Blue Brain Project

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Optimizing HPC resource allocation through monitoring

Alexandre Beche <alexandre.beche@epfl.ch>

Outlines

1. Context
2. Job monitoring
3. Beyond Slurm monitoring

Context

Blue Brain Project

Host Institution

Ecole Polytechnique Fédérale de Lausanne (EPFL)

Director

Henry Markram

Co-Directors

Sean Hill, Felix Schürmann

Team today

~85 scientists, engineers & staff

Timeline

2005 founded at EPFL
2011/2012 ETH Board funding
2013-2016 Swiss National Research Infrastructure

Main International Collaborations

Switzerland (CSCS, CERN)
Israel (HUJI)
USA (Yale, ANL, Allen Brain)
Spain (UPM)
Saudi Arabia (KAUST)
Europe (HBP)

Data-Driven Modeling & Simulation!

Blue Brain Project

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Neuronal anatomy

- ~ **2 mm** thick
- **55** morphological types
- **13** excitatory & **42** inhibitory m-types
- **31,000** neurons
- **111,700** neurons/mm^3
- Excitatory to inhibitory neuron ratio of **86:14** %
- **346** m of axon
- **211** m of dendrites
- Maximum branch order of m-types:

	Axon	Dendrites
Excitatory	24	35
Inhibitory	50	17

Neuronal physiology

- **11** electrical types
- **207** morpho-electrical types
- **13** HH type ion channel models
- **bAP** & **EPSP** attenuation for **207** morpho-electrical types
- Ion channel density distribution profiles:

	Axon	Soma	Dendrites
Uniform	8	6	6

Synaptic anatomy

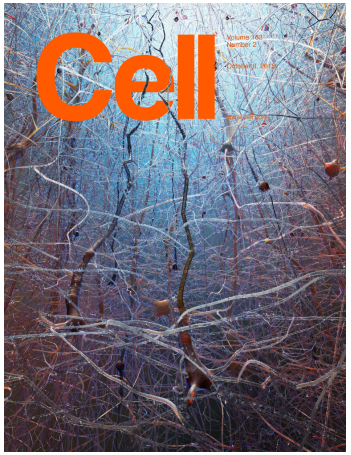
- **0.63** synapses/mm^3
- **Extrinsic** to **intrinsic** synapse ratio of **75:25** %
- **3025** possible synaptic pathways
- **2258** viable synaptic pathways
- **664** excitatory pathways
- **1594** inhibitory pathways
- **600** intra-laminar pathways
- **1658** inter-laminar pathways
- Mean synapses/connection

Exc.	Inh.
4.3	8.5

Synaptic physiology

- **6** synapse types
- **207** synaptomes
- **Space clamp** corrected **synaptic conductances** for **607** pathways
- The per synapse conductance of **1.5** nS **for connections between L5TTPCs** is the highest in the microrcircuit
- Mean conductance per synapse: **0.85** nS for excitatory & **0.66** nS for inhibitory synapses
- Total conductance in a single neuron is **971** nS

Excitatory	Inhibitory
697 nS	274 nS



Markram et al, Cell 2015

https://bbp.epfl.ch/nmc-portal

- 80 authors
- Joint effort between computer and neuro-scientists
- Reproducible work
- Extensible

HPC Today's Infrastructure

Elastic Compute

- Visualization and analysis
- Web services
- SW development
- Continuous Integration
- Continuous Deployment

IaaS
4 nodes

Viz & Analytics
36 nodes

IaaS
12 nodes

Viz & Analytics
14 nodes

GVA

CEPH

NetApp

CEPH

GPFS

Production HPC

- Model development
- Reconstruction
- Simulation
- SW development

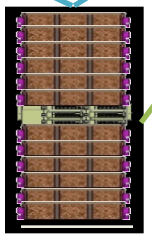
GPFS

4 Compute Racks
4096 Compute Nodes
5D CN torus
0.8 PF/s peak
64TB DRAM
4.2 PB GPFS storage

IBM BlueGene/Q

Systems Research

- Memory extension
- Application coupling
- Interactive supercomputing
- Reproducibility



64 IONodes
3D torus
128TB Flash
Linux

IBM BlueGene Active Storage

Blue Brain Project

HPC Resources usage

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Facts:

- 70 Users
- 35 Projects
- 3 Clusters

	Daily core hours available	Daily Job submitted	Active user*
BlueGene	1572864	103	20
Lugano cluster	13824	632	53
Geneva cluster	4512	259	20

* User who submitted at least one job over the last month

Blue Brain Project

HPC Resources usage

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Facts:

- 70 Users
- 35 Projects
- 3 Clusters

	Daily core hours available	Daily Job submitted	Active user*
BlueGene	1572864	103	20
Lugano cluster	13824	632	53
Geneva cluster	4512	259	20

* User who submitted at least one job over the last month

Observations:

- Clusters are scheduling 10x more jobs than bluegene
- Cluster nodes are shared (--exclusive is limited), not bluegene

