Architect of an Open World™

Slurm roadmap

SC-2012

Eric.Monchalin@bull.net

Head of Extreme Computing R&D

slurm workload manager

Largest Bull supercomputers powered by Slurm

cea

GENCI

PRACE

FUSION FOR ENERGY

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

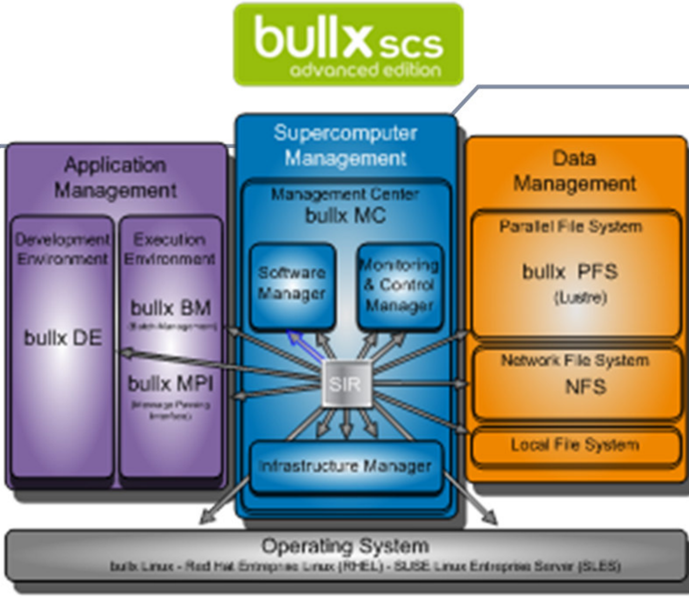
IFERC in figures

- 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 bm

- ❑ 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

