Towards energy consumption application profiling with Bull Energy software

Xavier Vigouroux and Ludovic Saugé

Tuesday, November 14th, PRACE booth 217

Trusted partner for your Digital Journey

Bull
atos technologies

PCP - **Whole-System Design for Energy Efficient HPC**

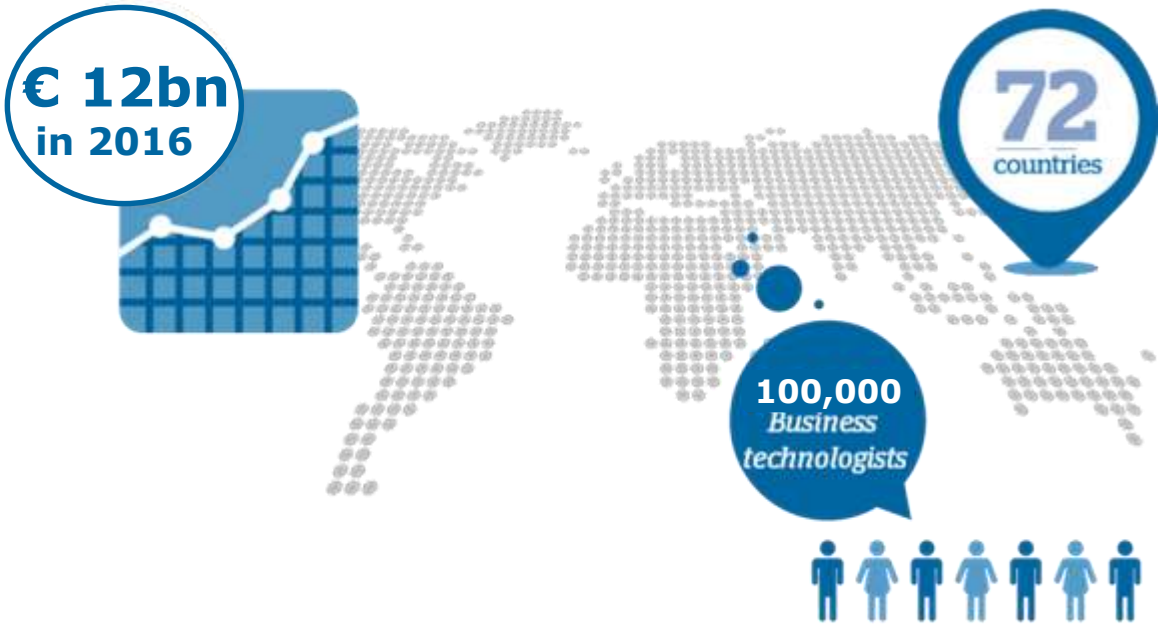
- Multi-country and multi-partner joint effort

- PCP aims to foster **innovation for economic growth** to **ensure sustainable high quality public services in Europe**
- **Addressing the major issue of the energy efficiency of large scale supercomputers.**
  - Improving energy efficiency of computing systems and reducing their environmental impact

