GEORGE MASON UNIVERSITY

Characterizing GPU Memory Errors: Insights from a Cross-supercomputer Study

Zhu Zhu (George Mason University)

Yu Sun (George Mason University)

Bo Fang (Pacific Northwest National Laboratory)

Steven Farrell (NERSC, Lawrence Berkeley National Laboratory)

Gregory H. Bauer (University of Illinois Urbana-Champaign)

Brett Bode (University of Illinois Urbana-Champaign)

Michael E. Papka (Argonne National Laboratory, University of Illinois Chicago)

Ian T. Foster (University of Chicago, Argonne National Laboratory)

William Gropp (University of Illinois Urbana-Champaign)

Zhao Zhang (Rutgers University)

Lishan Yang (George Mason University)

Errors Are Hurting LLM Training

OPT175B Model Training (Meta):

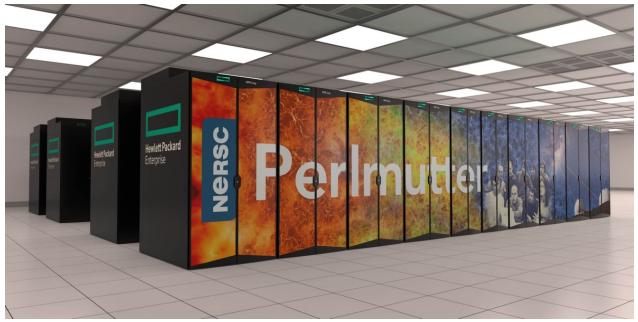
- ~922 A100 GPUs
- $2500 per hour
- Finished in 54 days
- Spent 18 days (33%) on errors

Error type	F	N
ECC errors	16	31
NCCL errors	12	33
CUDA errors	9	9
GPU lost errors	15	17
infoROM errors	9	19
Other GPU failures	7	10
IB errors	6	10
Software bugs	9	9
Other	16	17

F: Number of Failures
N: Number of Involved Nodes

https://github.com/facebookresearch/metaseq/blob/main/projects/OPT/chronicles/OPT175B_Logbook.pdf

Background



Delta

Nodes: 207

GPUs: 849

GPUs: NVIDIA A40 & A100

Polaris

Nodes: 560

GPUs: 2240

GPUs: NVIDIA A100

Perlmutter

Nodes: 1884

GPUs: 7534

GPUs: NVIDIA A100

Data Collection

Monitor GPU DRAM ECC

- SBEs: Single Bit Errors
- DBEs: Double Bit Errors
- ECC: Error Correction Codes, correct SBEs and detect DBEs
- Tool: NVIDIA dcgm (Data Center GPU Manager)

ClusterName	Error Type	Log Collection Dates	Log Length	Frequency
Delta	SBEs, DBEs	12/16/2022 – 01/07/2024	388 days	Every minute
Polaris	DBEs	10/01/2023 – 12/14/2023	75 days	Every 4 seconds
Perlmutter	SBEs, DBEs	11/01/2023 – 12/20/2023	50 days	Every 30 minutes

