UNIVERSITÉ LAVAL

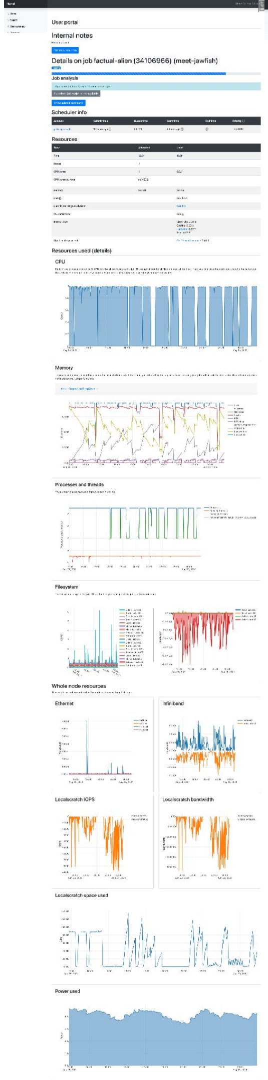
Calcul Québec

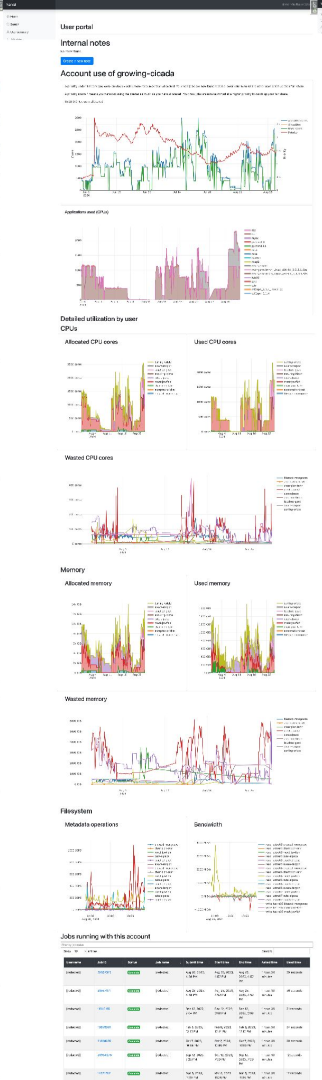
TrailblazingTurtle:
A Comprehensive Web Portal for Maximizing HPC Resource Utilization

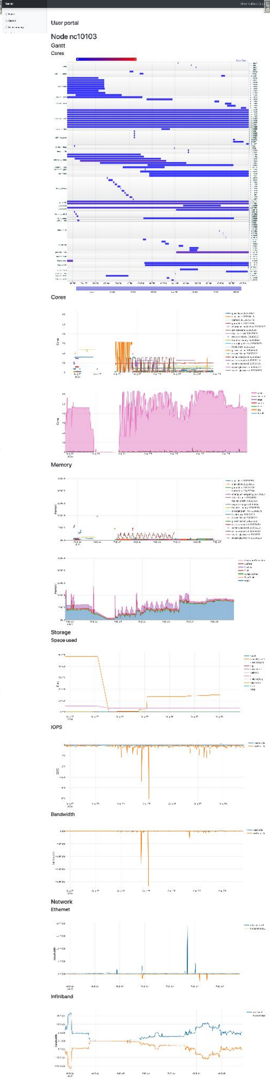
Simon Guilbault (simon.guilbault@calculquebec.ca)

Introduction

- HPC users struggle to fully use resources
  - Waste resources
  - Do not get the speedup they could have
- Solution
  - TrailblazingTurtle: A web portal to aggregate and present relevant information to users and staff
- Features
  - Job level monitoring
    - CPU, memory, GPUs, …
    - Lustre IO
  - Top users stats
  - Job table (squeue alternative)
  - Account stats
  - Long stats retention
  - Public view: portail.narval.calculquebec.ca







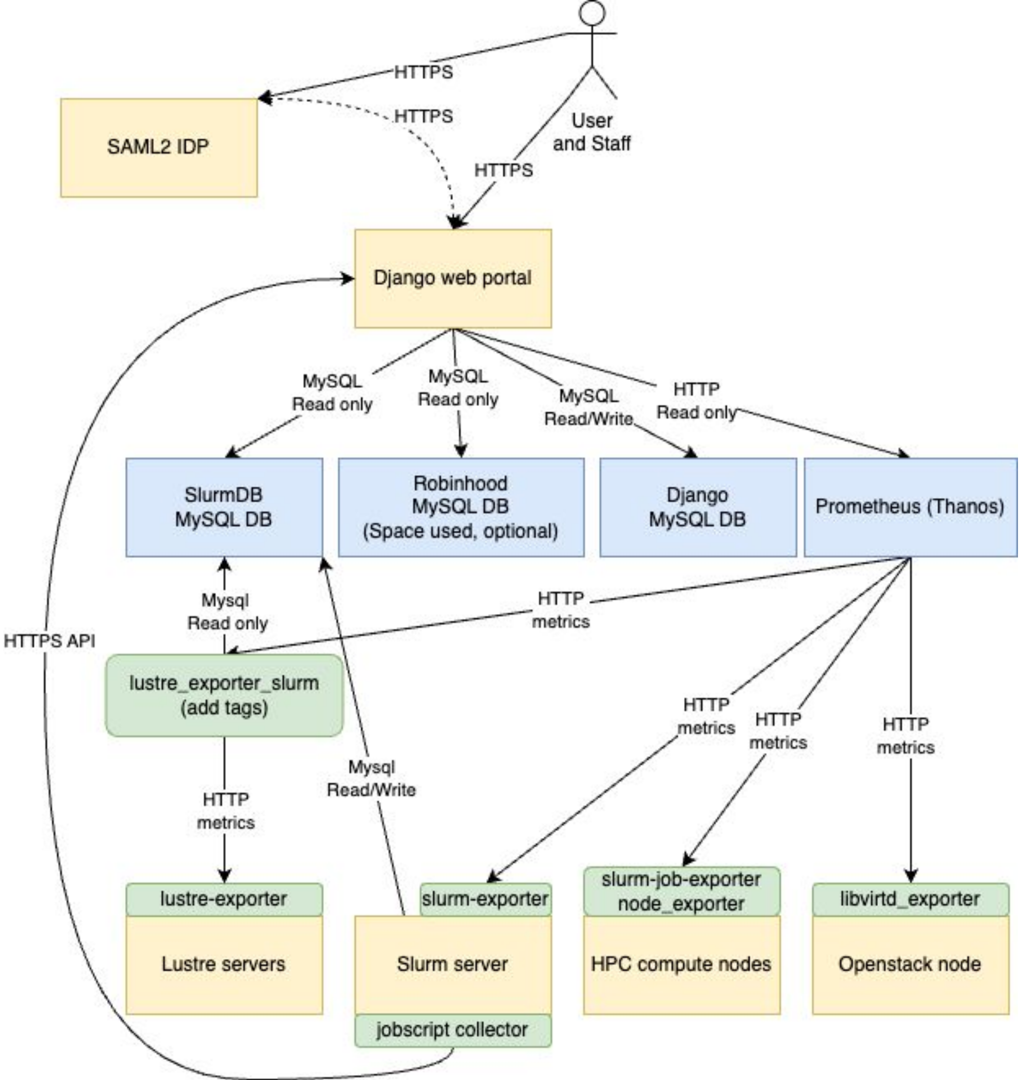
Overall design

- Django web portal
  - Can filter the view for each user and allow staff to see everything

