NSF/IUCRC CAC PROJECT

Profiling Power Consumption of Jobs with SLURM

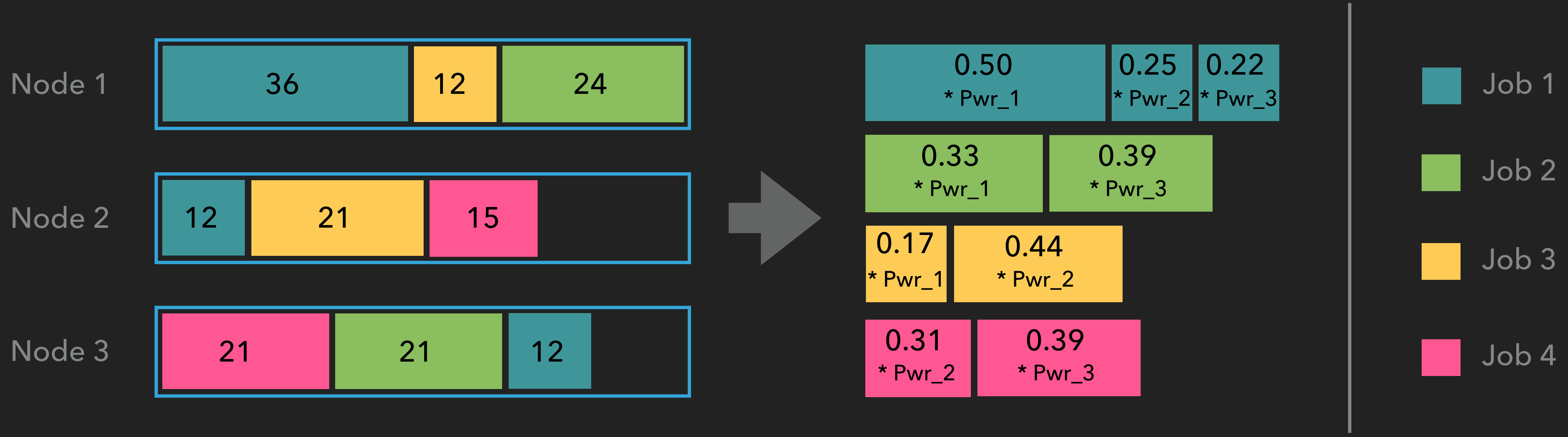
Jie Li
Doctoral Student, TTU
01/29/2020

Advisors:
Mr. Jon Hass, SW Architect, Dell Inc.
Dr. Alan Sill, Managing Director, HPCC, TTU
Dr. Yong Chen, Associate Professor, CS Dept, TTU

AGENDA

- Previous Proposal
- Background
- Methodology
- Summary & Future Work

PREVIOUS PROPOSAL



Node 1: 36 | 12 | 24

Node 2: 12 | 21 | 15

Node 3: 21 | 21 | 12

0.50 * Pwr_1 | 0.25 * Pwr_2 | 0.22 * Pwr_3

0.33 * Pwr_1 | 0.39 * Pwr_3

0.17 * Pwr_1 | 0.44 * Pwr_2

0.31 * Pwr_2 | 0.39 * Pwr_3

Job 1

Job 2

Job 3

Job 4

- Correlate the power consumption of nodes read from BMC with the jobs information fetched from UGE API

  Delay between UGE API and BMC API

- Assumption: power usage is proportional to the core usage

  Assumption is not applied in all situation

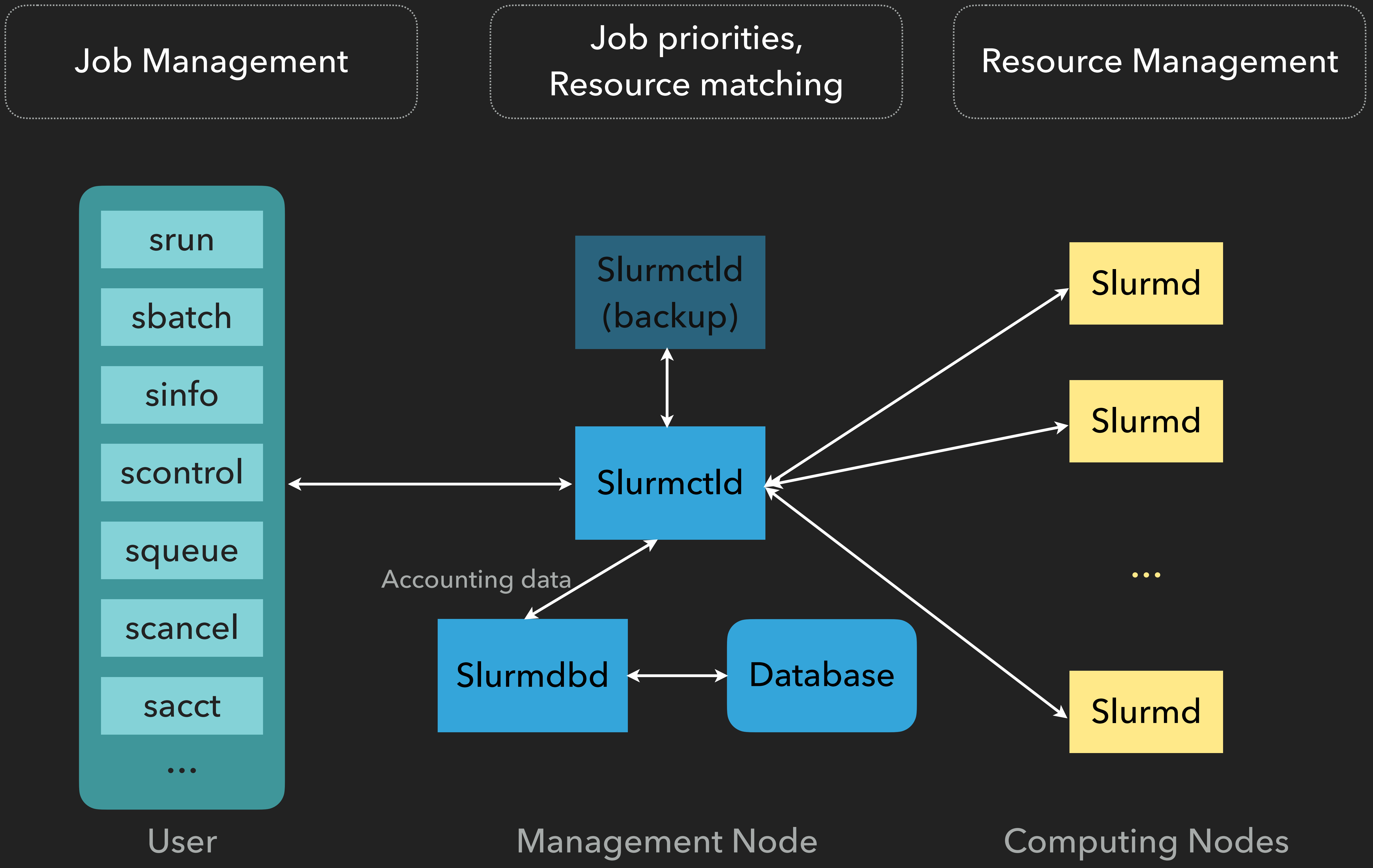
Ref: https://discl.cs.ttu.edu/lib/exe/fetch.php?media=talk:wiki:05062019seminarghazanfarali.pptx; https://discl.cs.ttu.edu/lib/exe/fetch.php?media=talk:wiki:09182019seminarjieli.pdf

SLURM FEATURES

- SLURM: Simple Linux Utility for Resource Management
- Open-source: freely available under the GNU General Public License
- Portable: written in C with a GNU autoconf configuration engine.
- Modular: support different kind of scheduling policies, interconnects, libraries, etc
- Scalable: designed to operate in a heterogeneous cluster with up to tens of millions of processors
- Power management: Power used by job is recorded; Idle resources can be powered down until needed

