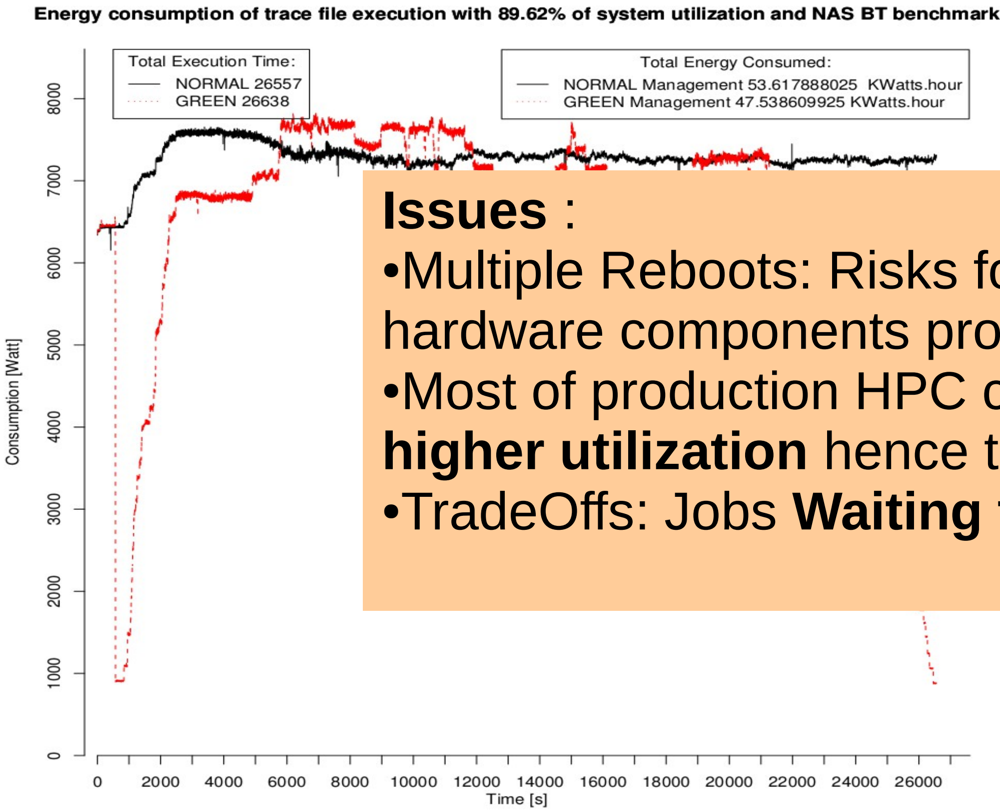
Energy Reduction Techniques

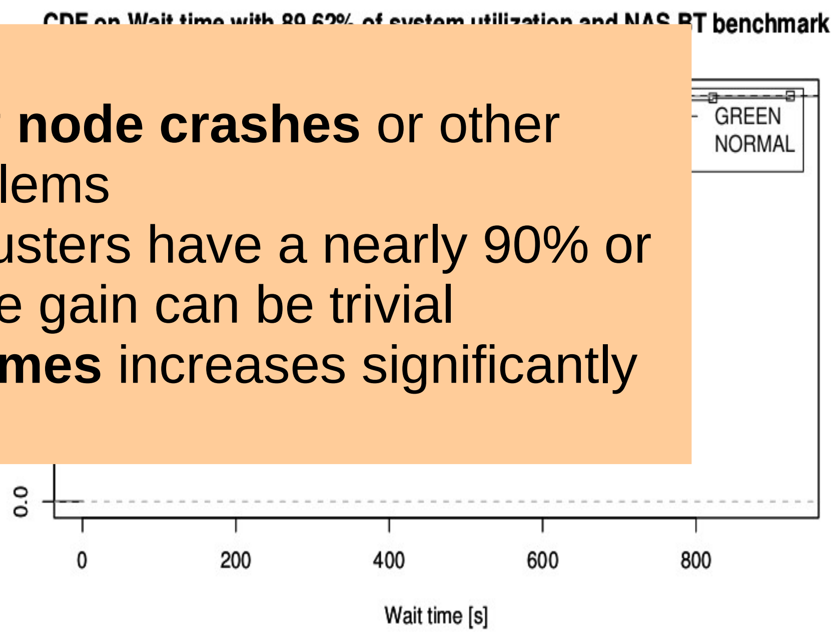




Yiannis Georgiou

Energy Reduction Techniques





Issues :

•Multiple Reboots: Risks for **node crashes** or other hardware components problems

•Most of production HPC clusters have a nearly 90% or **higher utilization** hence the gain can be trivial

•TradeOffs: Jobs **Waiting times** increases significantly

Power and Energy Management

Issues that we wanted to deal with:

- Attribute **power and energy data** to HPC components
- Calculate the **energy consumption of jobs** in the system
- Extract **power** consumption **time series of jobs**
- **Control** the Power and Energy usage of jobs and workloads

Power and Energy Measurement System

- Power and Energy monitoring per node
- Energy accounting per step/job
- Power profiling per step/job
- CPU Frequency Selection per step/job

How this takes place :

- In-band collection of energy/power data (IPMI / RAPL plugins)
- Out-of-band collection of energy/power data (RRD plugin)
- Power data job profiling (HDF5 time-series files)
- Parameter for CPU frequency selection on submission commands

Power and Energy Measurement System

- Power and Energy monitoring per node
- Energy accounting per step/job
- Power profiling per step/job

- **Overhead:** In-band Collection
- **Precision:** measurements and internal calculations
- **Scalability:** Out-of band Collection

How t

- In-band collection of energy/power data (IPMI / RAPL plugins)
- Out-of-band collection of energy/power data (RRD plugin)
- Power data job profiling (HDF5 time-series files)
- SLURM Internal power-to-energy and energy-to-power calculations