- Direct access to MySQL Slurmdb

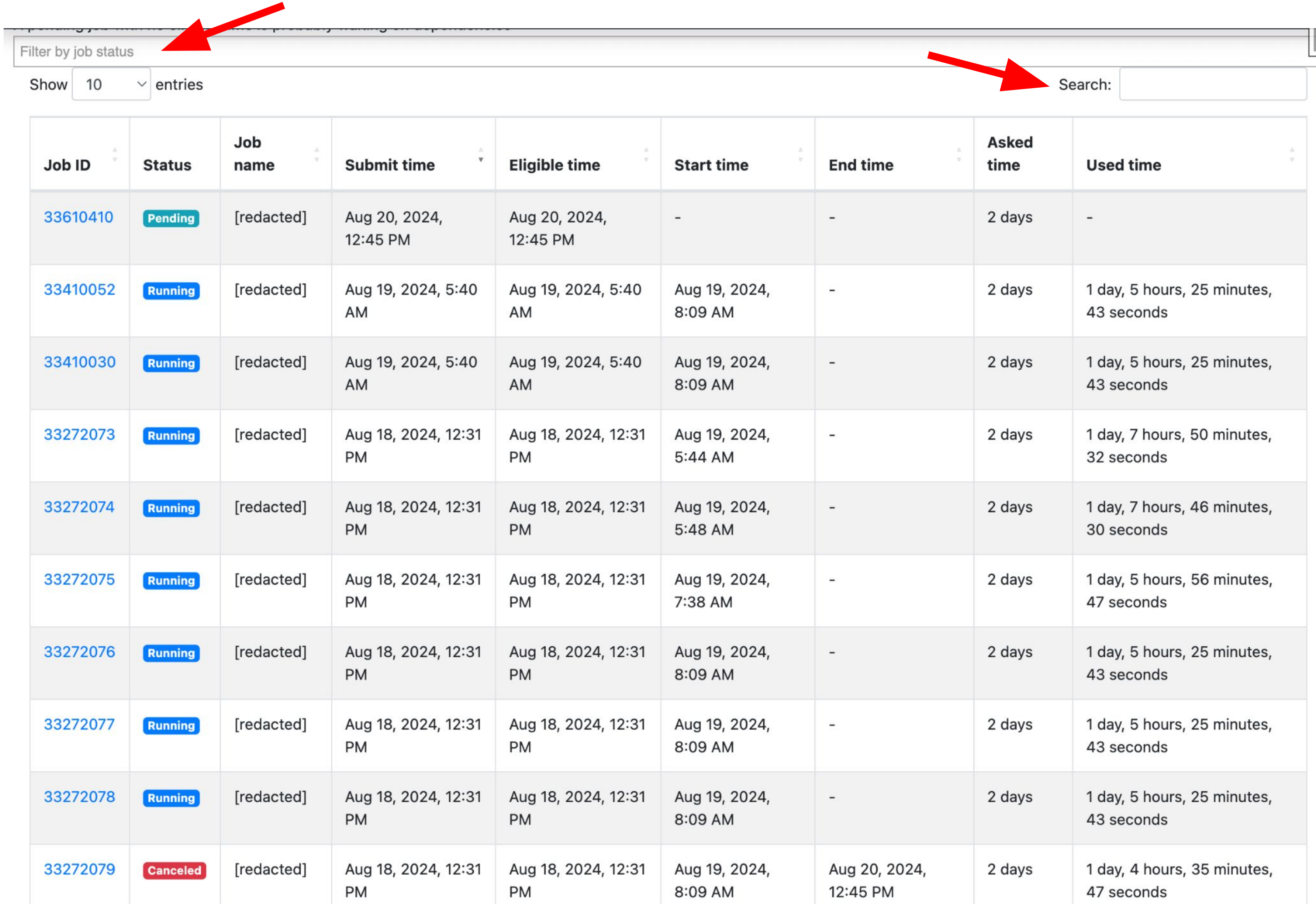
- Metrics in Prometheus
  - Thanos for retention in S3
  - Slurm-job-exporter
    - Other exporters are optional

- Collect job scripts with a REST API
  - 23.02 added this feature natively



Jobs table

- Direct access to database
- Full text search
- Filter by state
  - Pending
  - Running
  - Completed
  - OOM
  - Failed
  - Timeout
  - …
- Order by



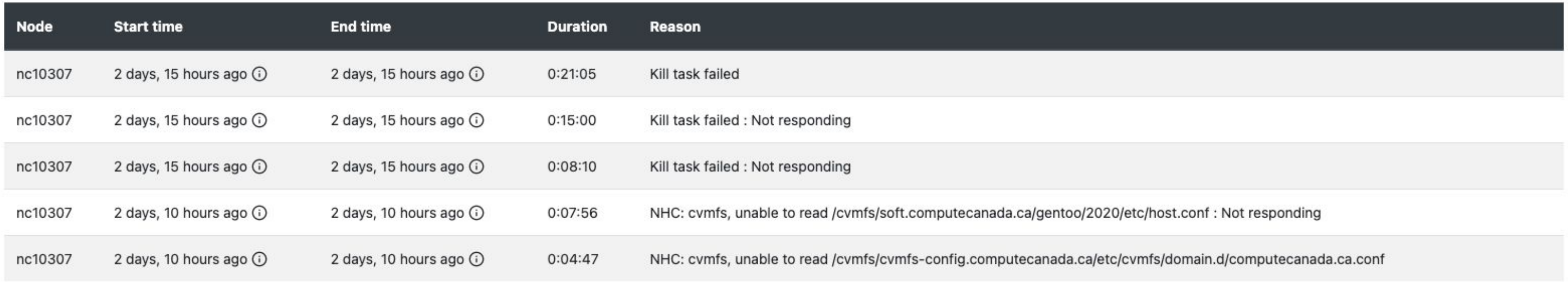
Filter by job status

Show 10 entries

Search:

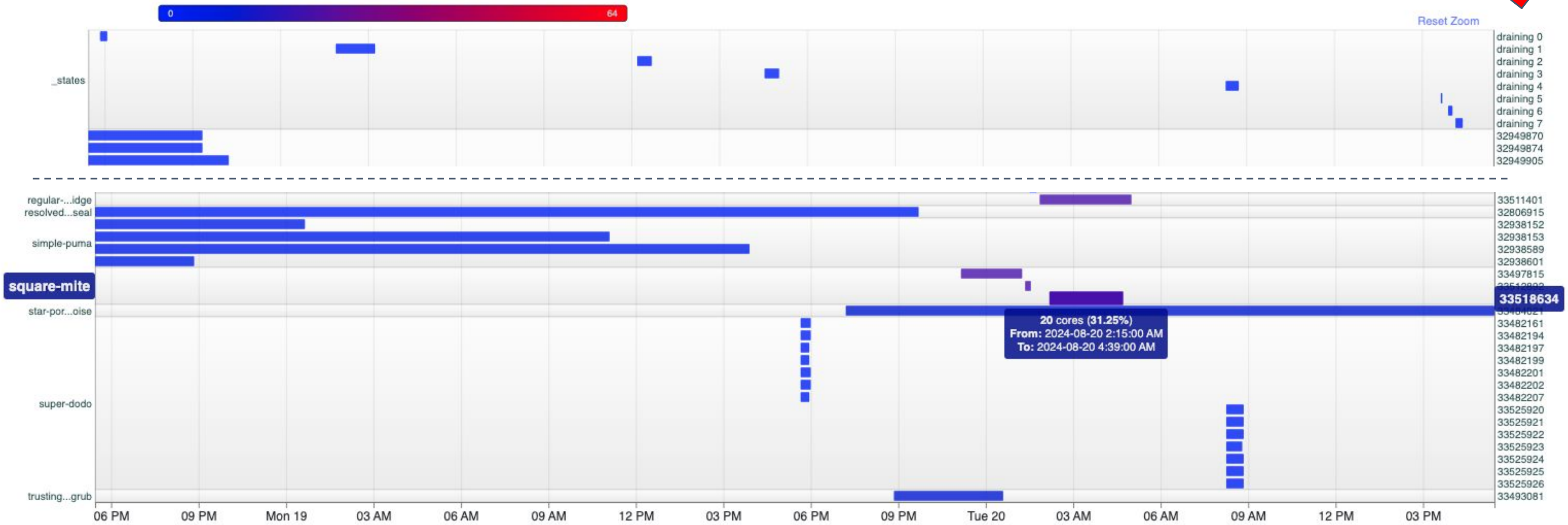
Job ID	Status	Job name	Submit time	Eligible time	Start time	End time	Asked time	Used time
33610410	Pending	[redacted]	Aug 20, 2024, 12:45 PM	Aug 20, 2024, 12:45 PM	-	-	2 days	-
33410052	Running	[redacted]	Aug 19, 2024, 5:40 AM	Aug 19, 2024, 5:40 AM	Aug 19, 2024, 8:09 AM	-	2 days	1 day, 5 hours, 25 minutes, 43 seconds
33410030	Running	[redacted]	Aug 19, 2024, 5:40 AM	Aug 19, 2024, 5:40 AM	Aug 19, 2024, 8:09 AM	-	2 days	1 day, 5 hours, 25 minutes, 43 seconds
33272073	Running	[redacted]	Aug 18, 2024, 12:31 PM	Aug 18, 2024, 12:31 PM	Aug 19, 2024, 5:44 AM	-	2 days	1 day, 7 hours, 50 minutes, 32 seconds
33272074	Running	[redacted]	Aug 18, 2024, 12:31 PM	Aug 18, 2024, 12:31 PM	Aug 19, 2024, 5:48 AM	-	2 days	1 day, 7 hours, 46 minutes, 30 seconds
33272075	Running	[redacted]	Aug 18, 2024, 12:31 PM	Aug 18, 2024, 12:31 PM	Aug 19, 2024, 7:38 AM	-	2 days	1 day, 5 hours, 56 minutes, 47 seconds
33272076	Running	[redacted]	Aug 18, 2024, 12:31 PM	Aug 18, 2024, 12:31 PM	Aug 19, 2024, 8:09 AM	-	2 days	1 day, 5 hours, 25 minutes, 43 seconds
33272077	Running	[redacted]	Aug 18, 2024, 12:31 PM	Aug 18, 2024, 12:31 PM	Aug 19, 2024, 8:09 AM	-	2 days	1 day, 5 hours, 25 minutes, 43 seconds
33272078	Running	[redacted]	Aug 18, 2024, 12:31 PM	Aug 18, 2024, 12:31 PM	Aug 19, 2024, 8:09 AM	-	2 days	1 day, 5 hours, 25 minutes, 43 seconds
33272079	Canceled	[redacted]	Aug 18, 2024, 12:31 PM	Aug 18, 2024, 12:31 PM	Aug 19, 2024, 8:09 AM	Aug 20, 2024, 12:45 PM	2 days	1 day, 4 hours, 35 minutes, 47 seconds