Ref: http://www.sc-camp.ogr/2015/SLURM_Toturial_SCcamp2015.pdf

SLURM ARCHITECTURE



Ref: https://slurm.schedmd.com/overview.html

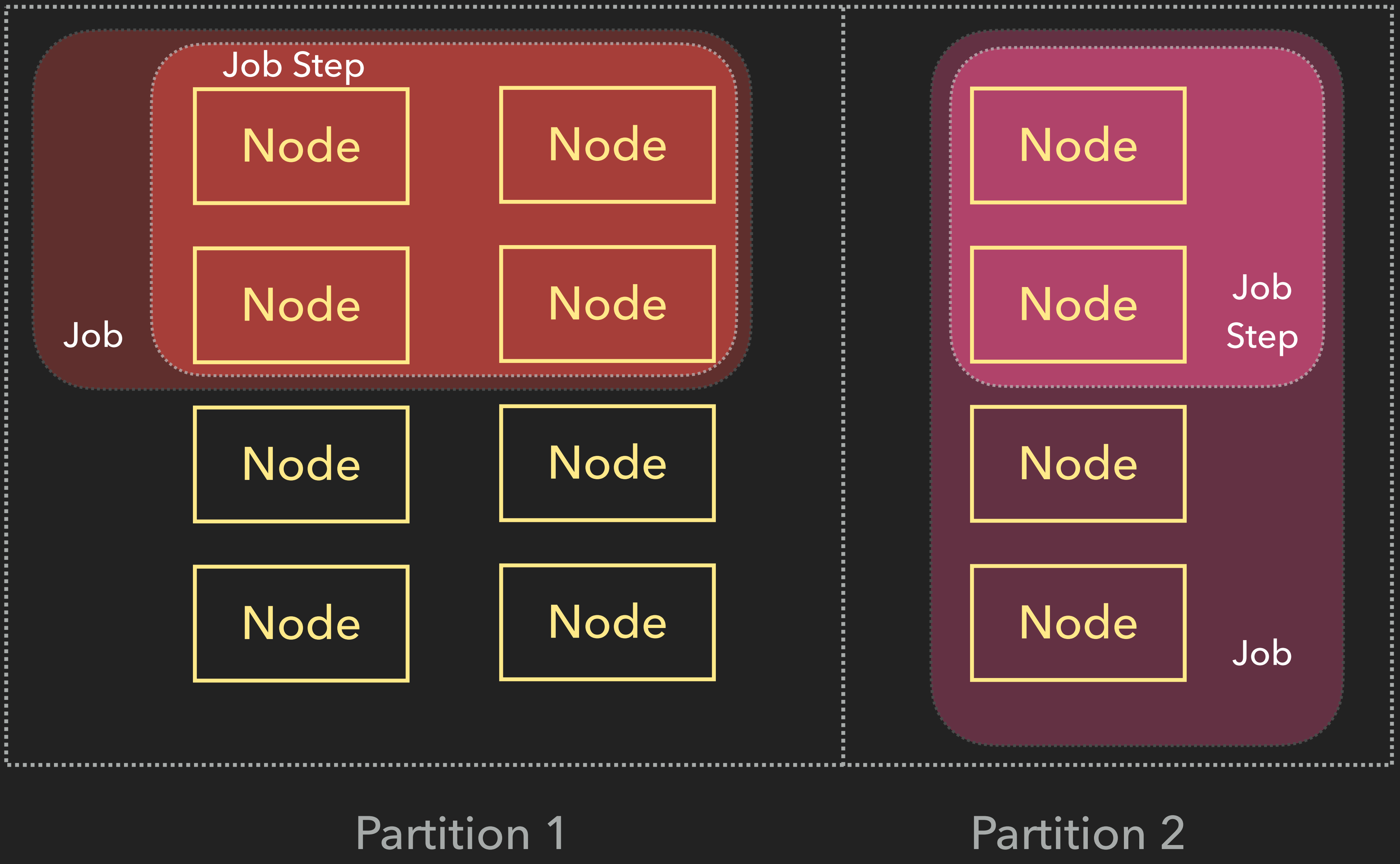
SLURM TERMS

Computing node

Partition

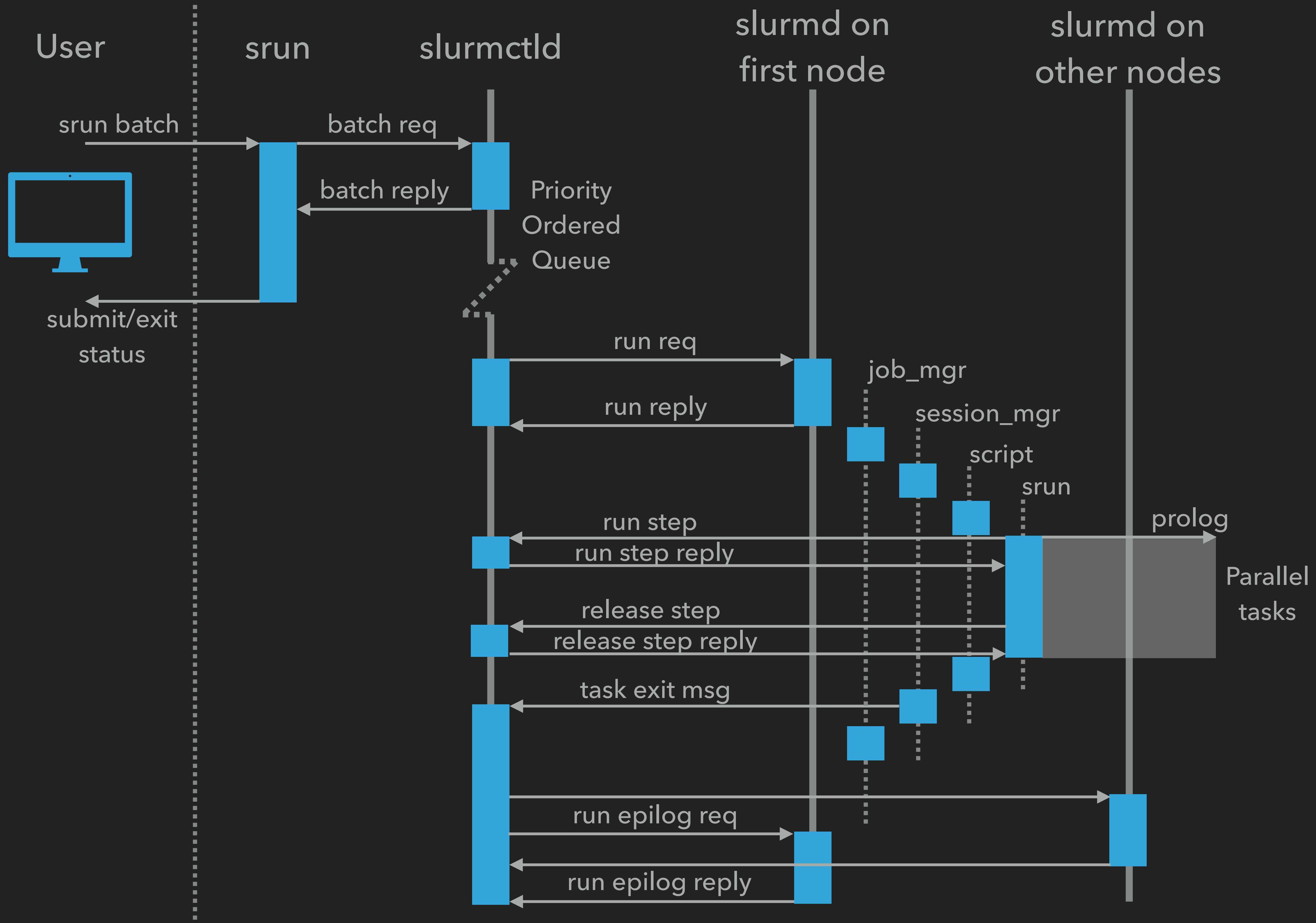
Job

Step



Ref: https://slurm.schedmd.com/overview.html

QUEUED JOB INITIATION



Ref: Yoo, Andy B., Morris A. Jette, and Mark Grondona. "Slurm: Simple linux utility for resource management."
In *Workshop on Job Scheduling Strategies for Parallel Processing*, pp. 44-60. Springer, Berlin, Heidelberg, 2003.

INTERMISSION

SLURM Architecture	SLURM Terms	Job Submission Model	Power data collection	Job Power Profiling
slurmctld	Computing node	Batch	IPMI	
slurmdbd	Partition		RAPL	
slurmd	Job			
	Step			

IPMI

- **I**ntelligent **P**latform **M**anagement **I**nterface(IPMI)
- Message-based, hardware-level interface specification
- Used to perform recovery procedures or monitor platform status (such as temperatures, voltages, fans, power consumption, etc)
- Hidden on the baseboard management controller(BMC) which collects data from various sensors
- Can be found in nearly all current Intel architectures