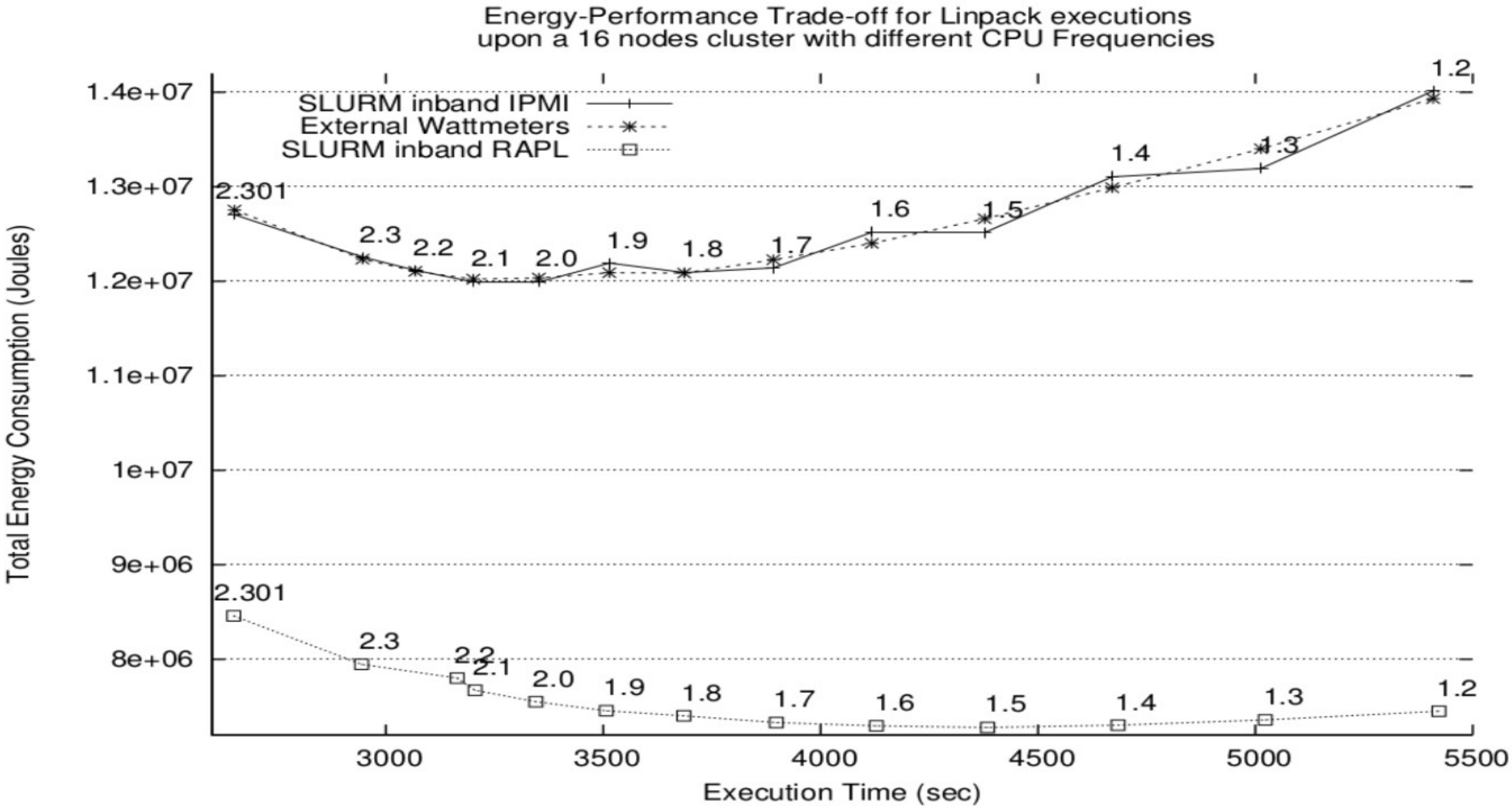
Power and Energy Measurement System

```
[root@cuzco108 bin]# sacct -o "JobID%5,JobName,AllocCPUS,NNodes
%3,NodeList%22,State,Start,End,Elapsed,ConsumedEnergy%9"
JobID    JobName  AllocCPUS NNodes              NodeList      State
           Start                   End    Elapsed ConsumedEnergy
----- ---------- ---------- --- ---------------------- ----------
------------------- ------------------- --------- ---------
  127    cg.D.32         32   4     cuzco[109,111-113]  COMPLETED 2013-
09-12T23:12:51 2013-09-12T23:22:03   00:09:12   490.60KJ

[root@cuzco108 bin]# cat extract_127.csv
Job,Step,Node,Series,Date_Time,Elapsed_Time,Power
13,0,orion-1,Energy,2013-07-25 03:39:03,0,126
13,0,orion-1,Energy,2013-07-25 03:39:04,1,126
13,0,orion-1,Energy,2013-07-25 03:39:05,2,126
13,0,orion-1,Energy,2013-07-25 03:39:06,3,140
13,0,orion-1,Energy,2013-07-25 03:39:07,4,140
13,0,orion-1,Energy,2013-07-25 03:39:08,5,140
```

