High definition power and energy monitoring support

Thomas Cadeau,
Yiannis Georgiou
BDS R&D, BULL AtoS.

27-09-2016

© Atos

Bull
atos technologies

Agenda

- Overview of Power and Energy Monitoring in Slurm
- Introducing HDEEM
- Experiments
- Conclusion and ongoing work

Bull
atos technologies

Power and Energy Monitoring

- What we would expect from a Resource and Job Management System
  - Attribute power and energy data to HPC components since they are resources characteristics
  - Calculate and report the energy consumption of jobs as new job characteristics
  - Extract and report power consumption time series of jobs for detailed profiling

Bull
atos technologies

Slurm Power and Energy Measurement System

- Expectations:
  - Power and Energy monitoring per node
  - Energy accounting per step/job
  - Power profiling 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)
  - Slurm internal power-to-energy and energy-to-power calculations

Bull
atos technologies

Slurm Power and Energy Measurement System

- Expectations:
  - Power and Energy monitoring per node
  - Energy accounting per step/job
  - Power profiling 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)
  - Slurm internal power-to-energy and energy-to-power calculations

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

Bull
atos technologies

Slurm Power and Energy Monitoring

- Slurm version 2.6 (September 2013)
  - Introduced monitoring plugins through IPMI and RAPL for in-band and ext-sensors for out-of-band
  - Introduced profiling plugins based on HDF5
- Slurm version 15.08 (August 2015)
  - Optimized the plugins to support the collection of multiple sensors (i.e. Blade, CPU, Memory)
  - Optimized the scalability and flexibility of HDF5 plugins
- Slurm version 17.02 (February 2017)
  - To introduce high definition power and energy monitoring support in order **to increase accuracy and minimize overhead**

Bull
atos technologies

Slurm Power and Energy Measurement System

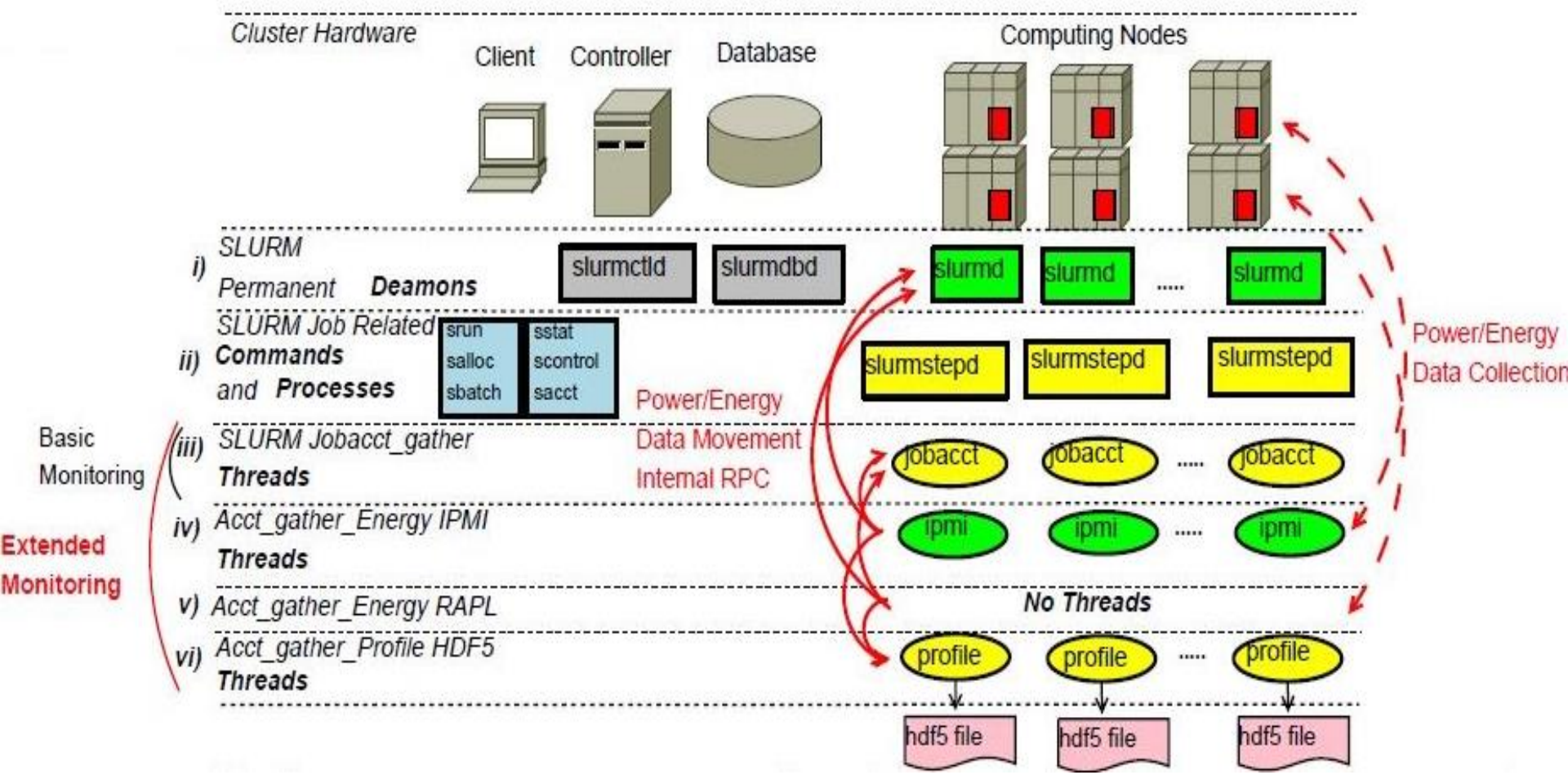
```
[root@cuzco108 bin]# $ scontrol show n=mo38 | grep ConsumedJoules
   CurrentWatts=105 LowestJoules=105 ConsumedJoules=17877

[root@cuzco108 bin]# sacct -o
"JobID%5,JobName,AllocCPUS,NNodes%3,NodeList%22,State,Start,End,Elapse
d,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
```

