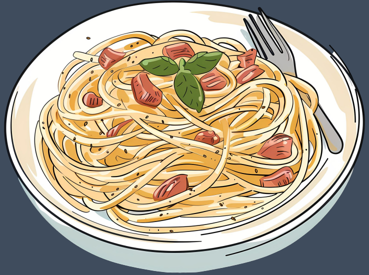
Carbonara



ACCRE
Advanced Computing Center
for Research & Education

Lou Jorgensen
Cavin Jacobson
Andrew Richard

Data provided by Vanderbilt University

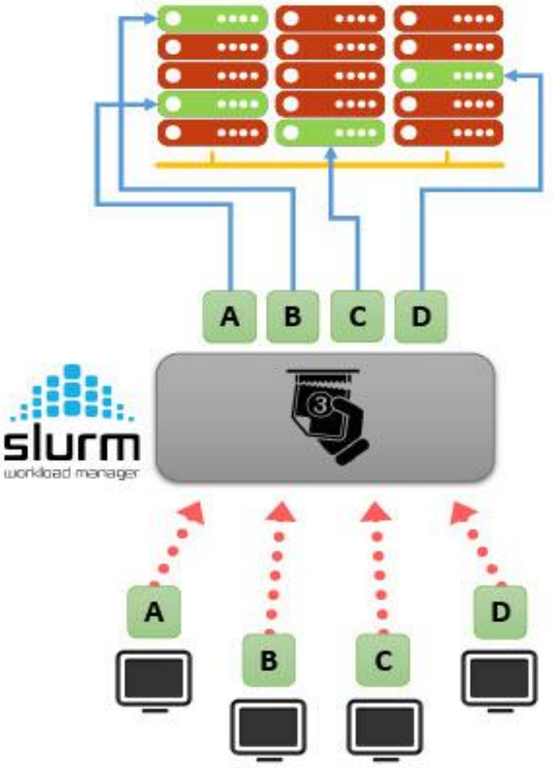
SLURM JOB SCHEDULER

Does the frequency of completed jobs at any given time lead to a crash in the Slurm Job Scheduler?

Are there other factors that contribute to the crashes?

What is Slurm?

ACCRE
Advanced Computing Center for Research & Education



Slurm is an open-source cluster management and job scheduling system for linux clusters

- ❖ Slurm keeps track of available resources on the cluster
- ❖ Collects users' resource requests for jobs
- ❖ Assigns priorities to jobs
- ❖ Runs jobs on assigned compute nodes

Our Data Encompasses Completed Jobs in a Year

373

Days: 10-01-2020 00:10:15 to 10-07-2021 20:41:11

137

Days without any Slurm Failures:

- At least 15 seconds for failure to occur

3,296

Slurm crashes

- Highest amount of 145 on 12-04-2020

19,825.5

Average jobs **completed** in a day

- Highest amount of 109,952 on 08-31-2021

3 h 48 m:

Average time to complete a job

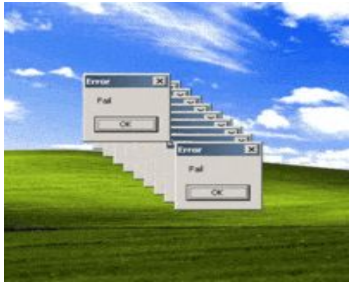
5317.5

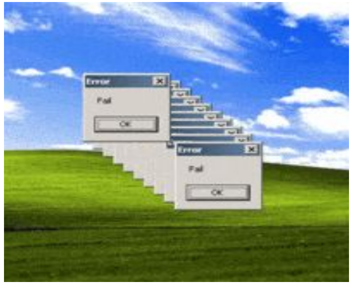
Average memory used in Megabytes per node

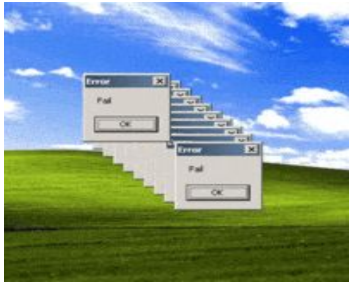
- Highest memory usage in a day at 25,480.79 on 07-17-2021

The data frame also included CPUS, Nodes, and Partitions: Nodes are grouped into the partitions and along with CPUS, determine how much memory is allocated to a job

What caused those 3,296 slurm crashes?