Node states, events and jobs gantt-chart



Node	Start time	End time	Duration	Reason
nc10307	2 days, 15 hours ago ⓘ	2 days, 15 hours ago ⓘ	0:21:05	Kill task failed
nc10307	2 days, 15 hours ago ⓘ	2 days, 15 hours ago ⓘ	0:15:00	Kill task failed : Not responding
nc10307	2 days, 15 hours ago ⓘ	2 days, 15 hours ago ⓘ	0:08:10	Kill task failed : Not responding
nc10307	2 days, 10 hours ago ⓘ	2 days, 10 hours ago ⓘ	0:07:56	NHC: cvmfs, unable to read /cvmfs/soft.computecanada.ca/gentoo/2020/etc/host.conf : Not responding
nc10307	2 days, 10 hours ago ⓘ	2 days, 10 hours ago ⓘ	0:04:47	NHC: cvmfs, unable to read /cvmfs/cvmfs-config.computecanada.ca/etc/cvmfs/domain.d/computecanada.ca.conf

NHC



Data sources

- Mysql slurmdb
  - Read only access

- Prometheus
  - node_exporter
  - slurm-job-exporter
    - Job level stats
- slurm-exporter
  - Account priorities, node down
- lustre_exporter + lustre_exporter_slurm
  - Collect stats by job, and add username/group based on slurmdb
- redfish_exporter
  - Power by node (Dell iDRAC)
- pcm-sensor-server
  - Intel only: L3 cache, IPC, NUMA and Memory bandwidth

Prometheus

- 1500 compute nodes (2000 nodes in total)
  - 250k metrics per second
  - 2 bytes per sample
  - 0.5MB/s -> 43 GB per day

- Aggregation using recorder rules
  - Sum per user, ...

- Production VM (1 per cluster)
  - ~6 cores
  - 70GB of ram
  - 350GB of disk, ~25 IOPS, 2 MB/s
  - 200Mb/s of network traffic
  - Local retention of a few days

Thanos

- Archival and aggregation of multiple clusters

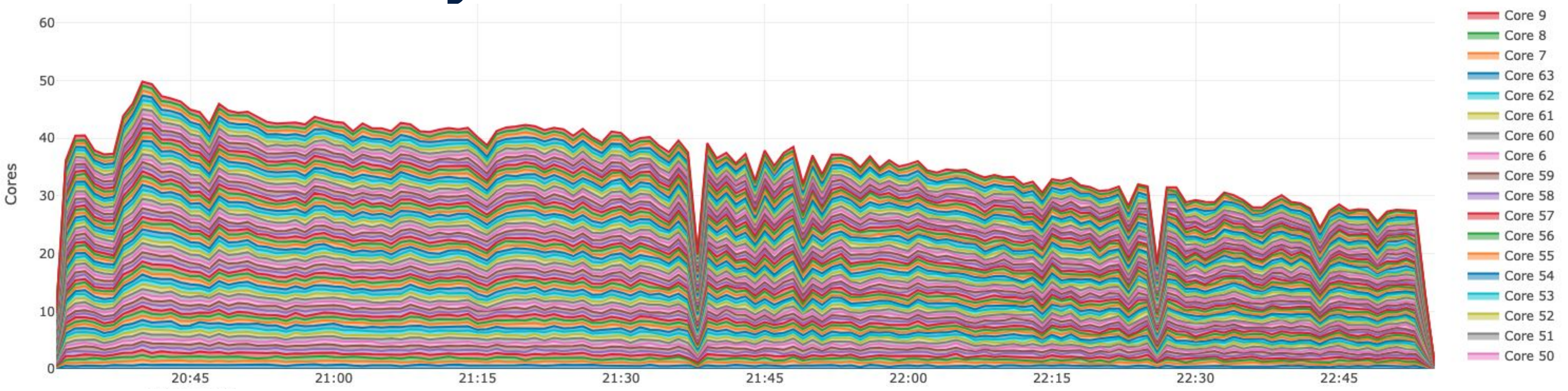
- Compact up to 14 days chunks
- S3 storage (Ceph)
  - 30 TB
- Removing some stats after 7 months
  - Rewriting blocks
  - 40 GB -> 10 GB

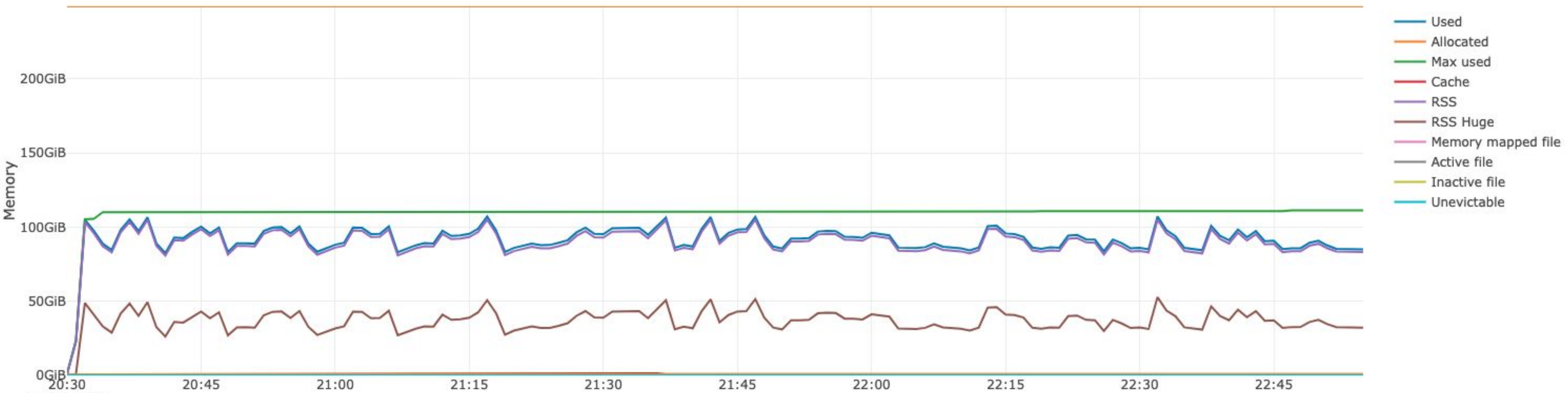
slurm-job-exporter

- Gather metrics within each cgroup created by Slurm to contain each job
  - `/sys/fs/cgroup/memory/slurm/uid_1000/job_42/`
  - `SLURM_JOB_ACCOUNT` to get the account
  - CPU
    - Nanoseconds counter per core
  - Memory
    - Can measure the absolute peak, regardless of sampling frequency
  - Nvidia GPUs
    - DCGM and NVML
  - Process/threads count and paths within each job

```
# HELP slurm_job_memory_usage Memory used by a job
# TYPE slurm_job_memory_usage gauge
slurm_job_memory_usage{account="group1",slurmjobid="1",user="user1"} 1.634453504e+010
slurm_job_memory_usage{account="group2",slurmjobid="2",user="user2"} 8.271761408e+09
# HELP slurm_job_core_usage_total Cpu usage of cores allocated to a job
# TYPE slurm_job_core_usage_total counter
slurm_job_core_usage_total{account="group1",core="1",slurmjobid="1",user="user1"} 1.165134620225e+012
slurm_job_core_usage_total{account="group1",core="2",slurmjobid="1",user="user1"} 1.209891619592e+012
slurm_job_core_usage_total{account="group2",core="3",slurmjobid="2",user="user2"} 5.711518455e+012
```