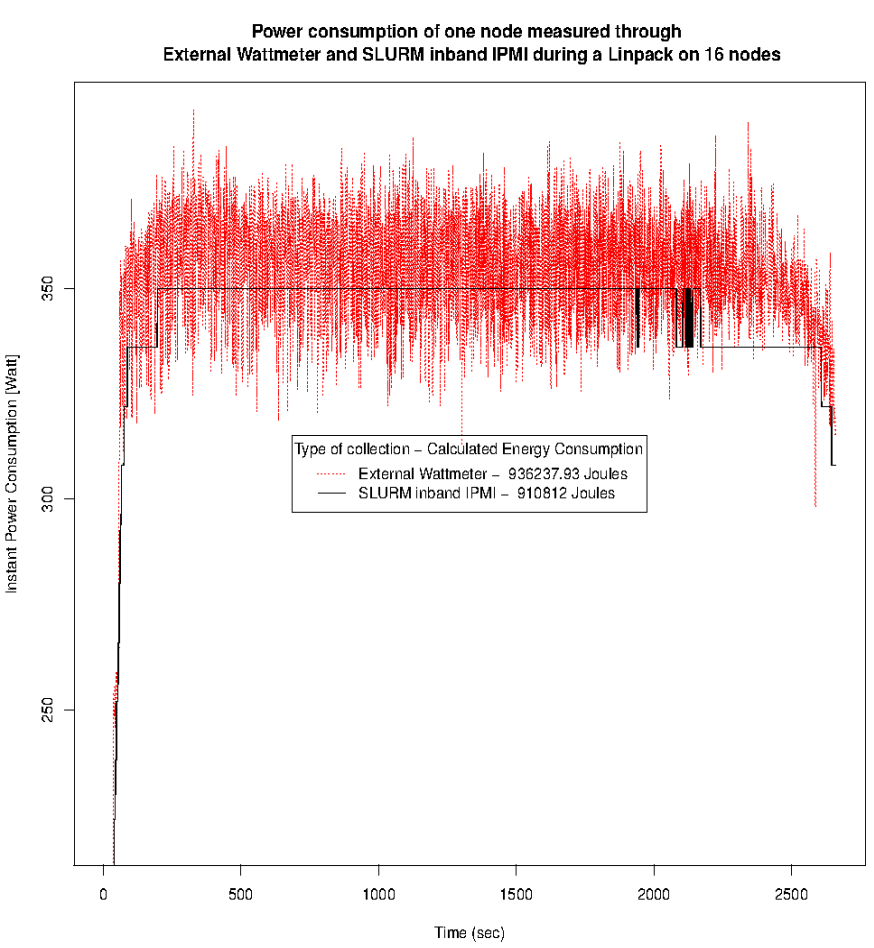

Bull
atos technologies

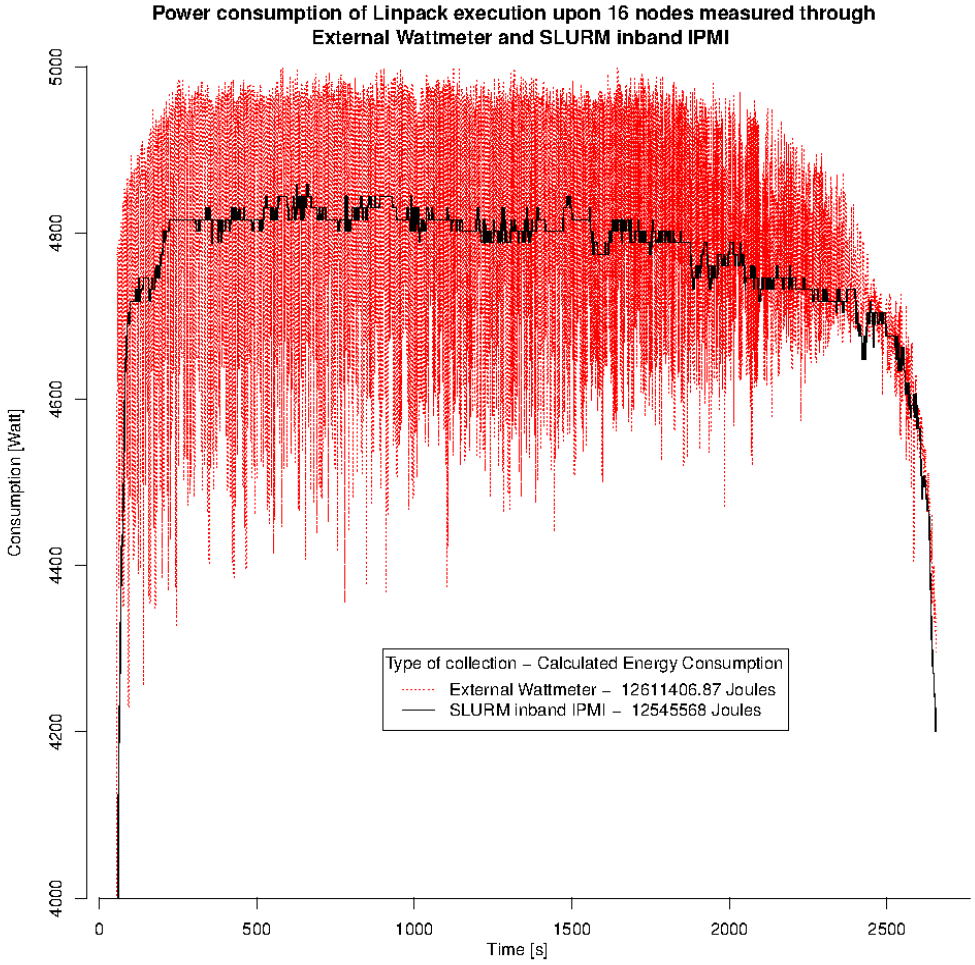
Energy Accounting and Power Profiling Architecture



