Understanding GPU Memory Corruption at Extreme Scale: The Summit Case Study

*Vladyslav Oles, **Anna Schmedding**, George Ostrouchov, Woong Shin, Evgenia Smirni, and Christian Engelmann*

WILLIAM & MARY
CHARTERED 1693

OAK RIDGE
National Laboratory

GPU Memory Corruption

Bit flips in GPU memory

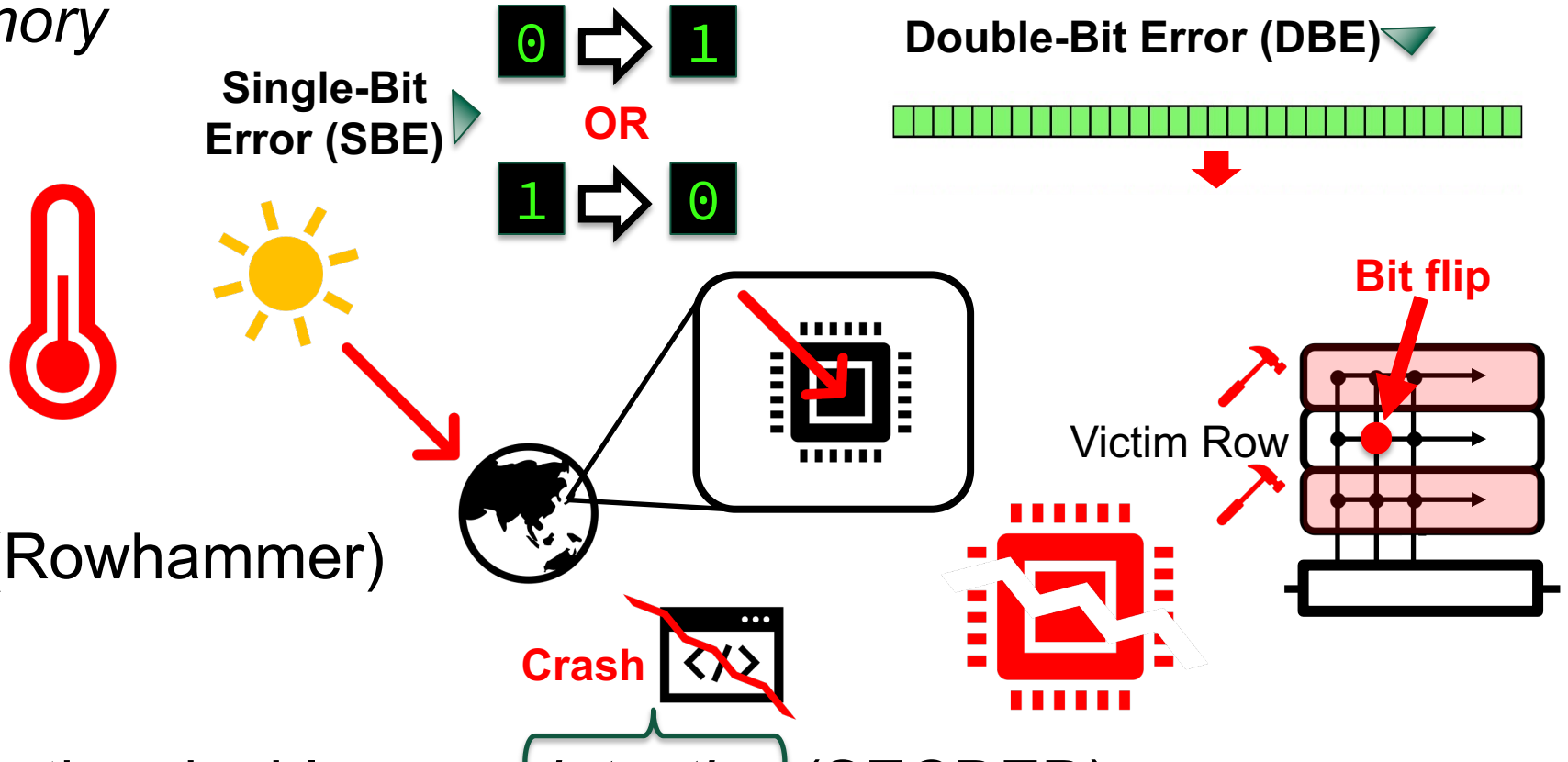
- Single-Bit
- Double-Bit
- …

Why? Common?

- Temperature
- Cosmic radiation
- Memory access (Rowhammer)
- …

Error correction?

- Single error correction double error *detection* (SECDED)