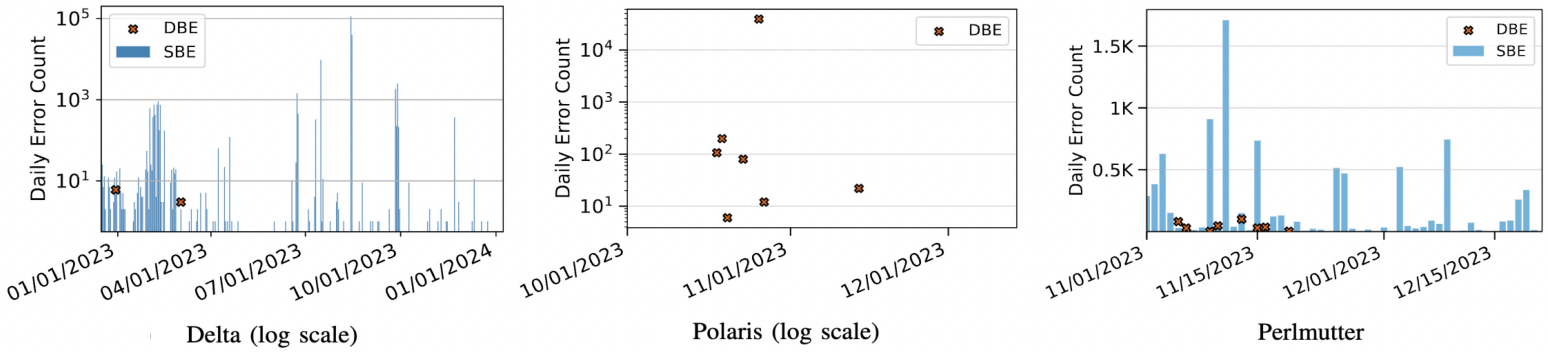
Error Overview: Error Rate

- *SBEs/ DBEs rate* : SBEs/ DBEs per GPU per day

Cluster Name	GPUs	SBEs Rate	DBEs Rate
Delta	849	0.528	0.000027
Polaris	2240	N/A	0.2355
Perlmutter	7534	0.0238	0.00087

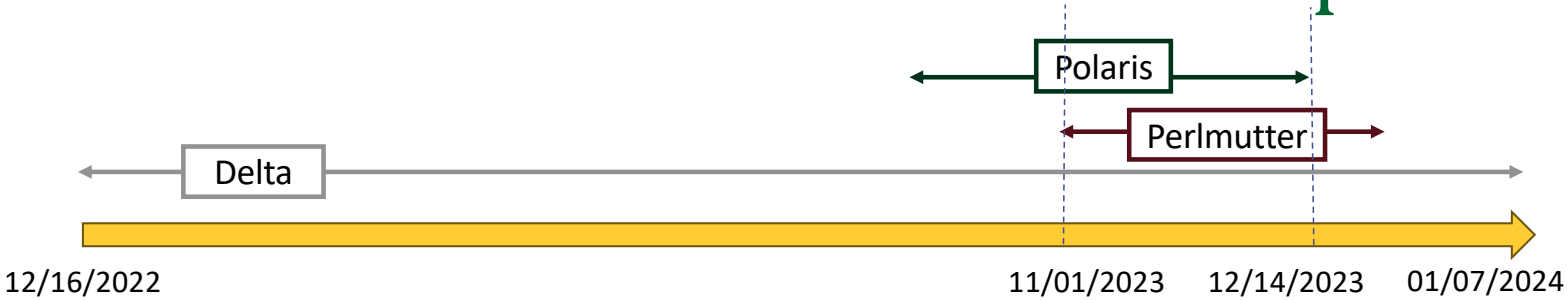
The observed error rates vary on different clusters.

Error Overview: Bursty Pattern



Bursty error patterns exist in GPU clusters.

Error Overview: Cross-cluster Comparison



All Time Rage

Cluster Name	Delta	Polaris	Perlmutter
# SBEs	173936	N/A	8964
# SBE Events	3324	N/A	327
SBE Rate (Per GPU Per Day)	**0.528**	N/A	**0.0238**
# DBEs	9	39568	322
# DBE Events	2	15	8
DBE Rate (Per GPU Per Day)	**0.000027**	**0.2355**	**0.00087**

Overlapping

Cluster Name	Delta	Polaris	Perlmutter
# SBEs	382	N/A	8183
# SBE Events	209	N/A	288
SBE Rate (Per GPU Per Day)	**0.010**	N/A	**0.025**
# DBEs	0	22	322
# DBE Events	0	2	8
DBE Rate (Per GPU Per Day)	**0**	**0.00022**	**0.00097**

Error characteristics are strongly biased by bursty error patterns.