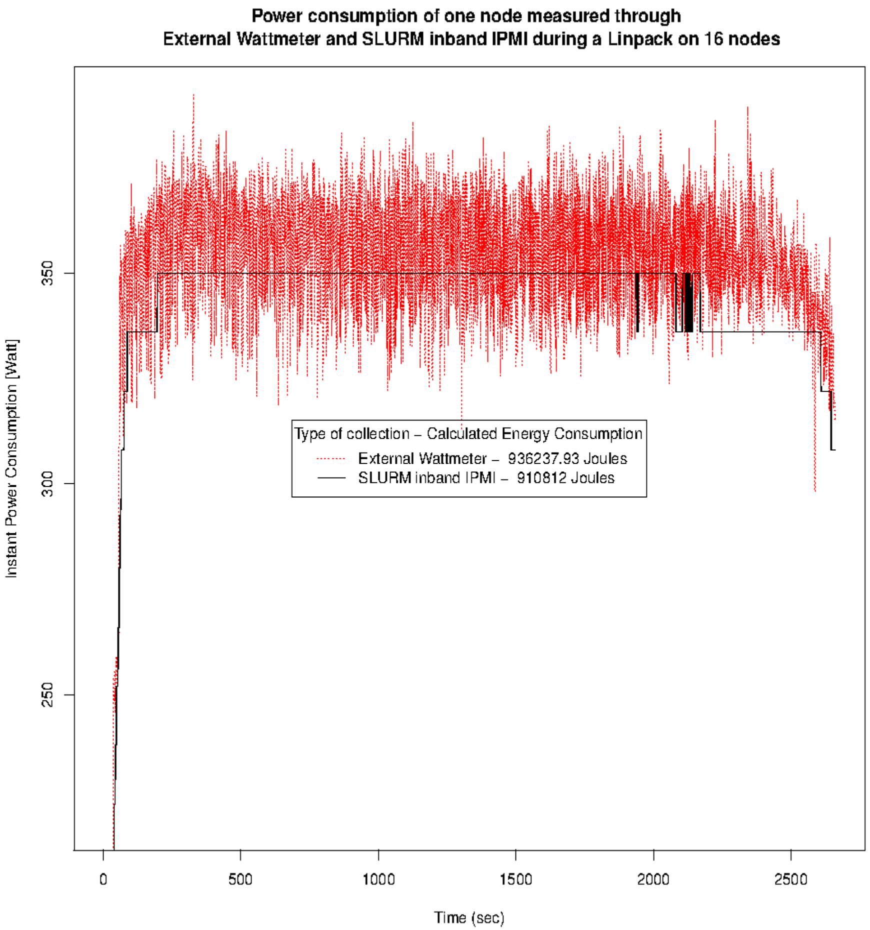

Yiannis Georgiou, Thomas Cadeau, David Glesser, Danny Auble, Morris Jette and Matthieu Hautreux
Energy Accounting and Control with SLURM Resource and Job Management System
(In proceedings of ICDCN 2014)

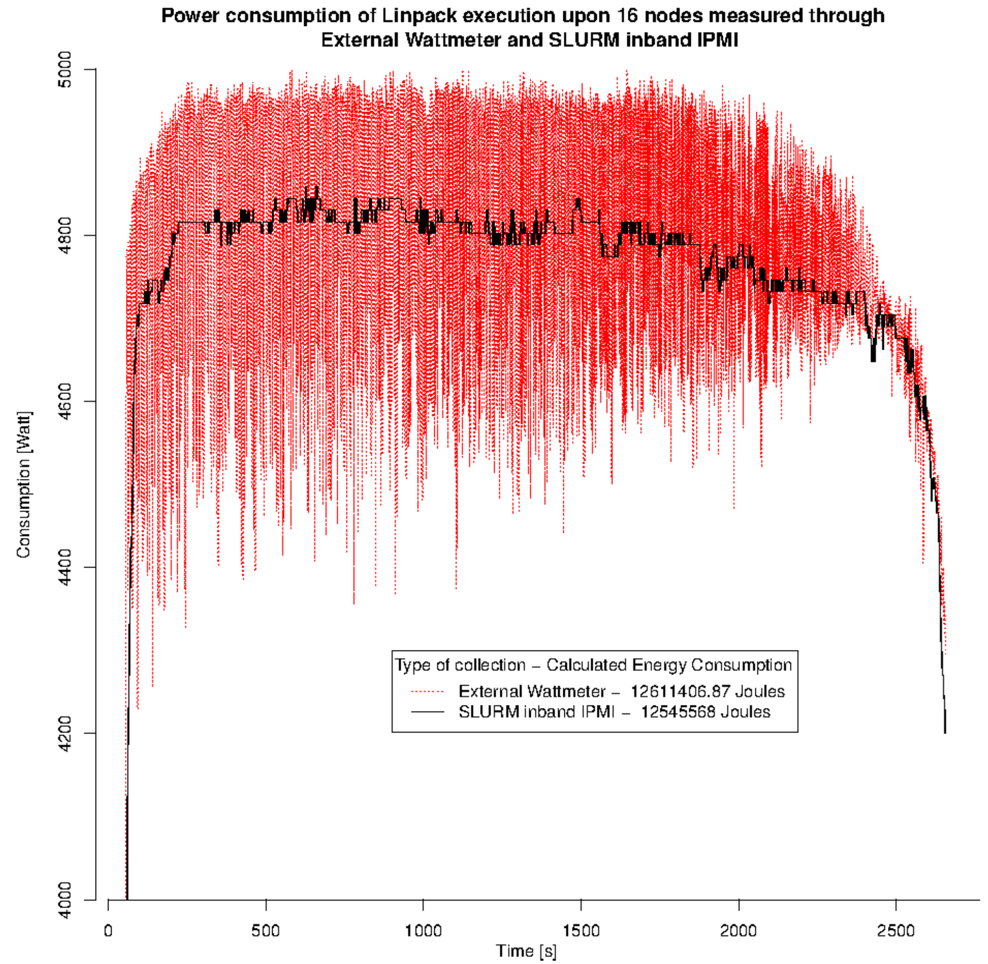
Power and Energy Measurement System



Yiannis Georgiou, Thomas Cadeau, David Glesser, Danny Auble, Morris Jette and Matthieu Hautreux
Energy Accounting and Control with SLURM Resource and Job Management System
(In proceedings of ICDCN 2014)