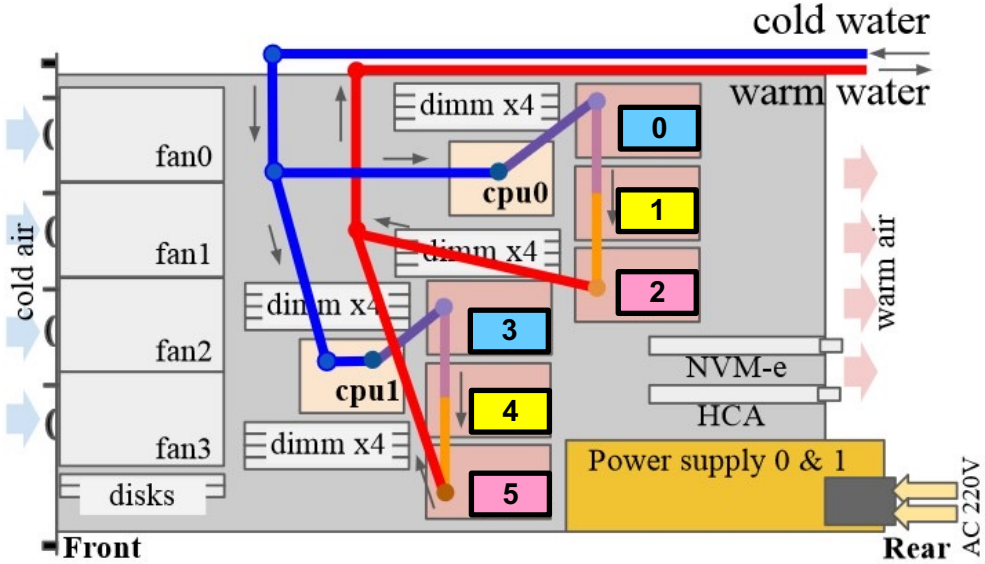


What is the relationship between GPU activity and bit flips in GPU memory *at scale*?

Summit Supercomputer

- Oak Ridge Leadership Computing Facility
- 122.3 Petaflops
- Layout
  - Incomplete 8 x 37 grid of *cabinets*
  - 18 *nodes* per *cabinet*
  - 2 CPUs, 6 Nvidia V100 GPUs per *node*
  - GPUs have 16GB modules of ***stacked HBM2 memory***

Node Layout



Data Sets

2.5 Years of Real-World Data
Jan 2020 - May 2022

Nvidia GPU XID error log — 600 MB
- GPU hardware and software errors

Summit node reboot log — 7 MB
- PCIe bus and serial #s

Location augmented GPU error logs

System snapshots

3M Rows

938K

88M

268B

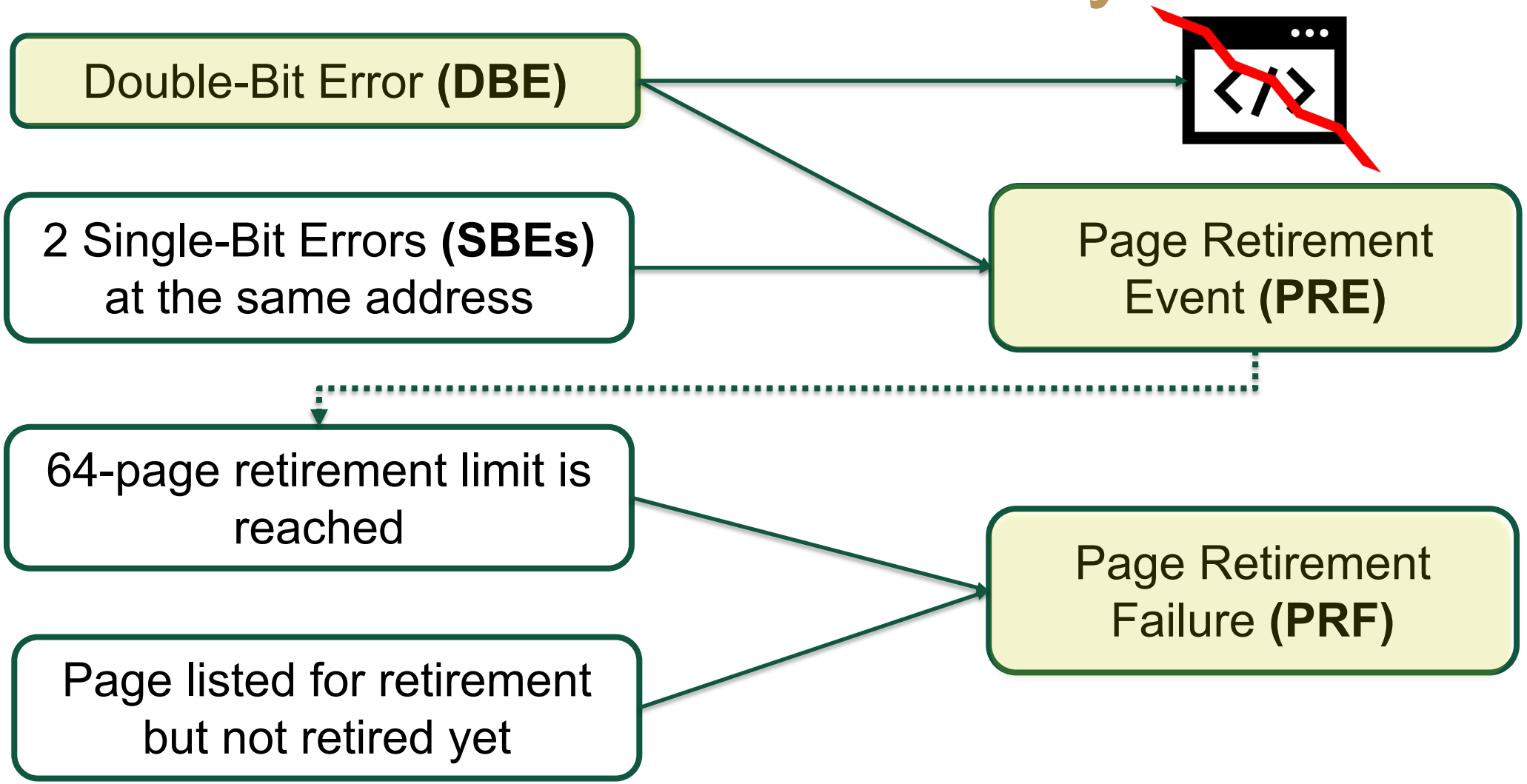
Job scheduler allocation history — 285 MB
- Project, user, parameters, start & end time