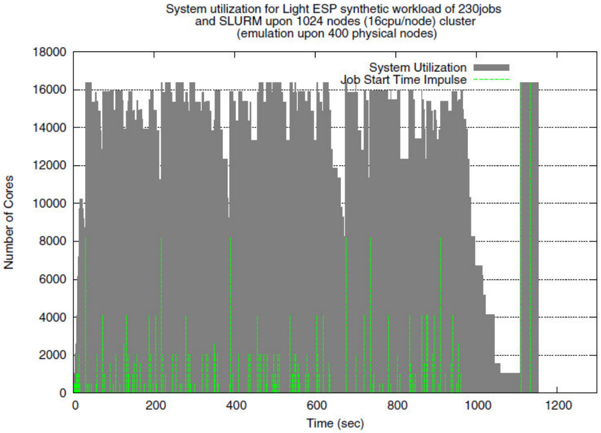
slurm
workload manager

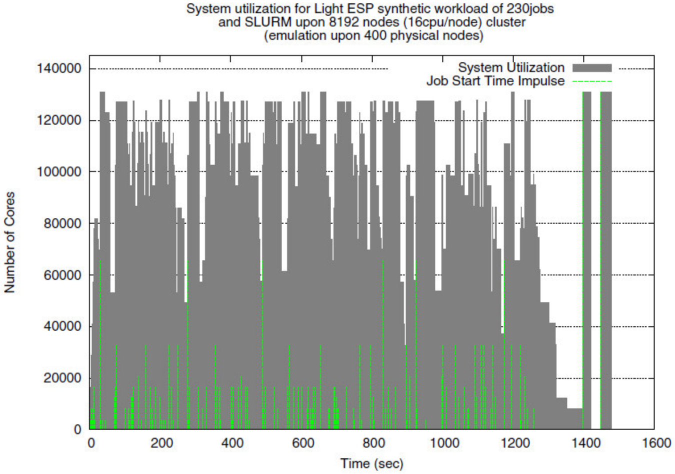
© Bull, 2012

Slurm demonstrates its scalability

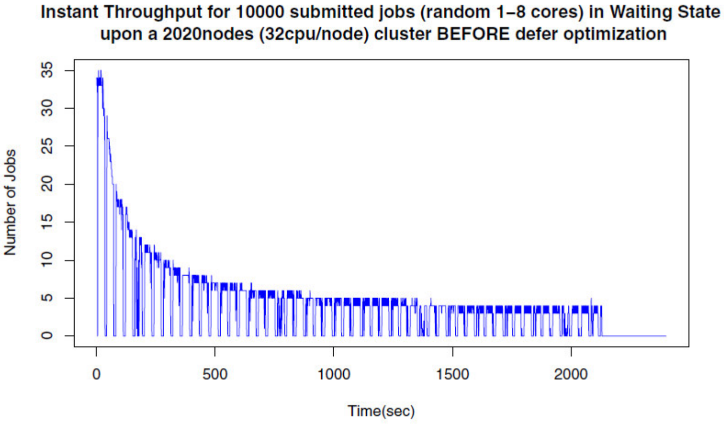
Scalability / High Throughput Study

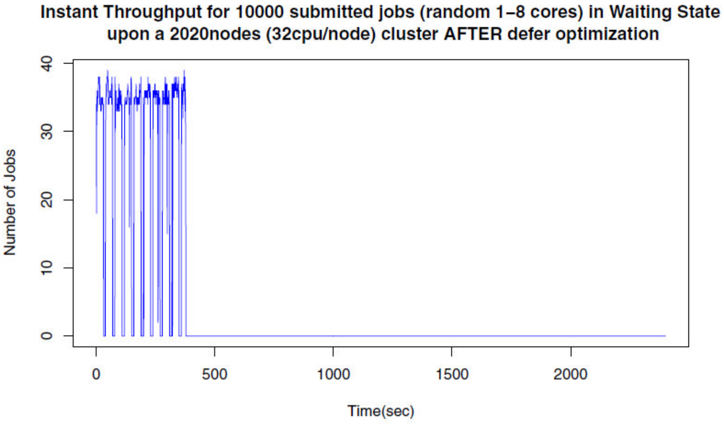
- Simulations up to 16K nodes (500K cores)


System utilization for Light ESP synthetic workload of 230jobs and SLURM upon 1024 nodes (16cpu/node) cluster (emulation upon 400 physical nodes)