CPU/Memory stats



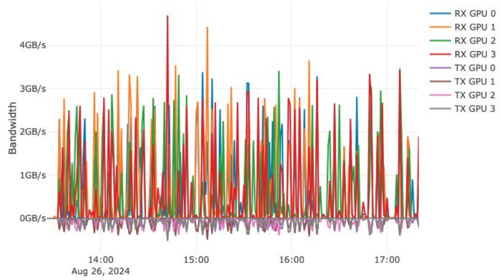


NVIDIA GPUs stats

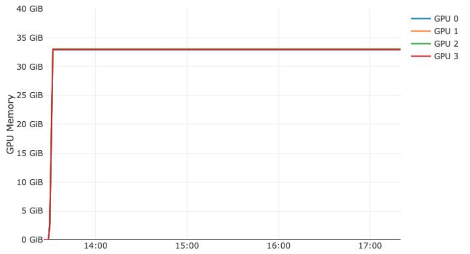
- Using DCGM (NVML fallback)
  - SM active, SM occupied
  - FP64, FP32, FP16, Tensor
  - Memory used and bandwidth
  - Nvlink/PCIe bandwidth
  - Power
- “nvidia-smi -L” in each cgroup to map each GPU to a job

GPU PCIe bandwidth

Note that this is from the perspective of the GPU

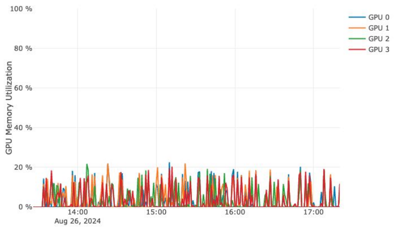


GPU Memory used



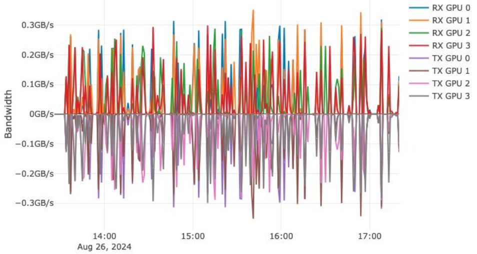
GPU Memory access cycle used

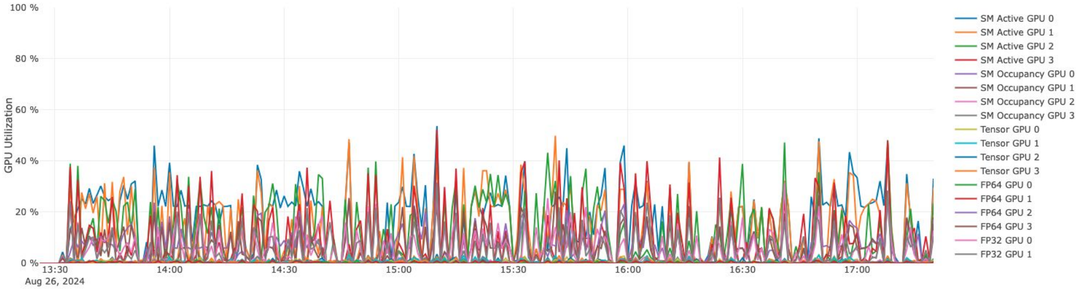
The ratio of cycles the device memory interface is active sending or receiving data