Per-node job schedule history — 14 GB
- Job allocation history and statistics

Per-GPU telemetry — 16 TB Compressed
- Power and temperature

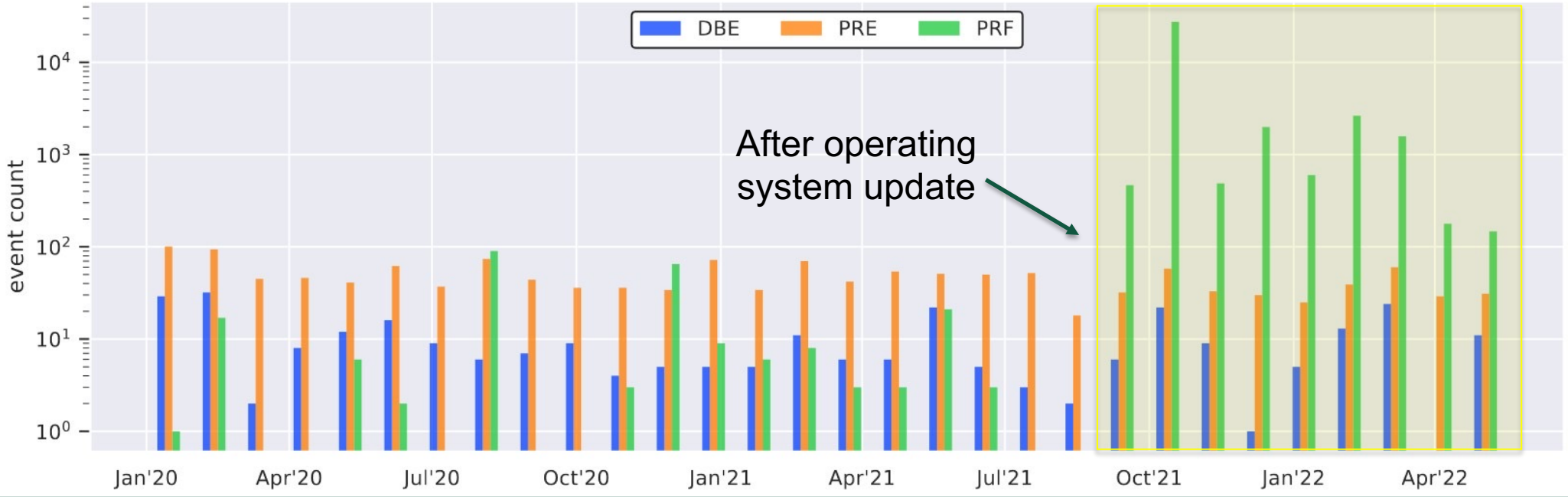
Publicly available! ***https://doi.org/10.13139/OLCF/1970187***

Memory Corruption Event Types: Rare but Deadly



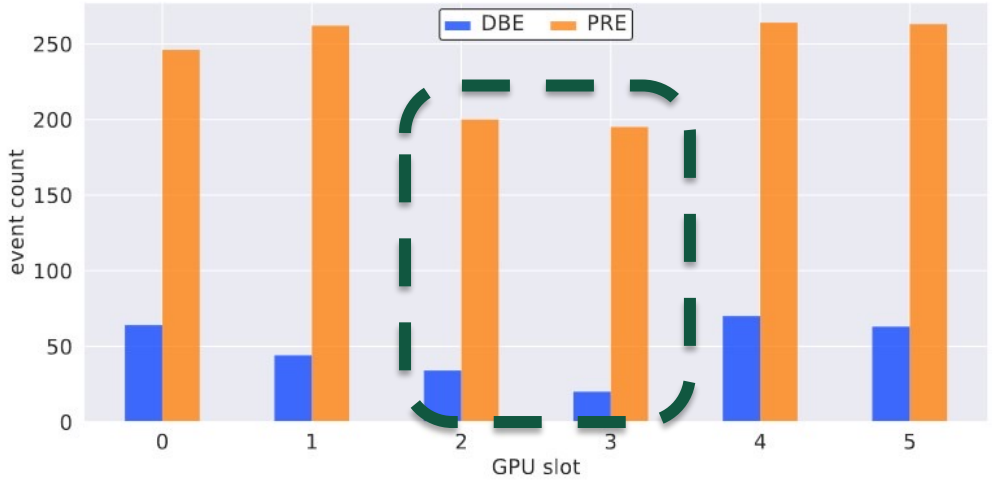
Temporal Trends

- PREs/DBEs consistent over time
- **Many PRFs occur repeatedly on the same GPU**
  - 97.2% of PRFs are from 27 jobs
  - Application repeatedly accesses memory page listed for retirement