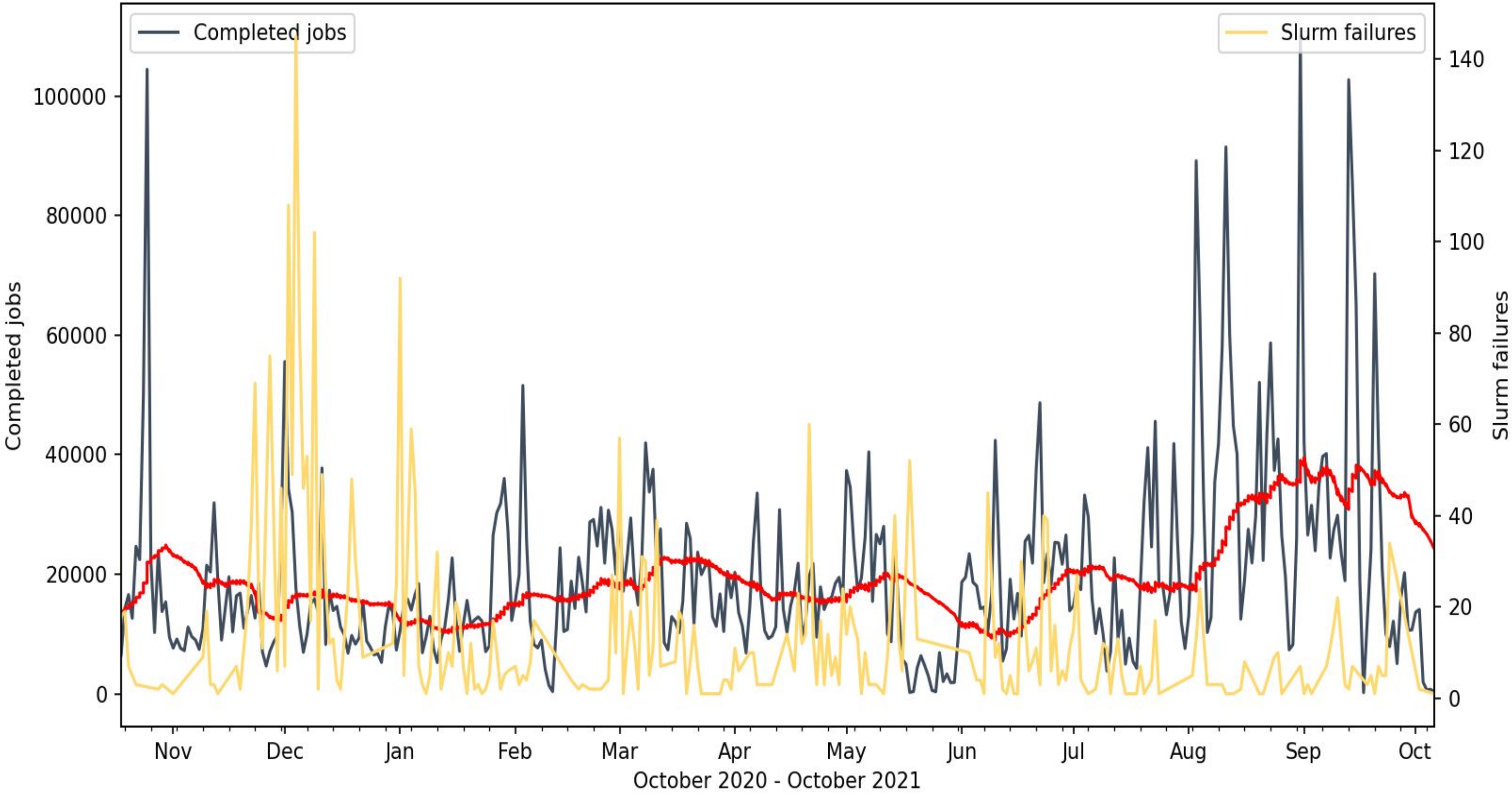




Null hypothesis: **no relationship** between the frequency of completed jobs and a failure of the Slurm system

Alternative Null hypothesis: **no relationship** between used memory, nodes or CPUs and a failure of the Slurm system

Frequency of Slurm Failures and Completed Jobs



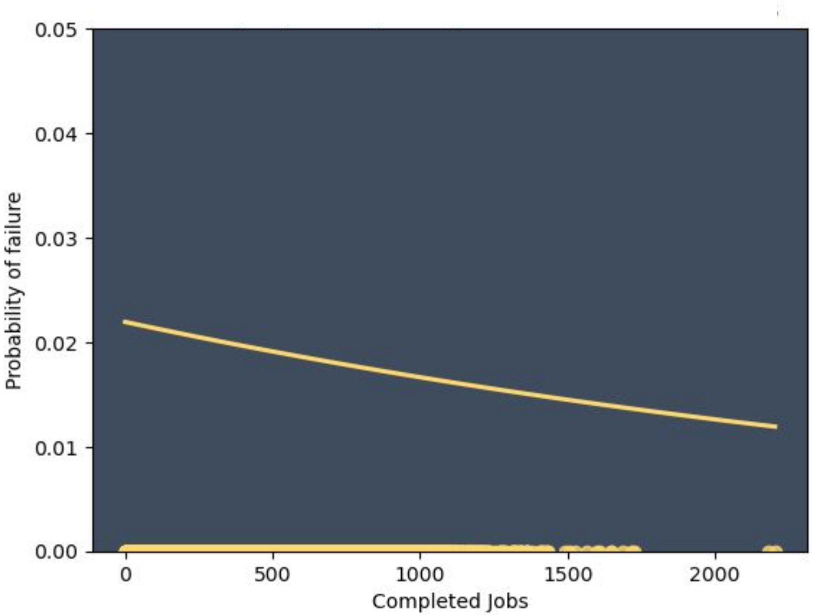
Monthly total of slurm crashes and completed jobs.

The red line indicates the **rolling average** of **completed jobs** over a period of 30 days.

A high frequency of job completions **does not appear to relate** to slurm crashes.