Bull
atos technologies

Slurm IPMI in-band plugin Monitoring

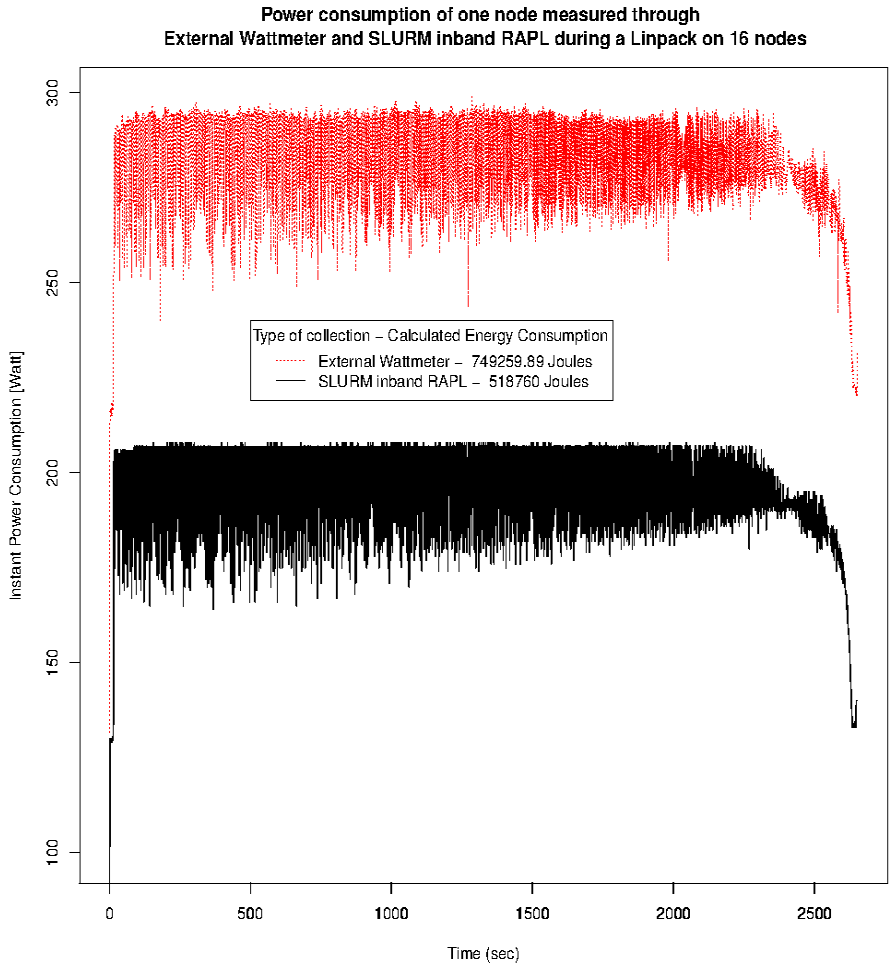


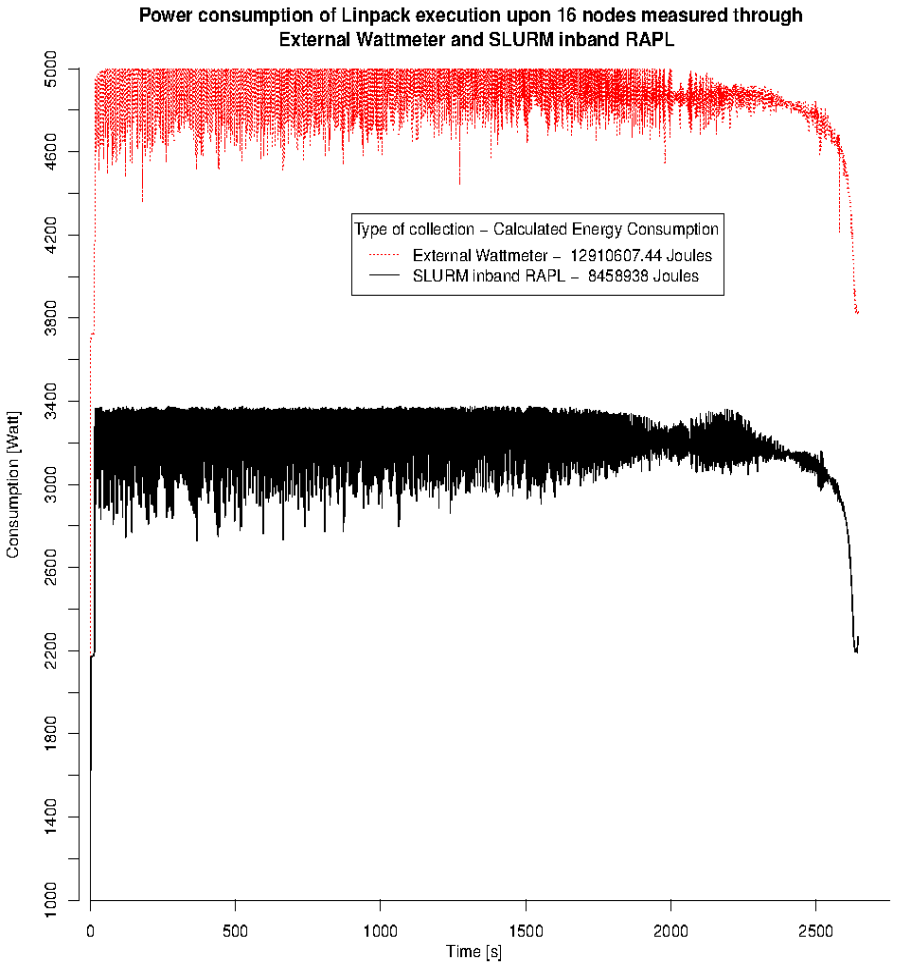


Yiannis Georgiou, Thomas Cadeau, David Glesser, Danny Auble, Morris Jette and Matthieu Hautreux
Energy Accounting and Control with SLURM Resource and Job Management System
(In proceedings of ICDCN 2014)