Bull atos technologies

We are THE European IT Leader

and a top 5 Digital services player worldwide



Atos

worldline
e-payment services

Bull
atos technologies

UNIFY

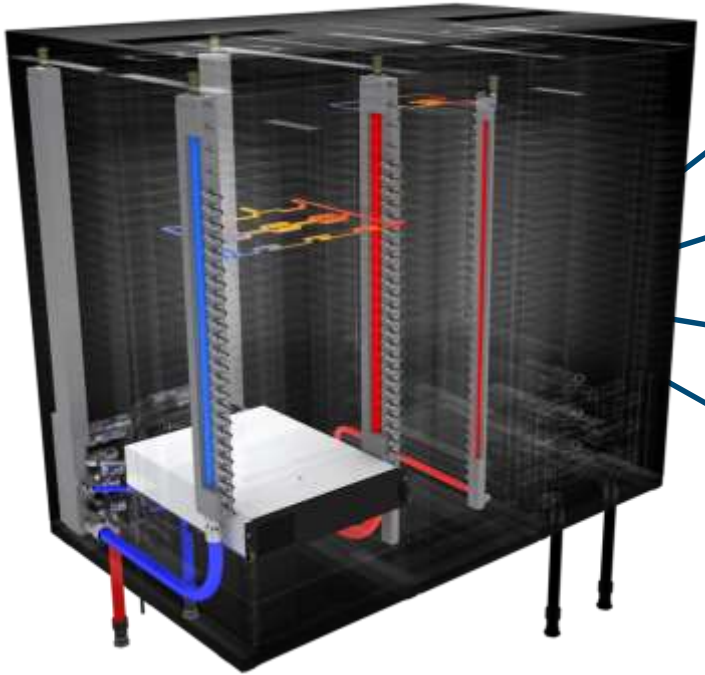
Bull
atos technologies

Energy efficiency and Power management
Building blocks

Based on Bull Sequana X1000 Exascale flagship

- **Open and modular** platform designed for the long-term
  - To preserve customer investments
  - To integrate current and future technologies
- **Ultra-energy efficient**
  - Enhanced DLC – up to 40°C for inlet water and ~98% DLC
- **Scales up to tens of thousands of nodes**

Energy efficiency and Power management
Building blocks



Enhanced version of the Bull Direct Liquid Cooling (DLC) solution
100% of the components water cooled, including PSUs

Dynamic Resource Reconfiguration for energy efficiency
Addon to bull SCS5 for Smart energy management sub-system

Energy-aware system
Fine-grain energy sensors
New generation of the HDEEM technology

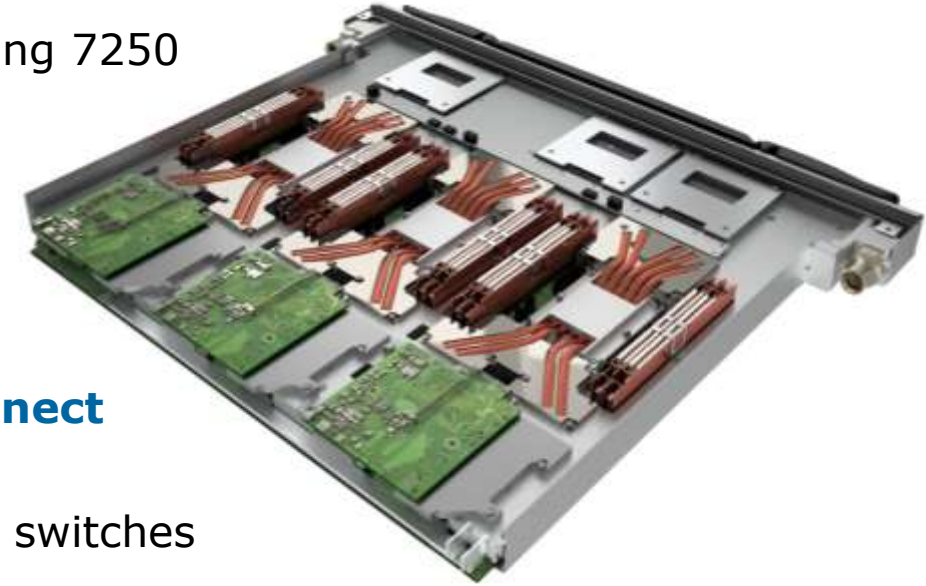
Application Co-design
Optimisation of a set of scientific applications

HDEEM
High Definition Energy Efficiency Monitoring

Prototype configuration

Based on Sequana platform – the open exascale class supercomputer

- **56x Bull sequana X1210 Intel® Xeon KNL blades (511 TFLOP/s)**
  - 3 nodes per blade and per node:
    - 1x Intel® Xeon-Phi® Knights Landing 7250 16GB GDDR5 215W Passive
    - 6x 16GB@2400MT/s DDR4 DIMMs
    - 1x 240GB 2.5" 7mm SATA3 SSD
    - 1x InfiniBand EDR connection



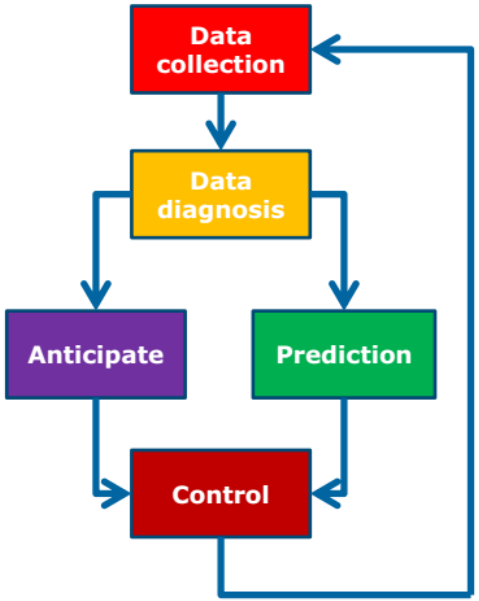
- **InfiniBand EDR High Speed Interconnect**
  - 2 levels fat-tree topology
  - Based on 36 ports per InfiniBand EDR switches
    - oversubscription ratio 2:1 (L1/L2)