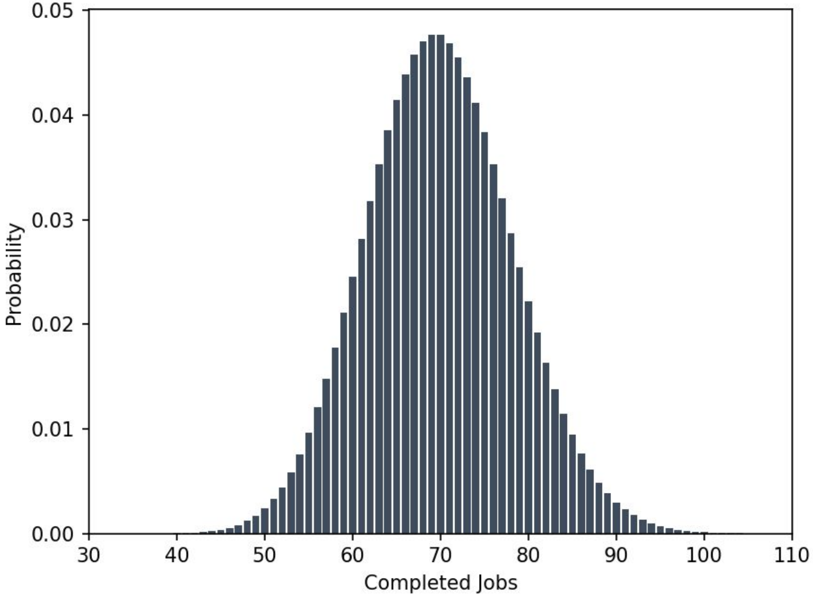
Data provided by Vanderbilt University

Probability of Jobs Impacting Fails in 5 Minute Intervals



There is a **2% chance** that a failure happens due to a completed job, and that goes towards a **1% chance** as the number of jobs increases

Poisson Distribution of Completed Jobs Every 5 Minutes

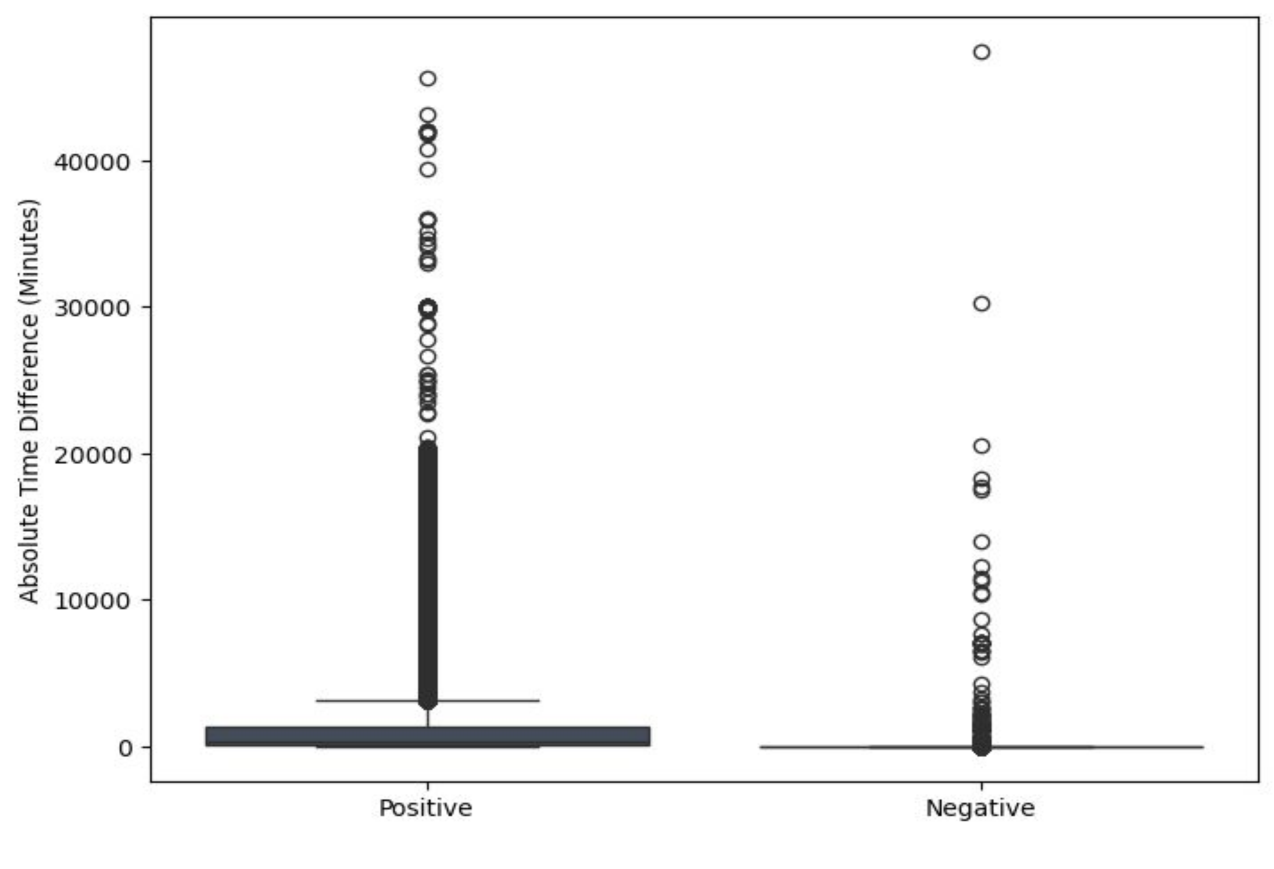


Within a 5 minute period of time, there will most likely be anywhere between **50** and **90 jobs** being completed simultaneously