Ref: https://en.wikipedia.org/wiki/Intelligent_Platform_Management_Interface

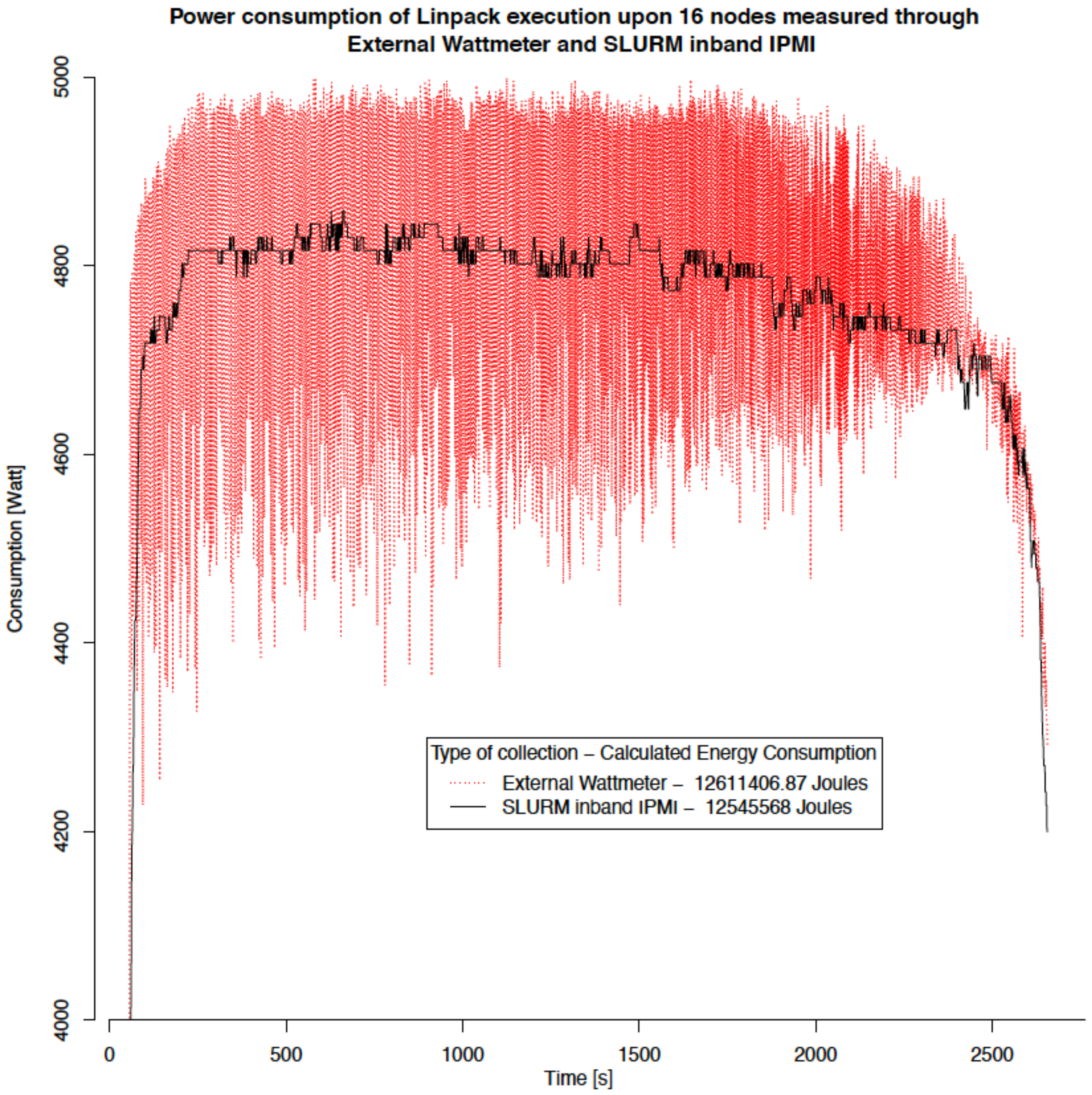
RAPL

- **R**unning **A**verage **P**ower **L**imit (RAPL)
- Introduced with the Intel Sandy Bridge processors and exits on all later Intel models
- Provides an operating system access to energy consumption information based on a software model driven by hardware counters
- Tracks the energy consumption of CPUs and DRAM but not that of the actually energy of the machine

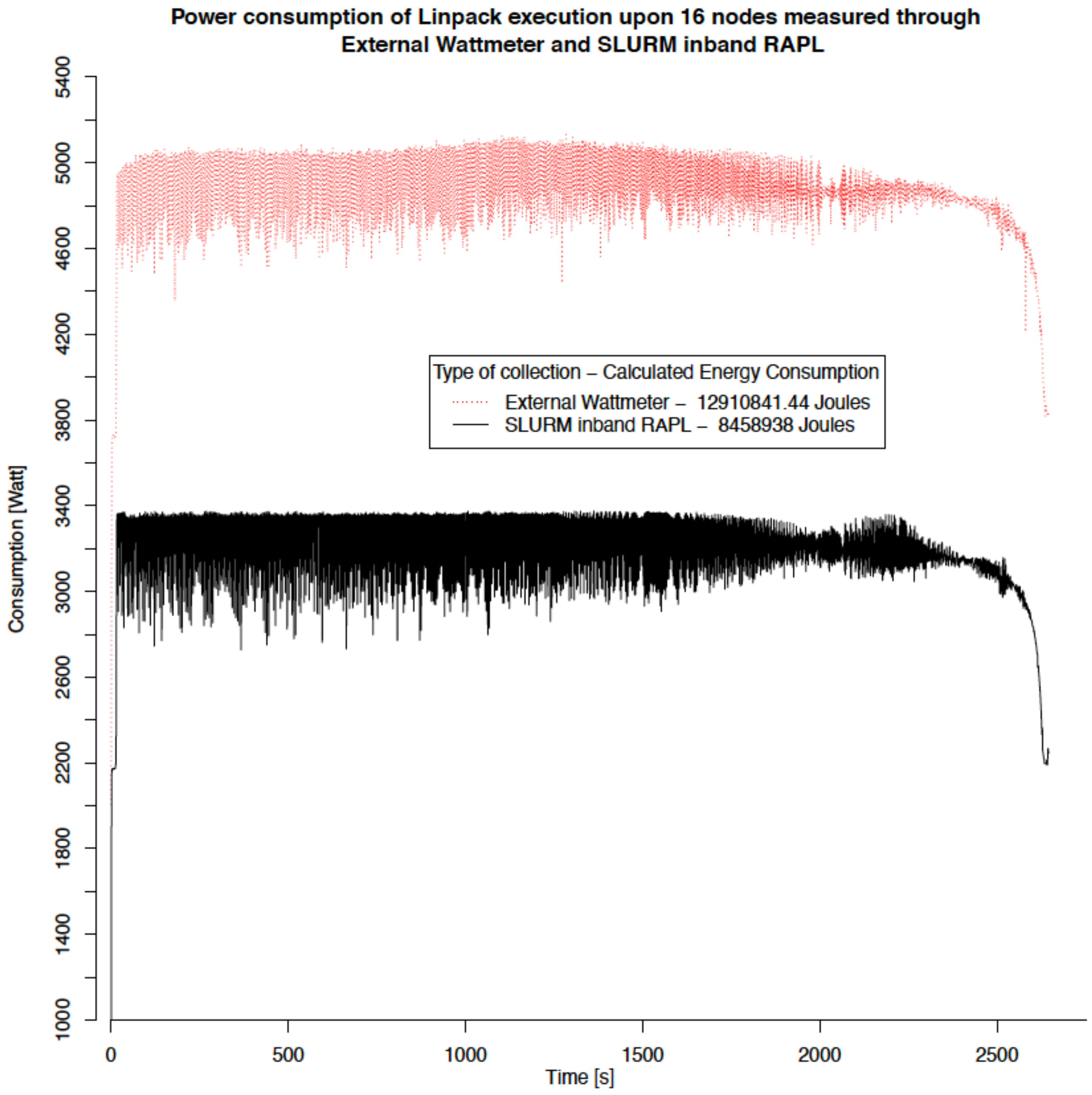
Ref: https://01.org/blogs/2014/running-average-power-limit-%E2%80%93-rapl