Energy efficiency and Power management

Building blocks

Bull Power Management Framework

- **HDEEM** (Hardware component)
- Bull Energy Optimizer (**BEO**)
- **Slurm** energy plugins
- High Definition Energy Efficiency VIsualiZation (**HDEEVIZ**)
- Bull Dynamic Performance Optimizer (**BDPO**)



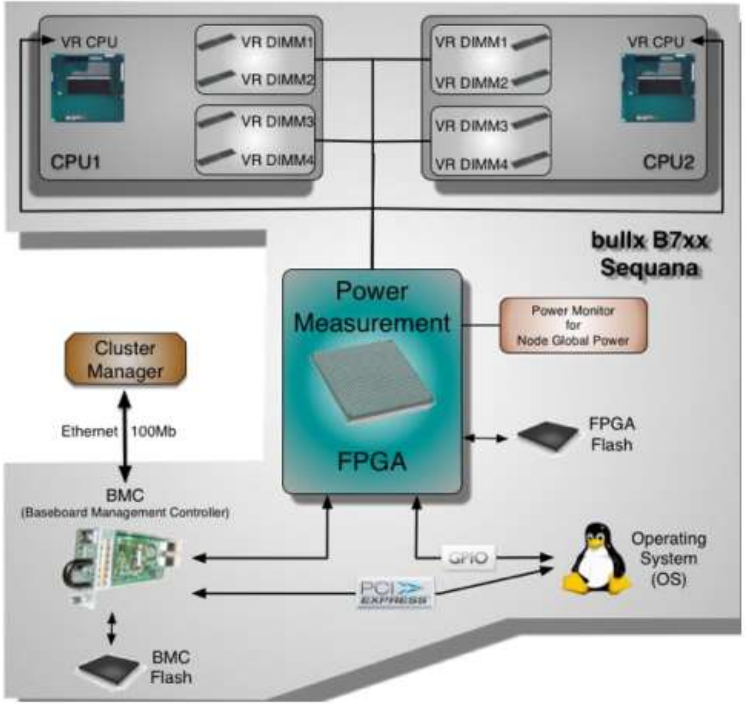
Energy efficiency and Power management

HDEEM

Bull Sequana X1000

2nd generation supporting a power measurement FPGA integrated in each compute node

- Provides a **sampling rate** up to:
  - **1 kHz for global power** including sockets, DRAM, SSD and on-board
  - **100 Hz for voltage regulators**
- **High accuracy with 2-5% of uncertainty** after calibration
  - 2% for blades
  - 5% for VR
- **Collection modes**
  - **Out-of-band** using BMC (4 Hz)
  - **In-band** through HDEEM API (eg SLURM)
- Time stamped measurements



Bull
atos technologies

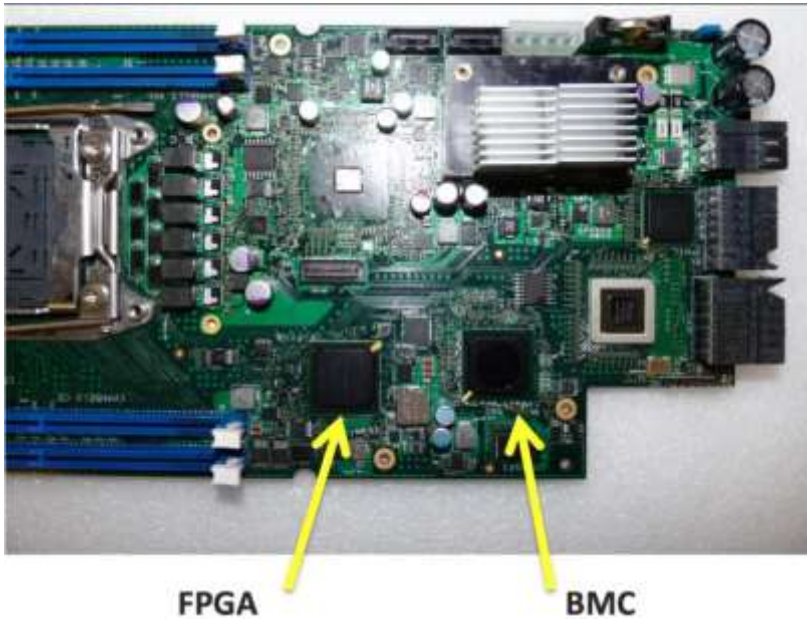
Energy efficiency and Power management

HDEEM

Bull Sequana X1000

2nd generation supporting a power measurement FPGA integrated in each compute node

