How Do ML Jobs Fail in Datacenters? Analysis of a Long-Term Dataset from an HPC Cluster

Xiaoyu Chu

Ph.D. Student @ VU Amsterdam, AtLarge Research

x.chu@vu.nl

https://atlarge-research.com/

Sacheendra Talluri

Laurens Versluis

Alexandru Iosup

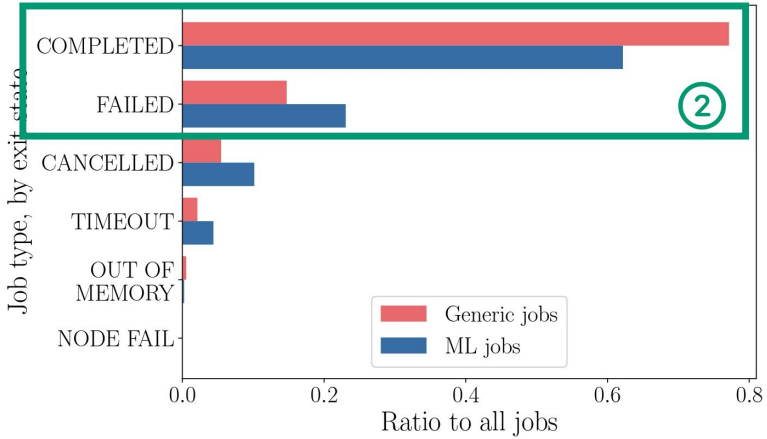
Why failures and ML job failures

Job Failures: Failed jobs in HPC datacenters waste users' time, compute resources.

ML jobs: ML workloads are fast-emerging workloads in HPC datacenters.

ML job failures: ML jobs have a 10% lower completion rate compared to generic jobs.

Dataset	Description
Source	SURF Lisa
Timespan	12 months
Year	2022
#Nodes	348
#Users	2662
1 #Jobs	2,301,128
%ML Jobs	13.32%
%Generic Jobs	86.68%

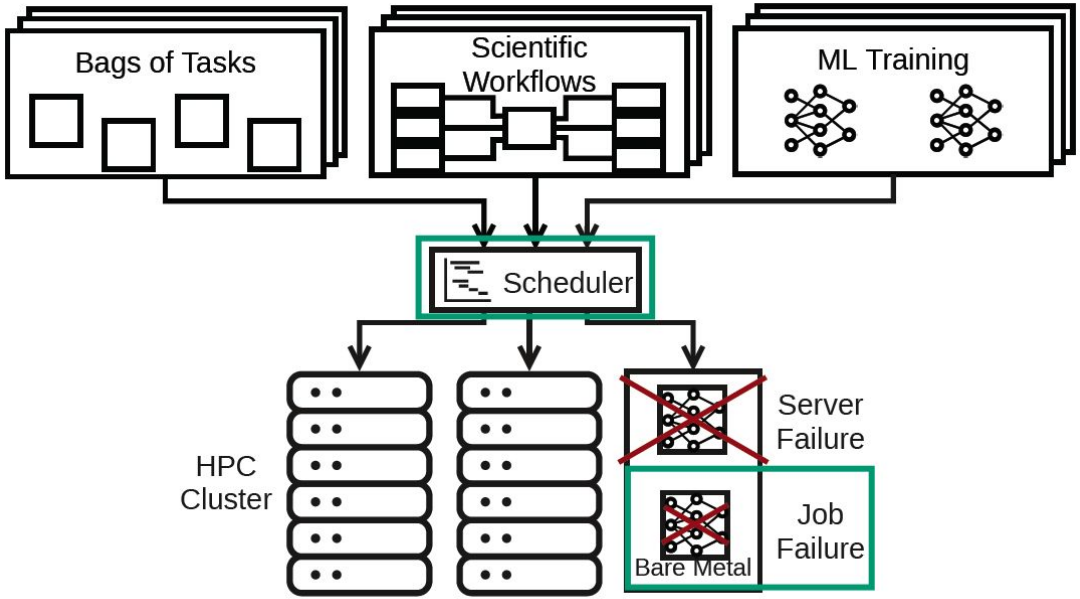


System Model

Scheduler: Different jobs are submitted and scheduled to different servers.