IPMI VS RAPL

Benchmark: Linpack



IPMI

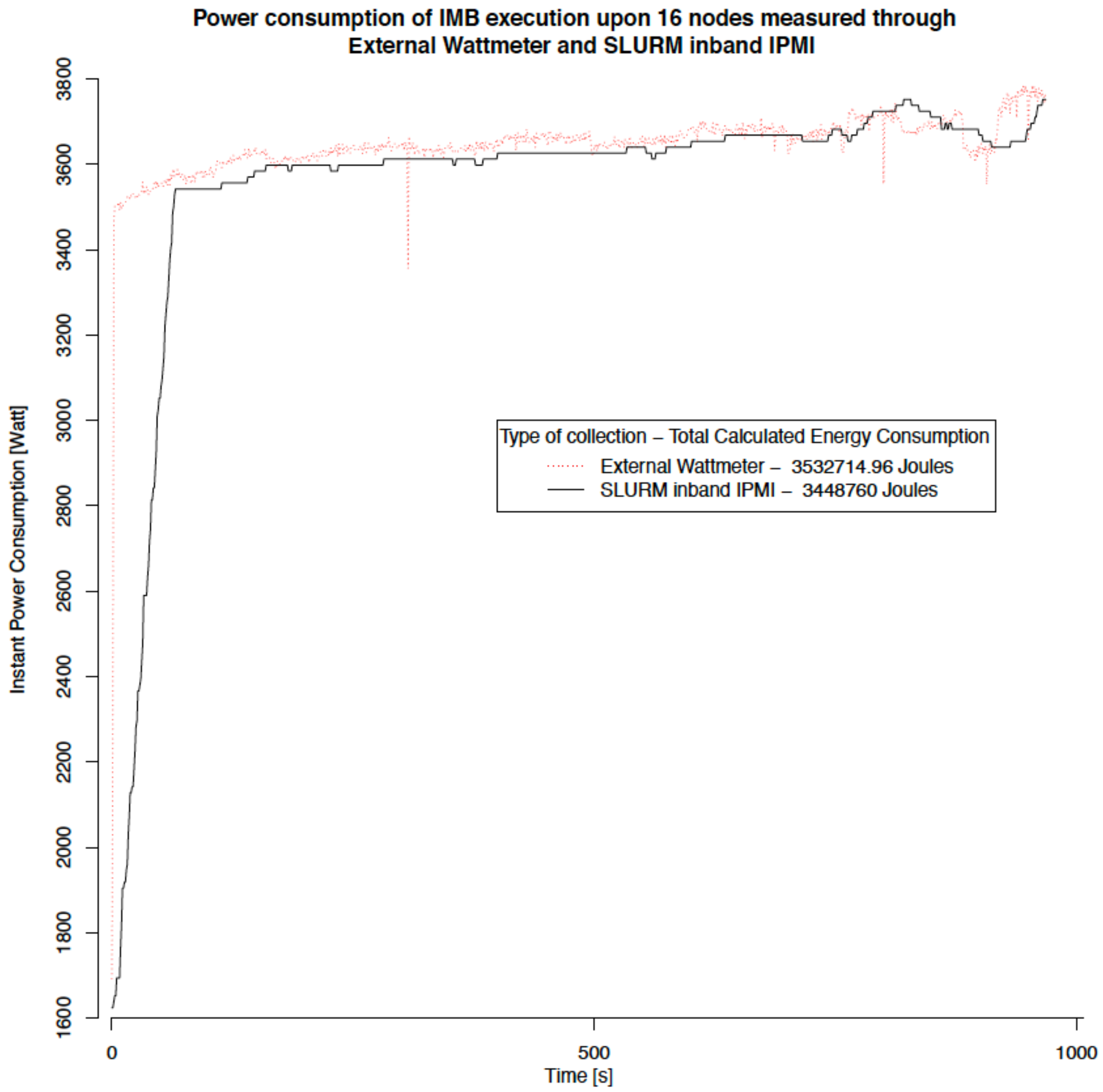


RAPL

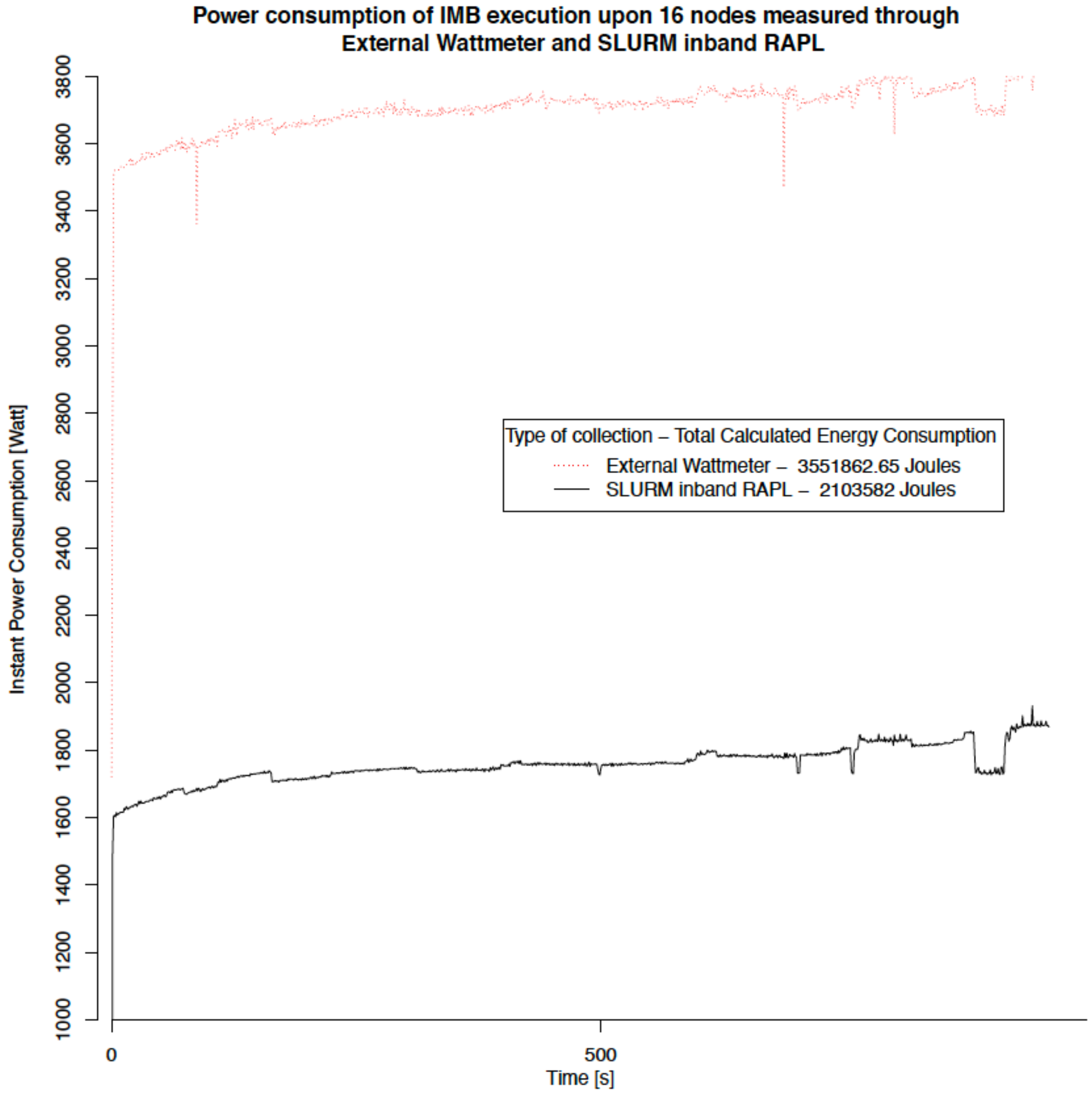
Ref: Georgiou, Yiannis, Thomas Cadeau, David Glesser, Danny Auble, Morris Jette, and Matthieu Hautreux. "Energy accounting and control with SLURM resource and job management system." In International Conference on Distributed Computing and Networking, pp. 96-118. Springer, Berlin, Heidelberg, 2014.