Bull
atos technologies

Slurm RAPL in-band plugin Monitoring

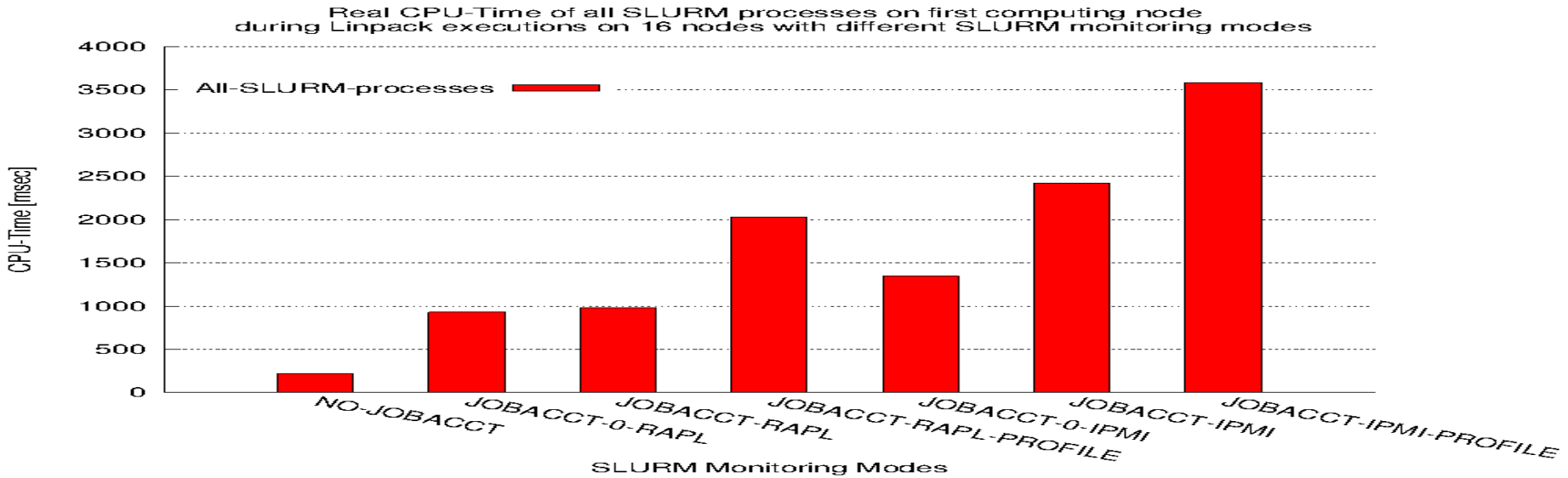


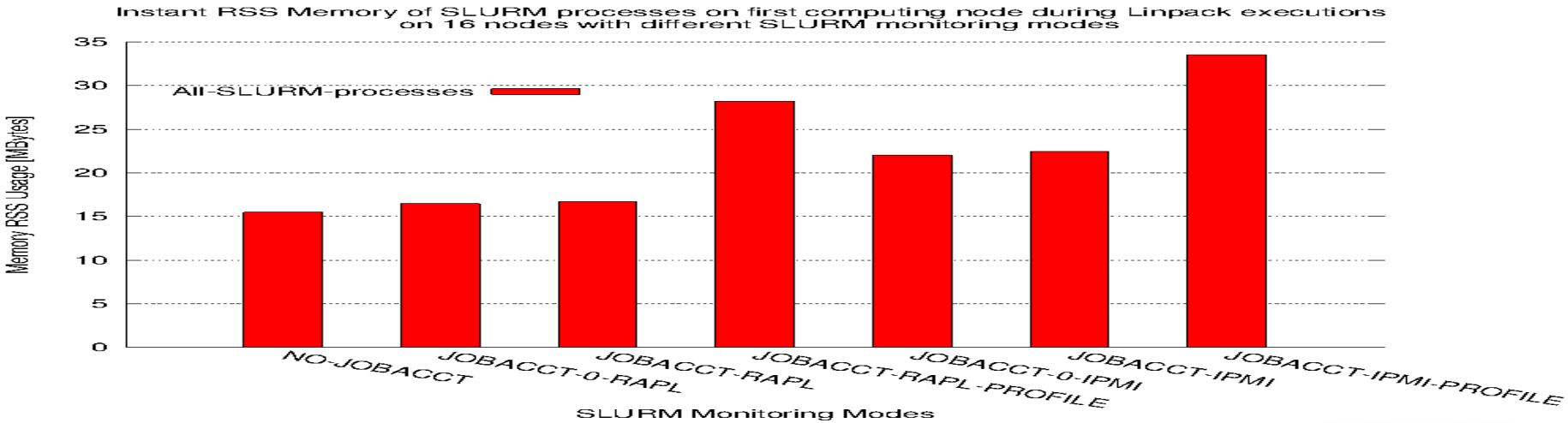


Yiannis Georgiou, Thomas Cadeau, David Glesser, Danny Auble, Morris Jette and Matthieu Hautreux
Energy Accounting and Control with SLURM Resource and Job Management System
(In proceedings of ICDCN 2014)

Bull atos technologies

Power and Energy Measurement System
CPU-Memory Overhead for IN-Band techniques





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)

Bull
atos technologies

Summary for accuracy and overhead of current Slurm version

- Accuracy for energy accounting in comparison with watt-meters
  - Good precision with IPMI
  - Good precision with RAPL (but only sockets + RAM)
- Accuracy for power profiling in comparison with watt-meters
  - Excellent precision with RAPL (but only sockets + RAM)
  - Very bad precision with IPMI
- Overhead
  - Low but not trivial overhead especially if profiling is activated
  - RAPL lower overhead than IPMI since no extra thread is needed

