



Architect of an Open World™

Slurm BOF SC13

Bull's Slurm roadmap

SC13

Eric Monchalin

Head of Extreme Computing R&D





bullx bm

- bullx MPI integration (runtime)
 - *Automatic Placement coherency*
 - *Scalable launching through PMI2*
- bullx DE integration (Development environment)
 - *Debuggers, Profilers, Scientific Libraries*
 - *bullx Prof*
- bullx MC integration (Management Center)
 - *Topology design generation*
 - *Global High Availability services*
 - *Infrastructure Energy collection*

- Slurm 2.6

Bull's contributions:

- Scalability improvements
- Performance improvements
- Power Management facilities
- Accounting facilities



Largest Bull supercomputers powered by Slurm



TERA100 – 2010

1st European PetaFlop-scale System

Rank #6



CURIE – 2011

1st PRACE PetaFlop-scale System

Rank #9



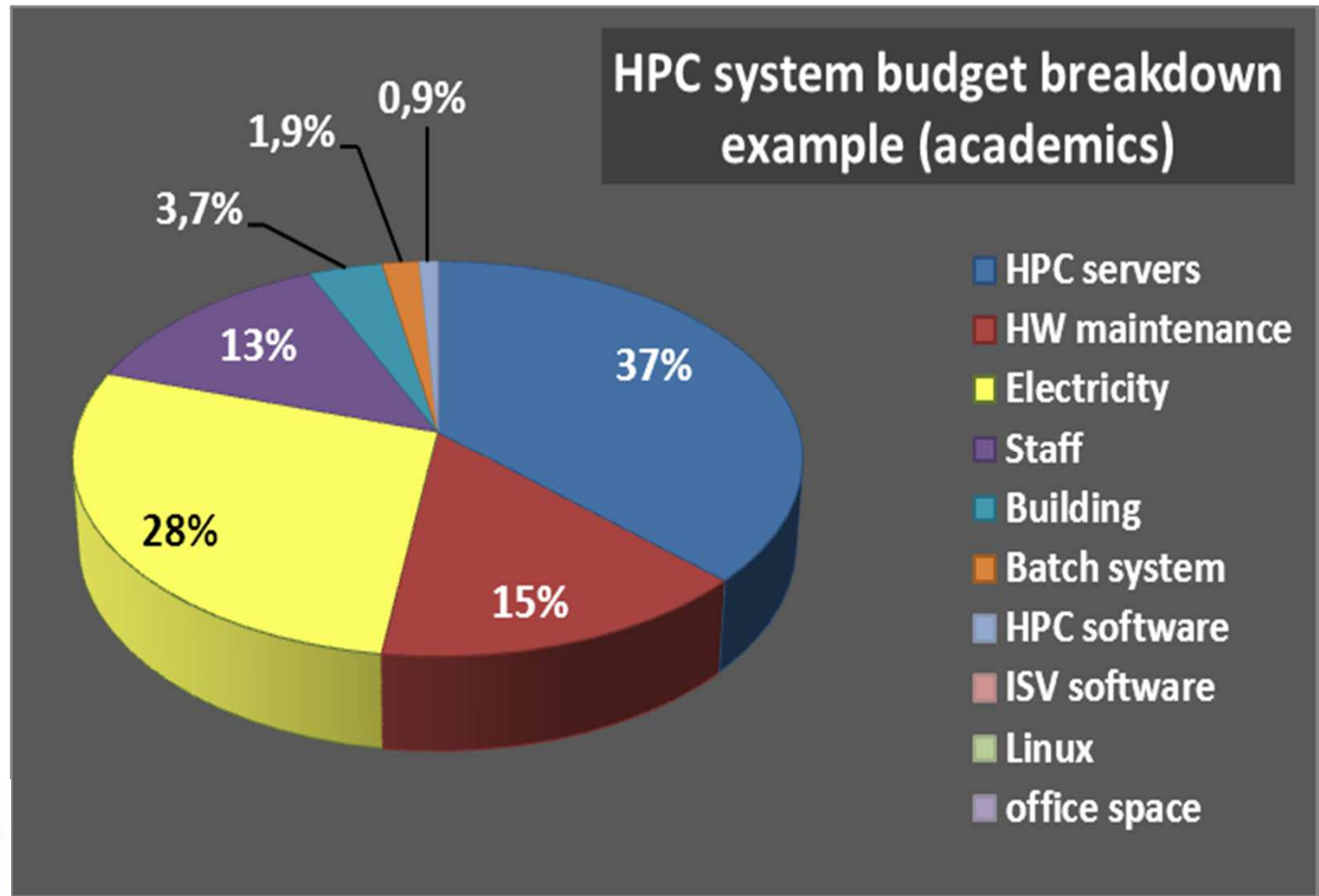
BEAUFIX – 2013

1st Intel Xeon E5-2600 v2 System

Direct Liquid Cooling Technology



2013: Stay focus on Power Management



Power Management

Accounting

- Users billed separately for CPU, IO, ... and Energy
- Keep compute center electricity bill within budget