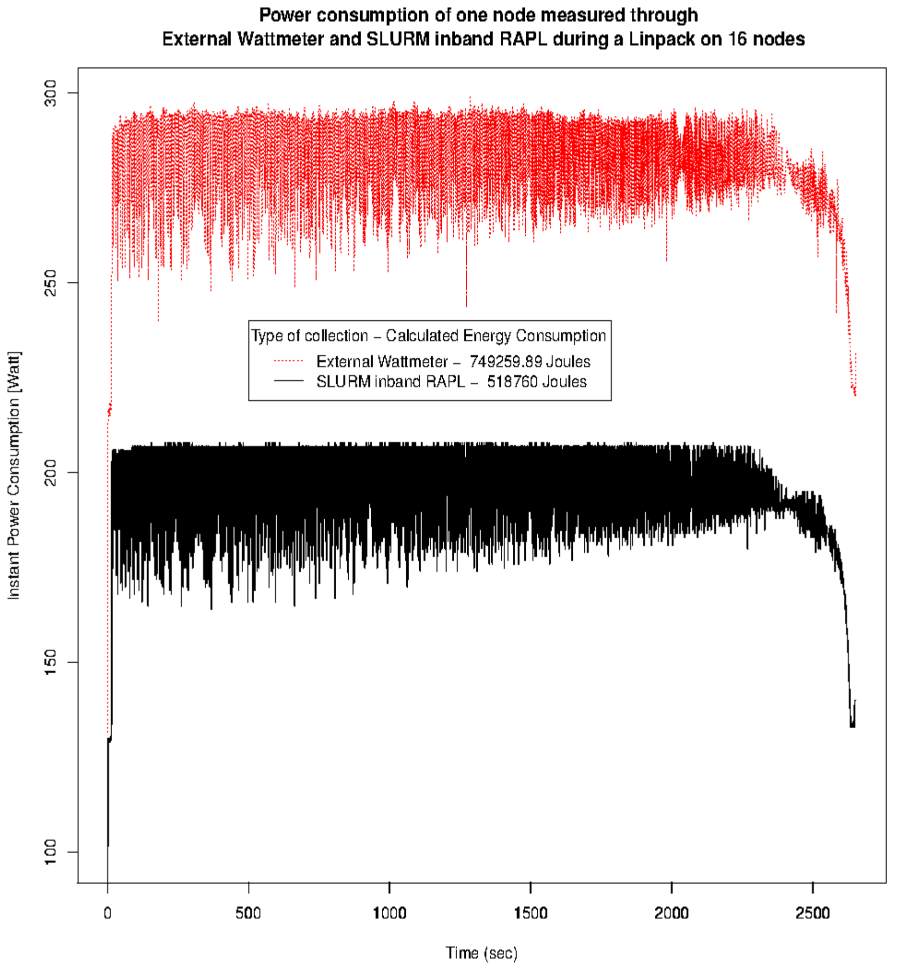
Power and Energy Measurement System

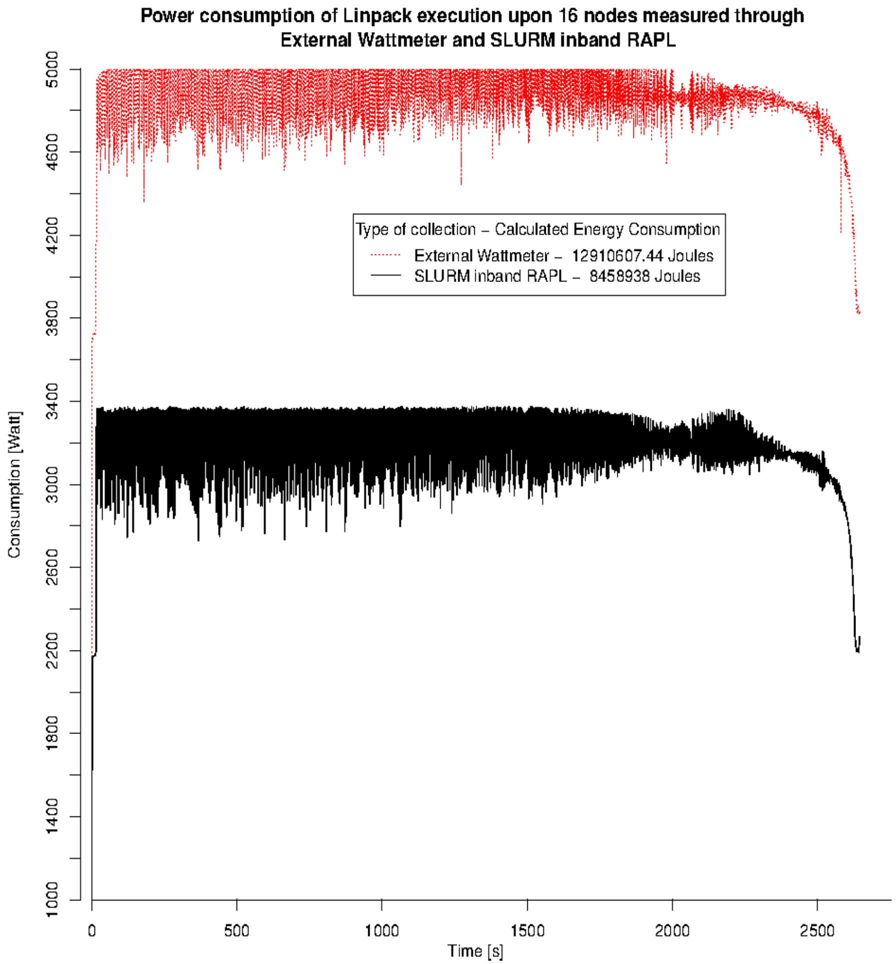




Yiannis Georgiou, Thomas Cadeau, David Glesser, Danny Auble, Morris Jette and Matthieu Hautreux
Energy Accounting and Control with SLURM Resource and Job Management System
(In proceedings of ICDCN 2014)