HDEEM: High Definition Energy Efficiency Monitoring

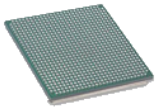
- Bull/Atos and TU Dresden collaborative project
  - 01/2013 - 12/2017
- Introduce novel power measurement tools (hardware and software)
- Allow high accuracy energy/power analyses of parallel HPC user codes

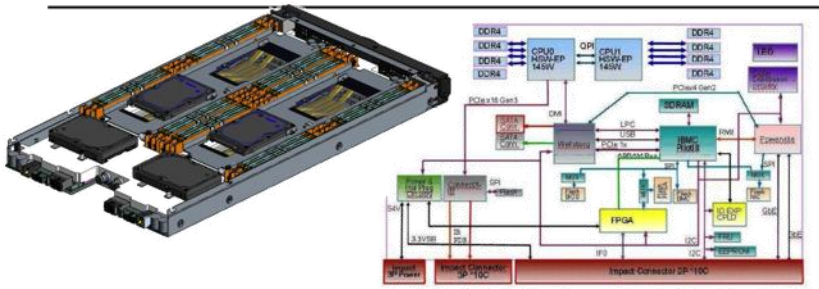
HDEEM
High Definition Energy Efficiency Monitoring

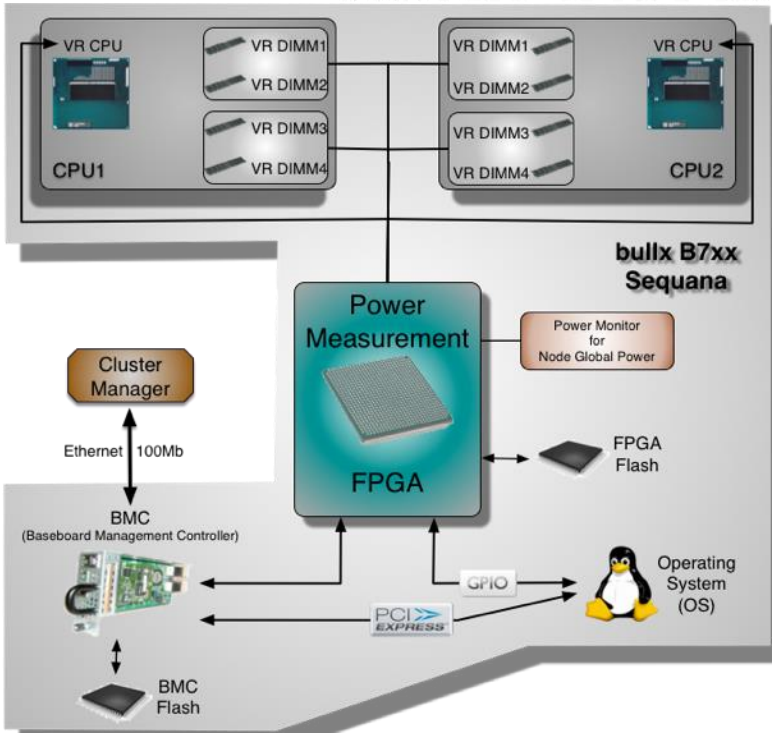
TECHNISCHE UNIVERSITÄT DRESDEN

Bull atos technologies

FPGA for power measurement



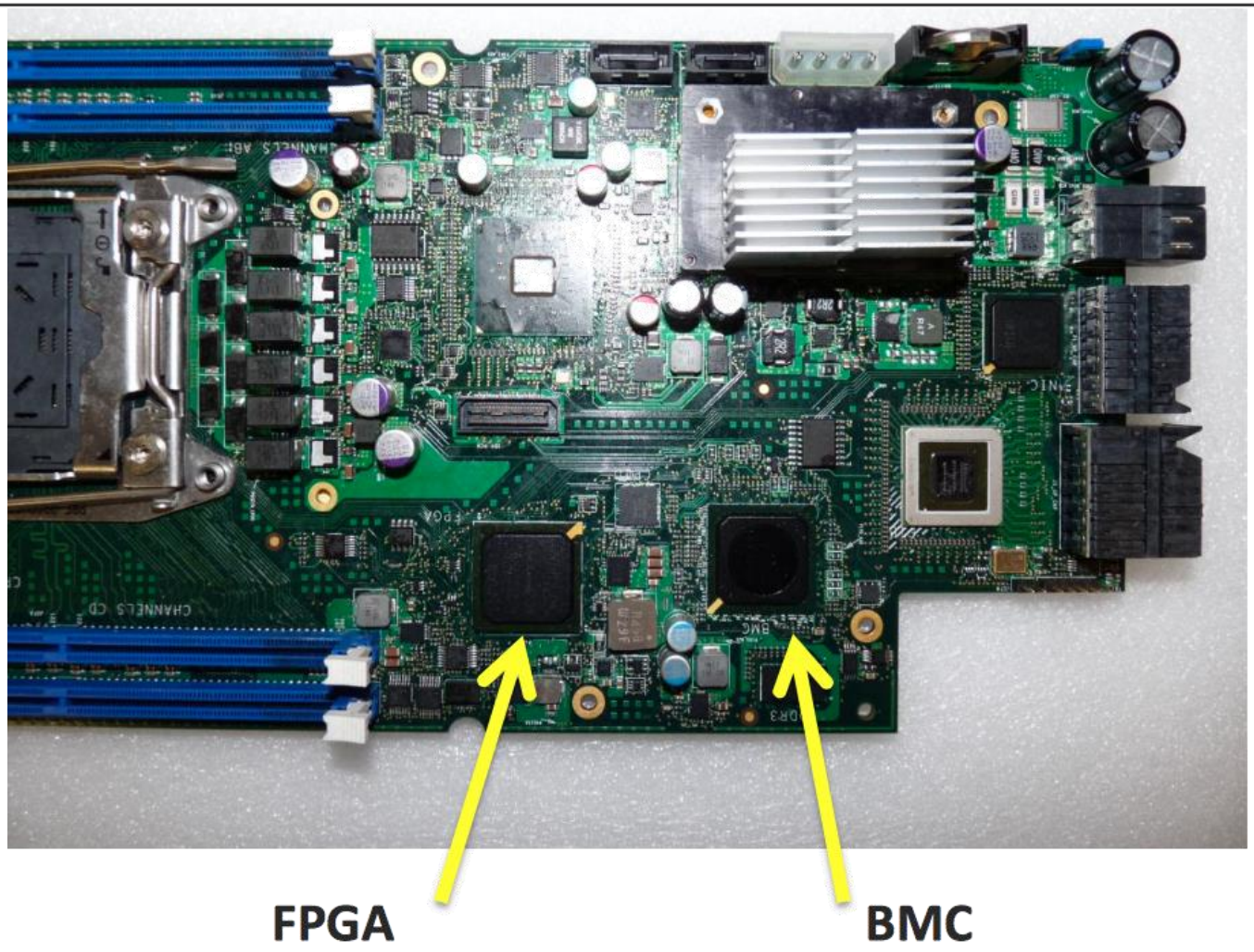




- On bullx B7xx and Bull Sequana platform a power measurement FPGA is integrated in each compute node
- Provides a sampling up to:
  - 1000 samples per second for global power including sockets, DRAM, SSD and on-board
  - 100 samples per second for voltage regulators (VR) – 6 VR: one per socket + 4 for DRAM (one / 2 lanes)
- High accuracy with 2-5% of uncertainty after calibration
  - 2% for blades
  - 5% for VR
- Time stamped measurements

Bull
atos technologies

Hardware implementation
bullx B7xx



Bull
atos technologies

HDEEM with FPGA power measurement

High Definition Energy Efficiency Monitoring

- C API ease to use to gather power data
  - Start / Stop / Print / Check / Clear
- Goal is to be able to integrate power measurement in application performance traces tool(s) and also in resource manager accounting and profiling without performance overhead
  - Measurement and buffering is done on BMC side for several hours

```
HDEEM_VERSION, 2.2.17ms

BMC address, localhost

==== HDEEM status ====          ,
Last start time for blade       , 2016-09-23 16:33:23.000
Last start time for vr          , 2016-09-23 16:33:22.970
Started by                      , GPIO
....
Total blade values              , 16038
Pending blade in BMC            , 16038
Total VR values                 , 1603
Pending VR in BMC               , 1603
----------
          BLADE,
     1,   172.750,
     2,   175.125,
     3,   158.000,
     4,   164.375,
     5,   164.750,
     6,   152.875,
     7,   156.250,
```

Power and Energy through Slurm HDEEM plugin

- High Definition energy efficiency monitoring based on new FPGA architecture supported through ipmi-raw