IPMI VS RAPL

Benchmark: IMB



IPMI

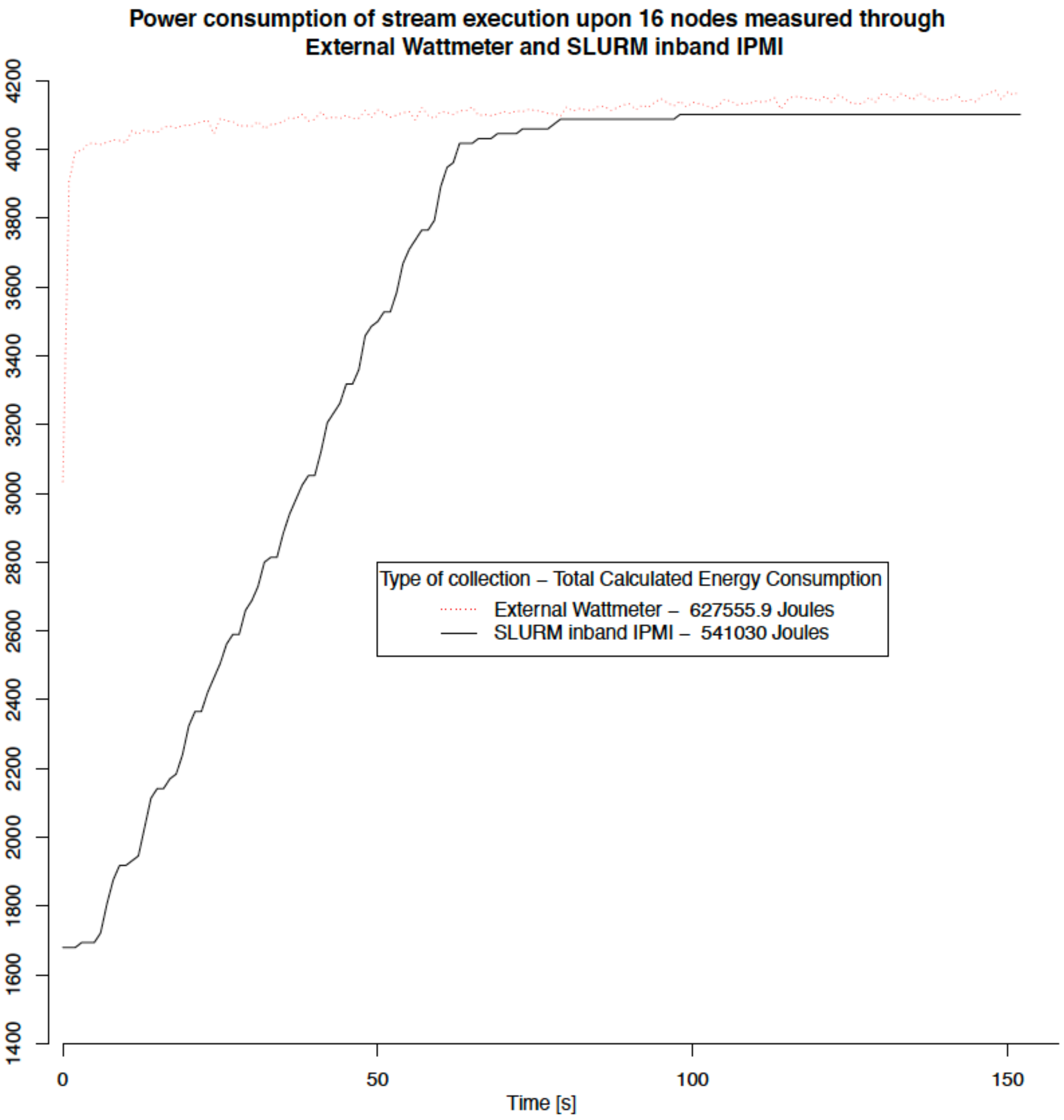


RAPL

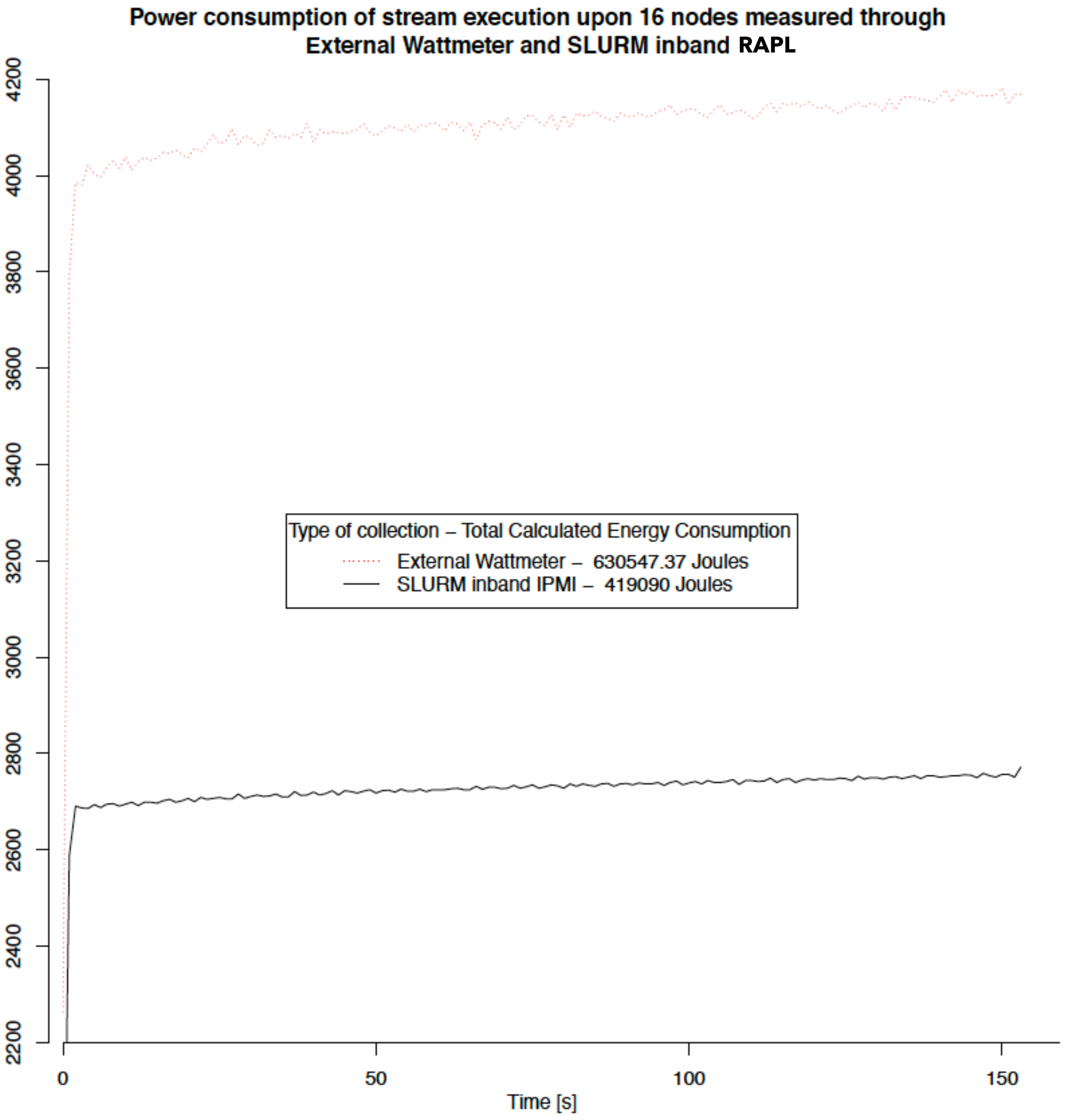
Ref: Georgiou, Yiannis, Thomas Cadeau, David Glesser, Danny Auble, Morris Jette, and Matthieu Hautreux. "Energy accounting and control with SLURM resource and job management system." In International Conference on Distributed Computing and Networking, pp. 96-118. Springer, Berlin, Heidelberg, 2014.

IPMI VS RAPL

Benchmark: Stream



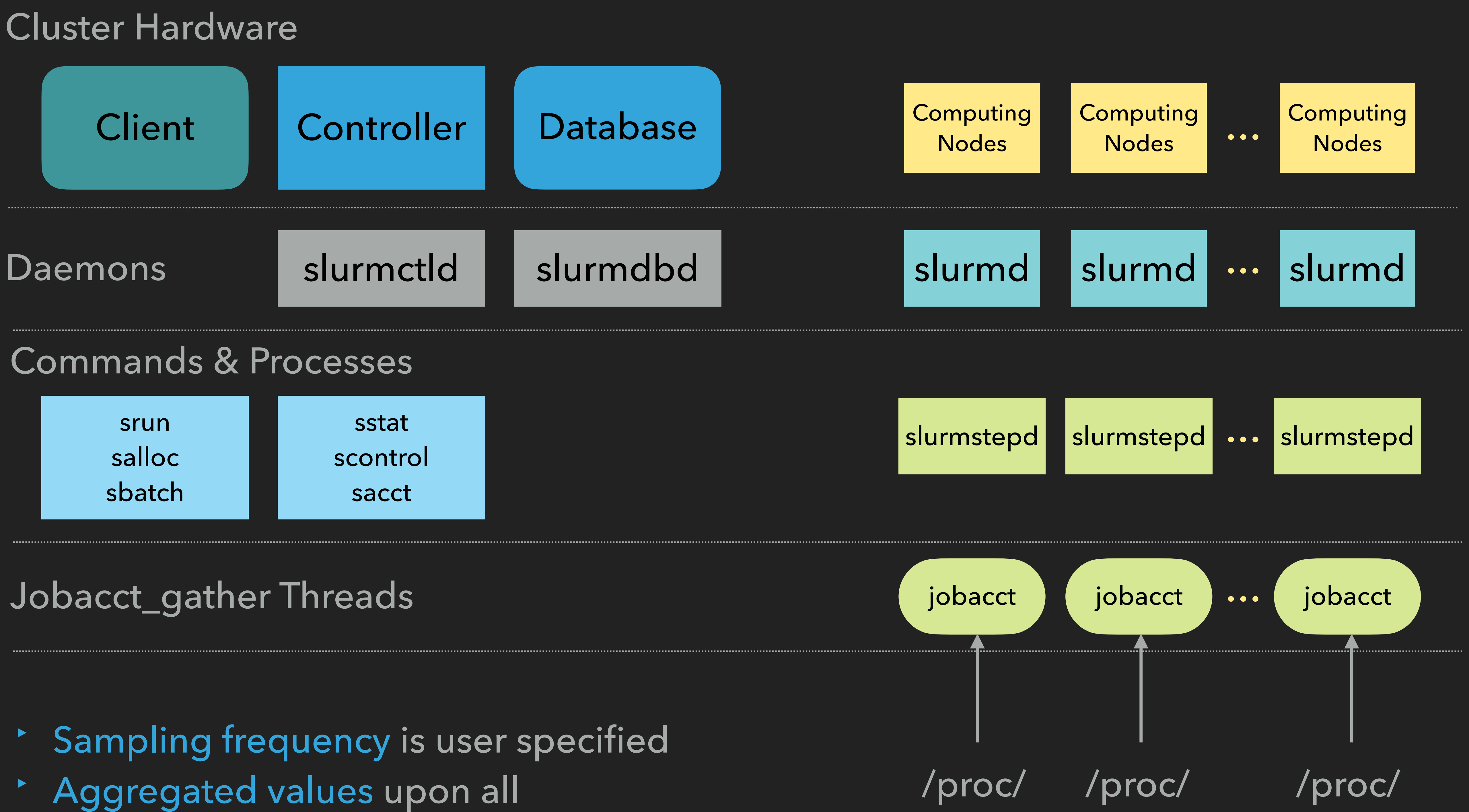
IPMI



RAPL

Ref: Georgiou, Yiannis, Thomas Cadeau, David Glesser, Danny Auble, Morris Jette, and Matthieu Hautreux. "Energy accounting and control with SLURM resource and job management system." In International Conference on Distributed Computing and Networking, pp. 96-118. Springer, Berlin, Heidelberg, 2014.