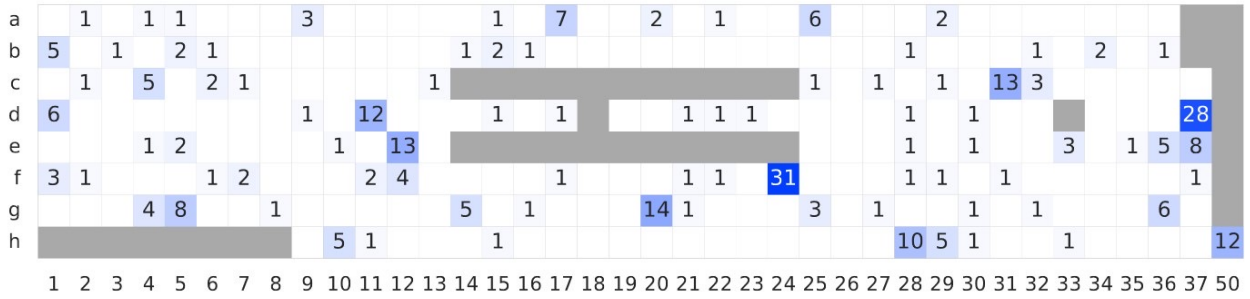
Spatial Distribution

- No correlation with cabinet location in the facility
- **No correlation with node height in cabinet**
- **GPU 2 & 3 on node are geometrically central with fewer errors**

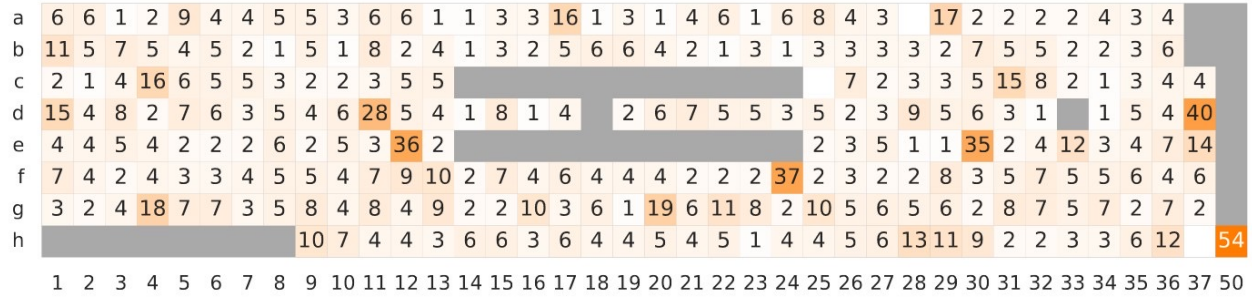


Errors per GPU slot

DBEs per Summit cabinet (top-down)