  - Improved accuracy for both power profiling per components (100Hz) and nodes (1000Hz)
  - Improved precision for energy consumption per job based on nodes (1000Hz) measurements
  - Decrease overhead on the application (CPU and Memory) since the collection is done internally within the FPGA

```
==== HDEEM statistics total from power awake ====
Time of total stats for blade      : 2016-11-26 20:41:10.373
Time of total stats for vr         : 2016-11-26 20:41:10.343
Blade values total                 : 5548045661
VR    values total                 : 554798998

          Average (W)       Energy (J)
  BLADE       87.578     485884493.940

   CPU0       25.959     144019836.416
   CPU1       22.141     122835530.217
 DDR_AB        1.287       7138687.771
 DDR_CD        1.313       7284187.843
 DDR_EF        0.851       4720172.055
 DDR_GH        2.349      13031283.751
```

HDEEM
High Definition Energy Efficiency Monitoring

Bull
atos technologies

Implementation for HDEEM in Slurm

- No specific algorithm for energy
  - Energy is used as return by BMC
  - Energy is the difference from current value and first value
  - No overhead since we only collect twice, once in the beginning and once in the end of job
- Power is calculated/extracted using the energy sensor value
  - We do not report events but real profile
- Multi session counter
  - slurmd (for node *ConsumedJoules*)
  - slurmstepd (for energy accounting and profiling)
  - any other usages (including user)
- Separation between the 2 counters
  - Any user/application can use the very high freqency power without conflicts

Bull atos technologies

Configuration/Usage

- slurm.conf
  - *AcctGatherEnergyType=acct_gather_energy/**hdeem***
  - *AcctGatherNodeFreq=25*
  - *AcctGatherProfileType=acct_gather_profile/hdf5*
  - *JobAcctGatherFrequency=10,energy=1*

- acct_gather.conf
  - *EnergyHDEEMItems=Node=BLADE;Cpus=CPU0,CPU1*
  - Commponents: CPU0, CPU1, DDR_AB, DDR_CD, DDR_EF, DDR_GH

- srun/sbatch/salloc
  - Nothing specific for accounting
  - Profiling: *--profiling=Energy --acctg-freq=energy=1*
  - ***Note:*** jobs are not disturbed by ipmi calls and there is no specific thread