Emphasis of the presentation will be put on HPC clusters

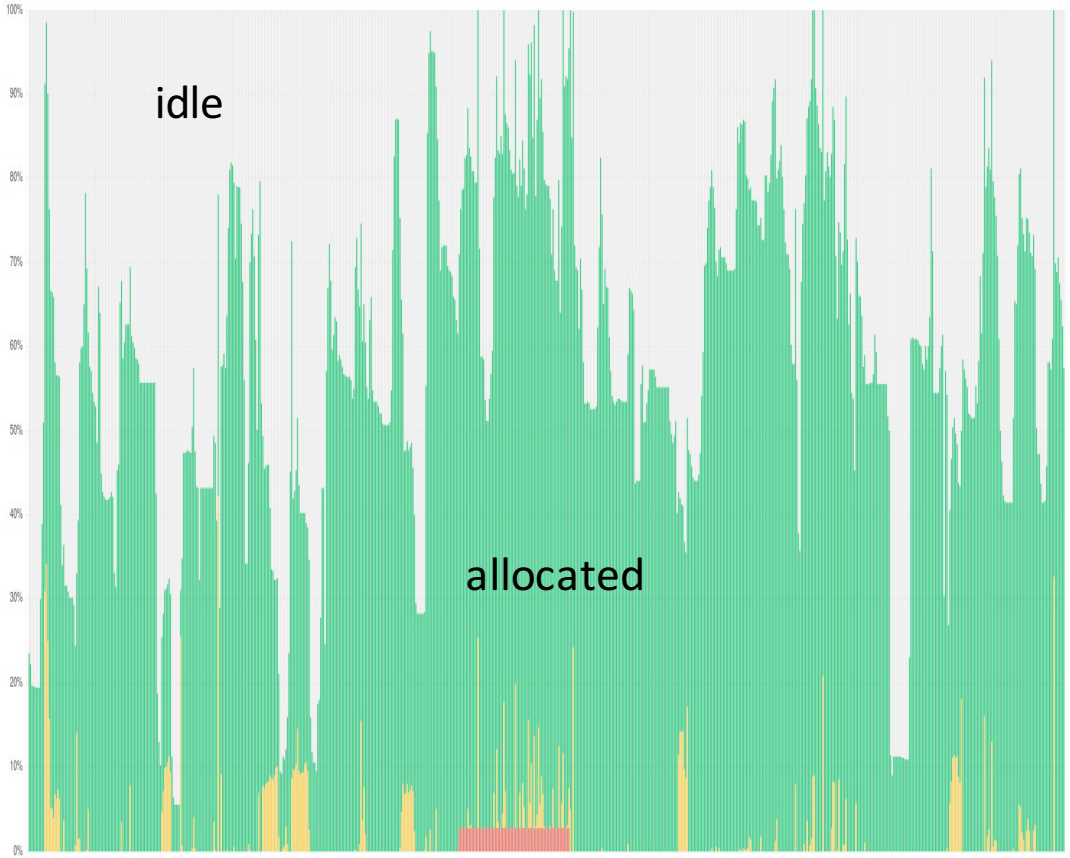
Job monitoring

Blue Brain Project

Problem description

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Interactive HPC Clusters usage over time



