



Introduction

- The FRESCO project, sourcing data from Purdue, the University of Illinois at Urbana-Champaign, and the University of Texas at Austin, focuses on improving computer system dependability through collection and curation of detailed system usage data, workloads, and outages.
- The data includes aspects like job submissions, resource allocation, and durations, essential for understanding computing system failures and utilization.

Questions

- How do jobs utilize cluster resources in university's centrally managed clusters?
- How do users use or do not use the options to share resources on a node?
- How often do resource demands exceed supply, and does this impact job failure rates?
- Can users estimate the time their jobs will need on the cluster?

FRESCO Data Tables Host Data Table

Job Data Table

COLUMN	туре
 jid	character varving(32)
submit_time	timestamp with time zone
start_time	timestamp with time zone
end_time	timestamp with time zone
runtime	real
timelimit	real
node_hrs	real
nhosts	integer
ncores	integer
ngpus	integer
username	character varying(64)
account	character varying(64)
queue	character varying(64)
state	character varying(64)
jobname	text
exitcode	text

- This table contains accounting information for each job.
- The 'exitcode' column allows us to identify failed jobs.

Column		Туре
+		
time	timestamp	with time zone
host	character	varying(64)
jid	character	varying(32)
type	character	varying(32)
event	character	varying(64)
unit	character	varying(16)
value	real	
diff	real	
arc	real	

- This table provides data for each host in the cluster.
- Represented as a timeseries such that each row represents a single value for a given event type.

Data Analysis

Please sele	t a statistic to calculate.	_		
Statistics	None	A		
	Mean			
	Median			
	Standard Deviation			
_	PDF	•		
Please prov:	Please provide the threshold if 'Ratio of Data Outside Threshold' was selected.			
Value:	0			
Please sele	Please select an interval type to use in the statistic calculation. If count is selec			
interval will correspond to a count of rows. If time is selected, the interval will be a				
window.				
Interval Type	Time	~		
If time was	selected, please select the	unit of time.		
Interval Unit	Minutes	~		
Please prov:	ide the interval count.			

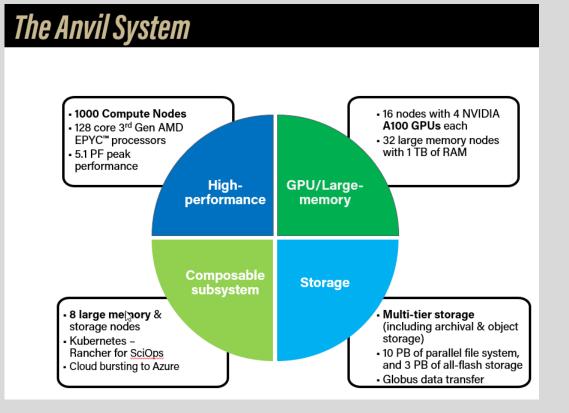
- After creating a dataset from the Host Data table, users can perform statistical calculations. The available metrics are:
- **cpuuser**: CPU user mode average %
- **gpu_usage**: GPU active time average %
- **block**: data transfer rate
- **memused**: total memory storage
- **memused_minus_diskcache**: physical memory usage excluding caches
- **nfs**: data transfer rate over NFS mounts

- 5.1 PF peak Query th Table Select start and Start Time: 1 End Time: 1 Times Va elect columns Choose sort co Order Bv: Direction: Set results limit Limit Results: Enter IN clause v IN Column: IN values: Add data filters: Column: Operator: Value: Add Condit Active filters Execute Qu

FRESCO Data Repository and Analytics Providing Public Large-Scale Computing System Usage and Failure Data

Saurabh Bagchi, Carol Song, Rajesh Kalyanam, Amiya Maji, Stephen Harrell, Aryamaan Dhomne, Joshua McKerracher