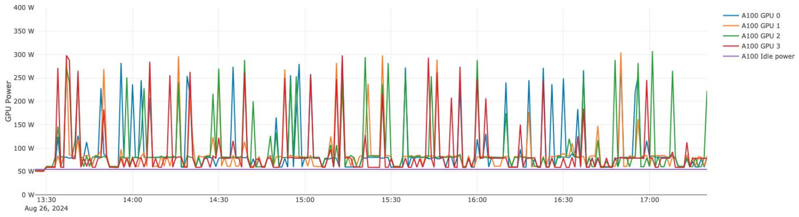
GPU Nvlink bandwidth

Note that this is from the perspective of the GPU

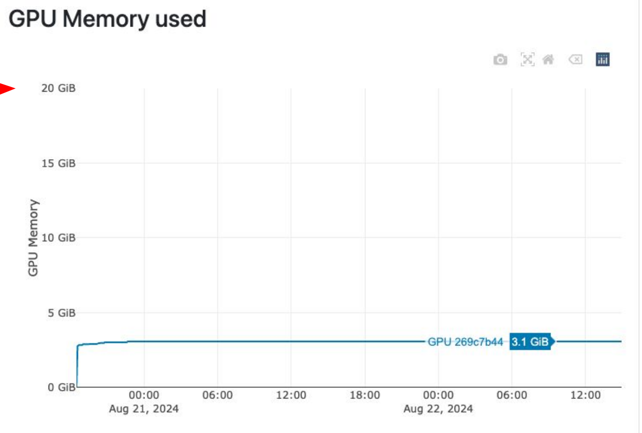




GPU Power

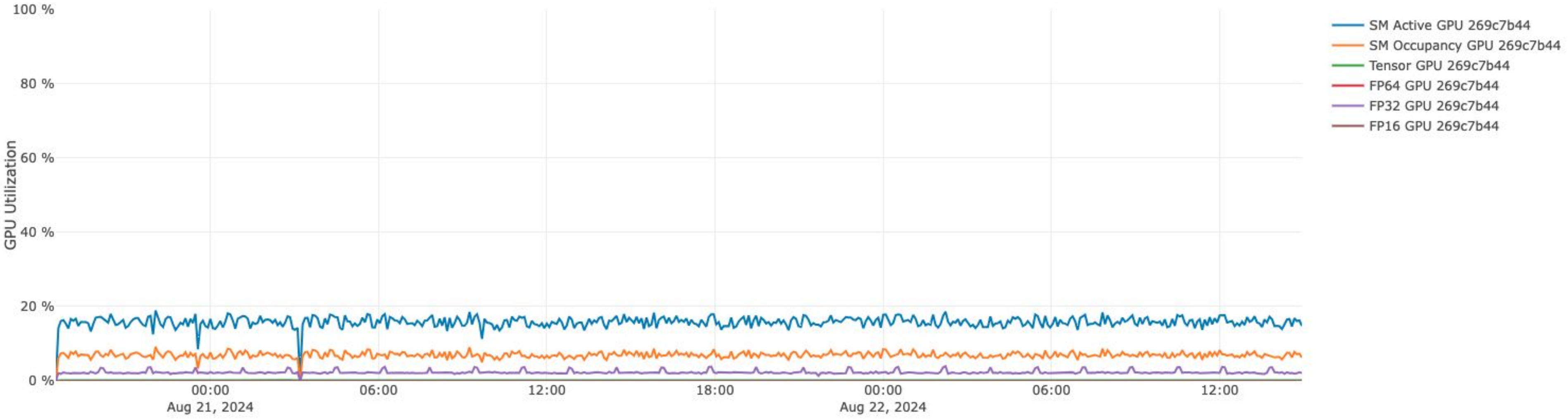


A100 40GB -> 3g.20gb

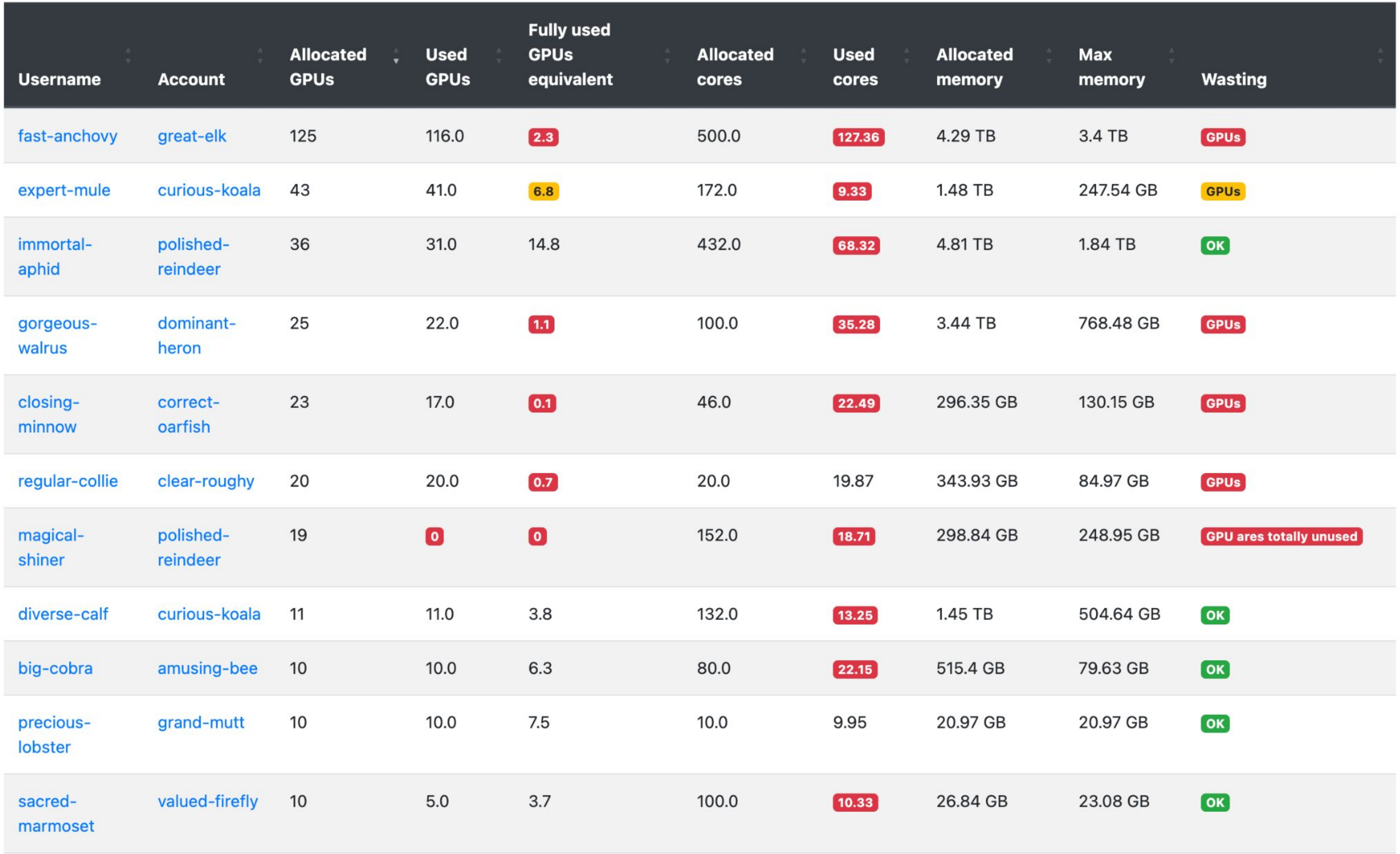

GPU Memory used

NVIDIA MIG support

- DCGM is required
- Detect and assign stats from MIG to jobs
- Using stats to evaluate how to split them in production
    - 20% of nodes are splitted with MIGs



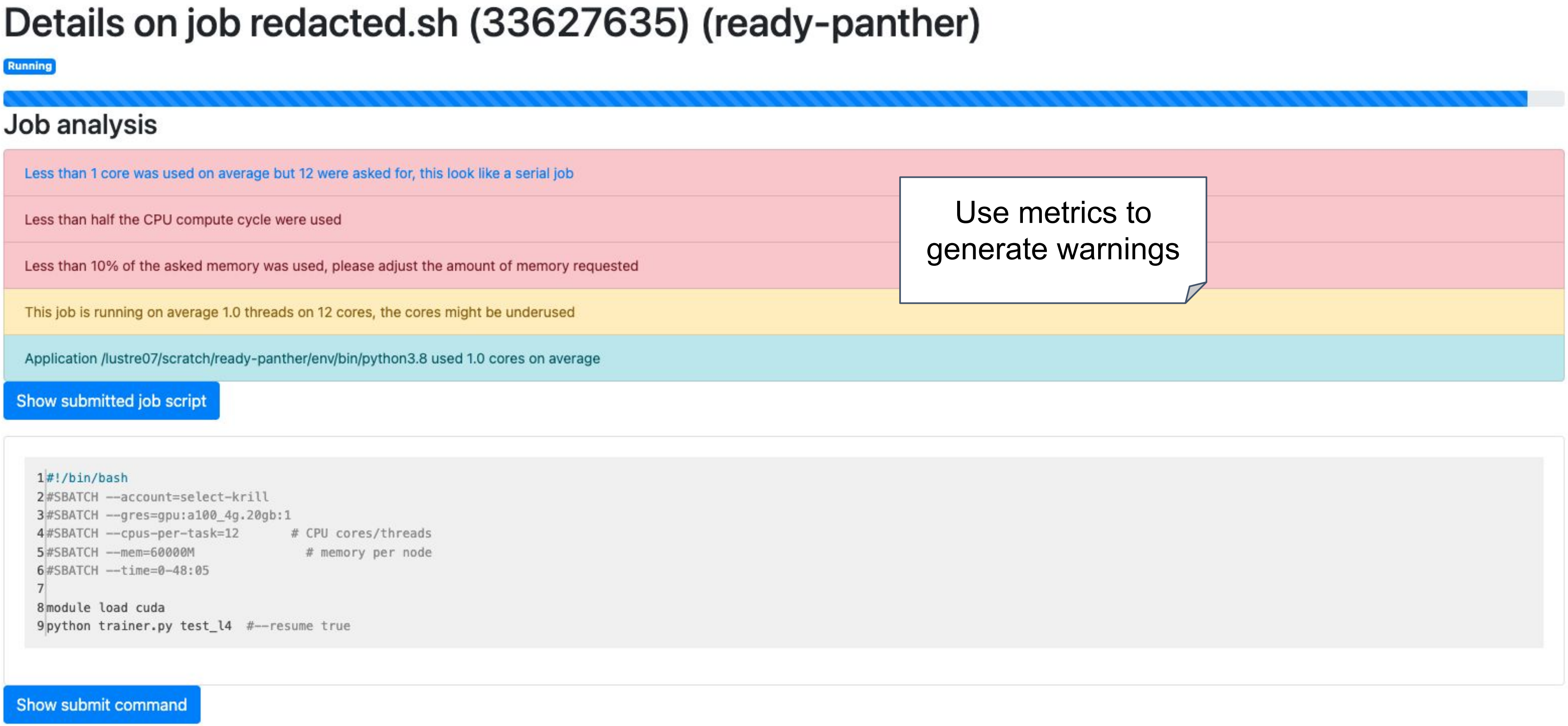
Top users



Username	Account	Allocated GPUs	Used GPUs	Fully used GPUs equivalent	Allocated cores	Used cores	Allocated memory	Max memory	Wasting
fast-anchovy	great-elk	125	116.0	2.3	500.0	127.36	4.29 TB	3.4 TB	GPUs
expert-mule	curious-koala	43	41.0	6.8	172.0	9.33	1.48 TB	247.54 GB	GPUs
immortal-aphid	polished-reindeer	36	31.0	14.8	432.0	68.32	4.81 TB	1.84 TB	OK
gorgeous-walrus	dominant-heron	25	22.0	1.1	100.0	35.28	3.44 TB	768.48 GB	GPUs
closing-minnow	correct-oarfish	23	17.0	0.1	46.0	22.49	296.35 GB	130.15 GB	GPUs
regular-collie	clear-roughy	20	20.0	0.7	20.0	19.87	343.93 GB	84.97 GB	GPUs
magical-shiner	polished-reindeer	19	0	0	152.0	18.71	298.84 GB	248.95 GB	GPU ares totally unused
diverse-calf	curious-koala	11	11.0	3.8	132.0	13.25	1.45 TB	504.64 GB	OK
big-cobra	amusing-bee	10	10.0	6.3	80.0	22.15	515.4 GB	79.63 GB	OK
precious-lobster	grand-mutt	10	10.0	7.5	10.0	9.95	20.97 GB	20.97 GB	OK
sacred-marmoset	valued-firefly	10	5.0	3.7	100.0	10.33	26.84 GB	23.08 GB	OK