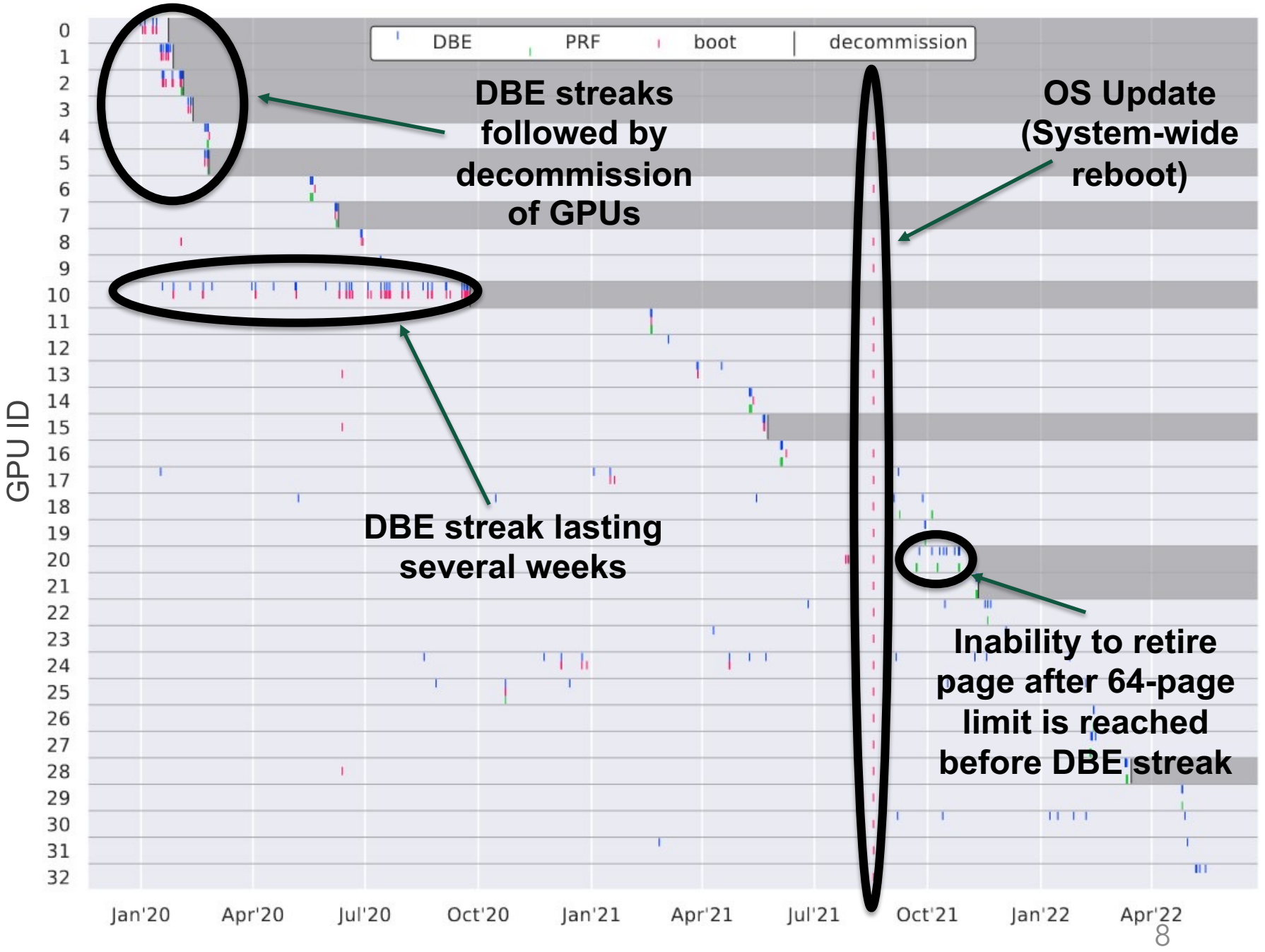


PREs per Summit cabinet (top-down)



Are There DBE Patterns?

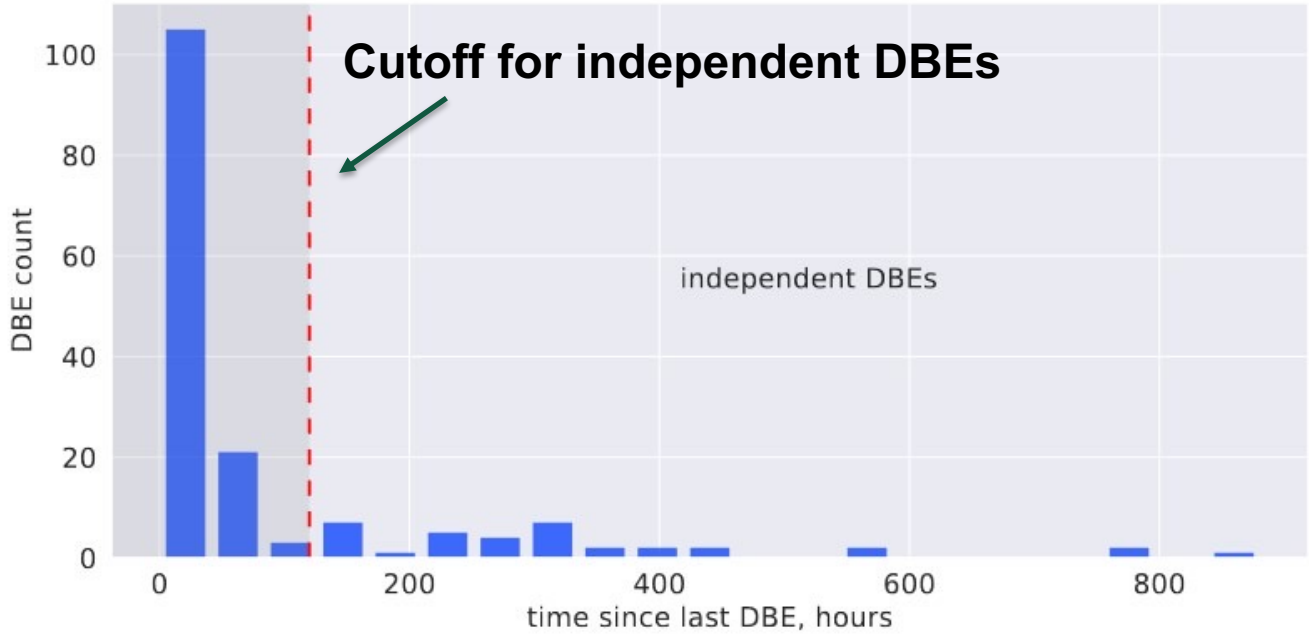
- **GPUs often experience DBEs in streaks**
- Time between DBEs on GPUs with >1 DBE:
  - Mean: 20 Days
  - Median: 20 Hours



Understanding GPU Trends

Snapshot (at a time stamp)

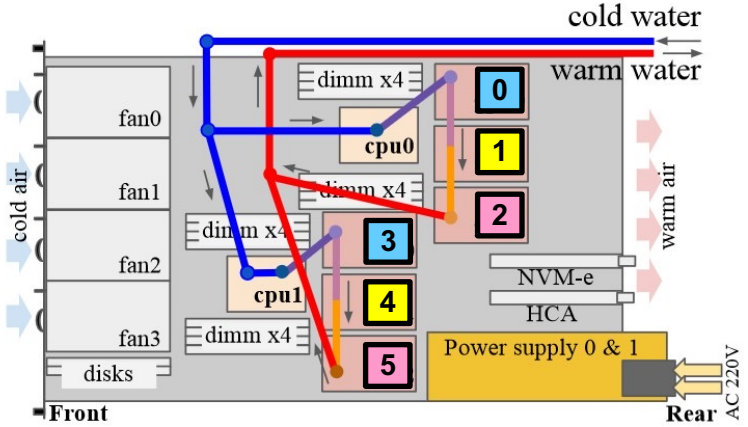
- Aggregates of power and temperature
- Job parameters
- Independent DBE points
- Infer GPU usage from power consumption