Bull atos technologies

Experiments with BMC 4Hz

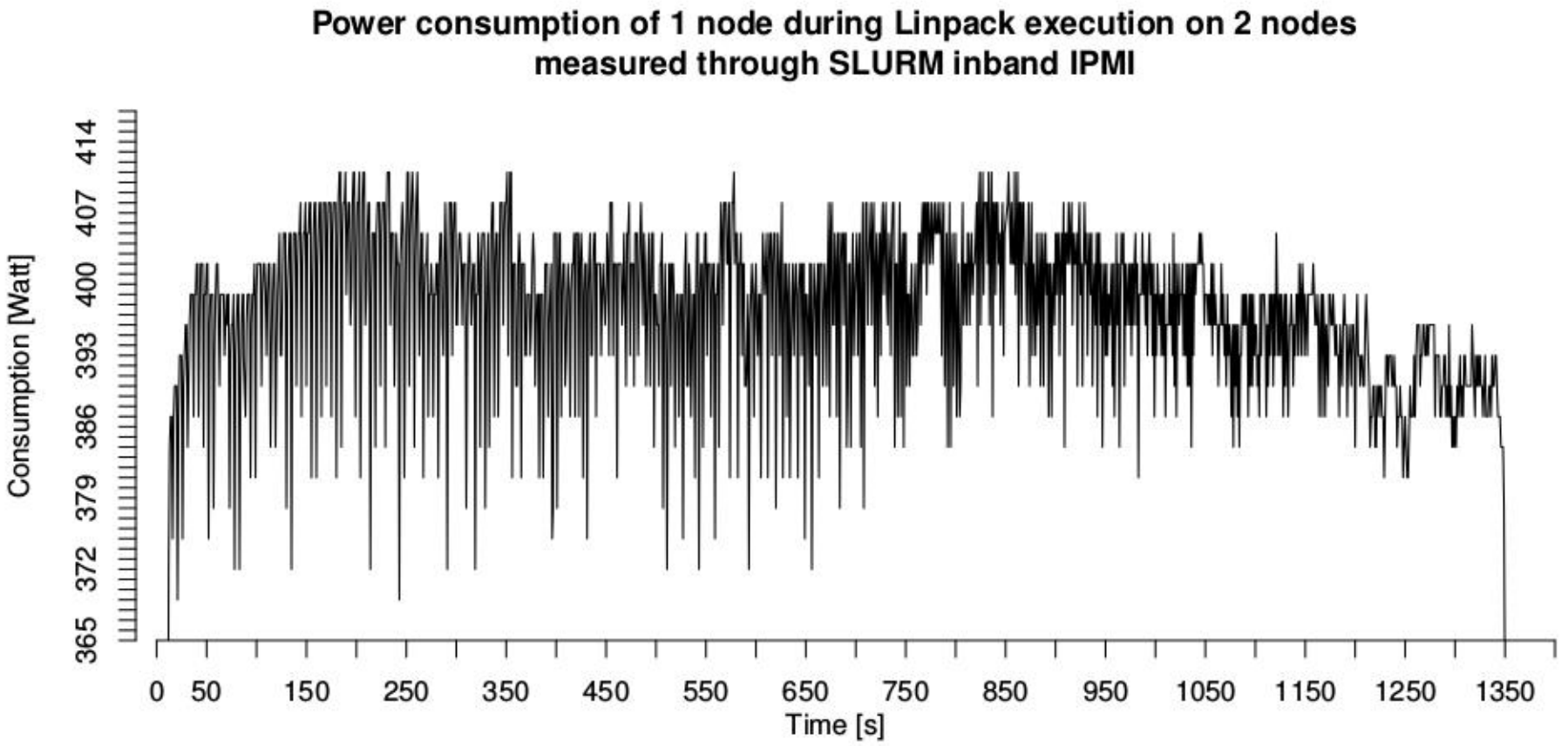
- Intermediate HDEEM version with only BMC 4Hz optimization (no FPGA) showed very promising results

Published in:

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

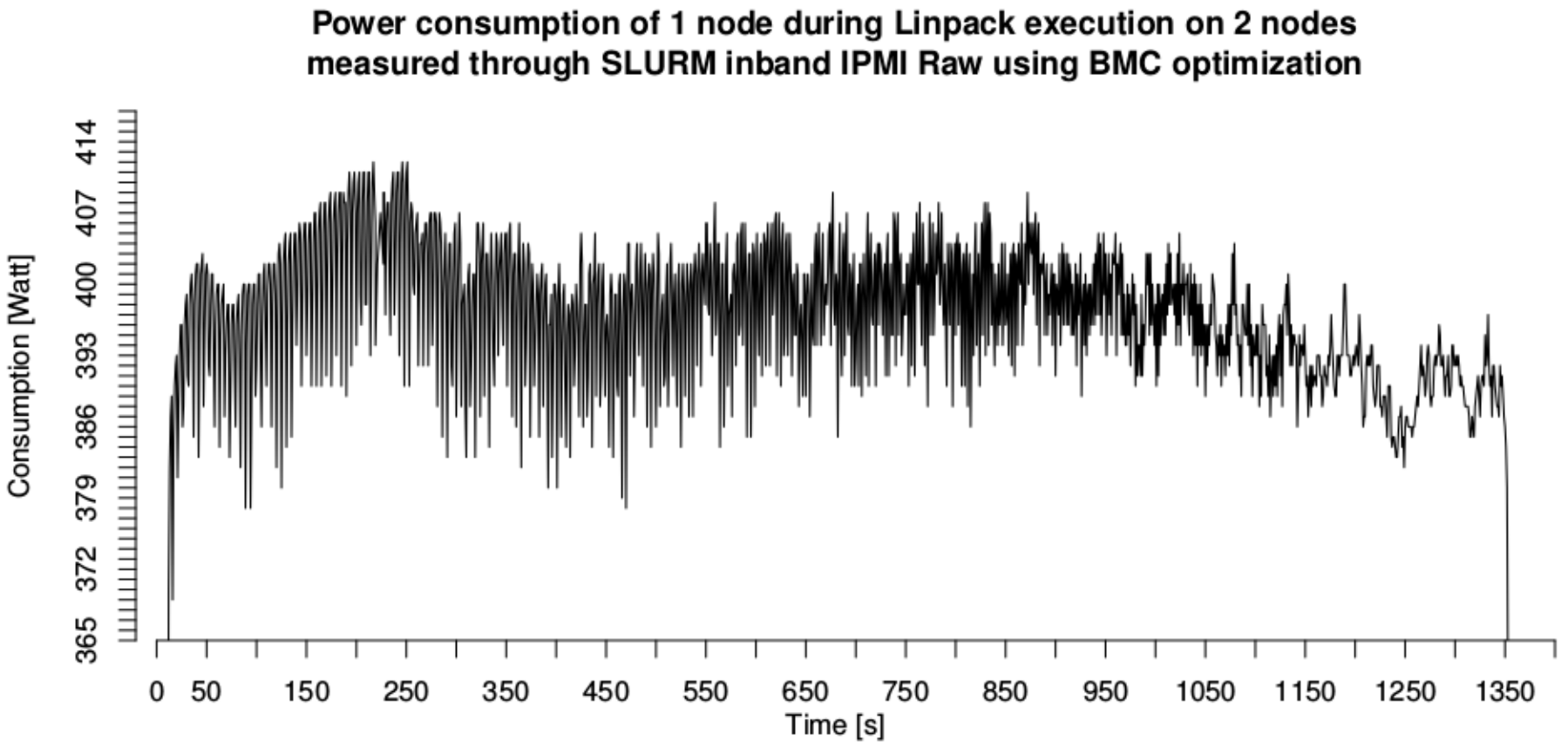
Bull atos technologies

Optimizations of Power and Energy Measurement System



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

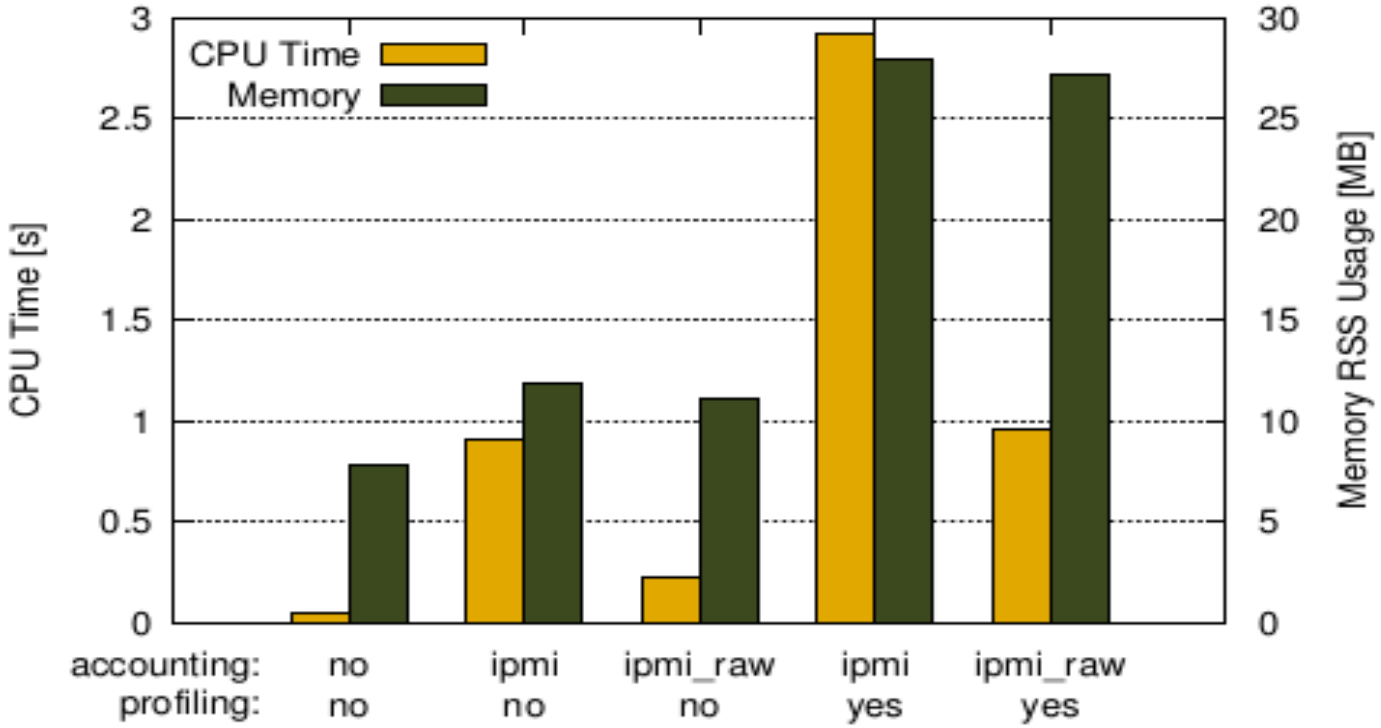
Optimizations of Power and Energy Measurement System



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

Optimizations of Power and Energy Measurement System

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



TECHNISCHE UNIVERSITÄT DRESDEN

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

Bull atos technologies

Conclusion

- Ongoing experiments for HDEEM with FPGA 1000Hz
  - Currently in test on Bull and TU Dresden HPC clusters
  - Promising first results (plan for publication)

- To be considered for Slurm version 17.02:
  - HDEEM library open source
  - Would work only for Bull hardware (with FPGA) but plugin can be used as base for similar optimizations since using Freeipmi calls to BMC which is a standard

- Very high precision on energy and power
- No overhead for application since no thread, polling within FPGA
  - information available for any other tool
  - For accounting and profiling (with no dependancy on Slurm frequencies)

Bull
atos technologies

THANKS

Atos, the Atos logo, Atos Consulting, Atos Worldgrid, Worldline, BlueKiwi, Bull, Canopy the Open Cloud Company, Yunano, Zero Email, Zero Email Certified and The Zero Email Company are registered trademarks of the Atos group. March 2016. © 2016 Atos. Confidential information owned by Atos, to be used by the recipient only. This document, or any part of it, may not be reproduced, copied, circulated and/or distributed nor quoted without prior written approval from Atos.

27-09-2016

Bull
atos technologies