Data provided by Vanderbilt University

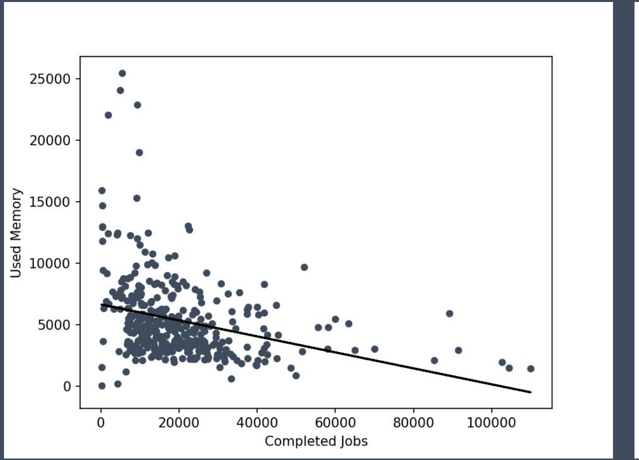
Time Differences Between Requested and Used Time

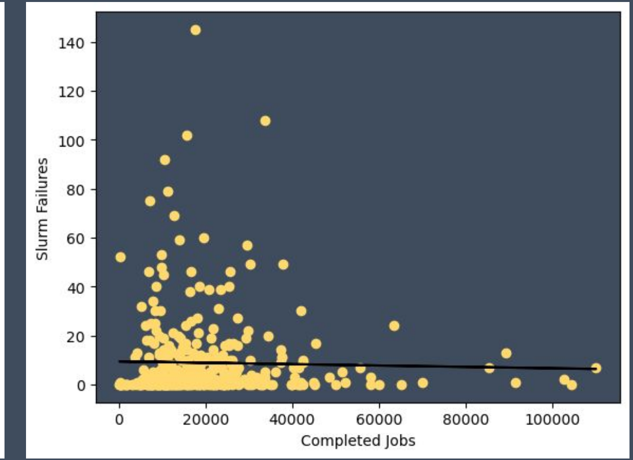
A positive time difference indicates using **less time** than originally requested, this scenario occurred **more frequently** than those which exceeded the requested time.

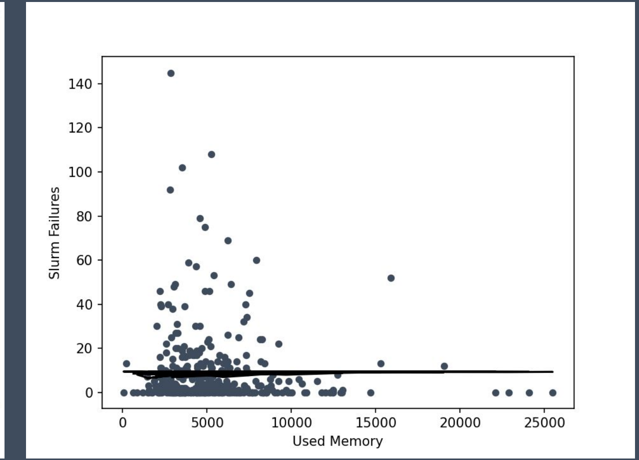


Data provided by Vanderbilt University

Pearson Correlations (daily averages)







Completed jobs and used memory have the highest (r) correlation at -0.15

Still a **weak** correlation

The correlation between completed jobs and Slurm failures is even **weaker** at -0.008