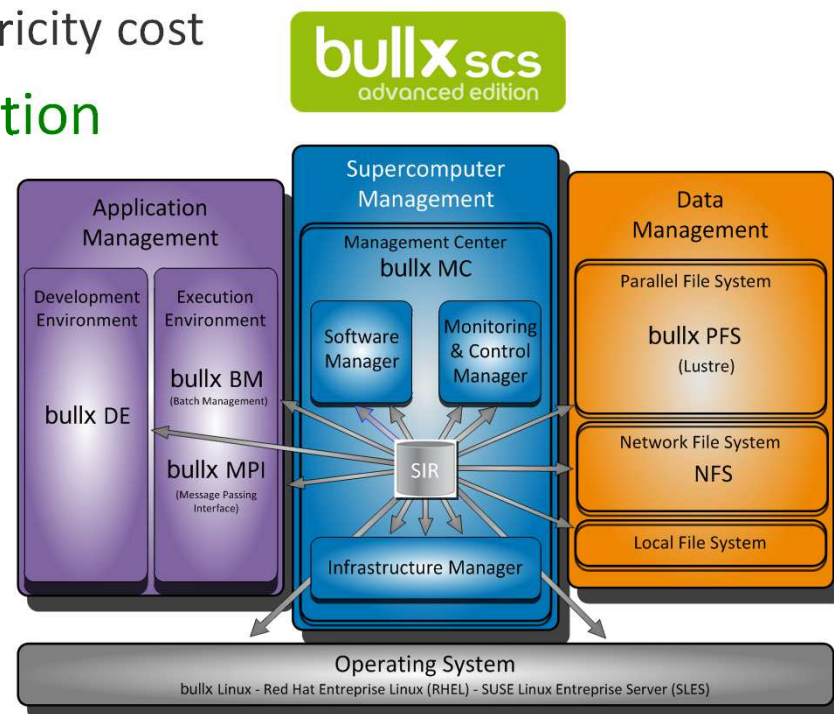
Control power

- Avoid running over capacity
- Allow for priority jobs
- Adjust power consumption with electricity cost

Energy consumption / cost optimization

- Fine & precise power monitoring
- Power data analysis
- Control all system resources power

... Enter in Software



Slurm Power Management

Monitoring

- New framework to allow per job energy consumption and node power monitoring

– With different capturing mechanisms

- RAPL for Sandy Bridge processors

```
scontrol show node=node1 | grep Consumed  
CurrentWatts=105 ConsumedJoules=9114853
```

- IPMI

- External RRD bases from bullx MC or external tools

```
scontrol show node=node2 | grep ExtSensors  
ExtSensorsJoules=7156821 ExtSensorsWatts=95 ExtSensorsTemp=72
```

– Precision of few seconds depending of capturing mechanisms

Slurm Power Management

Reporting

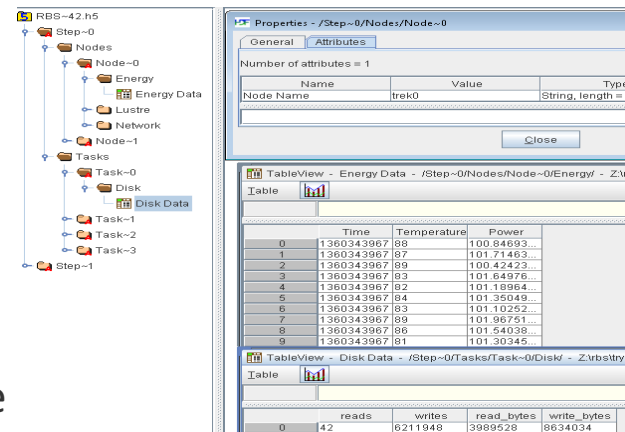
- Job, user and group accounting for general consumption information

```
>sacct [format options]
```

```
JobId Elapsed ConsumedEnergy
```

```
15 00:01:11 16302
```