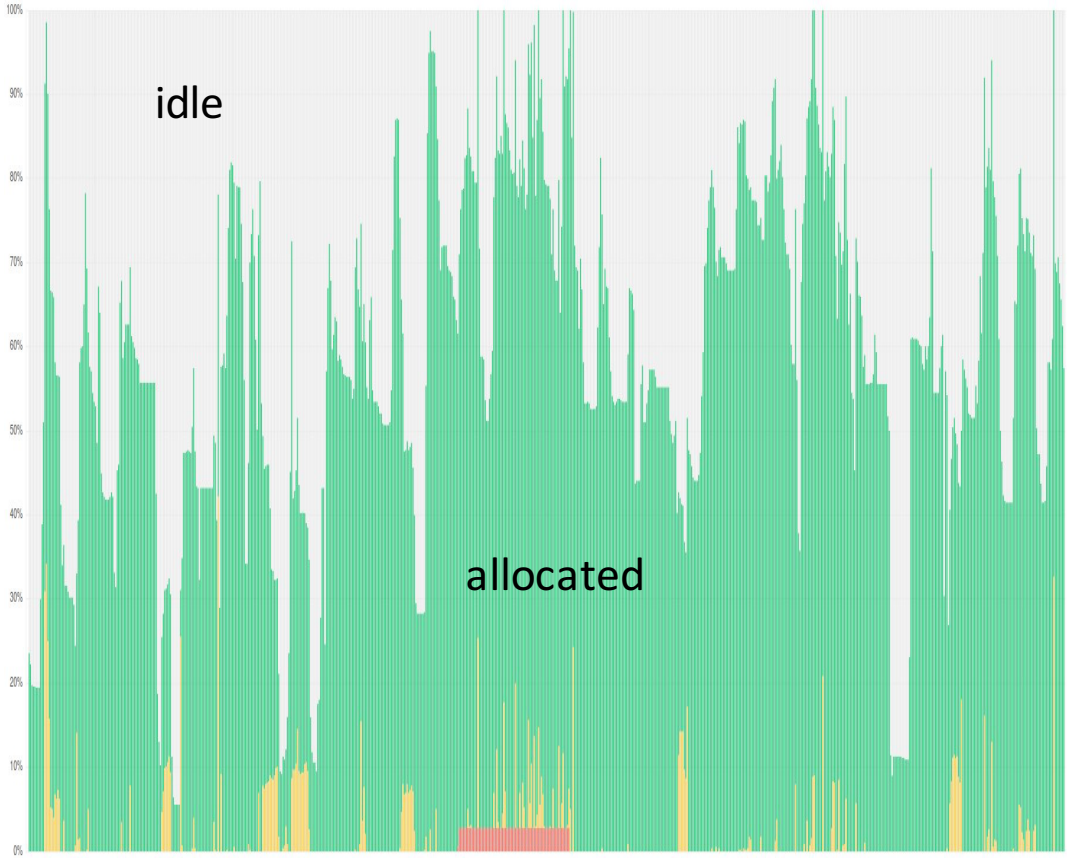
*Extracted from **sreport***

Blue Brain Project

Problem description

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Interactive HPC Clusters usage over time



*Extracted from **sreport***

Symptoms:

User can't get an allocation

Cause:

Cluster is fully allocated

Solution:

1) *Buy a bigger one* ☺
2) See if resources are optimally used

Blue Brain Project

Slurm accounting DB

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Knowledge of all jobs / step executed

- Average waiting time in the queue
- Submission rate by user / project

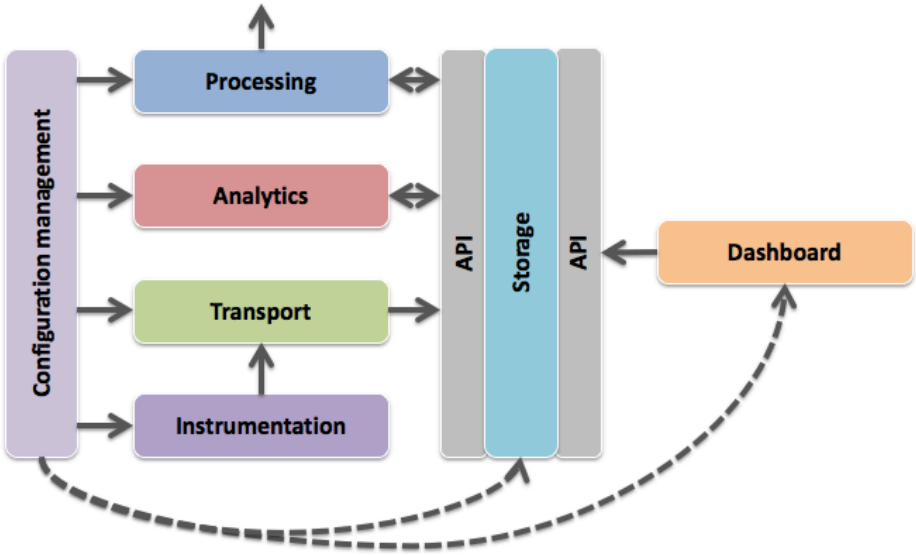
“sacct” data are indexed into ElasticSearch

- Near real-time (every 10 minutes)
- Analytics, web report generation

Limitations:

- Not natively aware of resources used by the job

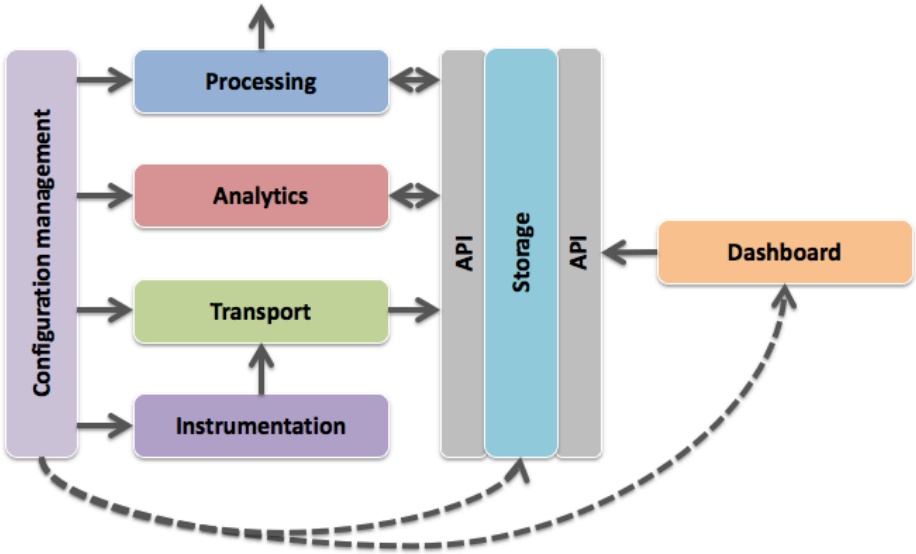
BBP Monitoring infrastructure



Scalable and extensible framework

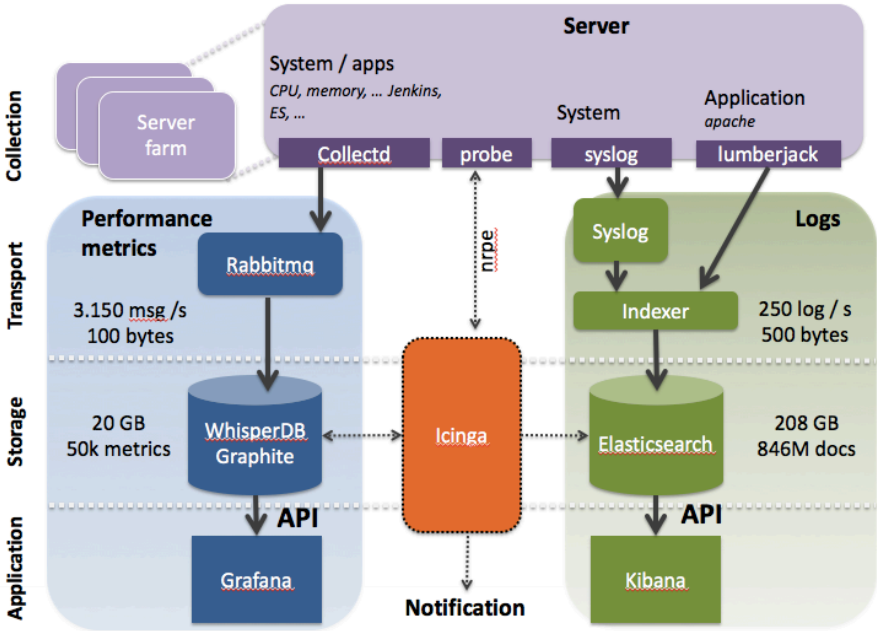
- Based on open source technologies
- Enable data collection and analysis

BBP Monitoring infrastructure



Scalable and extensible framework

- Based on open source technologies
- Enable data collection and analysis



Native host instrumentation

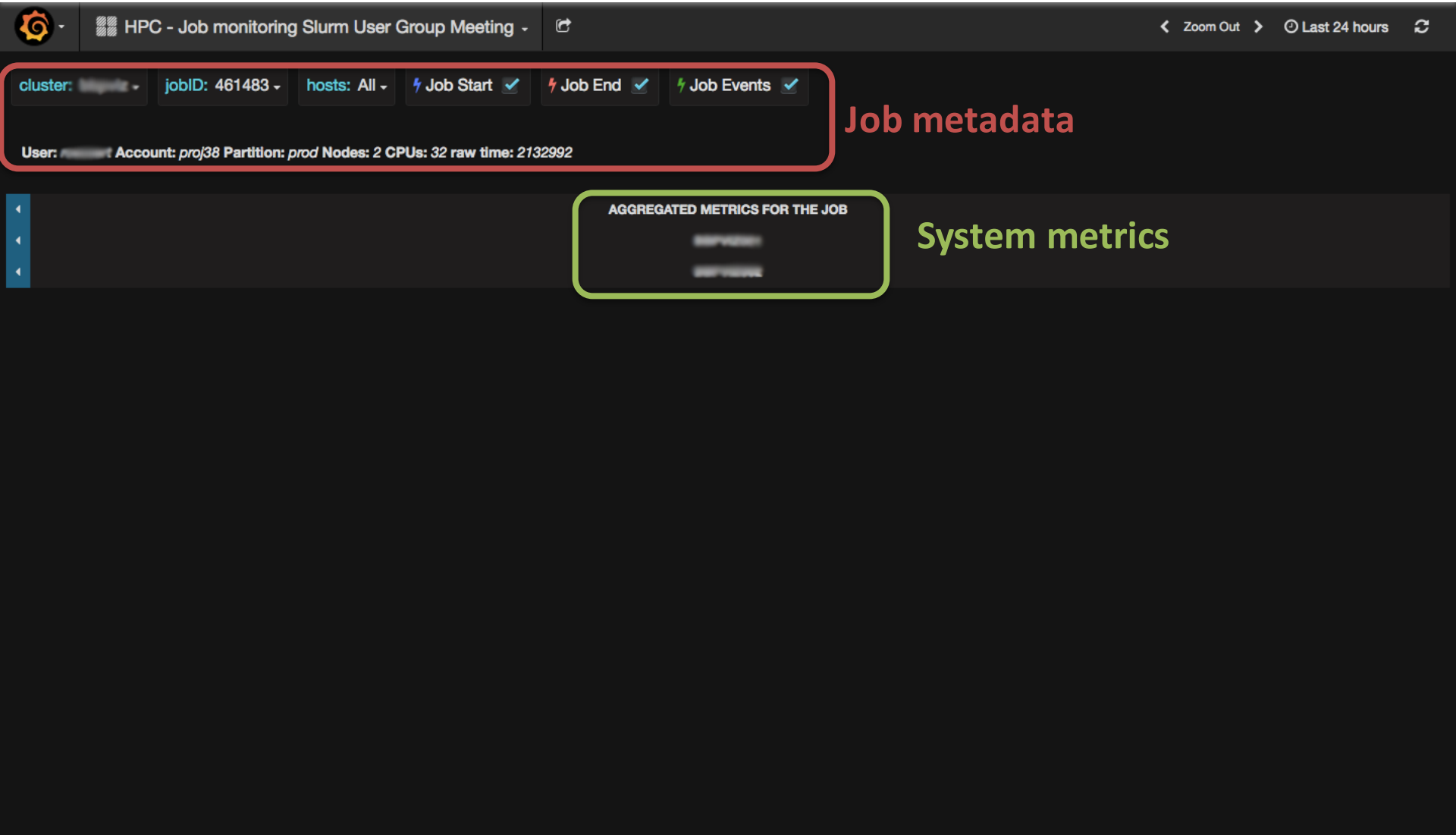
- **10** seconds resolution
- **250** metrics per node