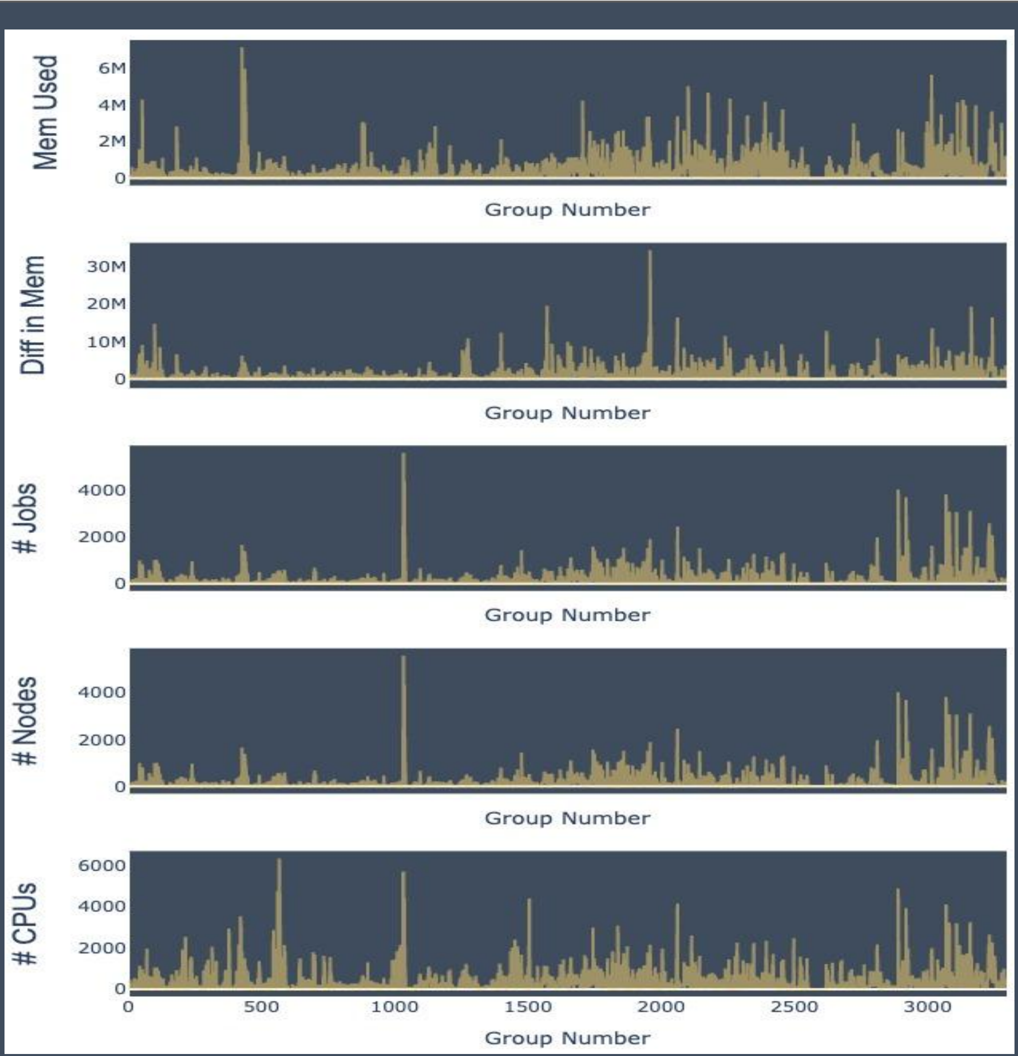
Used memory and Slurm failures correlate the **weakest** at -0.006

Data provided by Vanderbilt University

Exploratory Data Analysis

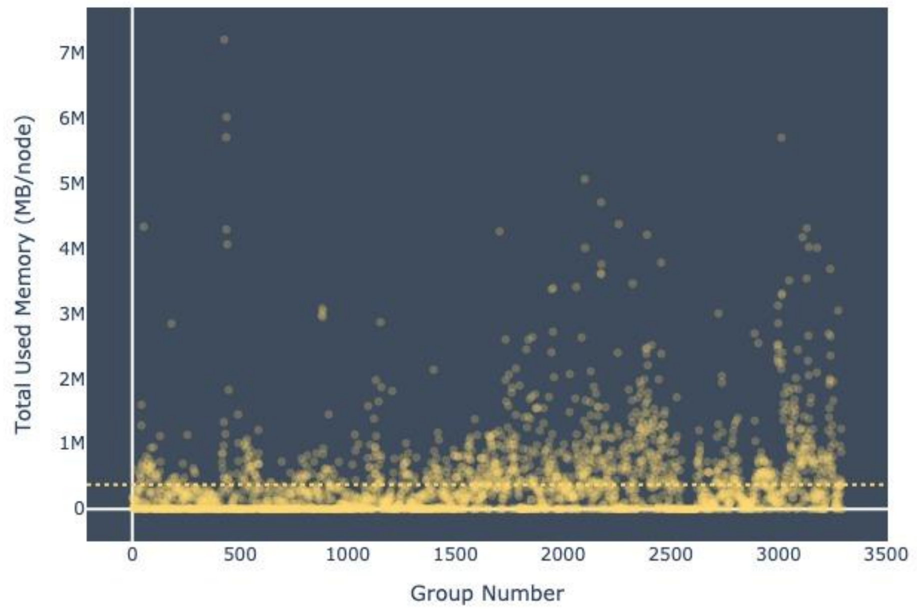
- Created groups containing the 20 minutes before each slurm fail (3926 groups)