MONITORING MECHANISM

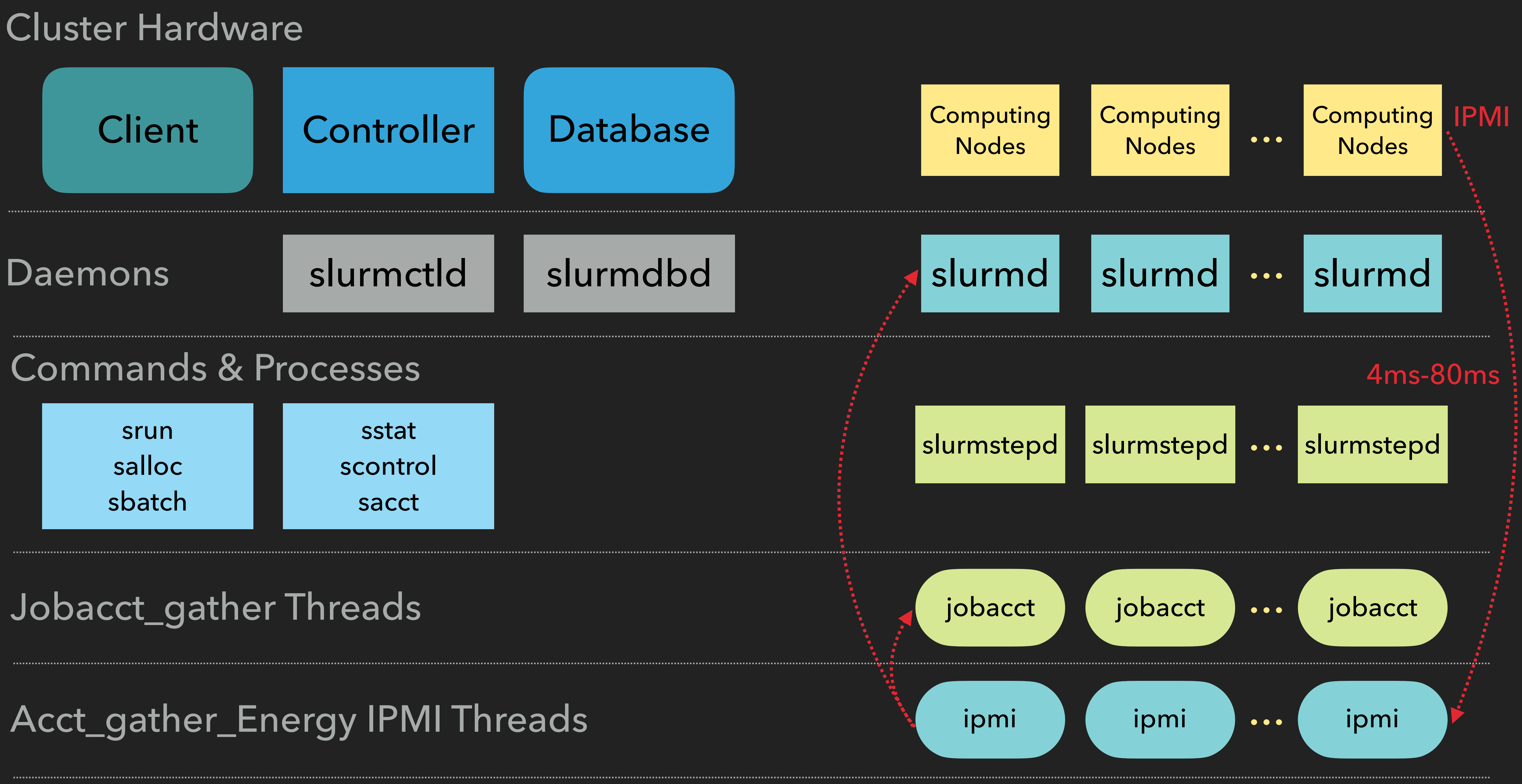


- Sampling frequency is user specified
- Aggregated values upon all nodes(average, max, etc) are stored in databases when the job is finished

Kernel data structure interface providing statistics upon various resources (CPU, Memory, etc)in the node

Ref: Georgiou, Yiannis, Thomas Cadeau, David Glesser, Danny Auble, Morris Jette, and Matthieu Hautreux. "Energy accounting and control with SLURM resource and job management system." In International Conference on Distributed Computing and Networking, pp. 96-118. Springer, Berlin, Heidelberg, 2014.

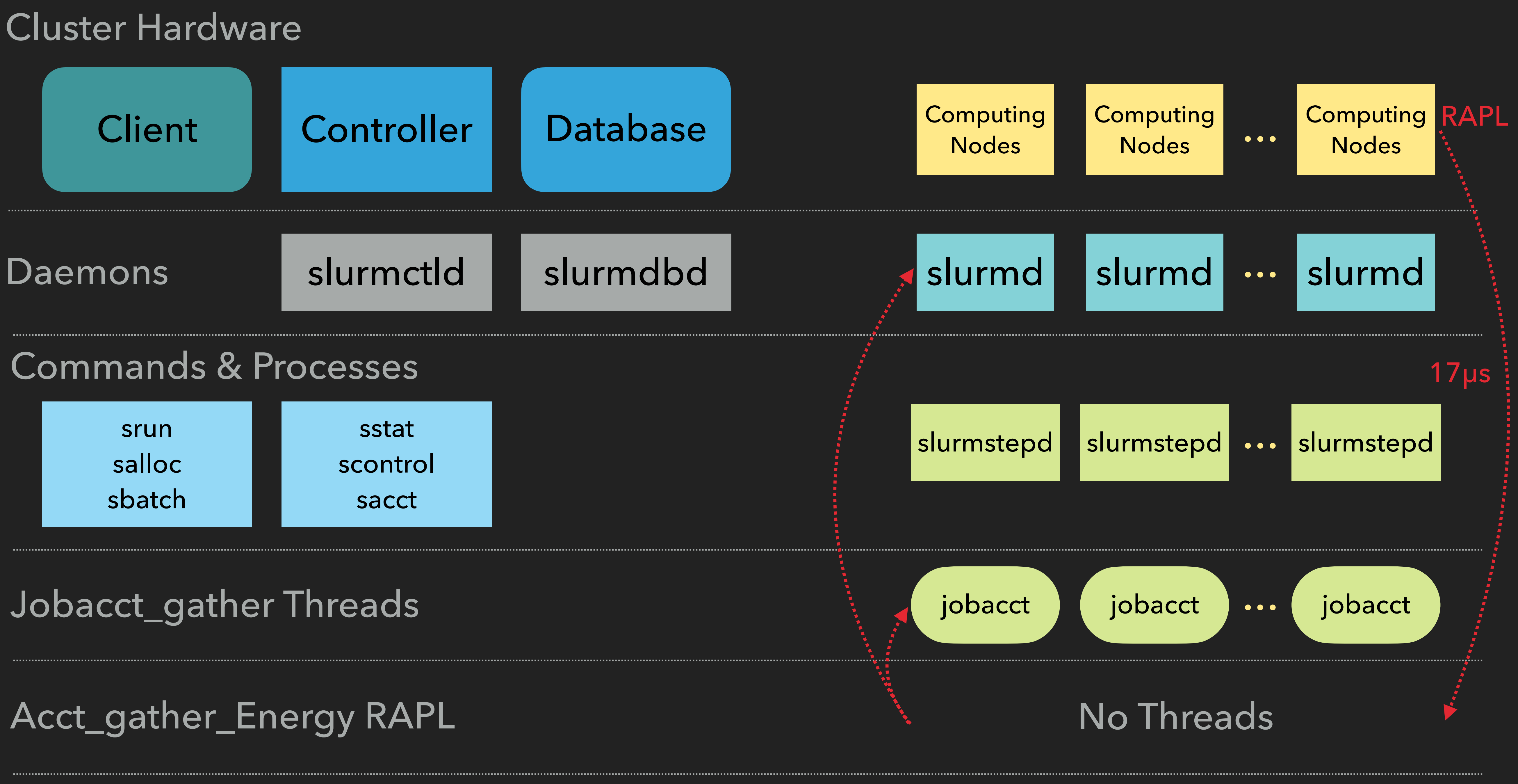
IPMI INTERFACE



- A particular algorithm is needed to calculate energy consumption per node
- Energy consumption data can be stored in databases when the job is finished

Ref: Georgiou, Yiannis, Thomas Cadeau, David Glesser, Danny Auble, Morris Jette, and Matthieu Hautreux. "Energy accounting and control with SLURM resource and job management system." In International Conference on Distributed Computing and Networking, pp. 96-118. Springer, Berlin, Heidelberg, 2014.

RAPL INTERFACE



Cluster Hardware

Client

Controller

Database

Computing Nodes

Computing Nodes

...