- Provides a **sampling rate** up to:
  - **1 kHz for global power** including sockets, DRAM, SSD and on-board
  - **100 Hz for voltage regulators**
- **High accuracy with 2-5% of uncertainty** after calibration
  - 2% for blades
  - 5% for VR
- **Collection modes**
  - **Out-of-band** using BMC (4 Hz)
  - **In-band** through HDEEM API (eg SLURM)
- Time stamped measurements



Energy efficiency and Power management

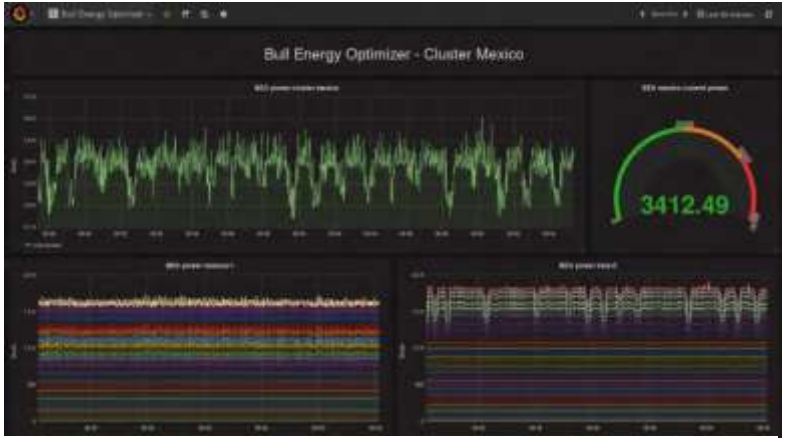
BEO (Bull Energy Optimizer) and Slurm Plugin

- **Infrastructure-centric**: **Bull Energy Optimizer** (BEO v1.0)
  - **Out of band**: no perturbation of running applications
    - IPMI for compute nodes
    - SNMP for switches
  - Multiple metrics are supported: power, energy and more
  - Follow the consumption of all cluster components
  - Follow the energy/power consumption of a job: **compute + network**

- **Job-centric: SLURM RJMS**
  - Follow the energy consumption of a job: **compute only**
  - Offers profiling capabilities to follow various metrics:
    - Compute: Time
    - Storage: I/O statistics
    - Communication: InfiniBand

Energy efficiency and Power management

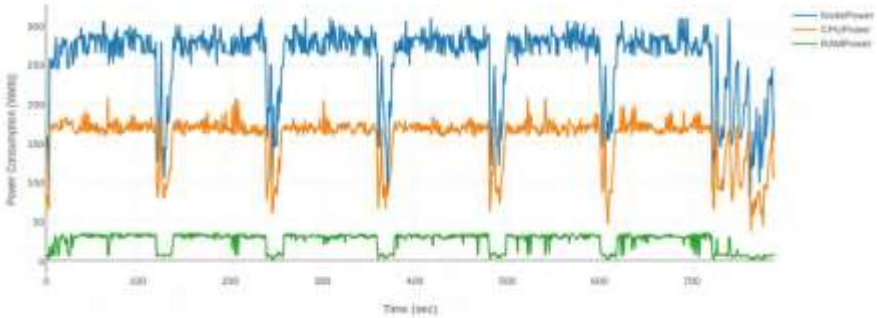
BEO (Bull Energy Optimizer) and Slurm Plugin



- **Infrastructure monitoring** based on **BEO**
  - BEO allows to follow power or energy of each:
    - Compute node (IPMI)
    - Switch (Ethernet + IB)

- **Energy job profiling** achieved through SLURM
  - Energy consumed by a job collected for
    - Compute node
    - CPUs
    - RAM

