Slurm roadmap

SC-2012  Eric.Monchalin@bull.net
Head of Extreme Computing R&D
Largest Bull supercomputers powered by Slurm

**TERA 100 in figures**
- 1.25 PetaFlops
- 140,000+ Xeon cores
- 256 TB memory
- 30 PB disk storage
- 500 GB/s IO throughput
- 580 m² footprint

**CURIE in figures**
- 2 PetaFlops
- 90,000+ Xeon cores
- 148,000 GPU cores
- 360 TB memory
- 10 PB disk storage
- 250 GB/s IO throughput
- 200 m² footprint

**IFERC in figures**
- 1.5 PetaFlops
- 70,000+ Xeon cores
- 280 TB memory
- 15 PB disk storage
- 120 GB/s IO throughput
- 200 m² footprint
bullx Batch Manager values

- **bullx** MPI
  - Automatic placement coherency
  - Scalable launching

- **bullx** Development Environment
  - Debuggers, Profilers,

- **bullx** Management Center
  - Topology design generation
  - Global High Availability services

- Slurm 2.5

- Bull’s contributions
  - Scalability
  - Resource management
  - Power Management
  - Usability

© Bull, 2012
Slurm demonstrates its scalability

**Scalability / High Throughput Study**

- Simulations up to 16K nodes (500K cores)
- Submission Burst up to 10K jobs
Framework to support the **capturing** of power/energy consumption from the computing nodes

- Scalable
- Modular
- Based on latest technology
Display

- **Job power usage**
- **Node Power usage**
- **Accounting DB**

- On-going power usage for a given job
- On-going power usage for a given node
- Job power consumption Saved in slum DB

© Bull, 2012
Fix the CPU frequency

```
$#srun --cpu-freq=2700000 --resv-ports -N2 -n64 ./cg.C.64
```

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Elapsed Time</th>
<th>Consumed Energy(J)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1200000</td>
<td>00:01:35</td>
<td>19366</td>
</tr>
<tr>
<td>1396460</td>
<td>00:01:23</td>
<td>19018</td>
</tr>
<tr>
<td>1780477</td>
<td>00:01:09</td>
<td>19353</td>
</tr>
<tr>
<td>1996186</td>
<td>00:01:05</td>
<td>19817</td>
</tr>
<tr>
<td>2200000</td>
<td>00:01:02</td>
<td>20494</td>
</tr>
<tr>
<td>2362500</td>
<td>00:00:59</td>
<td>21408</td>
</tr>
<tr>
<td>2653125</td>
<td>00:00:56</td>
<td>23125</td>
</tr>
</tbody>
</table>
Directions: on the road of the Exaflop

More resources
  → Scalability
  → Flexibility
  → Heterogeneity

New applications
  → Hybrid (MPI+X)
  → New HW optimization
  → Layer interop

Power Management
  → Optimize /Limit
  → App Power scheduling