Power and Energy Measurement System

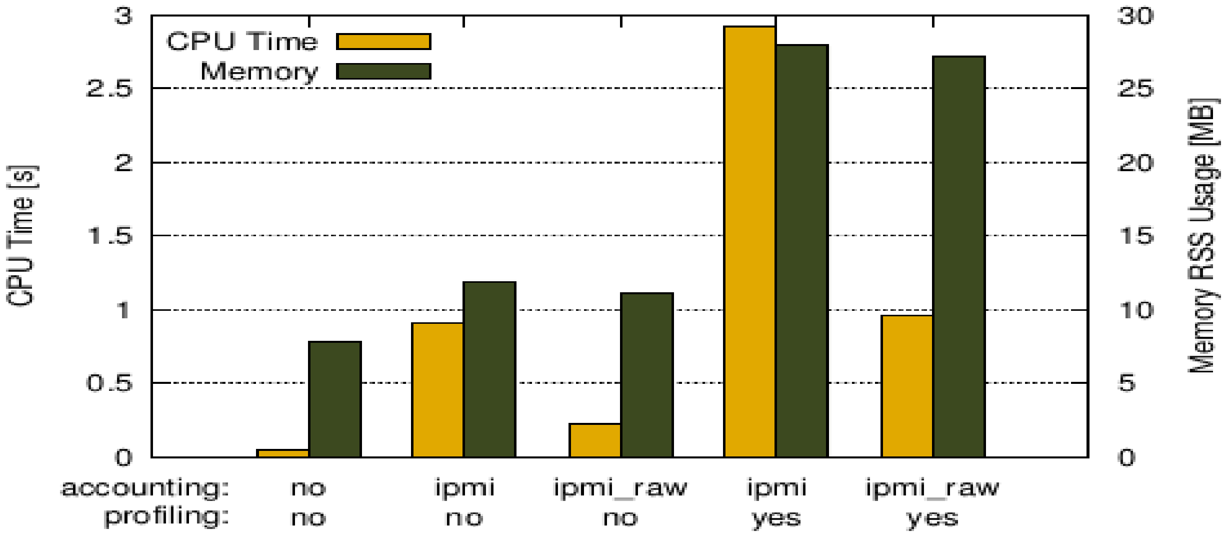




Yiannis Georgiou, Thomas Cadeau, David Glesser, Danny Auble, Morris Jette and Matthieu Hautreux
Energy Accounting and Control with SLURM Resource and Job Management System
(In proceedings of ICDCN 2014)

Optimizations of Power and Energy Measurement System

- Based on TUD/BULL - BMC firmware optimizations
  - sampling to 4Hz
  - No overhead for accounting



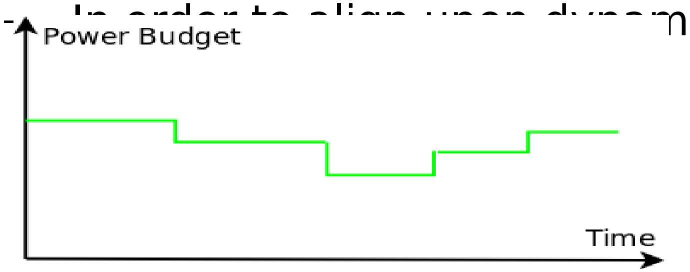
High Definition energy efficiency monitoring based on new FPGA architecture

- Sampling to 1000Hz
- Accuracy target to 2 % for energy and power

Daniel Hackenberg, Thomas Ilsche, Joseph Schuchart, Robert Sch¨ne, Wolfgang E. Nagel, Marc Simon, Yiannis Georgiou
HDEEM: High Definition Energy Efficiency Monitoring
In proceedings E2SC-2014

Power adaptive scheduling

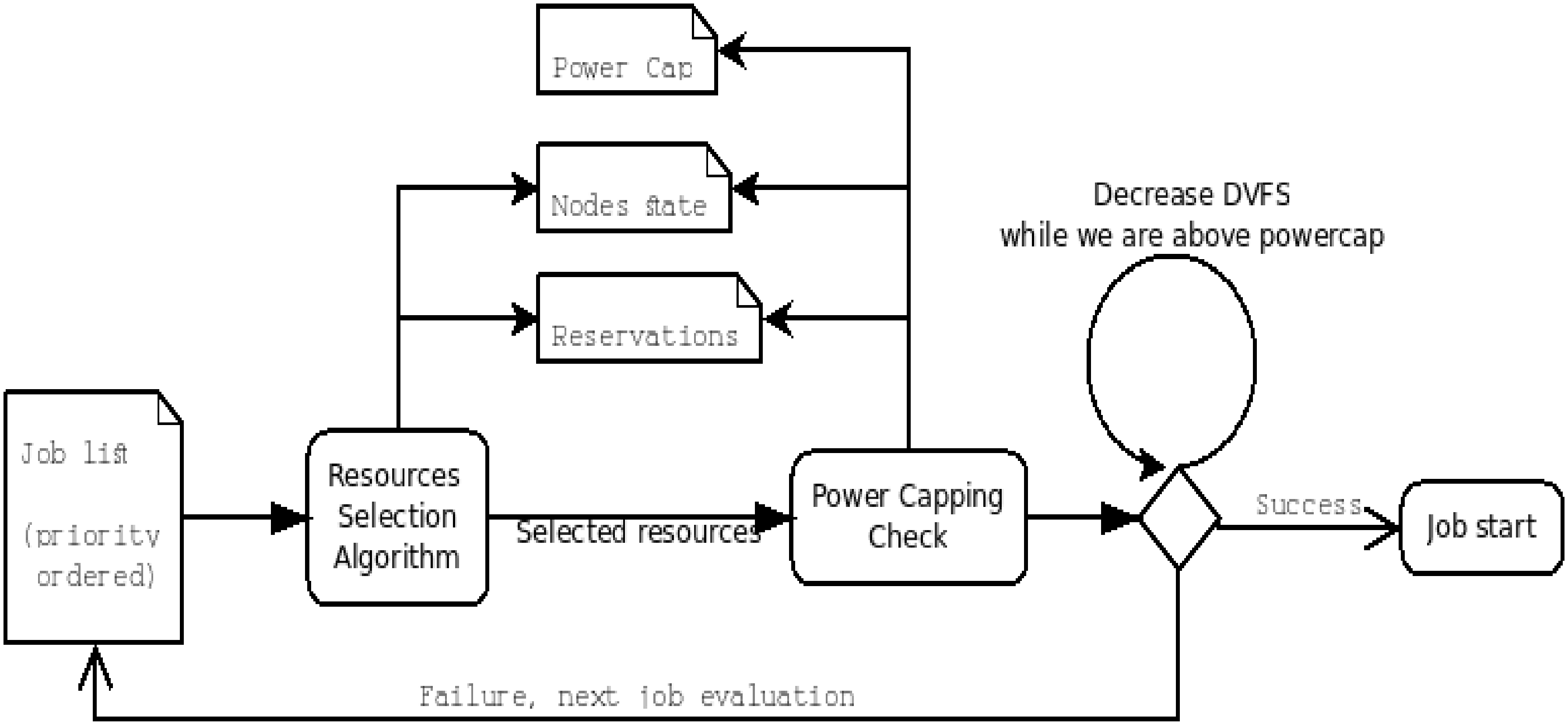
- Provide **centralized mechanism** to dynamically **adapt the instantaneous power** consumption of the whole platform
  - Reducing the number of usable resources or running them with lower power
- Provide technique to plan in advance for **future power adaptations**
  - In order to align upon dynamic energy provisioning and **electricity prices**