Interarrival Time of Errors

- Interarrival Time: time between errors
- MTBE: Mean Time Between Errors

Cluster Name	Delta	Polaris	Perlmutter
MTBE (SBE, Hour)	2.70 ±23.82	N/A	3.59 ±4.92
MTBE (DBE, Hour)	N/A	47.05 ± 114.81	47.71±19.17

Similar MTBEs observed in three clusters.

MTBE: Mean Time Between Errors

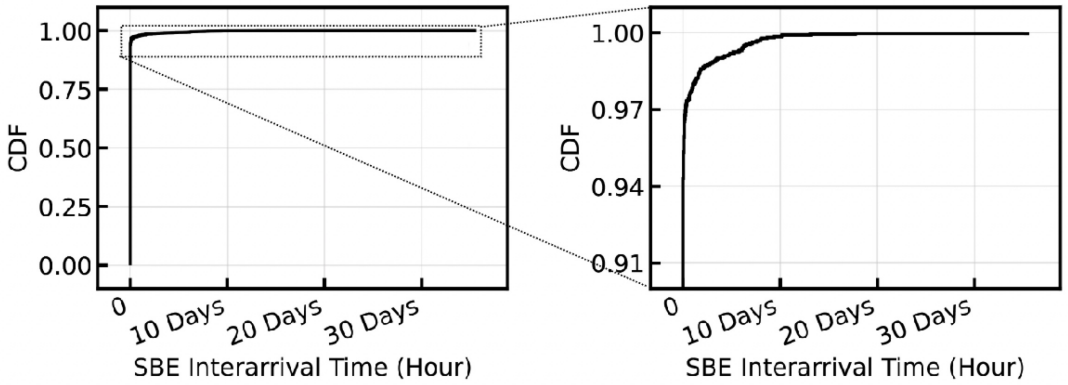
MTBE: Comparison with K20X GPUs

	K20X (Blue Waters)	K20X (Titan)	A100 (Polaris)	A100 (Perlmutter)
#GPUs	3072	18688	2240	7534
MTBE (DBE, Hour)	**768**	**160**	**47.05**	**47.71**

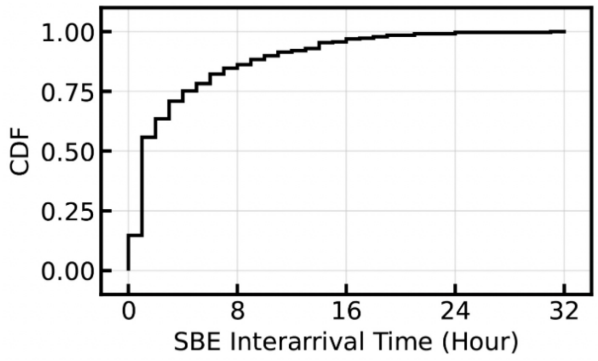
A100 GPUs are more vulnerable than K20X.

GEORGE MASON UNIVERSITY

Interarrival Time of Errors



Delta

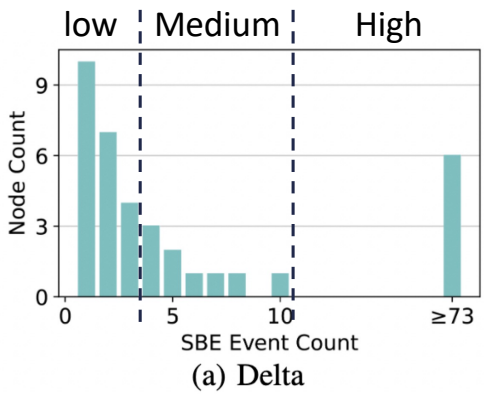


Perlmutter

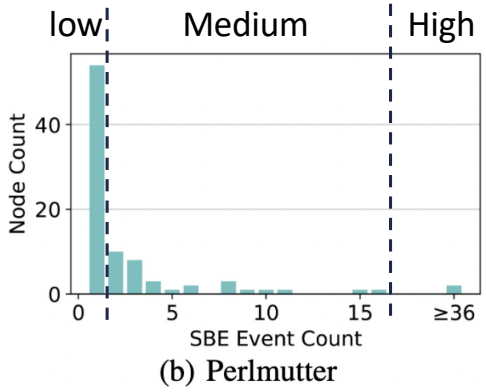
Bursty error patterns and the supercomputer scale can affect the characteristics of error interarrival times.