Limitation: System metrics does not have any knowledge about workload

Blue Brain Project

Dashboard 1/3

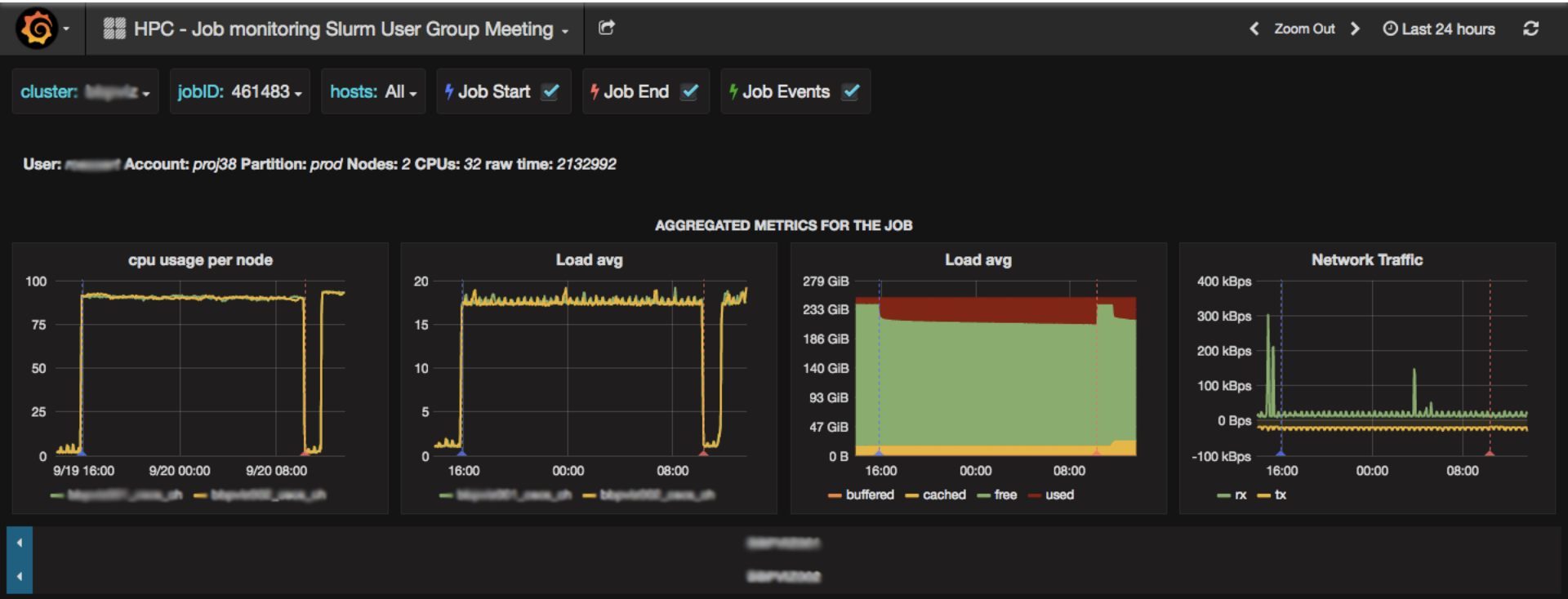
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE



Blue Brain Project

Dashboard 2/3

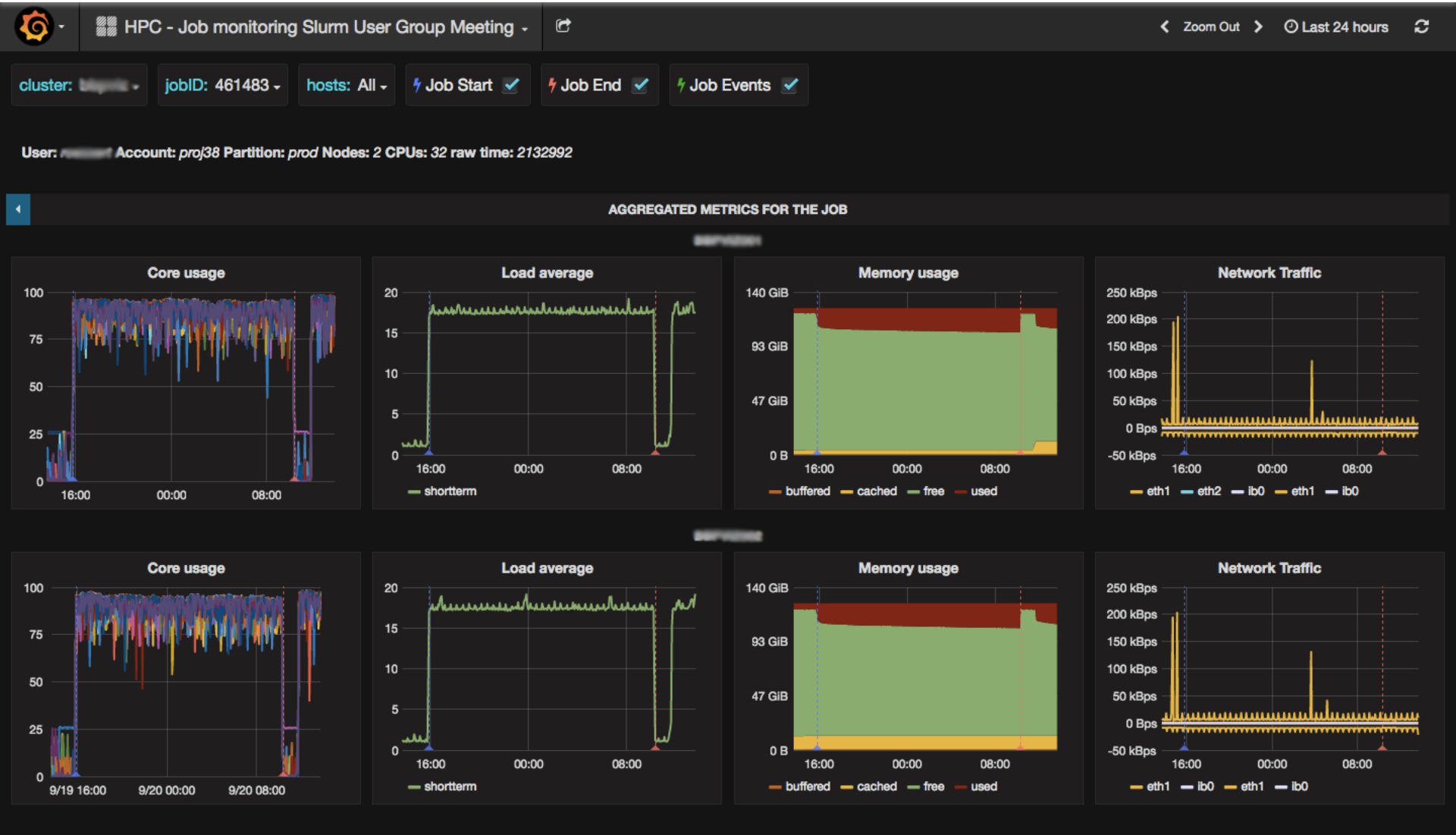
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE



Blue Brain Project

Dashboard 3/3

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE



Blue Brain Project

Non-optimal use case

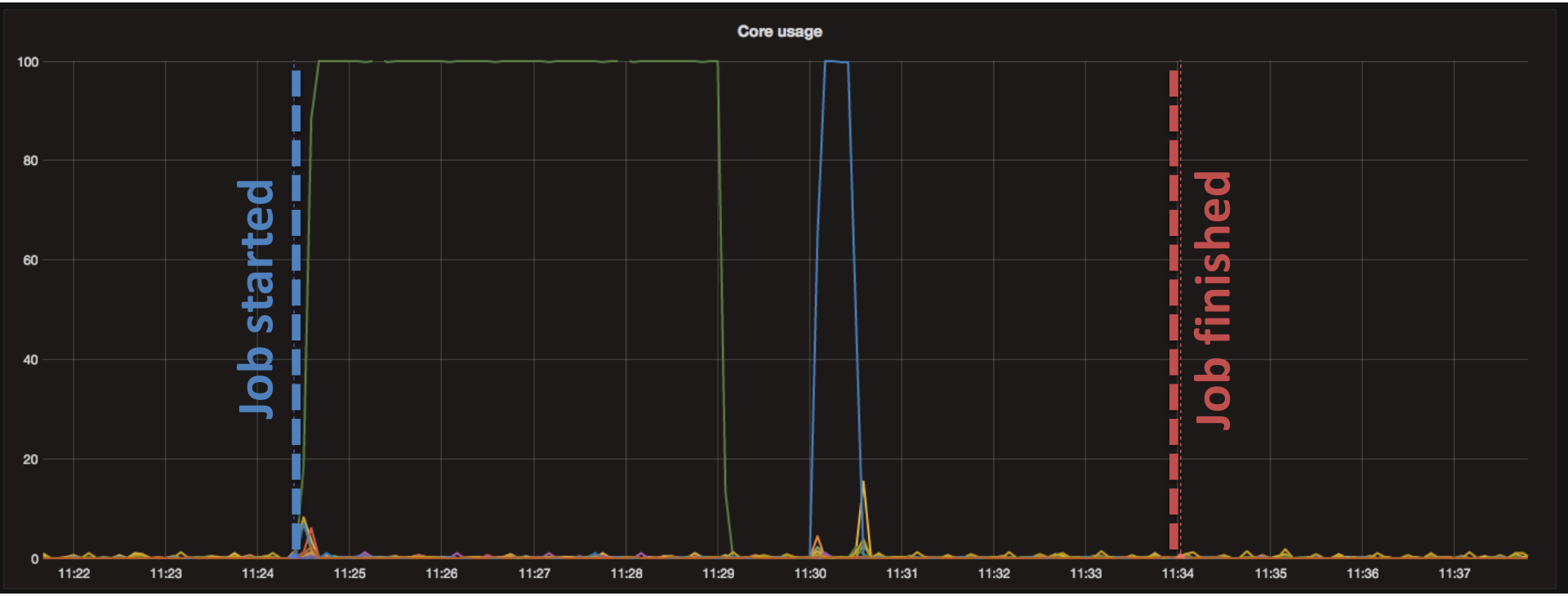
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Allocation details:

- Single nodes, all cores, batch partition

Job details:

- CPU bound



Single-thread used

Un-used allocation

Benefits

Developers

- Analyze code from a system perspective
  - Non-intrusive monitoring / negligible (perf) overhead
- Detect code inefficiency / limiting resource
  - Non optimal parallelization

Benefits

Developers

- Analyze code from a system perspective
  - Non-intrusive monitoring / negligible (perf) overhead
- Detect code inefficiency / limiting resource
  - Non optimal parallelization

System admins

- Analyze system metrics with workload context
- Detect non-optimal allocation
  - Allocation bigger than execution time

Holistic view enabling cross team (competencies) debugging

Ongoing work / limitation

Ongoing work

- Creating KPI out of the available metrics
  - Efficiency of a job (cpu seconds used /reserved)

Ongoing work / limitation

Ongoing work

- Creating KPI out of the available metrics
  - Efficiency of a job (cpu seconds used /reserved)

Limitations

- Mainly system metrics so far
  - Only memory are collected at cgroup level
- Missing infiniband metrics
- Job internals are hidden

Beyond Slurm monitoring

Blue Brain Project

Job workload context

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Focus has been put on monitoring the job in the infrastructure