- Reductions take place through following techniques coordinated by the scheduler:
  - Letting Idle nodes
  - Powering-off unused nodes
  - Running nodes in lower CPU Frequencies

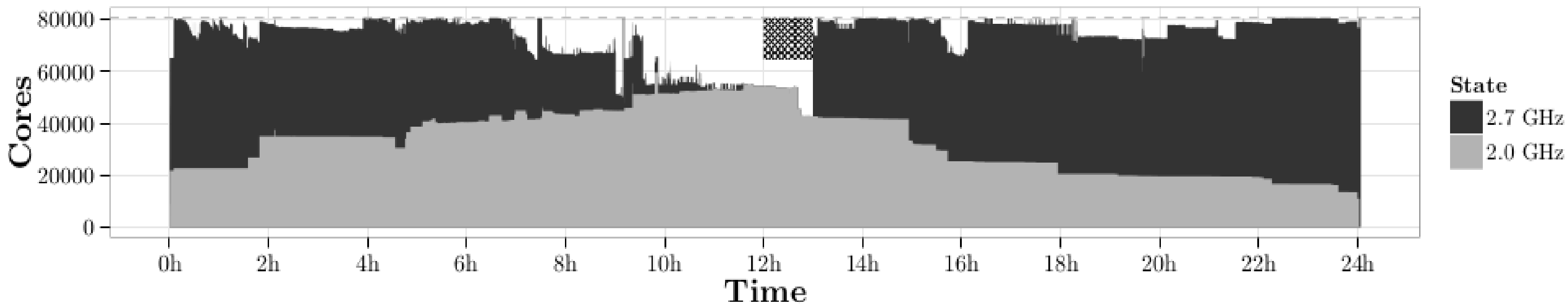
MOEBUS Project (http://moebus.gforge.inria.fr/)

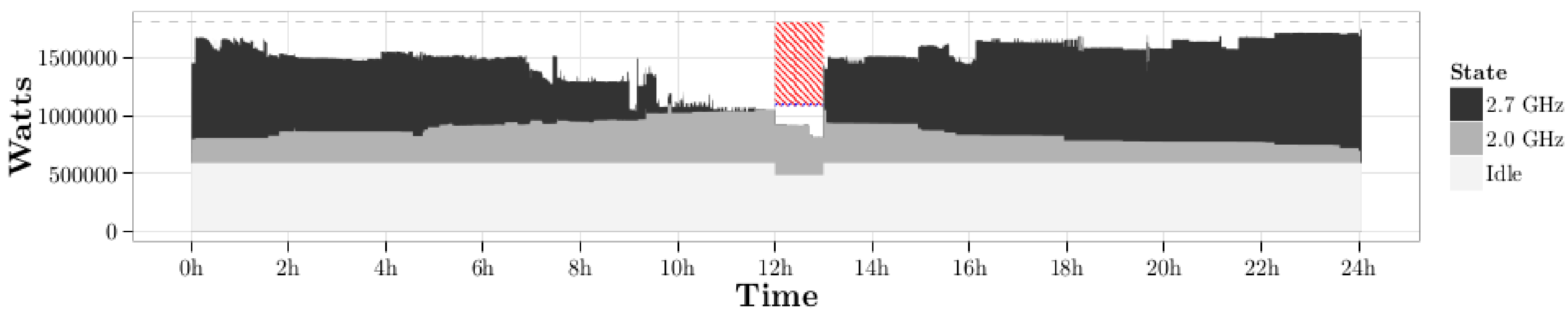
Power adaptive scheduling – algorithm



Power adaptive scheduling

System utilization in terms of cores (top) and power (bottom) for MIX policy during a 24 hours workload of Curie system with a powercap reservation (hatched area) of 1 hour of 40% of total power. Cores switched-off represented by a dark-grey hatched area.

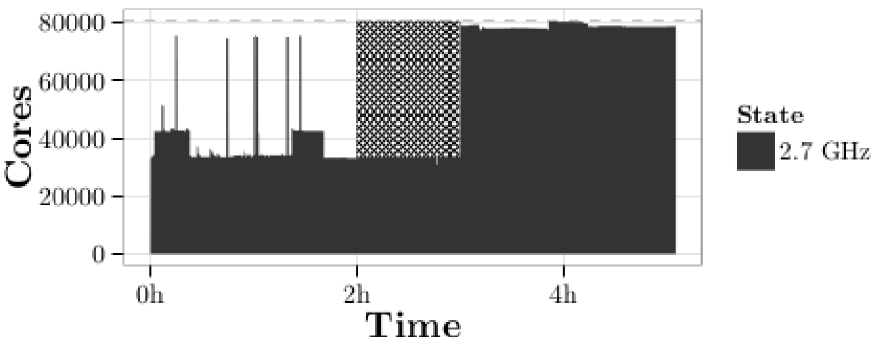


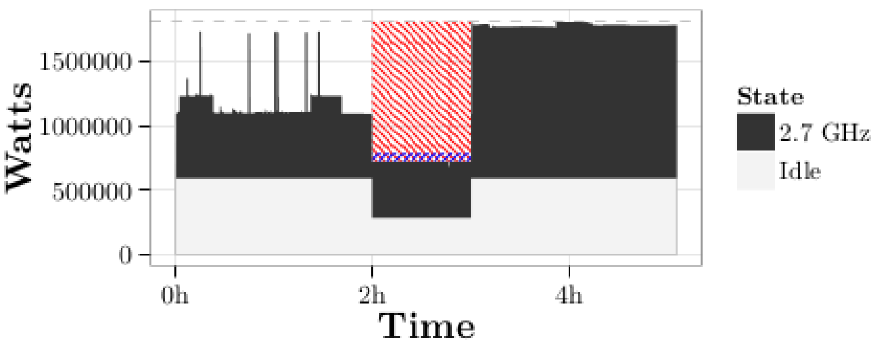


Yiannis Georgiou, David Glesser, Denis Trystram
Adaptive Resource and Job Management for limited power consumption
In proceedings of IPDPS-HPPAC 2015