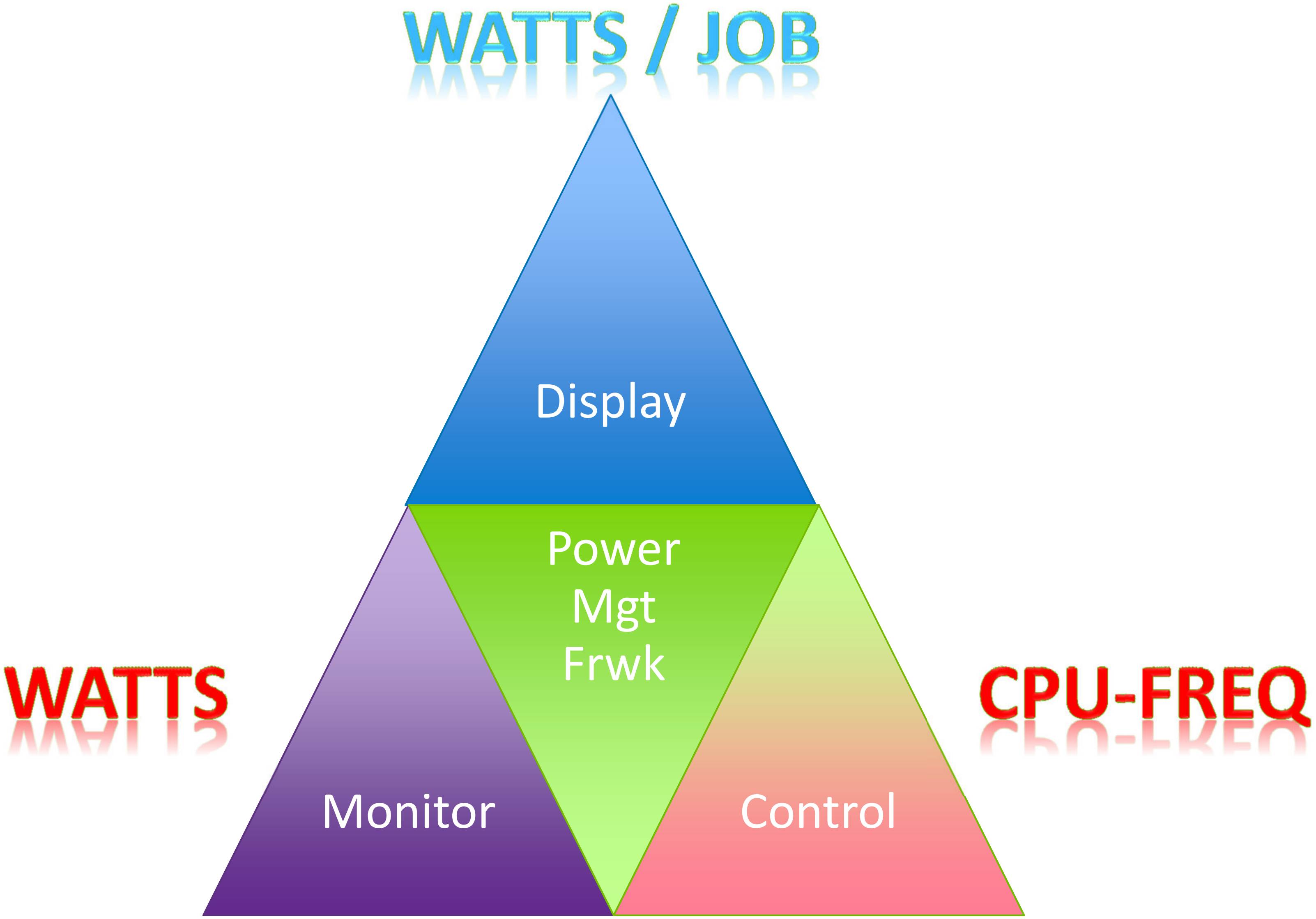

System utilization for Light ESP synthetic workload of 230jobs and SLURM upon 8192 nodes (16cpu/node) cluster (emulation upon 400 physical nodes)

- Submission Burst up to 10K jobs

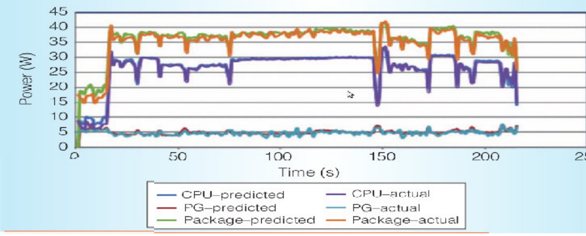

Instant Throughput for 10000 submitted jobs (random 1–8 cores) in Waiting State upon a 2020nodes (32cpu/node) cluster BEFORE defer optimization