Some are not even using the requested GPU

Job analysis using metrics



Details on job redacted.sh (33627635) (ready-panther)

Job analysis

Less than 1 core was used on average but 12 were asked for, this look like a serial job

Less than half the CPU compute cycle were used

Less than 10% of the asked memory was used, please adjust the amount of memory requested

This job is running on average 1.0 threads on 12 cores, the cores might be underused

Application /lustre07/scratch/ready-panther/env/bin/python3.8 used 1.0 cores on average

```
#!/bin/bash
#SBATCH --account=select-krill
#SBATCH --gres=gpu:a100_4g.20gb:1
#SBATCH --cpus-per-task=12       # CPU cores/threads
#SBATCH --mem=60000M               # memory per node
#SBATCH --time=0-48:05

module load cuda
python trainer.py test_l4  #--resume true
```

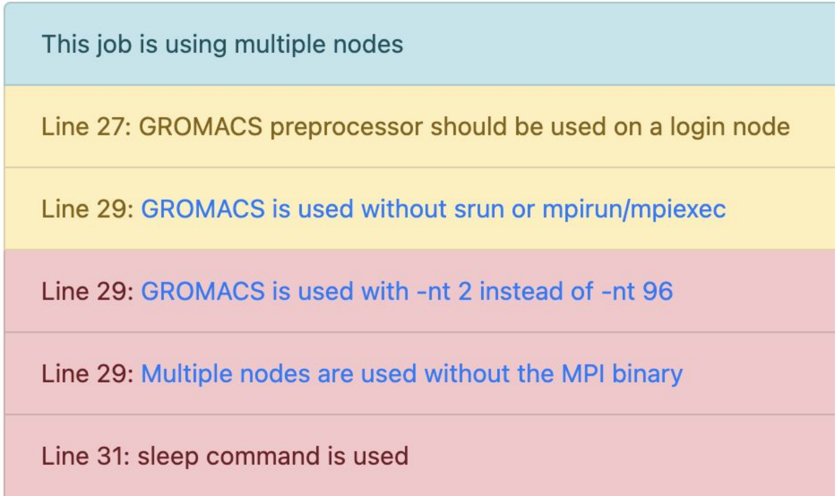
Use metrics to generate warnings

Job analysis using submitted script

- Regex templates to trigger messages

```
#SBATCH --ntasks=96        # number of MPI processes
module load  StdEnv/2020 gcc/9.3.0 openmpi/4.0.3 gromacs/2020.4
gmx grompp -f $mdp/emin_$LAMBDA.mdp
gmx mdrun -v -deffnm emin$LAMBDA -nt 2
sleep 10
```

Simplified script



This job is using multiple nodes

Line 27: GROMACS preprocessor should be used on a login node

Line 29: GROMACS is used without srun or mpirun/mpiexec

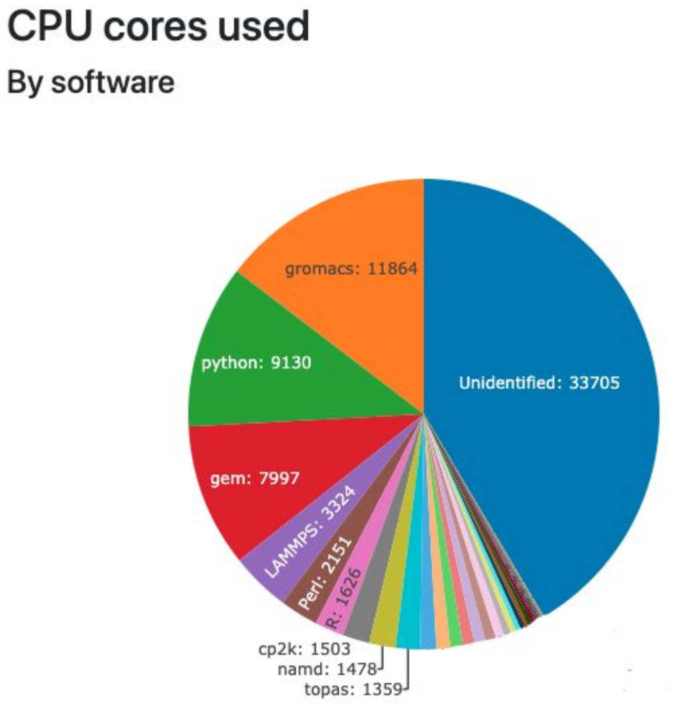
Line 29: GROMACS is used with -nt 2 instead of -nt 96