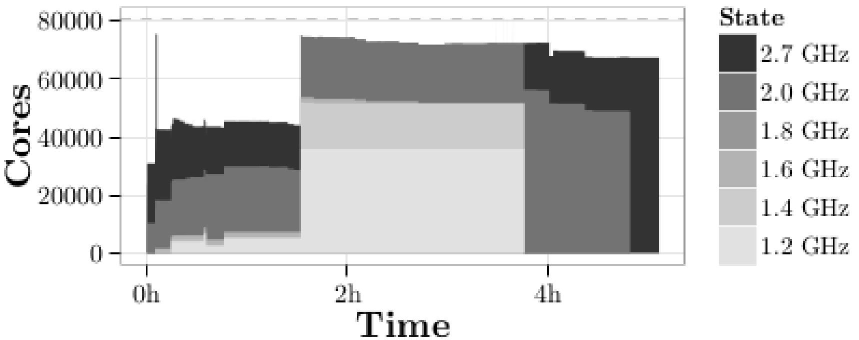
Power adaptive scheduling

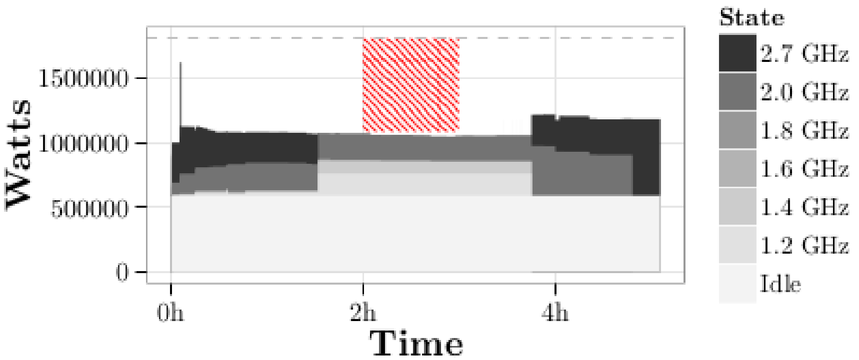
Powercap of 60% with mainly big jobs and SHUT policy





Powercap of 40% with mainly small jobs and DVFS policy





Yiannis Georgiou, David Glesser, Denis Trystram
Adaptive Resource and Job Management for limited power consumption
In proceedings of IPDPS-HPPAC 2015

Bull atos technologies

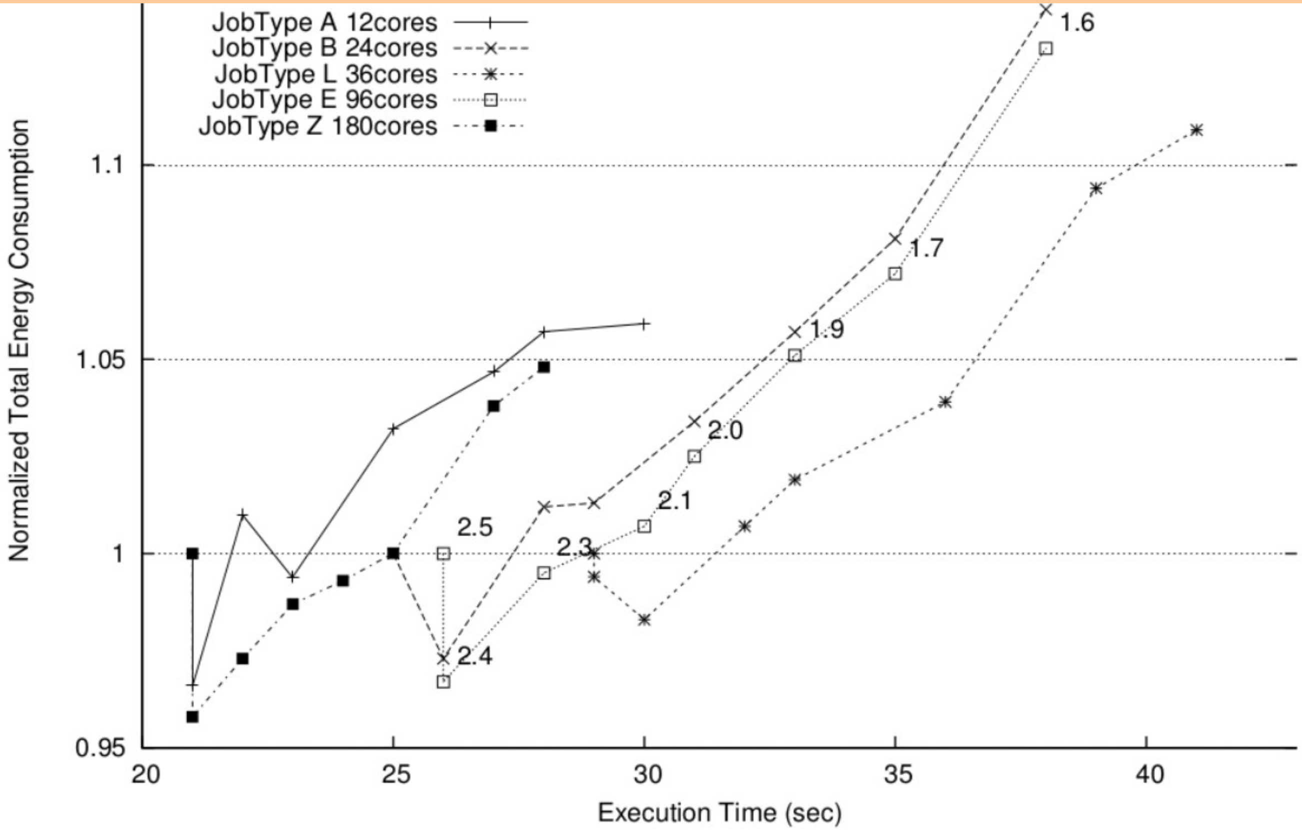
Energy Fairsharing

- Fairsharing is a common scheduling prioritization technique
- Exists in most schedulers, based on past CPU-time usage
- Our goal is to do it for **past energy usage**
- Provide **incentives to users** to be more energy efficient
  - Based upon the energy accounting mechanisms
  - Accumulate past jobs energy consumption and align that with the shares of each account
  - Implemented as a new multi-factor plugin parameter in SLURM

- Energy efficient users will be favored with lower stretch and waiting times in the scheduling queue

Energy Fairsharing

Performance vs. energy tradeoffs for Linpack applications as calibrated for different sizes and execution times running on an 180-cores cluster at different frequencies.