- Job detailed consumption information through detailed files reporting time/watts/energy per node
 - Based on **HDF5** data model which is a structured file format for storing and managing data
 - One collection for each node, One for each job
 - Collected data specifies through srun parameter
 - Compatible with all HDF5 tools
 - Lustre, CPU, and Memory profiling are also available

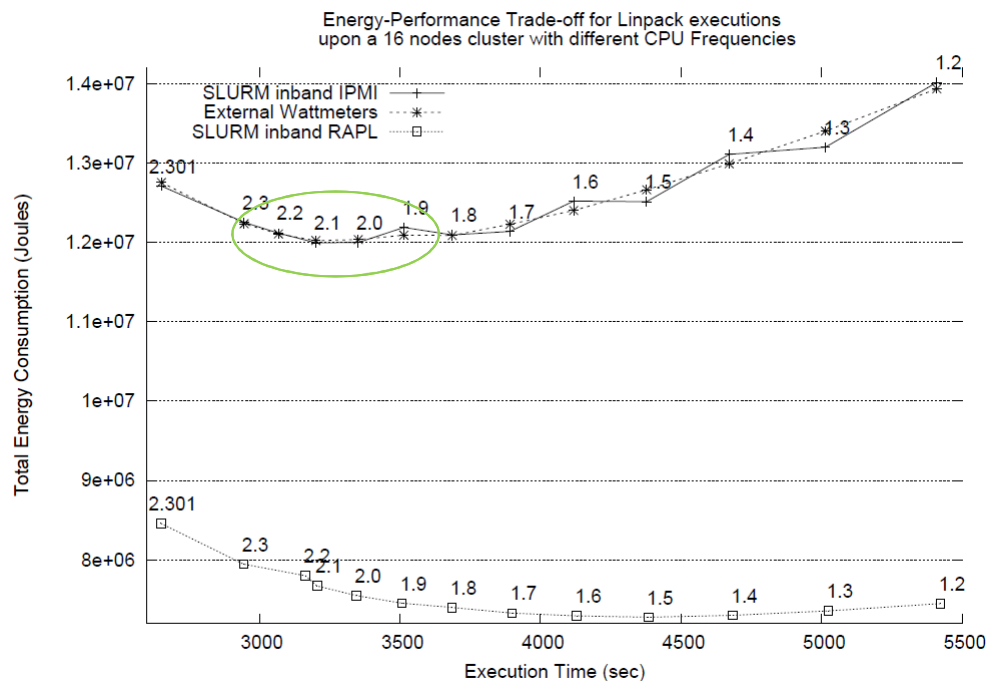


Controlling

- New srun parameter to allow CPU frequency scaling for job execution
 - Reporting of step's average CPU frequency and energy consumption
- Job energy consumption as a new factor in scheduling (Available in next version)