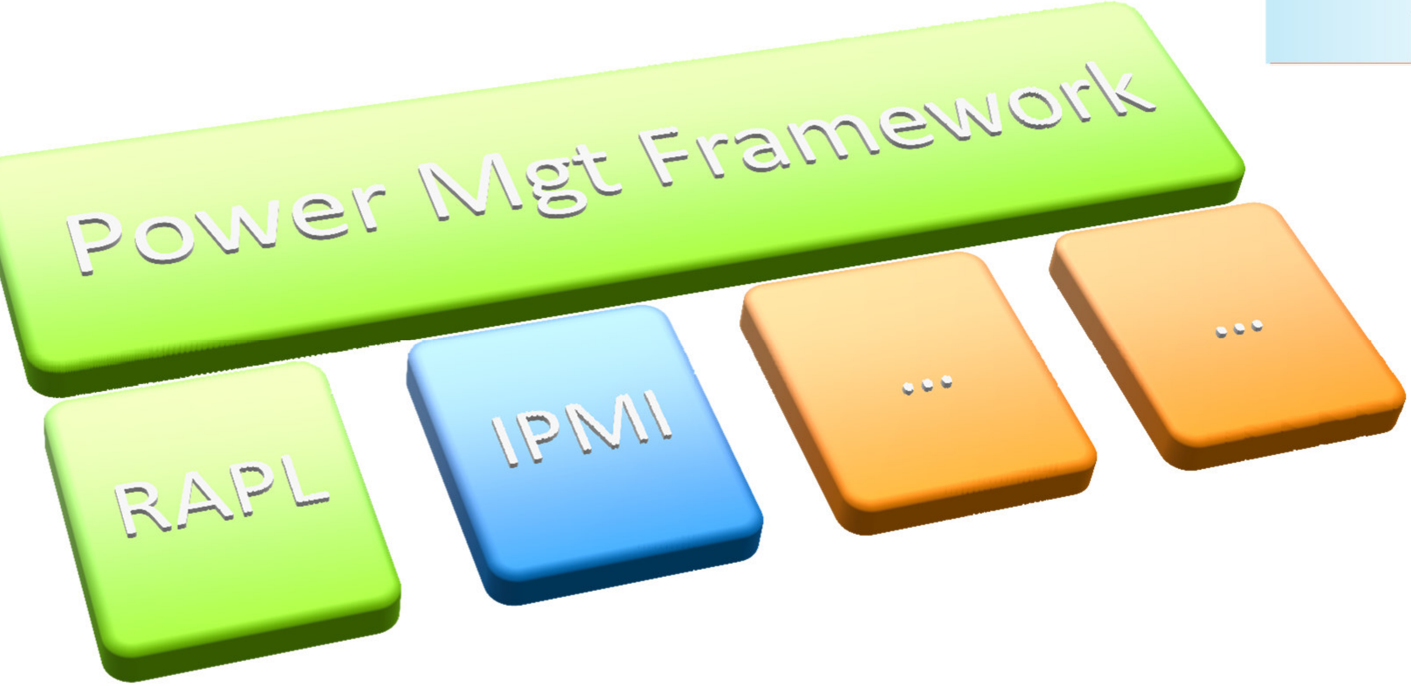

Instant Throughput for 10000 submitted jobs (random 1–8 cores) in Waiting State upon a 2020nodes (32cpu/node) cluster AFTER defer optimization