www.frescodata.xyz



Data Access

e Host Data	Query the Job Data Table Select start and end times (Max: 180 days).		
nd times (Max: 5 days).	Start Times	10/30/2022 07:02:26 PM	
/30/2022 07:02:26 PM 🗖			
/30/2022 08:02:26 PM			
73072022 00:02:20 PM	Times V	alid	
	Select column	s:	
	Columns:	queue	
A	columns.	state	
st		jobname exitcode	
e		host_list 🔻	
ent 👻		Select Distinct	
Select Distinct			
	Choose sort co	olumn and direction:	
nn and direction:	Order By:	None 🗸	
one 🗸			
	Direction:	ASC 🗸	
sc 🗸	Set results lim	it:	
	Limit Results:	0	
	Enter IN clause	a valuer:	
lues:	Enter IN Clause	e values.	
iues.	IN Column:	None 🗸	
one 🗸	IN values:	Enter values separated by	
nter values separated by		commas //	
ommas //			
	Add data filter	rs:	
	Column:	jid 🗸	
ost 🗸 🗸	Operator:	= 🗸	
~	Value:		
	Add Cond	dition Remove Condition	
n Remove Condition			
	Active filters:		
	Conditions:		
	Execute 0	Query	
ry			
	Current SQL q	uery:	
y:			

Parameters: [datetime.datetime(2022, 10, 3 30, 19, 2, 26) 0, 19, 2, 26), datetime.datetime(2022, 10,

• Users can interactively request data using a SQL query builder • Queried data can be exported as CSV or Excel spreadsheet

0, 19, 2, 26), datetime.datetime(2022, 12,

for further external analysis

Data Overview

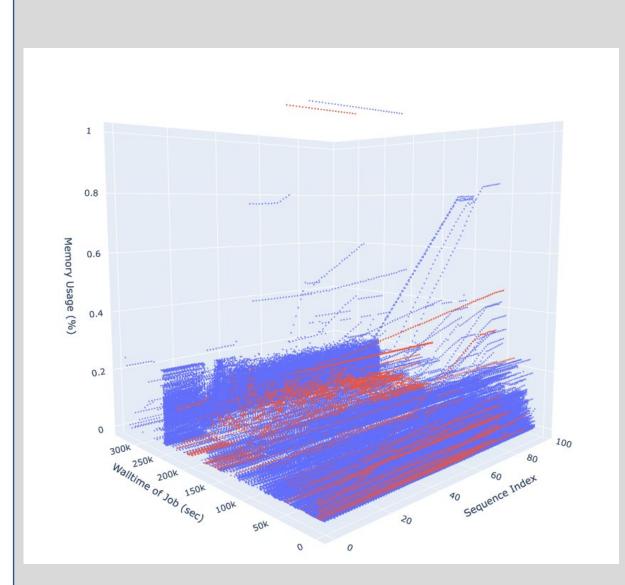
- Data comes from Anvil, an HPC cluster at Purdue – CPU/GPU/Large Memory jobs
- Data currently ranges from July 2022 to June 2023
- 1,469,223 total jobs; 302,096 failed or timed out jobs
- Tracks job lifecycle events including submission, start, and end times along with exit codes
- Provides detailed node-level resource usage metrics and host event data

Data Visualizations



• Plots can be generated interactively as the user queries different subsets of the dataset

Sample Findings

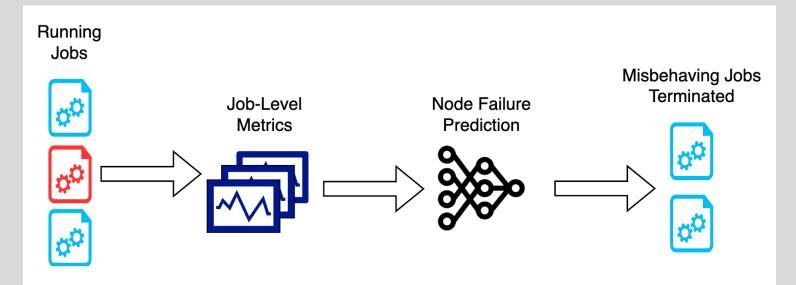


Develop a real-time monitoring solution that performs online inference with the node failure prediction model Explore using the real-time predictions of the models to terminate jobs that are likely to cause a node failure, preventing the loss of other jobs running on the same node Develop models to predict job failures and walltimes, providing quality of service improvements for cluster users



- This plot shows the mean memory usage across jobs on the Z axis, with the measurement index (relative to the measurements for each job) and the walltime of jobs on the X and Y axes respectively
- Blue points are successful jobs, red points are failed jobs
- Anomaly detection methods may be able to learn similar patterns across metrics to predict failed jobs

Future Plans



Acknowledgements

• FRESCO is supported by the National Science Foundation, CISE Community Research Infrastructure (CCRI) program, "Open Computer System Usage Repository and Analytics Engine". Project numbers CNS-2016704 and CNS-2016608.

• Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.