Computing Nodes

RAPL

Daemons

slurmctld

slurmdbd

slurmd

slurmd

...

slurmd

Commands & Processes

srun
salloc
sbatch

sstat
scontrol
sacct

17μs

slurmstepd

slurmstepd

...

slurmstepd

Jobacct_gather Threads

jobacct

jobacct

...

jobacct

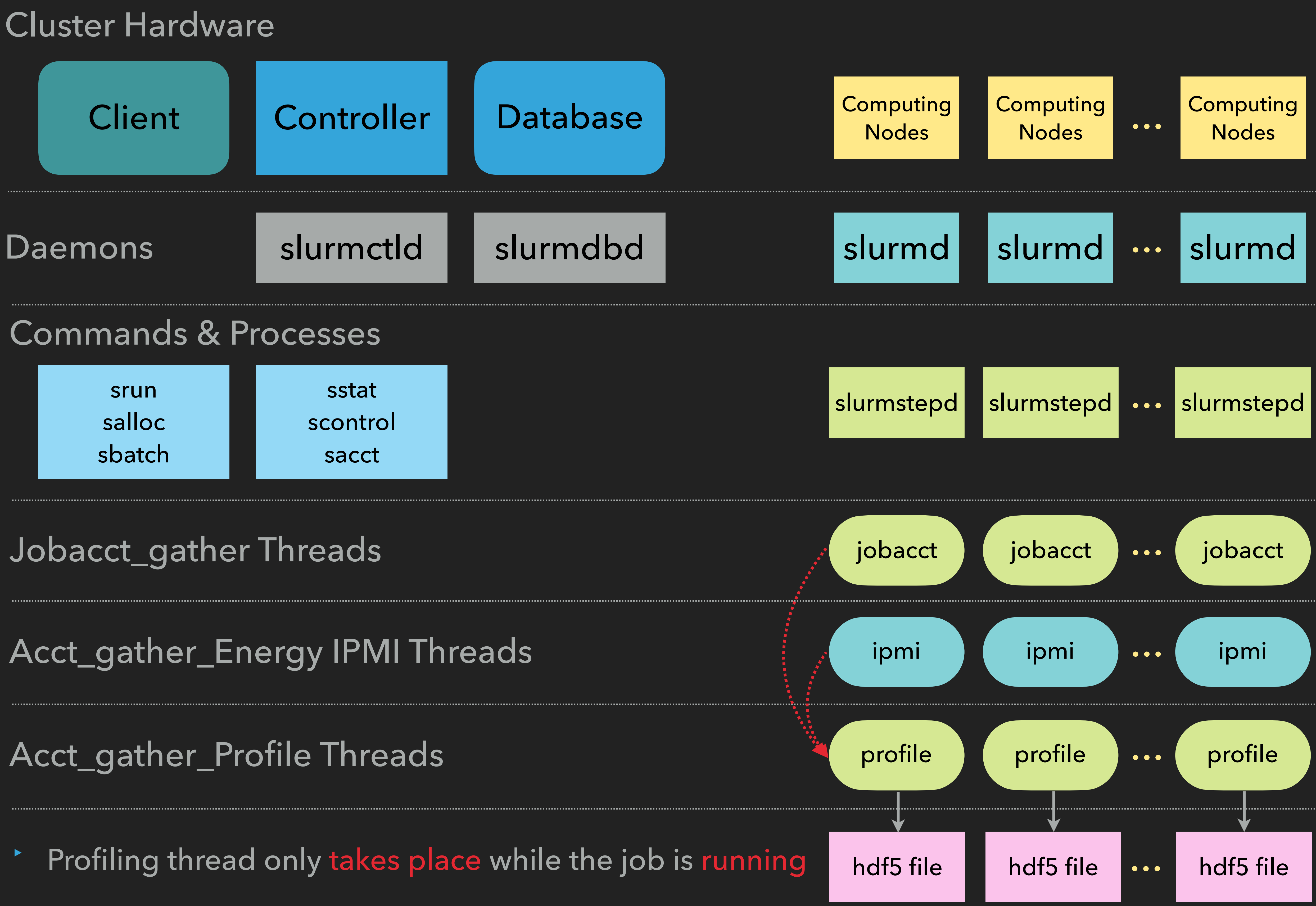
Acct_gather_Energy RAPL

No Threads

- Divide the energy consumption by the frequency of the sampling to get power consumption

Ref: Georgiou, Yiannis, Thomas Cadeau, David Glesser, Danny Auble, Morris Jette, and Matthieu Hautreux. "Energy accounting and control with SLURM resource and job management system." In International Conference on Distributed Computing and Networking, pp. 96-118. Springer, Berlin, Heidelberg, 2014.

PROFILING TYPE - HDF5 FILE



Cluster Hardware

Client | Controller | Database | Computing Nodes | Computing Nodes | ... | Computing Nodes

Daemons

slurmctld | slurmdbd | slurmd | slurmd | ... | slurmd

Commands & Processes

srun
salloc
sbatch

sstat
scontrol
sacct

slurmstepd | slurmstepd | ... | slurmstepd

Jobacct_gather Threads

jobacct | jobacct | ... | jobacct

Acct_gather_Energy IPMI Threads

ipmi | ipmi | ... | ipmi

Acct_gather_Profile Threads

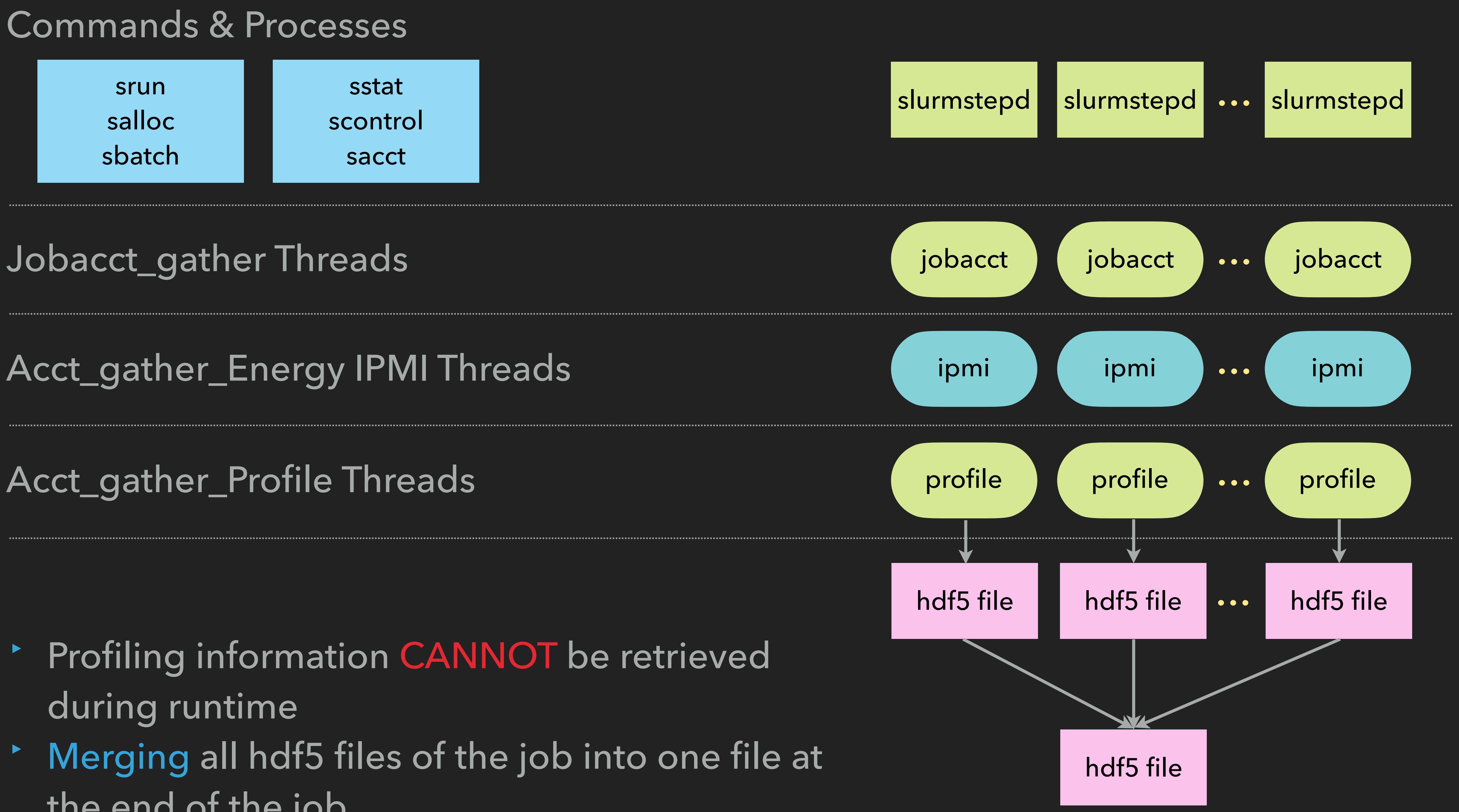
profile | profile | ... | profile

hdf5 file | hdf5 file | ... | hdf5 file

- Profiling thread only takes place while the job is running

Ref: Georgiou, Yiannis, Thomas Cadeau, David Glesser, Danny Auble, Morris Jette, and Matthieu Hautreux. "Energy accounting and control with SLURM resource and job management system." In International Conference on Distributed Computing and Networking, pp. 96-118. Springer, Berlin, Heidelberg, 2014.

PROFILING TYPE - HDF5 FILE



- Profiling information CANNOT be retrieved during runtime
- Merging all hdf5 files of the job into one file at the end of the job

Ref: Georgiou, Yiannis, Thomas Cadeau, David Glesser, Danny Auble, Morris Jette, and Matthieu Hautreux. "Energy accounting and control with SLURM resource and job management system." In International Conference on Distributed Computing and Networking, pp. 96-118. Springer, Berlin, Heidelberg, 2014.