Patterns of Lifetime Utilization

Power

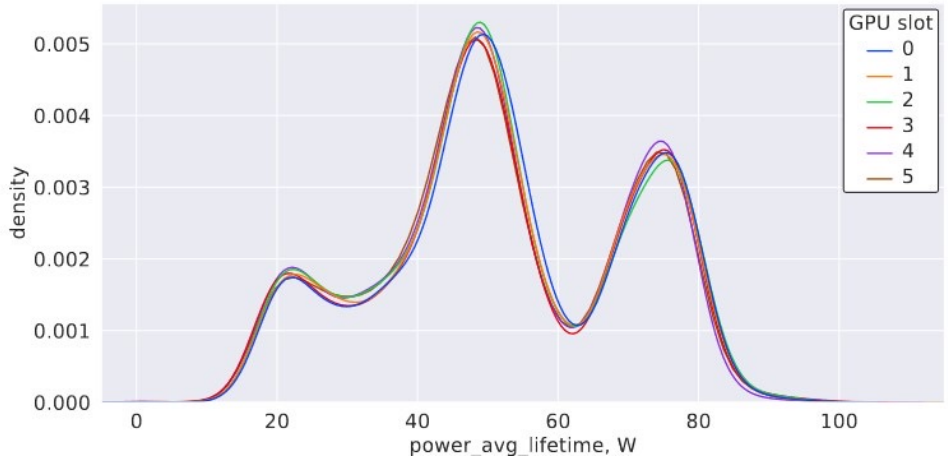
- Nodes fall into 3 groups with uneven workloads
- Identical per GPU slot
- **Caused by job scheduler policies**



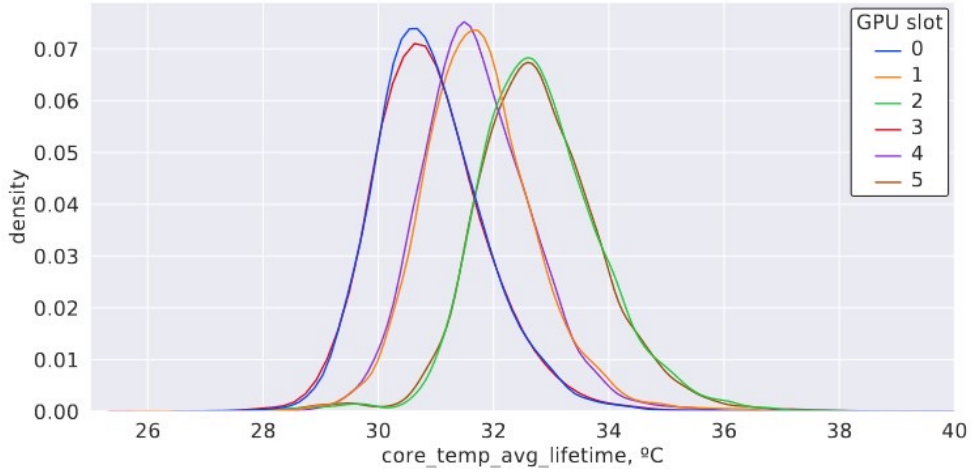
Temperature

- GPU temperature depends on **the order coolant reaches each slots**

Average lifetime power



Average lifetime core temp

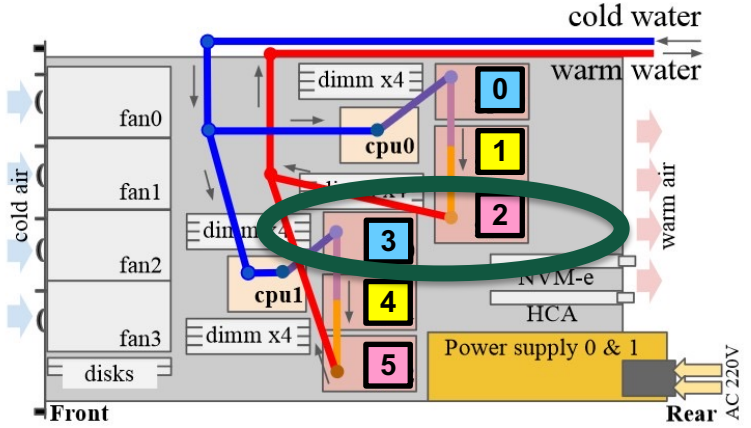


Does GPU Slot Affect DBEs?

Are snapshot type and GPU slot independent?