- Looking for some consistency to indicate failure



Data provided by Vanderbilt University

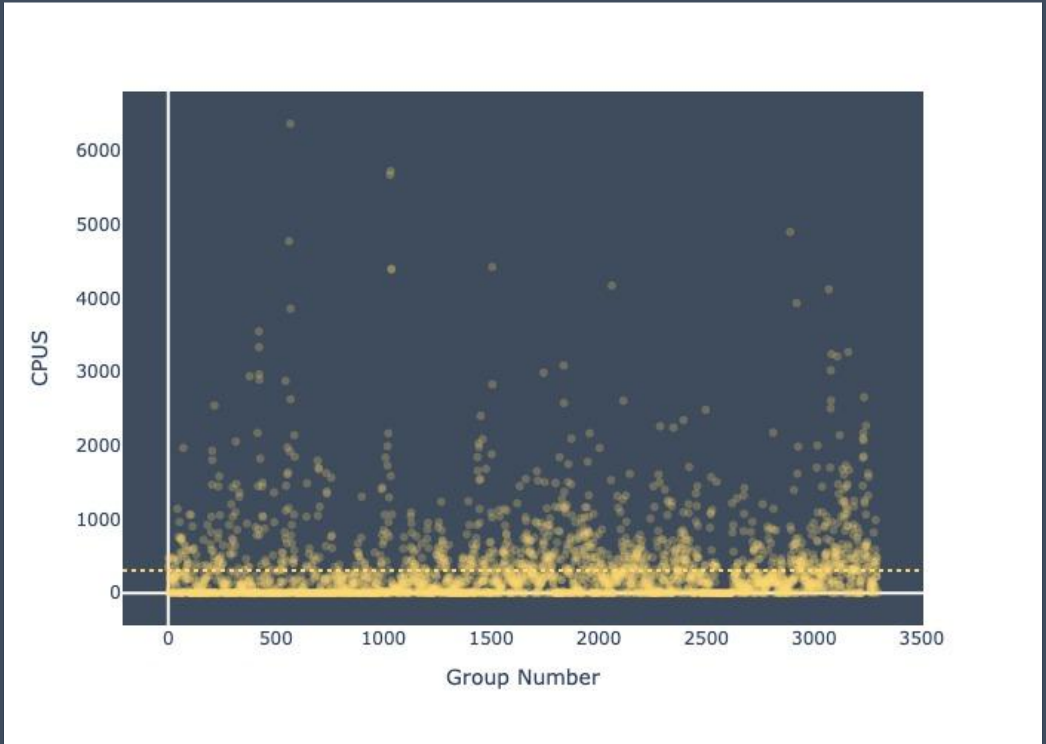
Exploratory Data Analysis



- ❖ No major spikes, but the back half used more memory than the front half
- ❖ Inconclusive, let’s look at CPUs

Data provided by Vanderbilt University

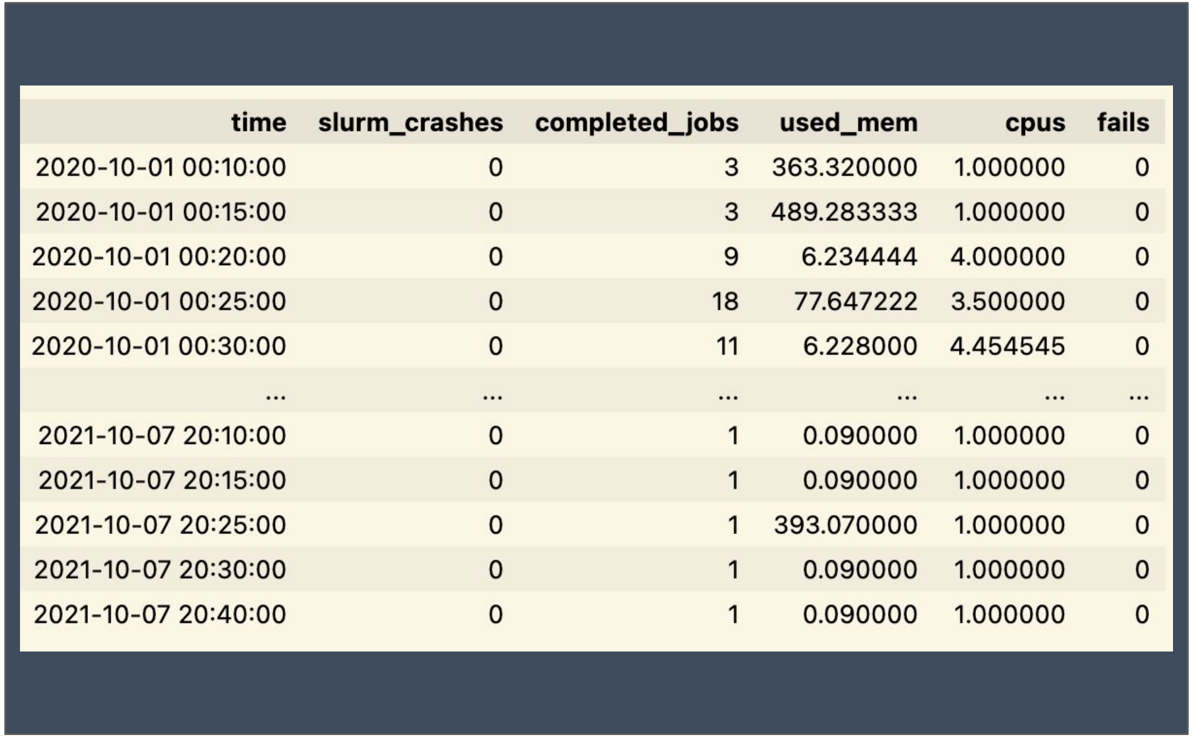
Exploratory Data Analysis



- ❖ More even distribution across all groups
- ❖ May be a candidate

Data provided by Vanderbilt University

Logistic Exploration



time	slurm_crashes	completed_jobs	used_mem	cpus	fails
2020-10-01 00:10:00	0	3	363.320000	1.000000	0
2020-10-01 00:15:00	0	3	489.283333	1.000000	0
2020-10-01 00:20:00	0	9	6.234444	4.000000	0
2020-10-01 00:25:00	0	18	77.647222	3.500000	0
2020-10-01 00:30:00	0	11	6.228000	4.454545	0
...	...	...	...	...	...
2021-10-07 20:10:00	0	1	0.090000	1.000000	0
2021-10-07 20:15:00	0	1	0.090000	1.000000	0
2021-10-07 20:25:00	0	1	393.070000	1.000000	0
2021-10-07 20:30:00	0	1	0.090000	1.000000	0
2021-10-07 20:40:00	0	1	0.090000	1.000000	0