PROFILING TYPE - INFLUXDB

Commands & Processes

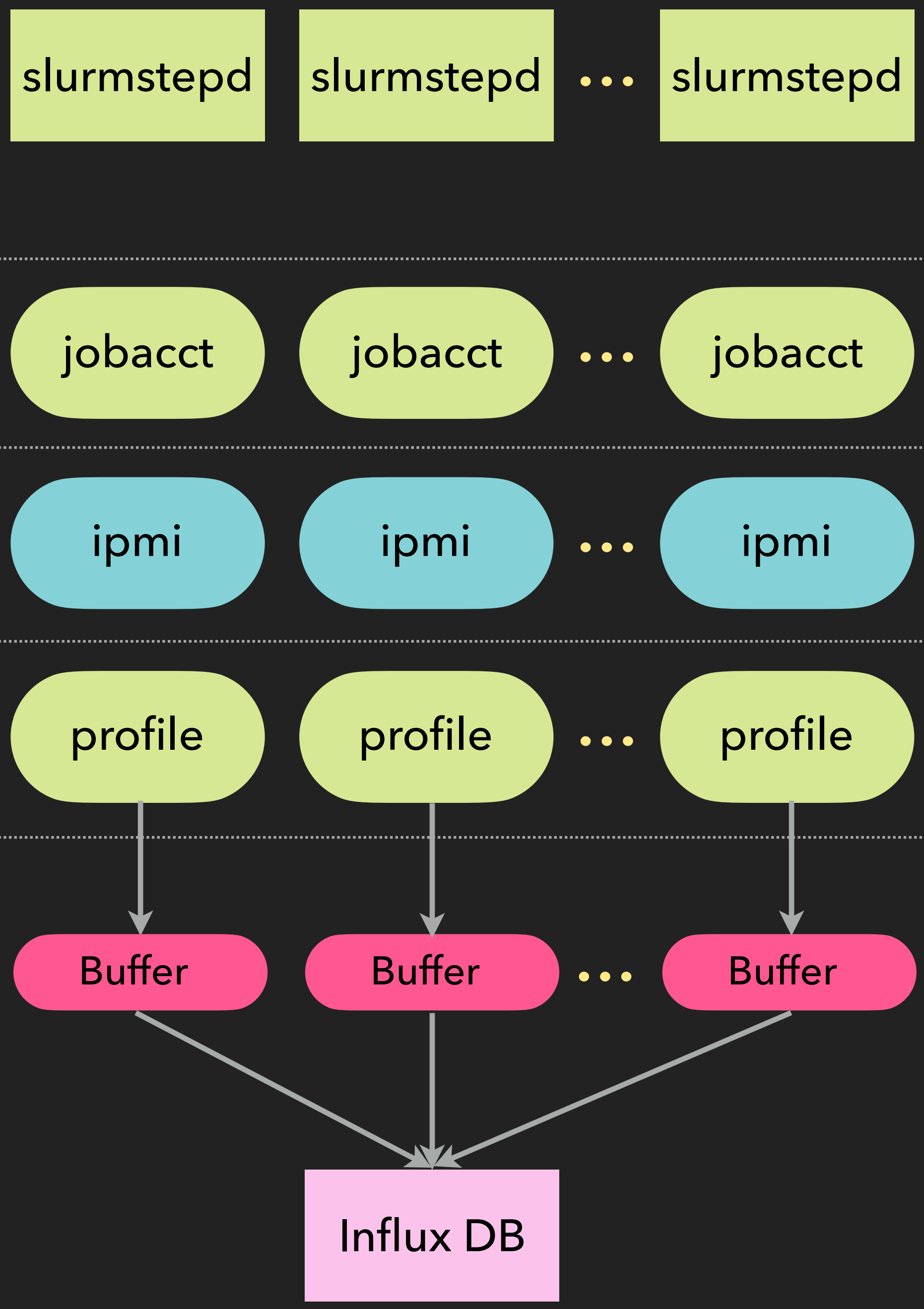
srun
salloc
sbatch

sstat
scontrol
sacct

Jobacct_gather Threads

Acct_gather_Energy IPMI Threads

Acct_gather_Profile Threads



- Profiling information written into influxDB
- Data include:
  - Energy
  - File system(Lustre)
  - Network(InfiniBand)
  - Task(I/O, Memory,…)
- Use internal buffer to avoid overloading the influxd instance with incoming connection requests

Ref:https://slurm.schedmd.com/acct_gather.conf.html

CPU FREQUENCY SCALING

- Parameter in srun command through which static CPU Frequency Scaling is supported
- The setting of CPU Frequency is made with manipulation of cpufreq/scaling_cur_freq under /sys/ along with particular governor drivers
- The mechanism sets the demanded frequency on the allocated CPUs when the job is started and set them back to their initial value after the job is finished

```
$ srun --cpu-freq=medium:conservative ….
$ srun --cpu-freq=performance …
$ srun --cpu-freq=2400 ...
```

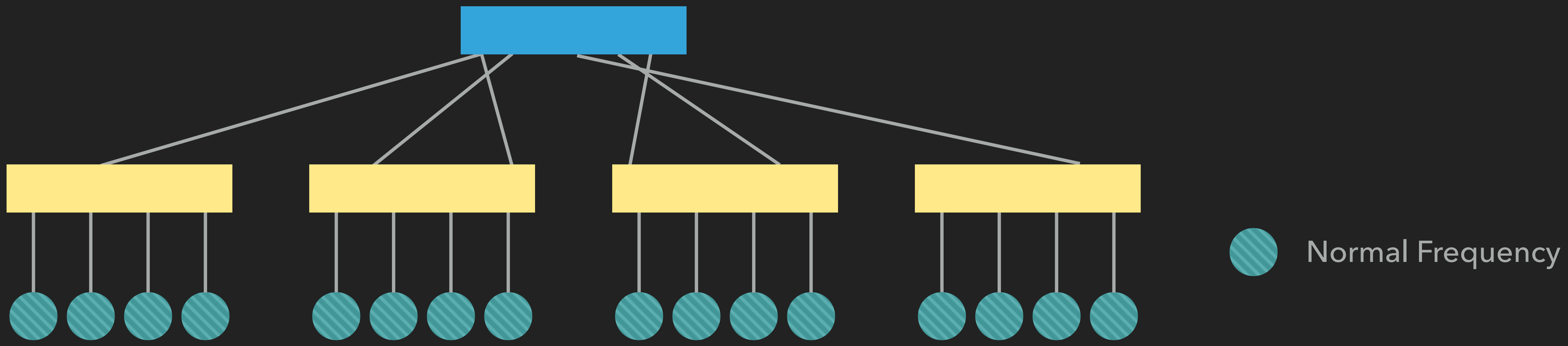
SUMMARY

- SLURM architecture
- Mechanism of profiling power consumption of jobs
- CPU Frequency Scaling supported in SLURM

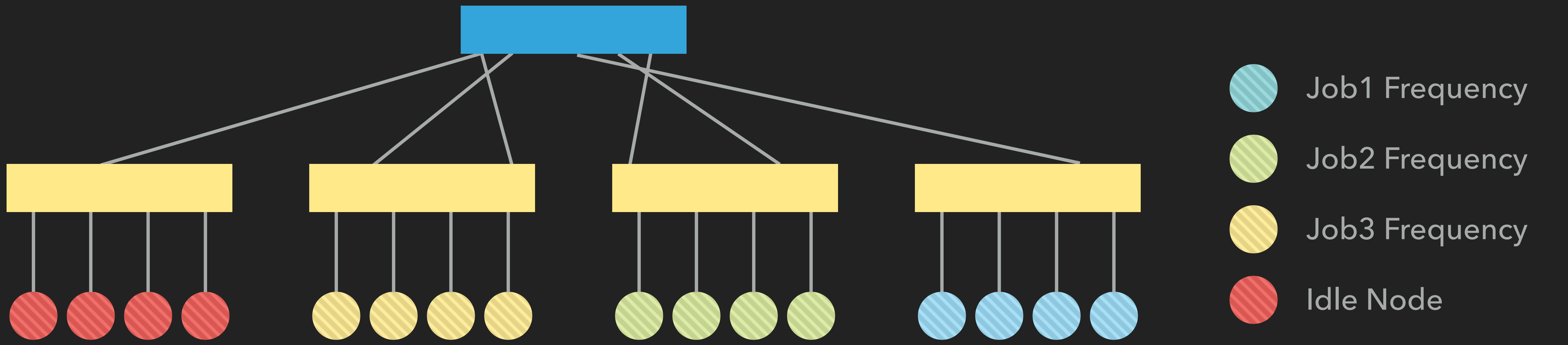
POSSIBLE DIRECTIONS OF FUTURE WORK

- Combine IPMI with RAPL to implement a high-sensitivity and high-accuracy power profiling model
- Explore the Energy-performance tradeoff of different jobs (CPU bound, memory bound or network bound)
    - Auto tune the CPU frequency for specific jobs to achieve best energy-performance tradeoff – similar to Ghazanfar’s project
- Energy aware scheduling: consider the energy consumption of submitted job in scheduling – result in important energy benefits for small performance losses

ENERGY AWARE SCHEDULING



Current Job Scheduling



Energy Aware Job Scheduling

QUESTIONS?/COMMENTS?