Line 29: Multiple nodes are used without the MPI binary

Line 31: sleep command is used

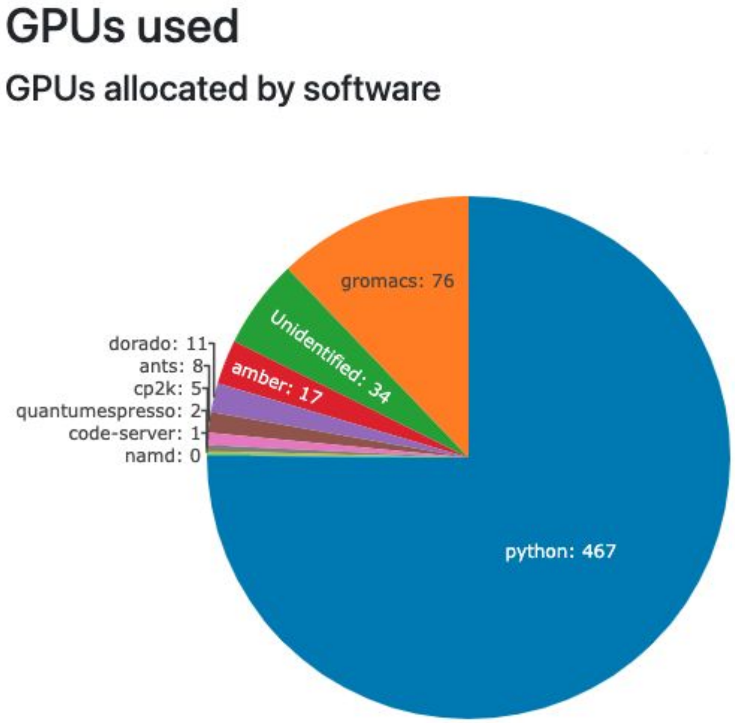
Software used on the cluster

- Gather all process in the cgroup
    - Regex to extract software used



CPU cores used

By software

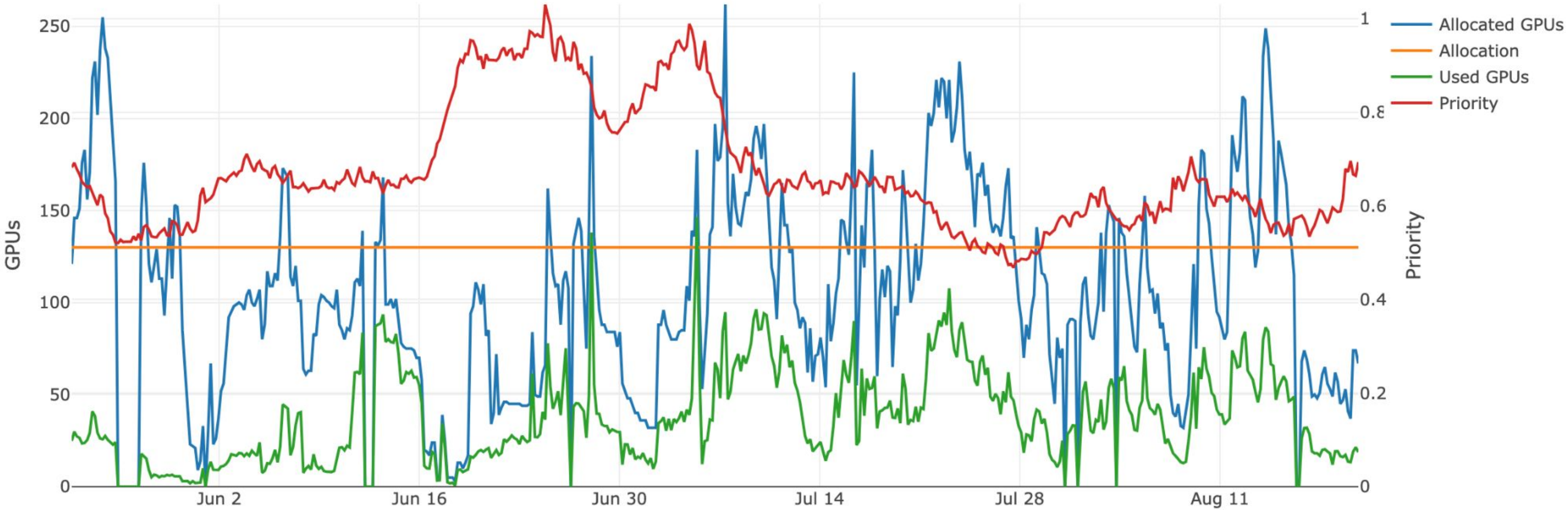


GPUs used

GPUs allocated by software

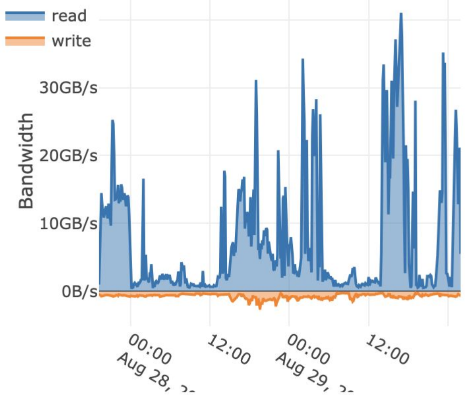
slurm-exporter

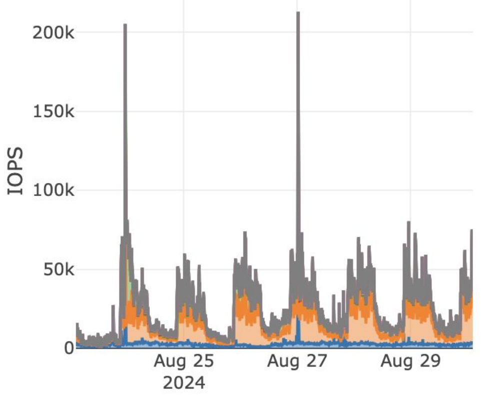
- Account priorities (levelFS)
  - See impacts of previous jobs on priority
- Resources requested and used
- Node states

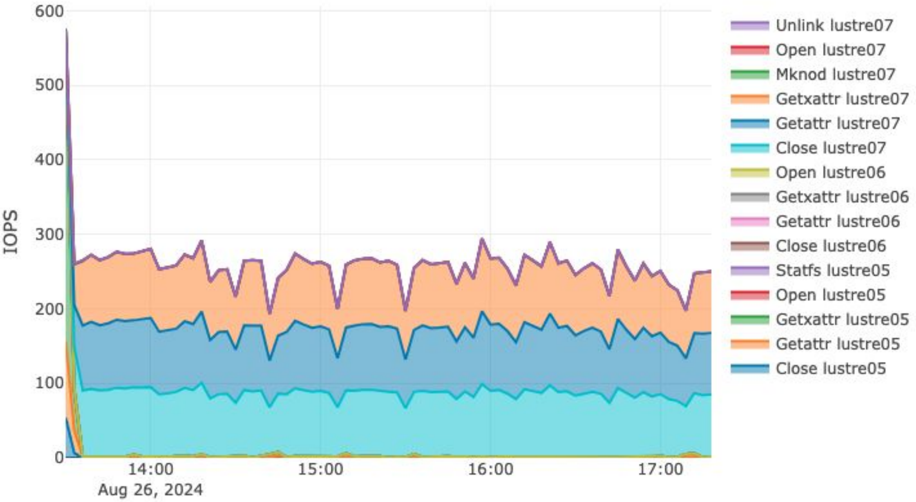


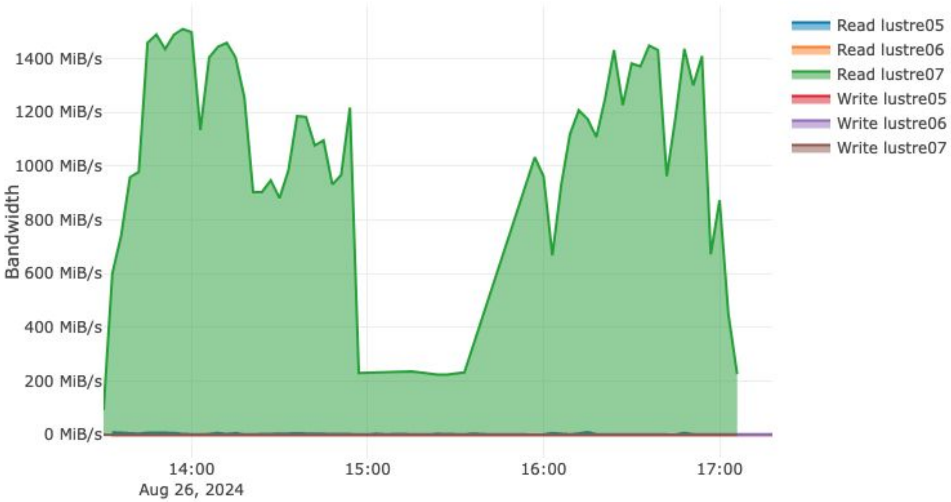
Lustre stats

- Aggregation per
  - Cluster (public view)
  - User
  - Group
  - Job

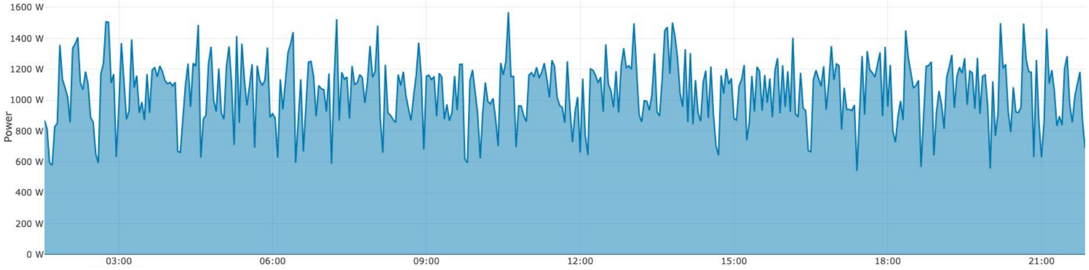




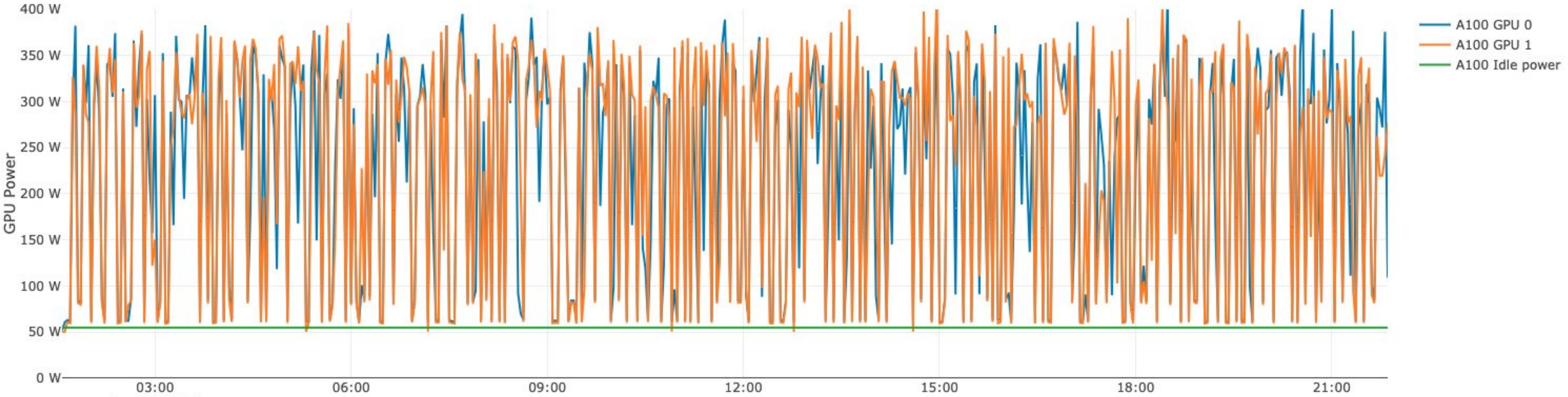




Power measurement



- Entire node with iDRAC
    - Include fans, networking cards and other components
    - Power spread among jobs on the same node
- By GPUs
    - Power assigned to the corresponding job