- Dependent
- If streaks excluded, independent

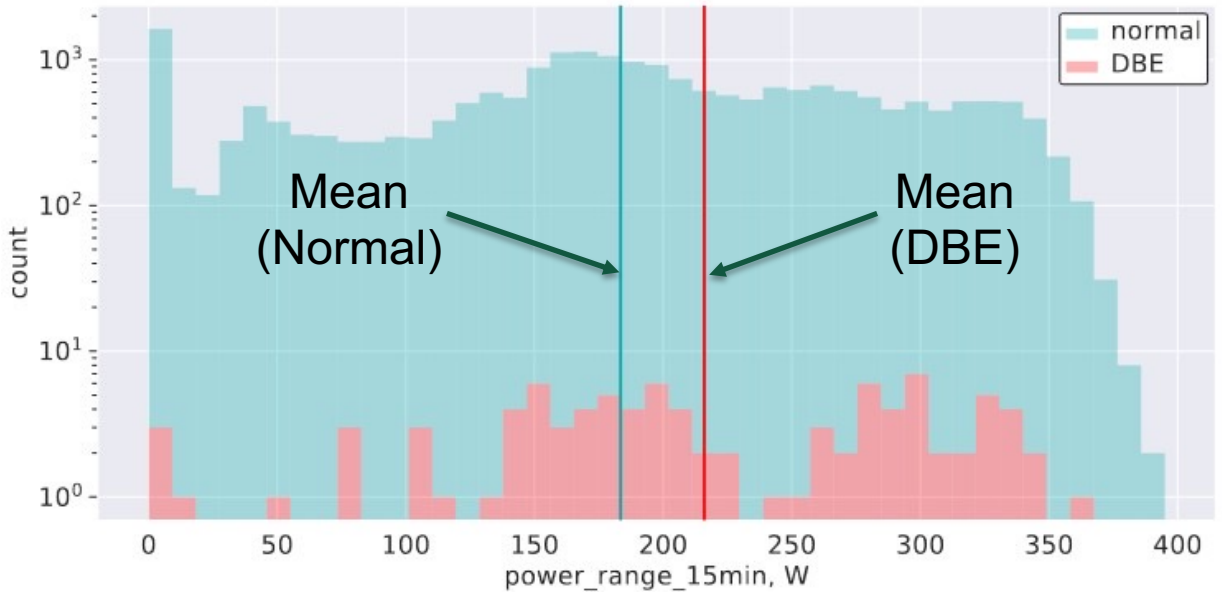
- **Geometrically central GPUs are more resilient to DBE streaks**
- Not linked to patterns in GPU *utilization* or *telemetry*



Thermal & Power Effects on DBEs?

Short term power range

- High variation in short-term power may increase susceptibility



Higher baseline activity increases susceptibility

Frequent utilization increases susceptibility

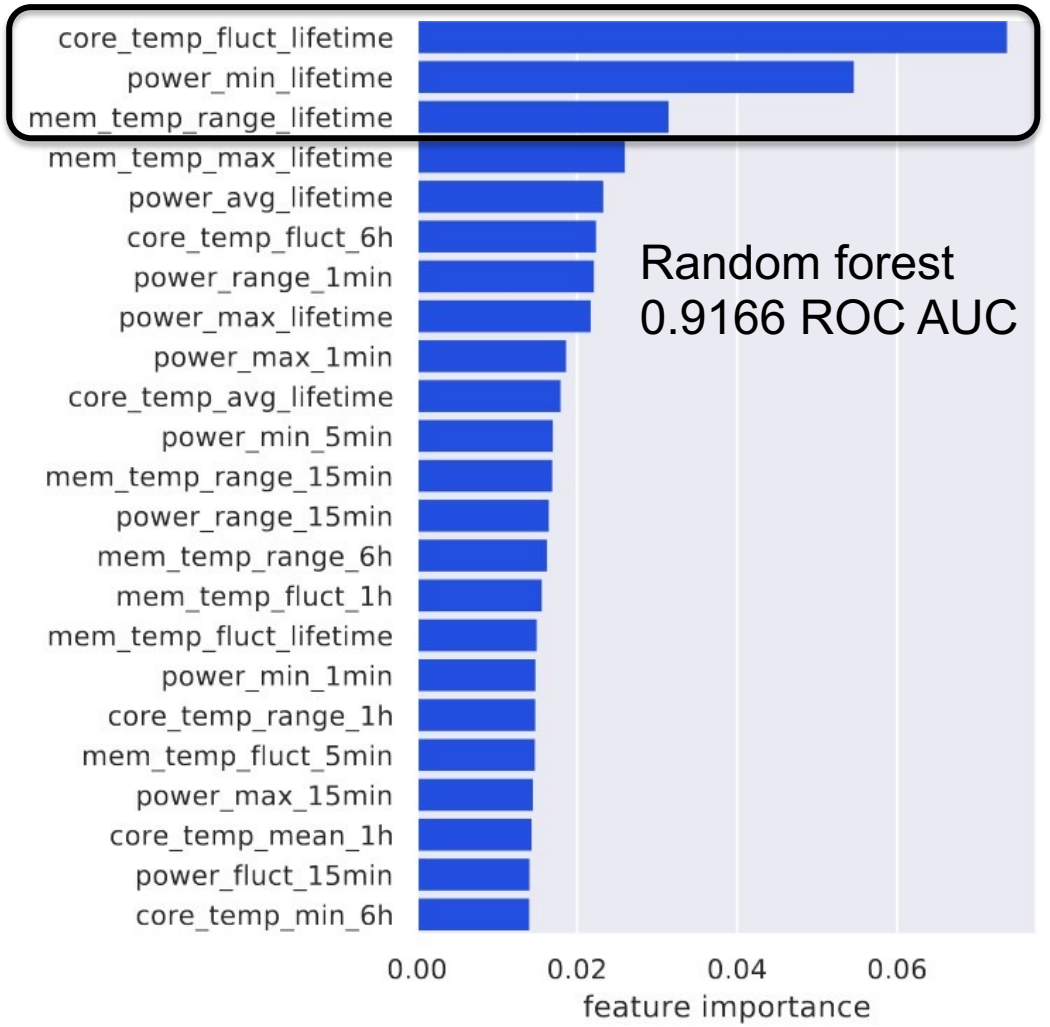
Most significant difference

Variable	p-value	DBE ≶ normal
power_min_lifetime	0.01618	>
power_avg_lifetime	0.01588	>
power_fluct_lifetime	0.01413	>
core_temp_min_lifetime	0.01033	>
core_temp_fluct_lifetime	0.01120	<
power_max_6h	0.01196	>
power_range_6h	0.01106	>
mem_temp_fluct_1h	0.04880	<
power_min_1h	0.00361	<
power_max_1h	0.00927	>
power_range_1h	0.00409	>
core_temp_max_1h	0.02344	>
core_temp_range_1h	0.00239	>
mem_temp_max_1h	0.02848	>
mem_temp_range_1h	0.00803	>
power_min_15min	0.01722	<
power_max_15min	0.00180	>
power_range_15min	**0.00056**	>
core_temp_max_15min	0.04717	>
core_temp_range_15min	0.00381	>
mem_temp_range_15min	0.02683	>
power_max_5min	0.00274	>
power_range_5min	0.00124	>
core_temp_range_5min	0.02453	>
power_max_1min	0.00271	>
power_range_1min	0.00239	>
core_temp_max_1min	0.04593	>