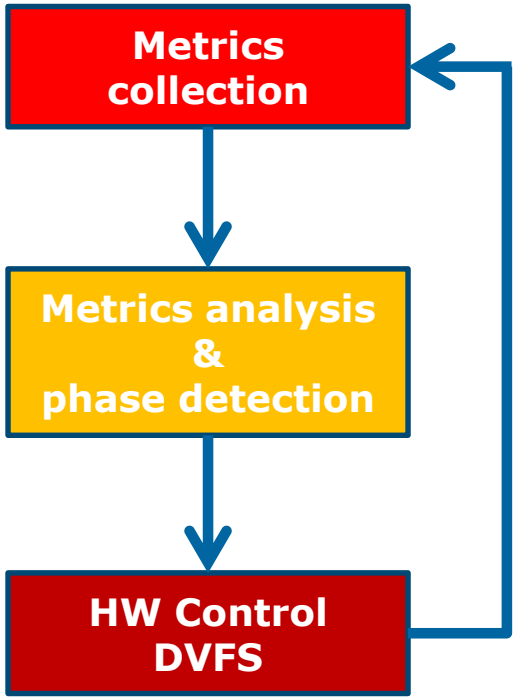
Node, CPU and RAM power consumption for 1 node during a NEMO execution upon 8 nodes (Default Run)



Energy efficiency and Power management

BDPO: Bull Dynamic Performance Optimizer

- **Extensible framework balancing performance and energy consumption**
  - Target real HPC applications
  - Avoid application intrusiveness: no source code modification
  - No/limited performance degradation



Energy efficiency and Power management

HDEEVIZ

- **HDEEVIZ uses HDEEM to retrieve data consumption at fine grain**
  (in-band collection)

- **HDEEVIZ is a tool for power consumption data vizualisation for end users to:**
  - **get global energy of a job**
  - **detect power consumption phases**
  - **get CPU and memory power consumption** % per phase
  - **compare** similar jobs (CPU freq. dependency, parameters tuning for energy...)

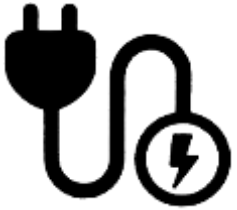
Energy efficiency and Power management

HDEEVIZ – Next step ?

Optimize performance metrics

Time-to-solution

Power-to-solution



Application set

- four scientific applications
  - BQCD
  - NEMO
  - Quantum Espresso
  - SpecFEM3D
- High Performance Linpack

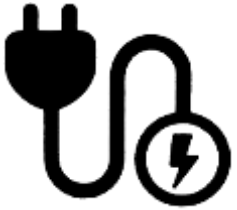
Preserve end user perspective

- No changes of the algorithms are allowed.
- Maximum of 10% of the source code may be modified.

Optimize performance metrics

Time-to-solution

Power-to-solution



Code Modernization

Vectorization

Memory usage

- Hybrid programming, mixing MPI and OpenMP
- Vectorization, through AVX-512 instruction set
- Optimize memory usage and especially the MCDRAM (fast memory)

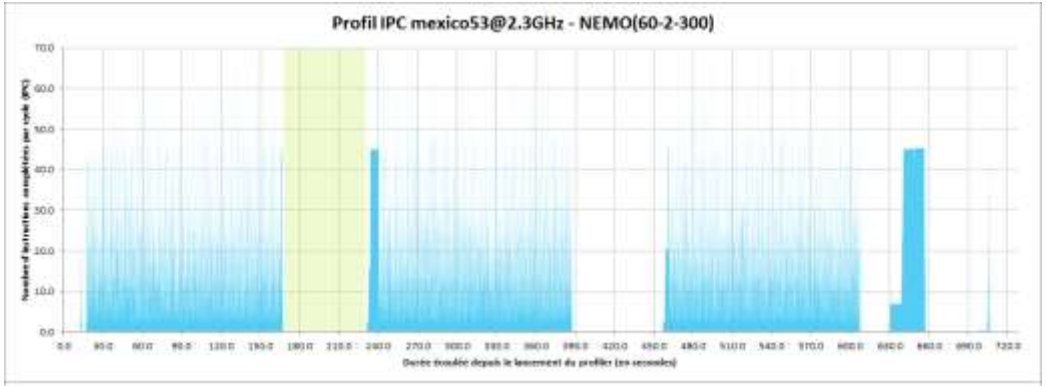
BDPO first results | NEMO test case

- What is NEMO?
  - Nucleus for European Modelling of the Ocean
  - Oceanic simulation framework
  - Developed by France, Italia and the UK (CNRS, CMCC, Nerc ...)
  - Open source