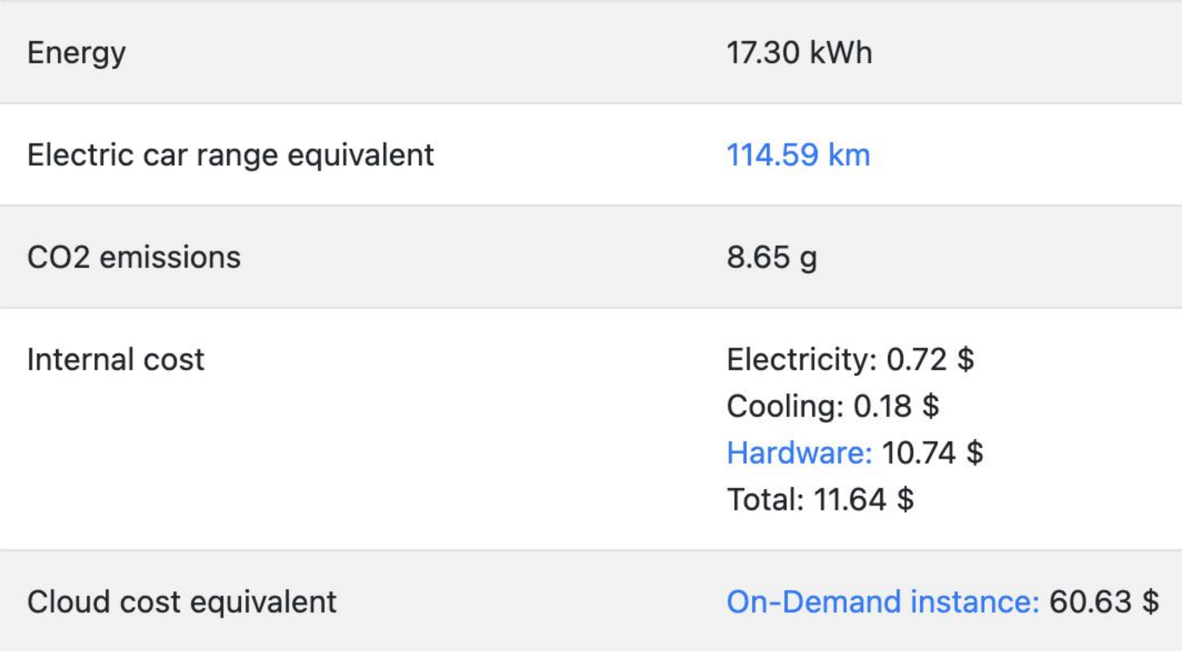


Energy and cost

- Configurable with local price and CO2 per kWh

- Power used (Hydro electricity)
    - CO2
    - Electricity cost
- Time used
    - Hardware Amortization
    - Cloud cost equivalent

- Total cost per job

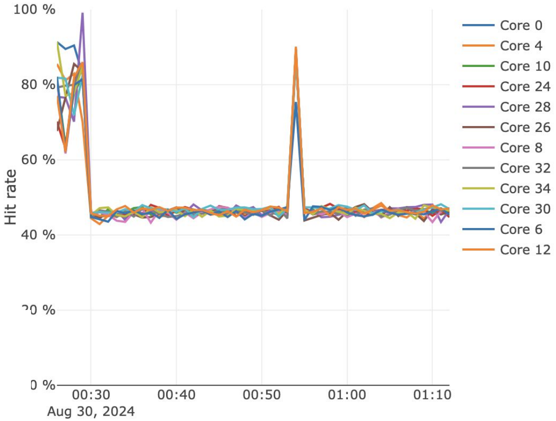


Energy	17.30 kWh
Electric car range equivalent	114.59 km
CO2 emissions	8.65 g
Internal cost	Electricity: 0.72 $ Cooling: 0.18 $ Hardware: 10.74 $ Total: 11.64 $
Cloud cost equivalent	On-Demand instance: 60.63 $

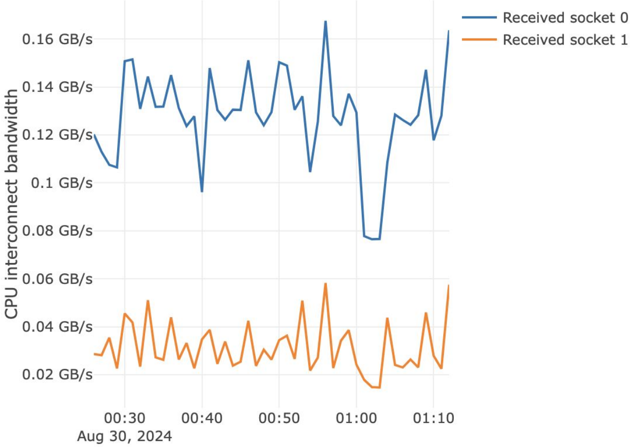
Intel PCM

- Low level CPU metrics
- Bandwidth bottleneck

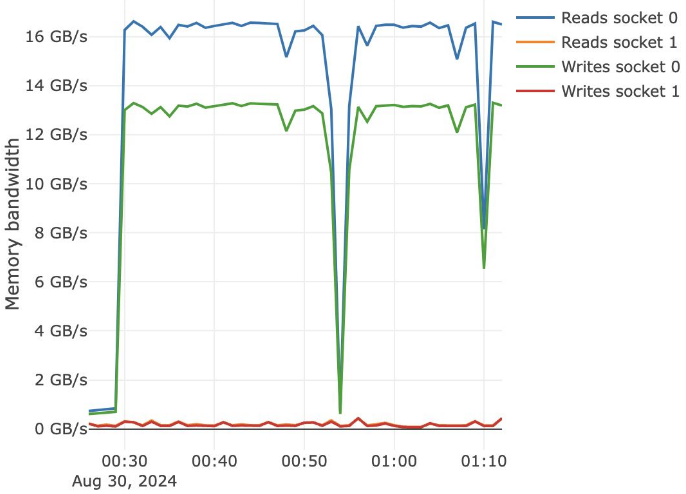
L3 cache hit rate



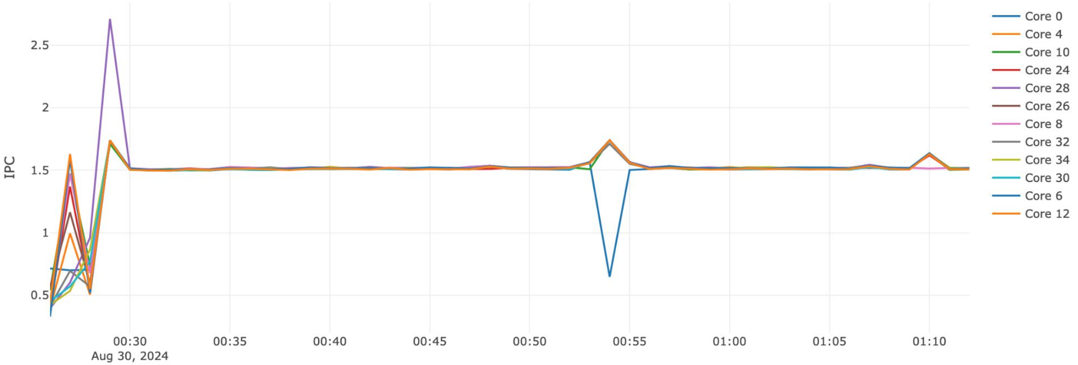
CPU interconnect bandwidth



Memory bandwidth



Instructions per cycle



Conclusion and future developpement

- Automatic email when resources are wasted
  - Analysts are periodically checking the “top” pages and help/warn users as required
- MIG automatic recommandation
  - We have about 20% of the GPUs currently splitted in half
- Using stats to change priorities of users/groups