Snapshot Classification

Classification of DBE vs. Normal

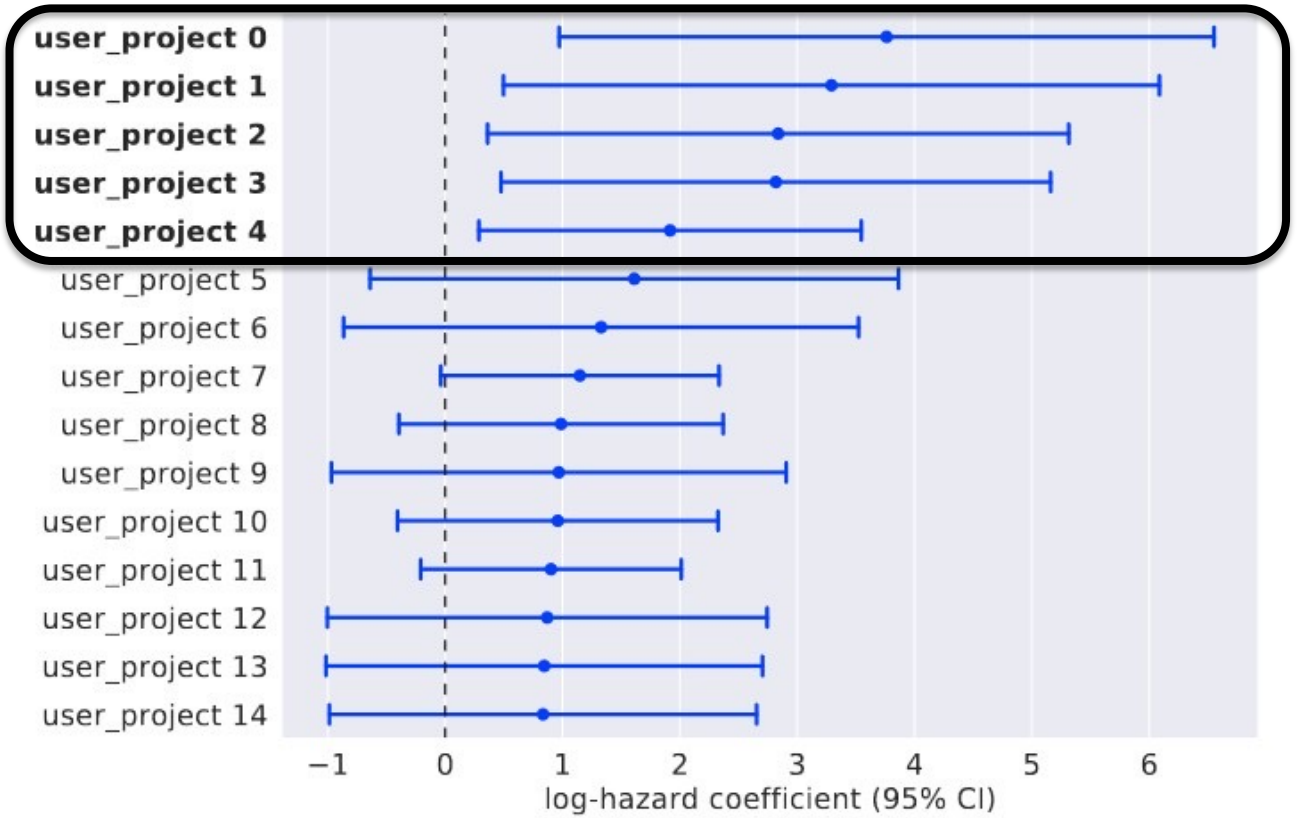
- Explainable ML methods (random forest, SVM, …)
  - Poor performance on entire data set
  - Good performance on GPUs with a prior DBE
- **Stress factors on predisposed units are distinct from the general population**



Workload Patterns and DBEs

Survival analysis

- 5 workloads have a higher chance of DBEs
  - 4 use **mixed precision arithmetic**



Conclusions

- Operational patterns affect increased DBEs in Summit
  - Physical node placement
  - HPC operation stresses
  - Individual predisposition
- Challenges
  - Unavailability of temporal SBE data
  - Unavailability of intensity of memory operations
- Root causes of DBEs are still an open problem
  - Findings here may be useful for other HPC systems

WILLIAM & MARY
CHARTERED 1693

OAK RIDGE
National Laboratory

QUESTIONS?

Thank you to my collaborators:

Vladyslav Oles, George Ostrouchov, Woong Shin, Evgenia Smirni, and Christian Engelmann

Anna Schmedding | akschmedding@wm.edu |
https://www.cs.wm.edu/~akschmed/