- Data grouped into 5 min increments
- Logistic regression fit to this dataset

Data provided by Vanderbilt University

Logistic Exploration

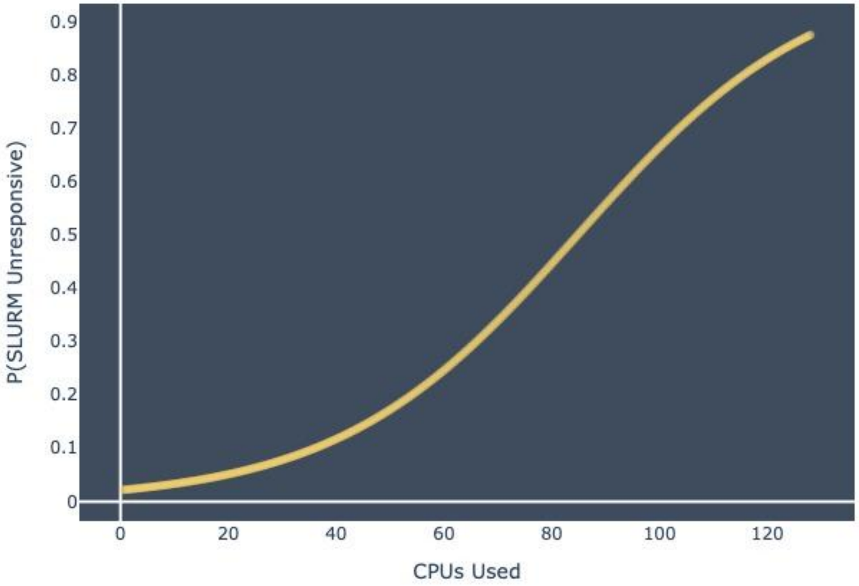
3 most likely candidates for logistic regression variables:

- ❖ Cpus used
- ❖ Used memory
- ❖ # of Completed jobs

$$\begin{aligned} \text{logit}(p) = {} & -3.8232 \\ & + 0.0451(\text{cpus}) \\ & - 2.123\text{e}^\wedge{-05}(\text{used\_mem}) \\ & - 0.0002(\text{completed\_jobs}) \end{aligned}$$

Data provided by Vanderbilt University

Logistic Exploration

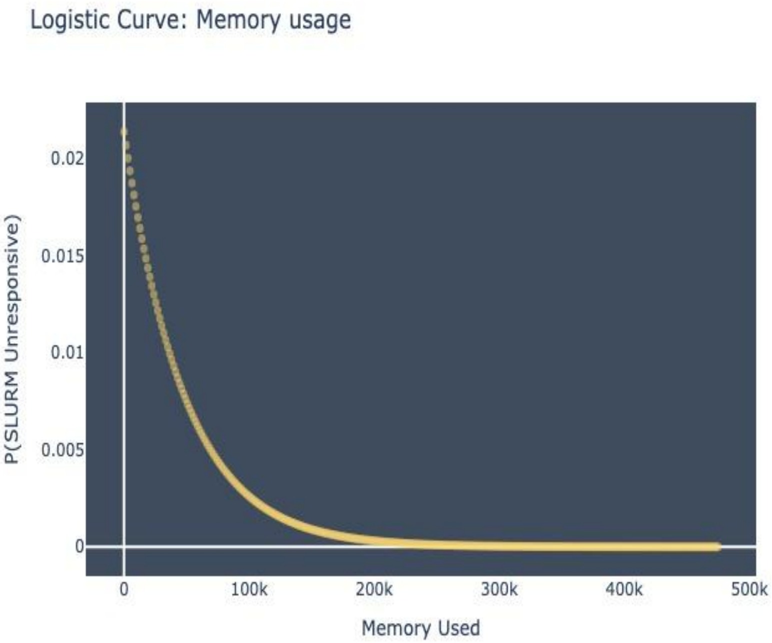


- ❖ Shape as expected
- ❖ As more cpus are used, the probability for slurm unresponsive increases

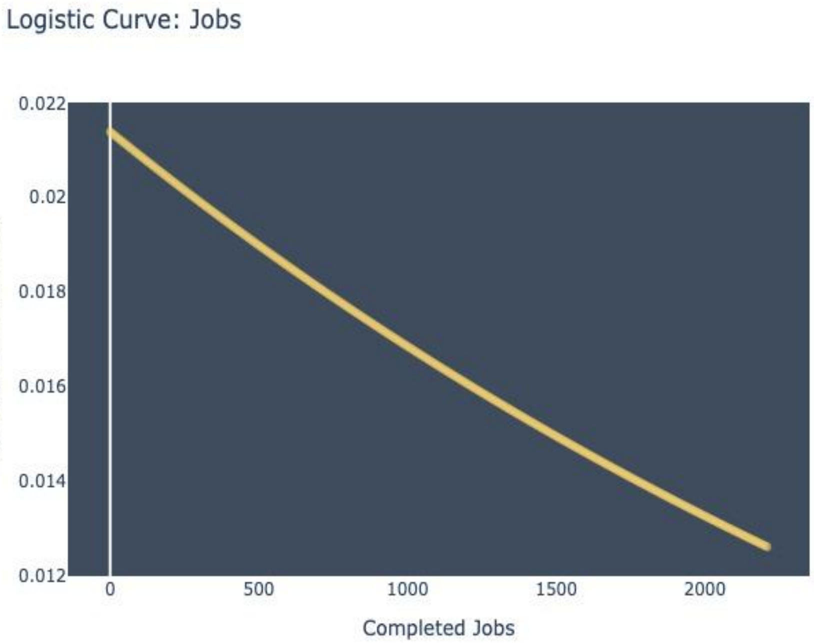
Data provided by Vanderbilt University

Logistic Exploration

- Consistent with negative coefficients
- Logistic regression likely not the best tool to evaluate this data, too much interplay between variables