- Tunable I/O checkpoints
- Size and frequency of checkpoints
- Regular compute-I/O phases alternation

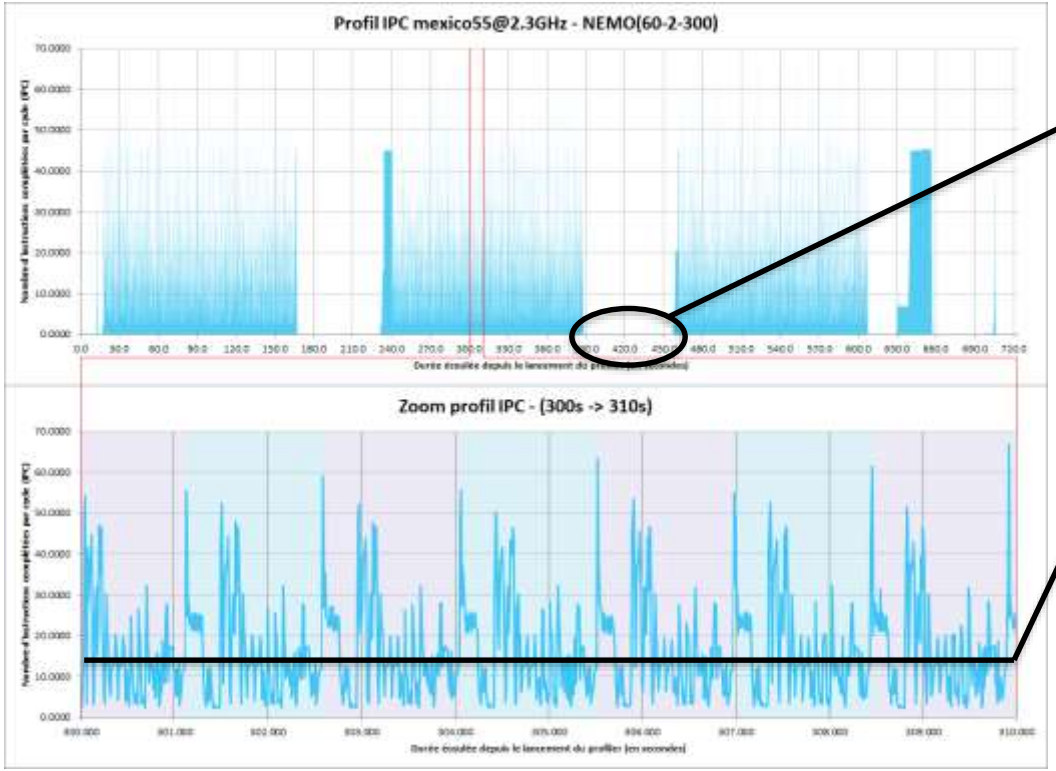
BDPO first results | NEMO test case

IPC and Power



BDPO first results | NEMO test case

IPC level identification



No gain on IO Phases:
→HW does the job
(moving CPU on idle C-State)

Focus on CPU/Memory phases
→ A low IPC level indicates a memory phase
→ CPU freq can be decreased

BDPO first results | NEMO test case

Energy saving with BDPO v1 on NEMO

- IPC level is set to 12.5
- under 12.5, f_{low} = 1.5 GHz and and upper 12.5, f_{high} = 2.0 GHz

	NEMO+BDPO	CPU governor Max=2,3GHz	CPU governor ondemand
Energy consumption - BEO (in Joules)	635 964	693 704	718 520
Execution time (in seconds)	521	516	519

BDPO first results | NEMO test case

Energy saving with BDPO v1 on NEMO

- IPC level is set to 12.5
- under 12.5, f_{low} = 1.5 GHz and and upper 12.5, f_{high} = 2.0 GHz

	CPU governor Max=2,3GHz	CPU governor Ondemand
Energy consumption (BEO) gain with NEMO+BDPO compared to …	**8.32 %**	**11.49 %**
Execution time loss with NEMO+BDPO compared to …	1.00 %	0.42 %

As a matter of conclusion …

Bull Power Management Framework

INTEGRATED IN ROADMAP

- BEO
- Slurm plugins
- HDEEVIZ
- BDPO

As a matter of conclusion …

Thanks

For more information please contact:
ludovic.sauge@atos.net
xavier.vigouroux@atos.net

Bull
atos technologies