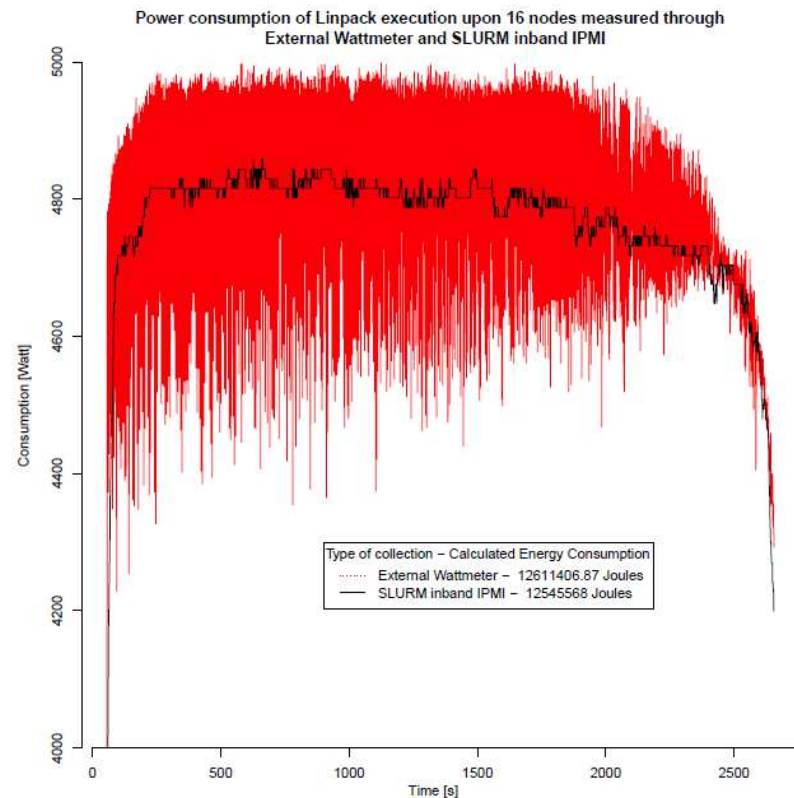
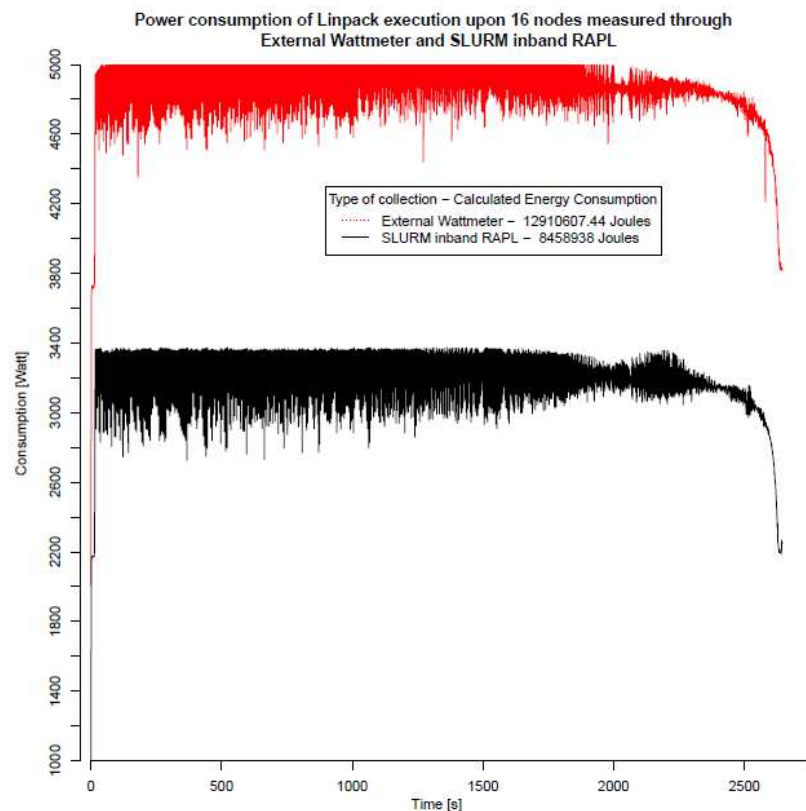
Slurm Power Management results

Find the Cpu frequency that leads to the minimum consumption



- Application Linpack
 - 80% of available memory
- Environment: 16 nodes
 - 2x Intel Xeon E5-2630 2.3GHz
- Slurm 2.6
 - Sampling period of 1 second
 - IPMI and RAPL plugins
- External Wattmeter

Slurm Power Management precision



- Power consumption variation is mainly due to CPU and memory
- RAPL plugin is highly sensitive to variation
- IPMI plugin provide more integrated results

Bull 's and slurm: next steps

Deal with hardware heterogeneity

- Extend Resource Management to support heterogeneous resources
 - MIC, IO, Energy ... through layout
 - License accounting and integration with license manager
 - Multi-parameter Scheduling
- Support for hybrid programming models

Improve scalability

- Network communication scalability optimizations
- Launching enhancement through PMI2 infrastructure

Increase Energy efficiency

- Power capping technics in scheduling
- Energy Fair Sharing

Directions: Go to the exaflop !!

More resources

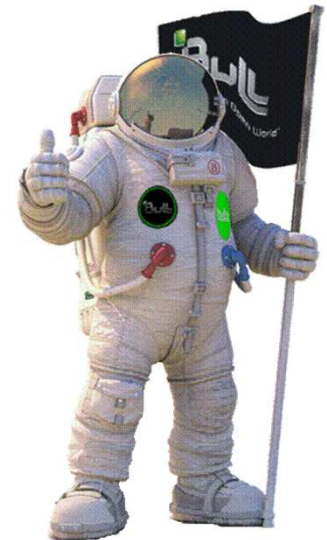
- Scalability
- Flexibility
- Precision

New applications

- Hybrid (MPI+X)
- New HW optimization
- Layer interop

Power Management

- Optimize /Limit
- App Power scheduling





Architect of an Open World™