Logistic Curve: Memory usage



Logistic Curve: Jobs

Data provided by Vanderbilt University

Conclusions

Based on our observations we found ***no significant relationship*** between frequency of jobs at completion and slurm failures

CPU and memory usage had the closest relationship but that did not show to have a significant bearing on Slurm failure

Thus, we accept both our Null hypothesis, as well as our alternative Null hypothesis

Possibilities for Future Research:

- Obtaining **Node failure data:**

  Taking an in-depth look at when nodes failed could give better insight into slurm crashes.

- More robust observations into the **partitions** of jobs

  Looking into partitions and determining where slurm crashes happen per partition.

- Encompassing **entirety of ce5 and ce6 data**:

  The slurm crashes for user 9204, which was a test user, may not give us the entire picture of when slurm crashes happen.

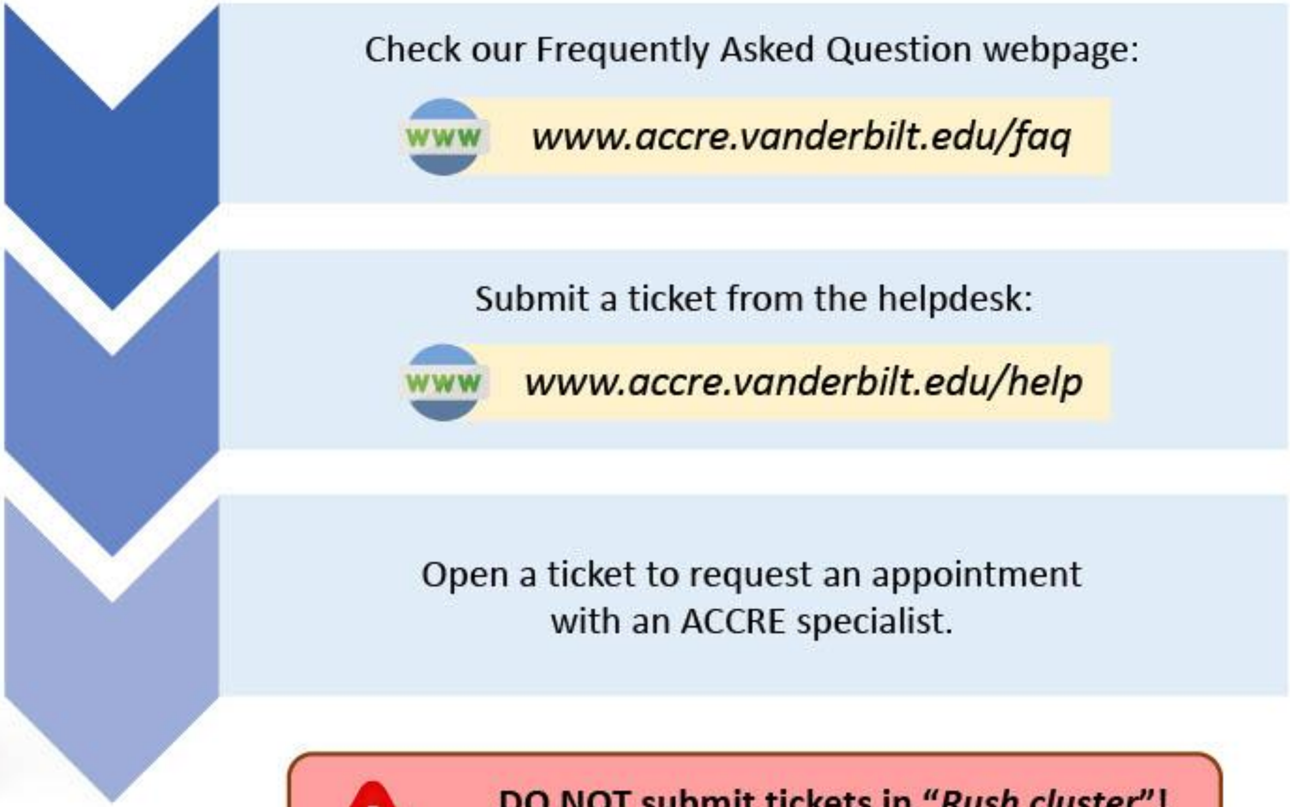
- **External factors**:

  Investigating more variables may help in determining the cause of slurm failures

  -E.g. **power outages** or **temperature** of the server rooms

NEED MORE HELP? Feel free to ask questions!

ACCRE
Advanced Computing Center for Research & Education



Check our Frequently Asked Question webpage:

www *www.accre.vanderbilt.edu/faq*

Submit a ticket from the helpdesk:

www *www.accre.vanderbilt.edu/help*

Open a ticket to request an appointment
with an ACCRE specialist.

! **DO NOT submit tickets in “*Rush cluster*”!**
Rush tickets are for cluster-wide issues only.

Thank you!