Power Management with bullx BM & Slurm



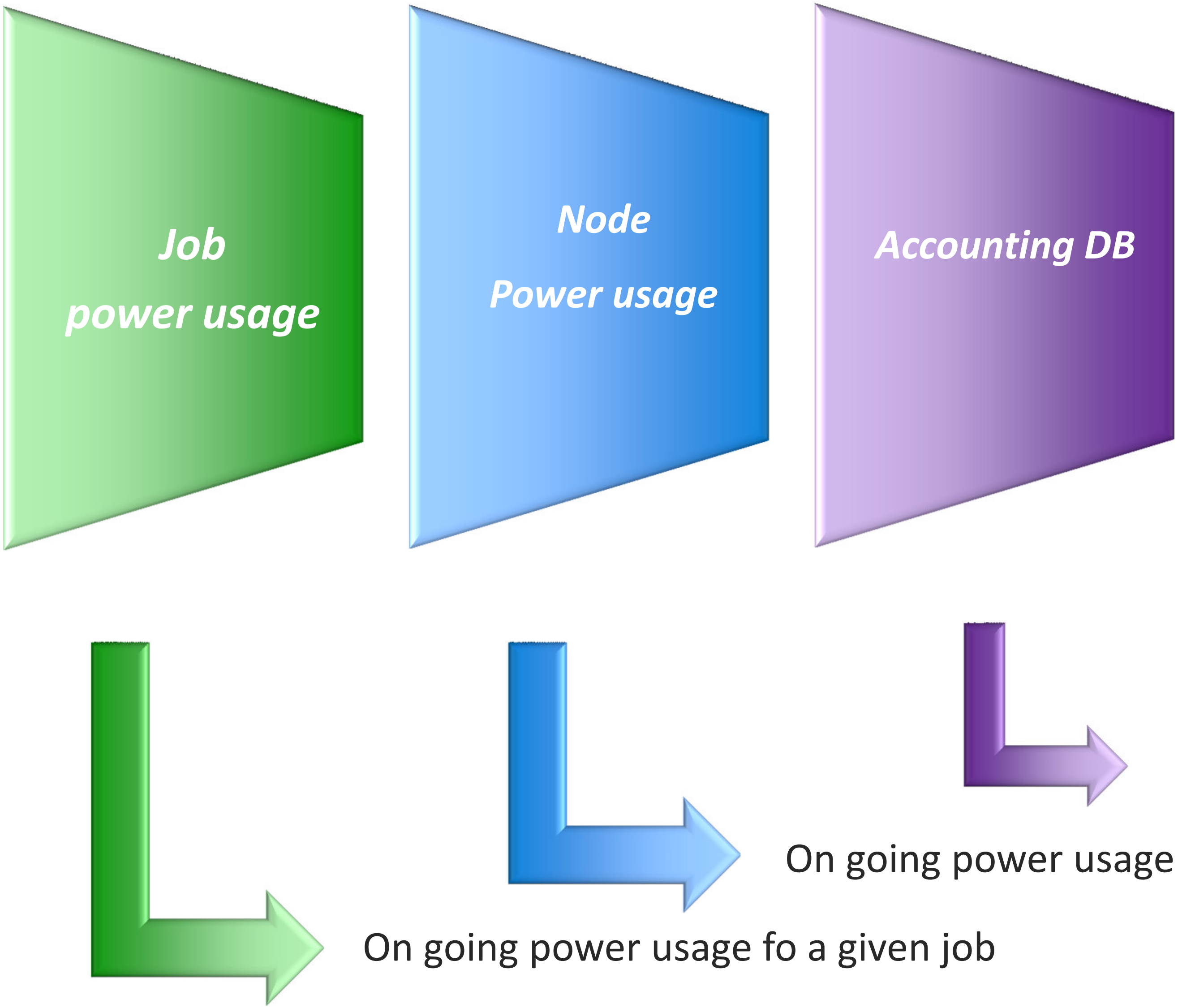
Monitor





- 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

Job power consumption
Saved in slum DB

On going power usage fo a given node

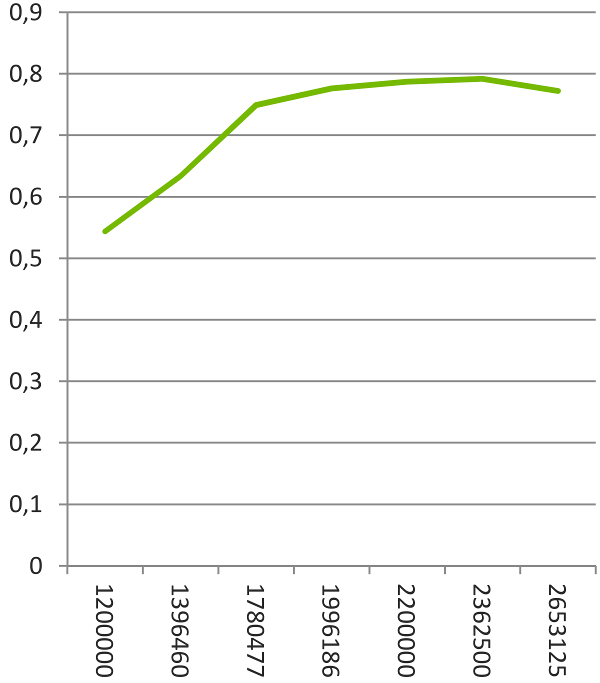
On going power usage fo a given job

Control

Fix the CPU frequency

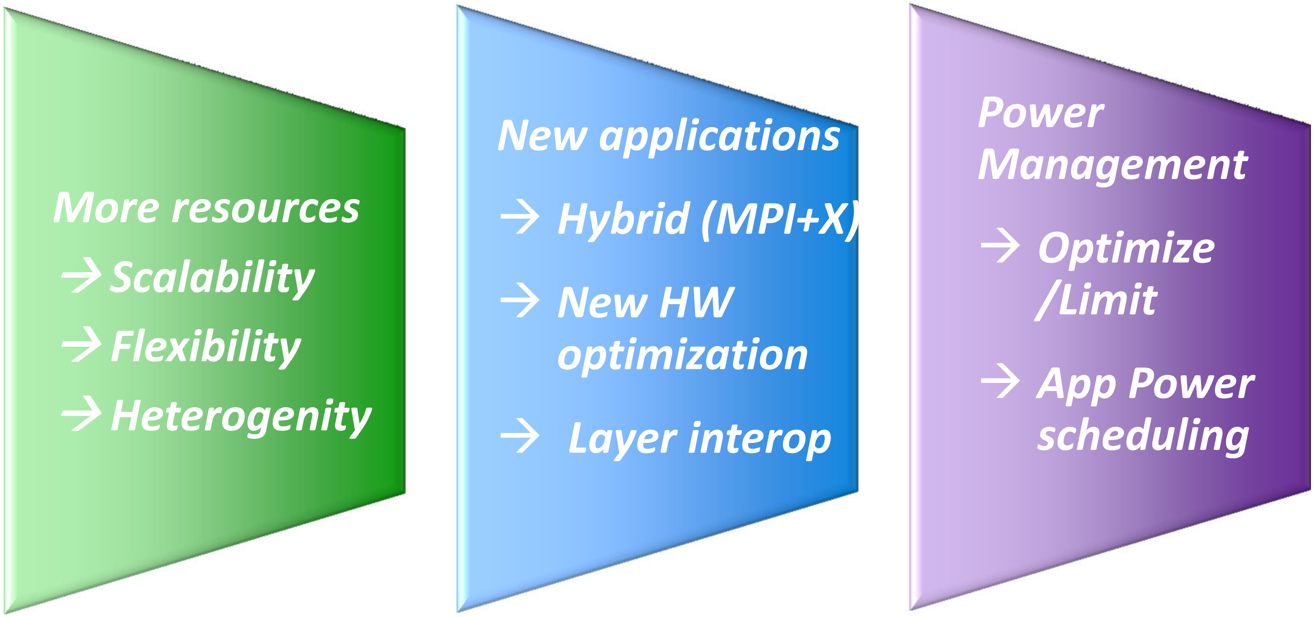
```
$#srun --cpu-freq=2700000 --resv-ports -N2 -n64 ./cg.C.64
```

```
AverageCPU Elapsed Time     Consumed Energy(J)
Frequency

1200000      00:01:35            19366
1396460      00:01:23            19018
1780477      00:01:09            19353
1996186      00:01:05            19817
2200000      00:01:02            20494
2362500      00:00:59            21408

2653125      00:00:56            23125
```



Ratio Time / Energy

Directions: on the road of the Exaflop



More resources

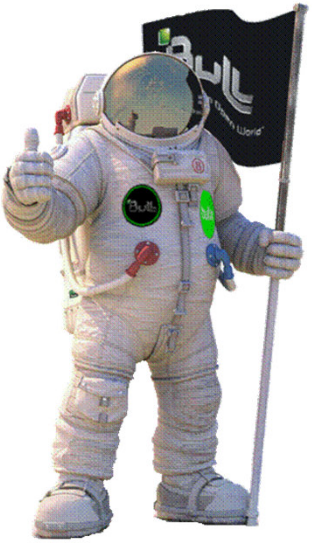
- → *Scalability*
- → *Flexibility*
- → *Heterogenity*

New applications

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

Power Management

- → *Optimize /Limit*
- → *App Power scheduling*



Architect of an Open World™