No hint is given to the job internals

- Job entering in a given phase

Job workload context

Focus has been put on monitoring the job in the infrastructure

No hint is given to the job internals

- Job entering in a given phase

Providing a library to enhance job context

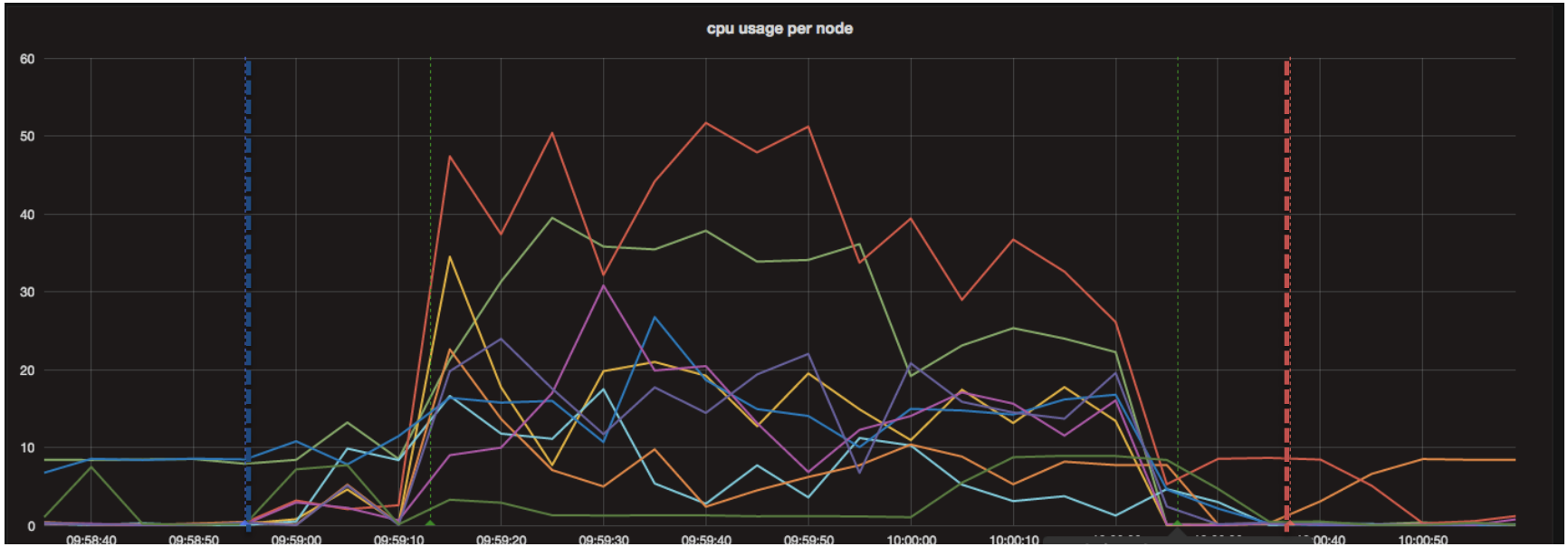
- Workload manager agnostic
- Not a profiling tool
- Lightweight way of sharing information from the job
- Allows to ship user-defined metadata

Blue Brain Project

BBP_instrumentations

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

```
def count_element(self):
    lines = self._data.count()
```

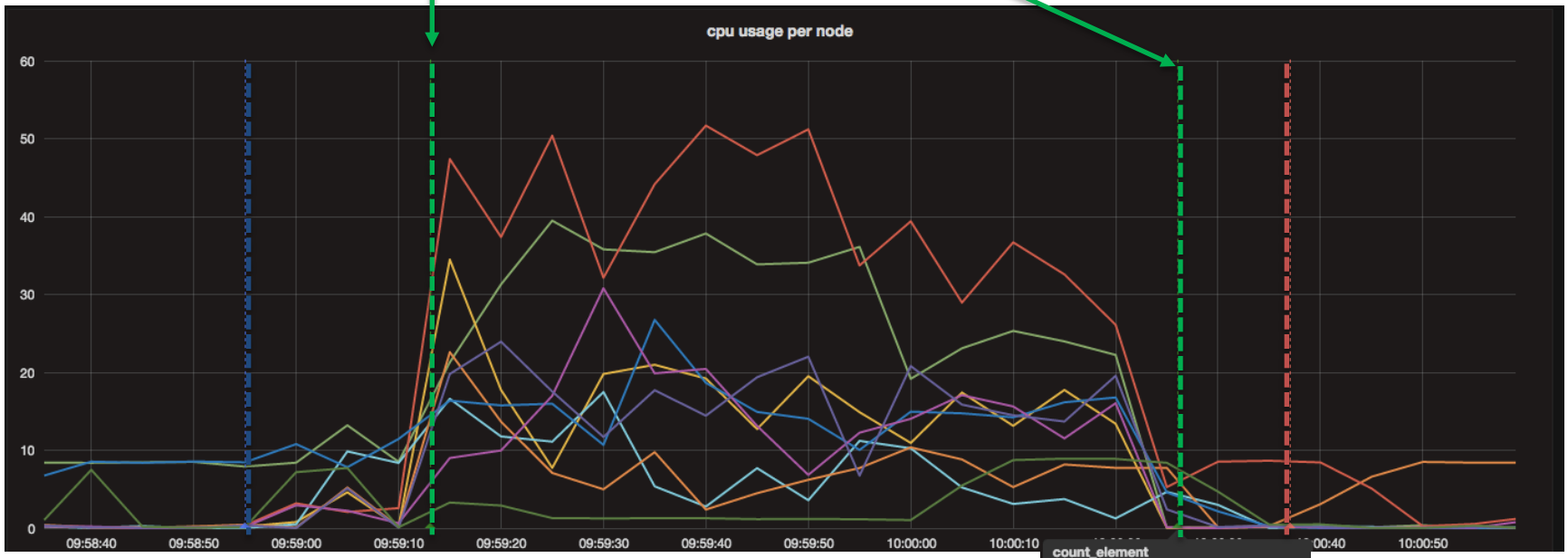


Blue Brain Project

BBP_instrumentations

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

```
@log_function_events
def count_element(self):
    lines = self._data.count()
```



Arbitrary tags

Report metrics

Summary

Detection of non-optimal usage

- Un-used allocation
- Developers now have tools to understand job behaviors

Internal job monitoring

- Allows understanding which resources are consumed by section of job through user-defined metadata

Correlation of scattered information enable powerful analysis

Blue Brain Project

Acknowledgements

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

BBP core services & HPC teams