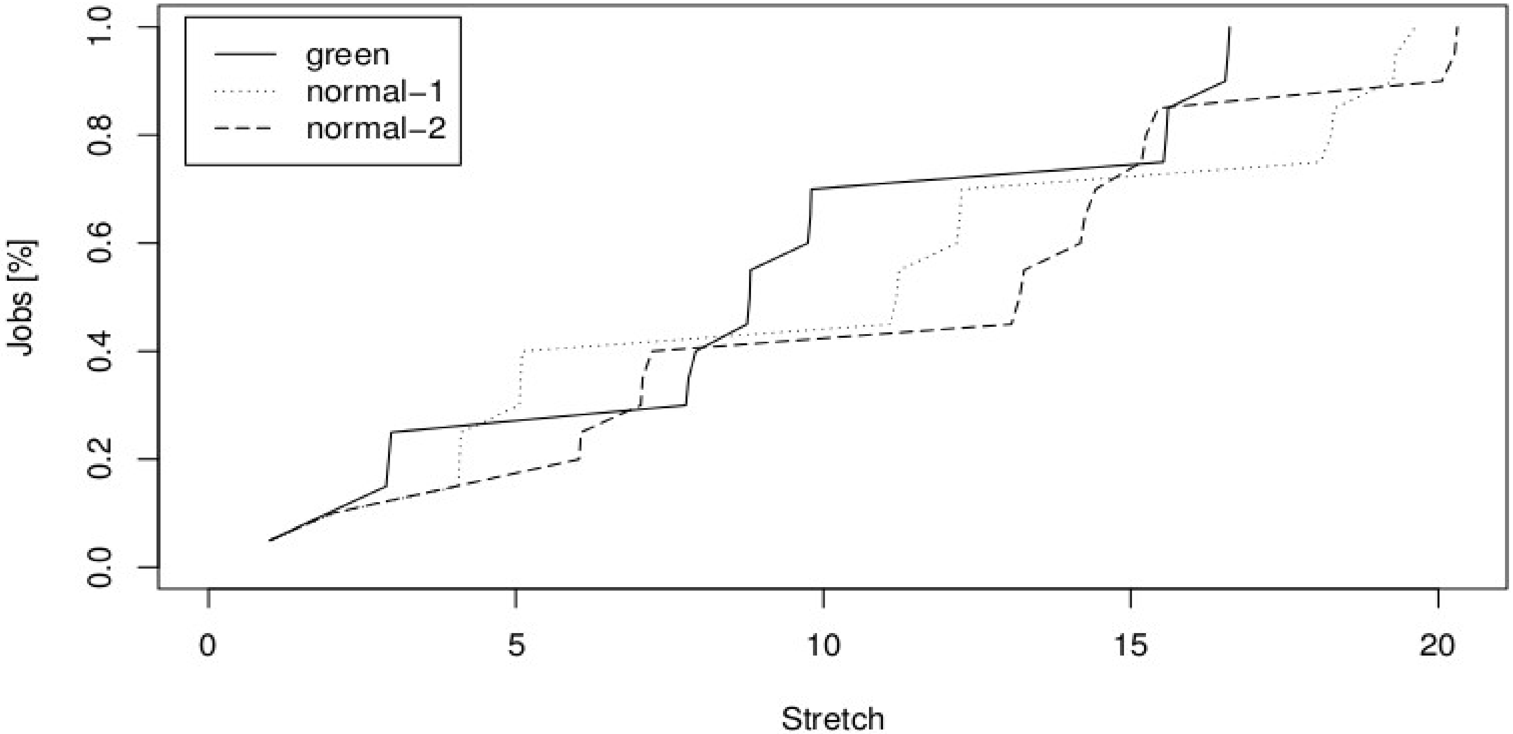


Yiannis Georgiou, David Glesser, Krzysztof Rzadca, Denis Trystram
A Scheduler-Level Incentive Mechanism for Energy Eciency in HPC
(In proceedings of CCGRID 2015)

Bull
atos technologies

Energy Fairsharing

Cumulated Distribution Function for Stretch with EnergyFairShare policy running a submission burst of 60 similar jobs with Linpack executions by 1 energy-efficient and 2 normal users (ONdemand and 2.3GHz)



Yiannis Georgiou, David Glesser, Krzysztof Rzadca, Denis Trystram
A Scheduler-Level Incentive Mechanism for Energy Eciency in HPC
(In proceedings of CCGRID 2015)

Ongoing Works – Energy Aware Scheduling

- **Workload Scheduling**
  - Consider groups of jobs and schedule those that will keep the **energy consumption stable**

- **Resources Selection**
  - Select the **best adapted resources** for lower energy consumption depending on the application profiles (data aware, topology aware, etc)
  - **Pack jobs** in order to leave parts of the cluster unused for powering off
  - Select resources based on **temperature** depending of the scope of scheduling

Summary

- **Power aware scheduling** important for your data center to adapt on the electricity prices and your energy budget
- **Energy fairsharing**: incentive for users to be more energy efficient
- Energy aware scheduling: ongoing works
- Research is published, developments open-source within SLURM
  - CPU Frequency selection parameters since SLURM 2.6 version
  - Energy measurement system plugins since 2.6 version
  - Power aware scheduling to appear in 15.08 version
  - Energy aware scheduling to appear in 16.03 version

Thanks

For more information please contact:
T+ 33 1 98765432
F+ 33 1 88888888
M+ 33 6 44445678
firstname.lastname@atos.net

12-05-2015