Reliability Behavior of Nodes

- Error Occurrence Event: An increased error count are observed
- Error Count: Number of errors in each error occurrence event

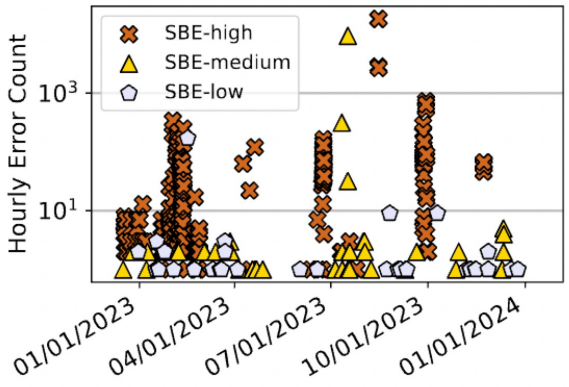


(a) Delta

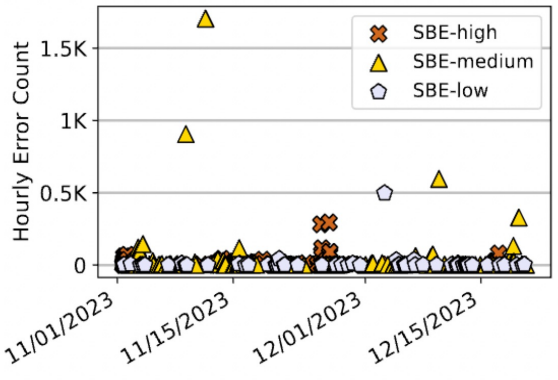


(b) Perlmutter

Reliability Behavior of Nodes



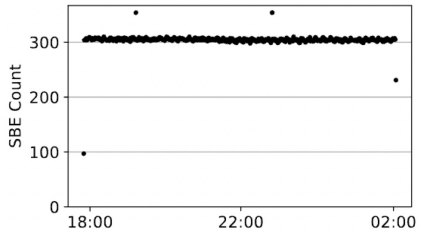
Delta (log scale)



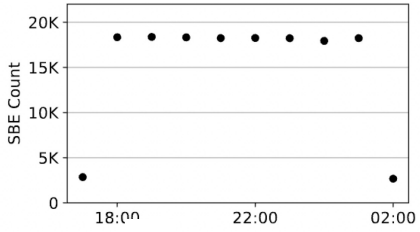
Perlmutter

The frequency of SBE events does not directly correlate with the number of errors per event.

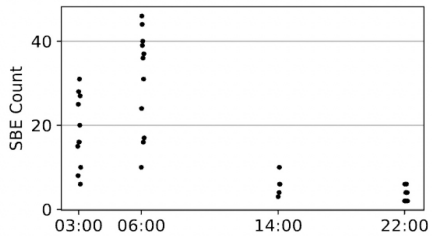
Discussion: Error Monitoring Frequency



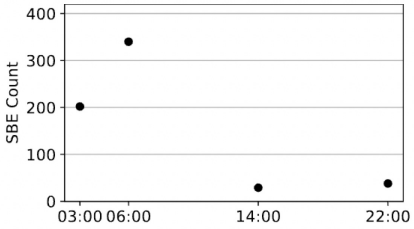
(a) Per Minute



Per Hour



Per Minute



Per Hour

Coarser-level error monitoring does not suffer much information loss, yet monitoring error at a finer level enables faster responses to errors.

Conclusions:

- Bursty error patterns have a significant impact on the characteristics of error rate and MTBE
- Cluster scale affects MTBE but the relationship is not linear
- A100 GPUs are more vulnerable than previous generation (K20X)