ML jobs: In this work, jobs on GPU nodes are seen as ML jobs.

Job failures: We define a job to have failed when it fails with an error, or is cancelled by the user, or runs out its reserved time.

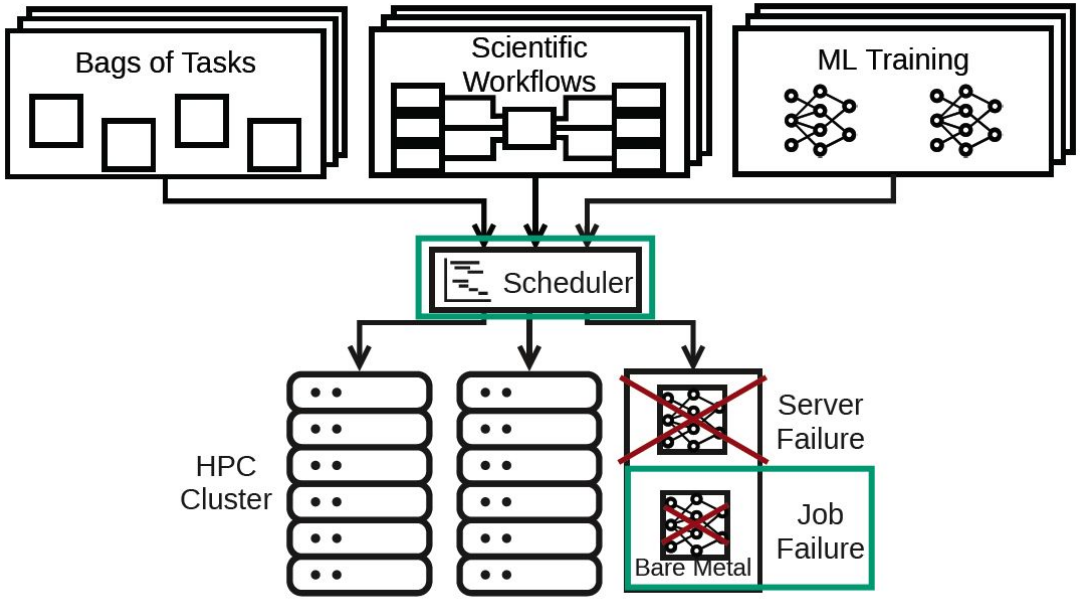


System Model

Scheduler: Different jobs are submitted and scheduled by the scheduler.

ML jobs: In this work, jobs on GPU nodes are seen as ML jobs.

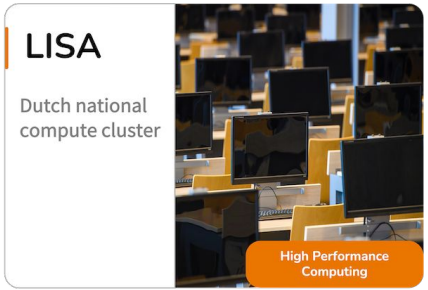
Job failures: We define a job to have failed when it fails with an error, or is cancelled by the user, or runs out its reserved time.



Research Question and Experiment Setup

Main Research Question: What are the characteristics of failed ML jobs in HPC datacenters?

Environment and Dataset:



HPC Datacenter

slurm
workload manager

Scheduler

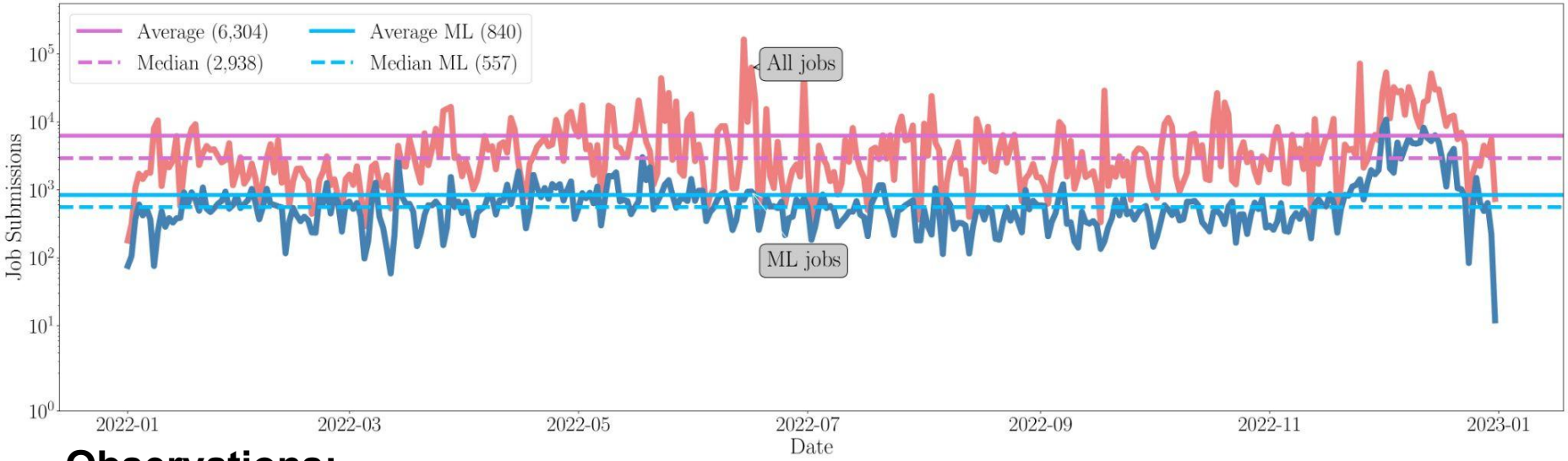
Dataset	Description
Source	SURF Lisa
Timespan	12 months
Year	2022
#Nodes	348
#Users	2662
#Jobs	2,301,128
%ML Jobs	13.32%
%Generic Jobs	86.68%

Dataset

Job Type
Generic completed jobs
Generic failed jobs
ML completed jobs
ML failed jobs

Job Types

Jobs submitted by date



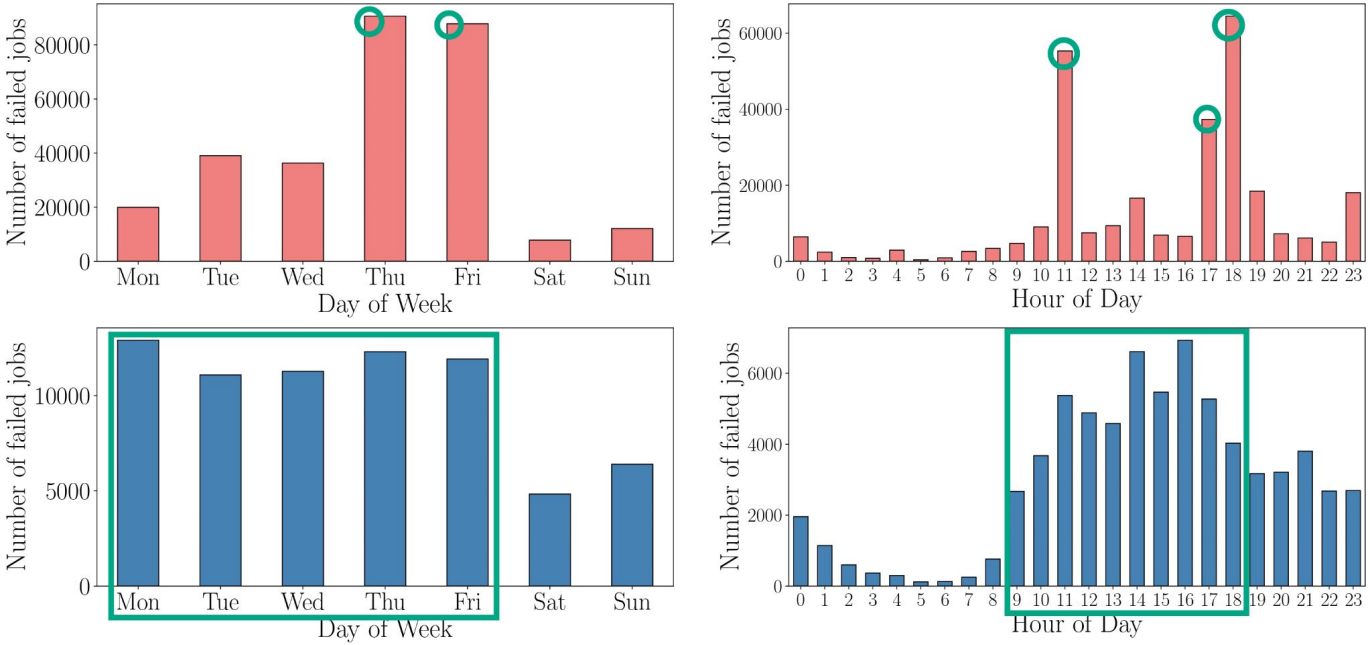
Observations:

O1. The number of submitted jobs per day is **highly variable**.

O2. The median amount of **ML jobs** increased from **329** to **557**, compared to data in 2020 [1].

[1] Laurens Versluis, Mehmet Cetin, Caspar Greeven, Kristian Laursen, Damian Podareanu,Valeriu Codreanu, Alexandru Uta, and Alexandru Iosup. 2021. A Holistic Analysis of Datacenter Operations: Resource Usage, Energy, and Workload Characterization–Extended Technical Report. arXiv preprint arXiv:2107.11832 (2021).

Job arrival patterns



Observations:

O3. ML job failures have a **workday pattern**.

O4. Generic jobs shows **anomaly peaks**.

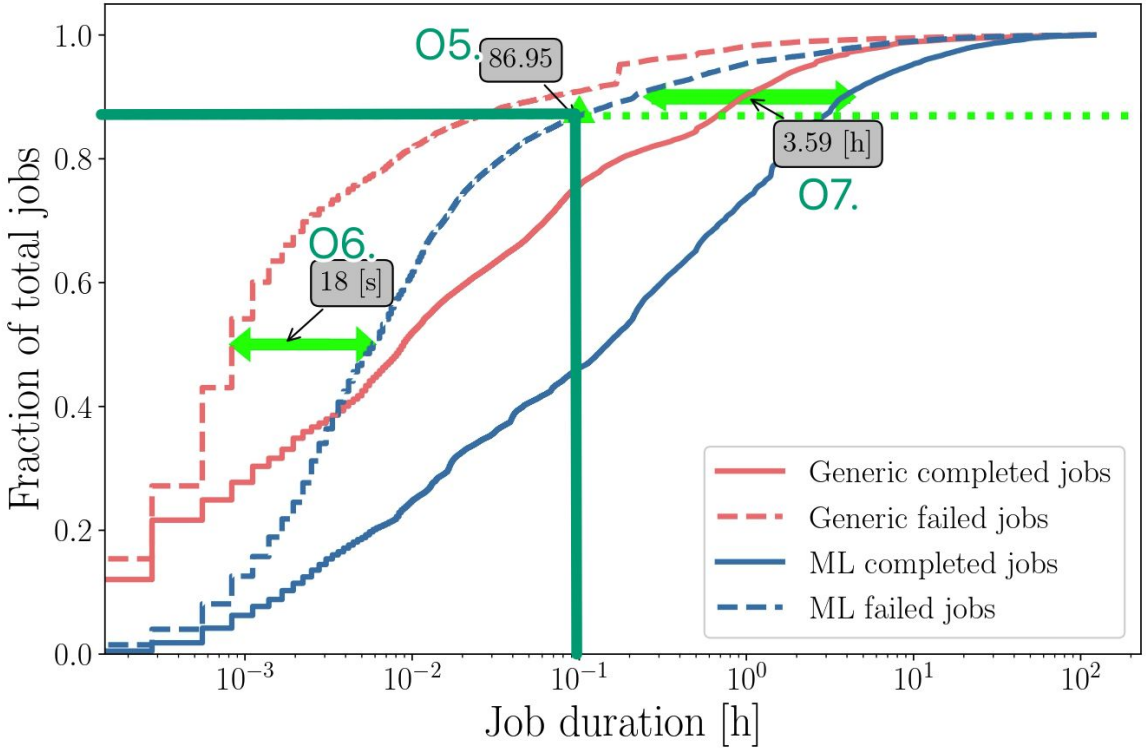
Duration of Jobs, CDF Plot

Observations:

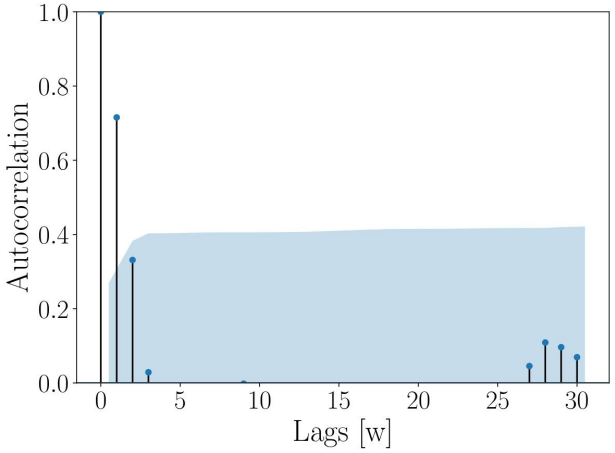
O5. ML jobs failed quickly: 86.95% of ML jobs failed in less than 6 minutes.

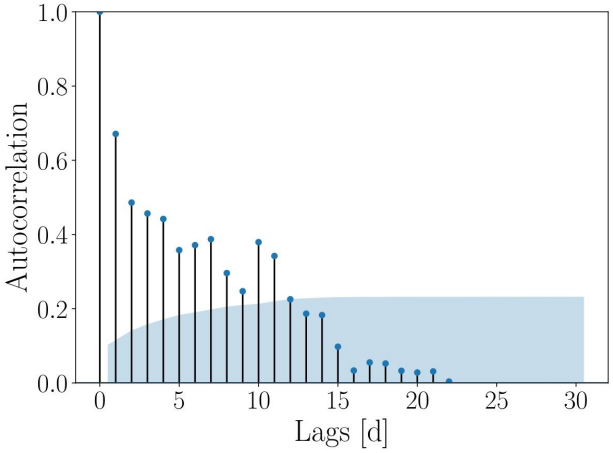
O6. The median runtime of ML failures is 18 seconds longer than generic failures.

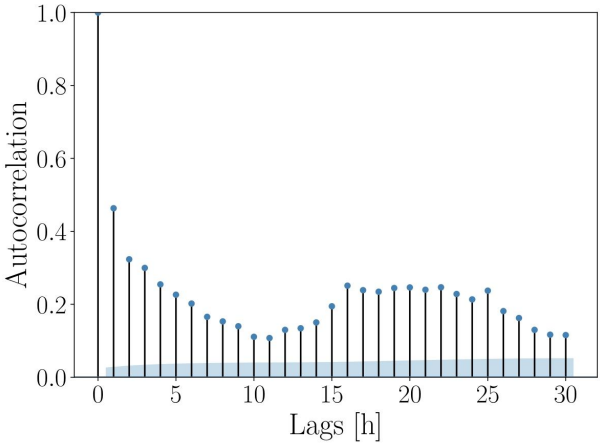
O7. 90% Successful ML jobs are 3.59 hours longer than failed jobs.



Autocorrelations with data aggregated by week, day, and hour







Observations:

O9. High autocorrelation at the **week** granularity for **small time lags**.

O10. The autocorrelation at the **day** granularity **declines steadily.**

O11. There is **little** autocorrelation at the **hour** granularity.

Conclusion

Take home messages:

- **ML jobs fail more than generic jobs:** ML jobs have a 10% lower completion rate compared to generic jobs in our dataset.
- **ML jobs failed quickly**: 86.95% of ML jobs failed less than 6 minutes.
- **The fail rate of ML jobs are autocorrelated at weeks and days.**

Future work:

- We will public our dataset and code after the workshop.
- We will continue to explore the connections between **job failures** and resource, energy, thermal data of **nodes.**