OAK RIDGE National Laboratory

GPU Lifetimes on Titan Supercomputer: Survival Analysis and Reliability

George Ostrouchov✱, Don Maxwell ⁺,
Rizwan A. Ashraf ✱, Christian Engelmann ✱,
Mallikarjun Shankar ⁺, James H. Rogers ⁺

✱ Computer Science and Mathematics Division
⁺ National Center for Computational Sciences Division
Oak Ridge National Laboratory

ORNL is managed by UT-Battelle, LLC for the US Department of Energy

U.S. DEPARTMENT OF ENERGY

Effects of Architecture and Scheduling on GPU Reliability

The Science

- Over 100,000 collective years of GPU operation
- Advanced statistical methods typical in biomedicine
- Data and analysis codes made publicly available for reproducibility

Findings

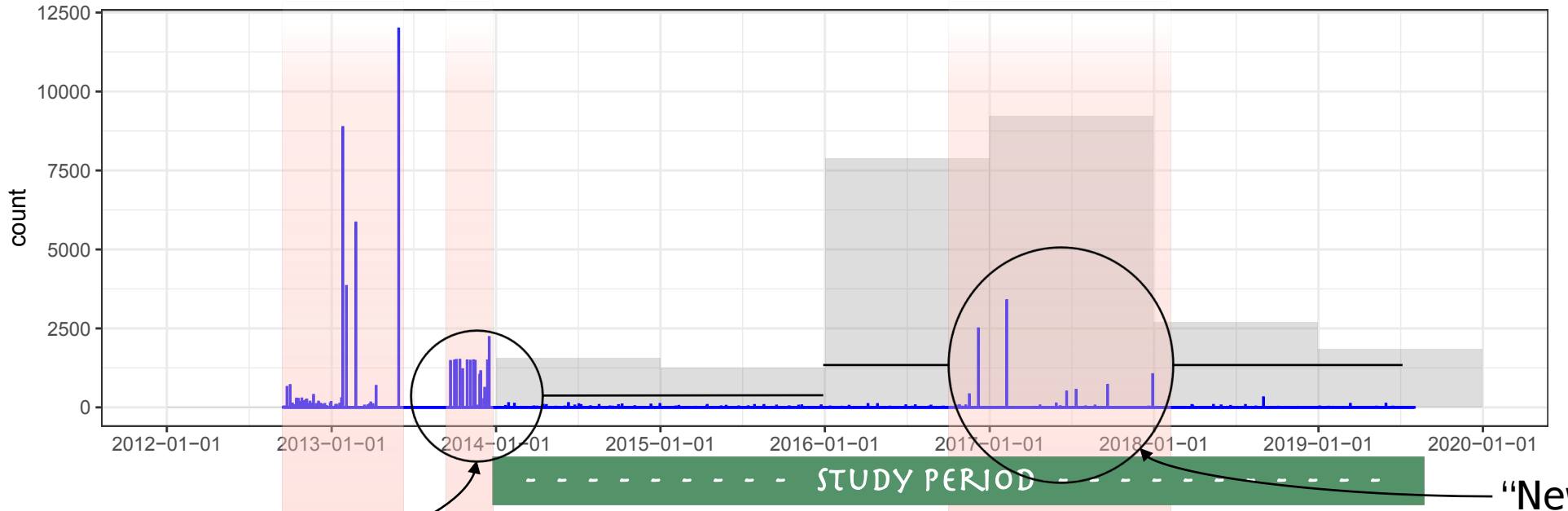
- Failure mode provided strong reliability signal
- Correlations explainable by cooling air transport
- Correlations explainable by job scheduling

Overview

- GPU-focused history of Titan
- Unexpected mid-life failures
- Titan GPU architecture
- Data collection and curation
- Traditional MTBF analysis
- Statistical survival analysis
- Conclusions

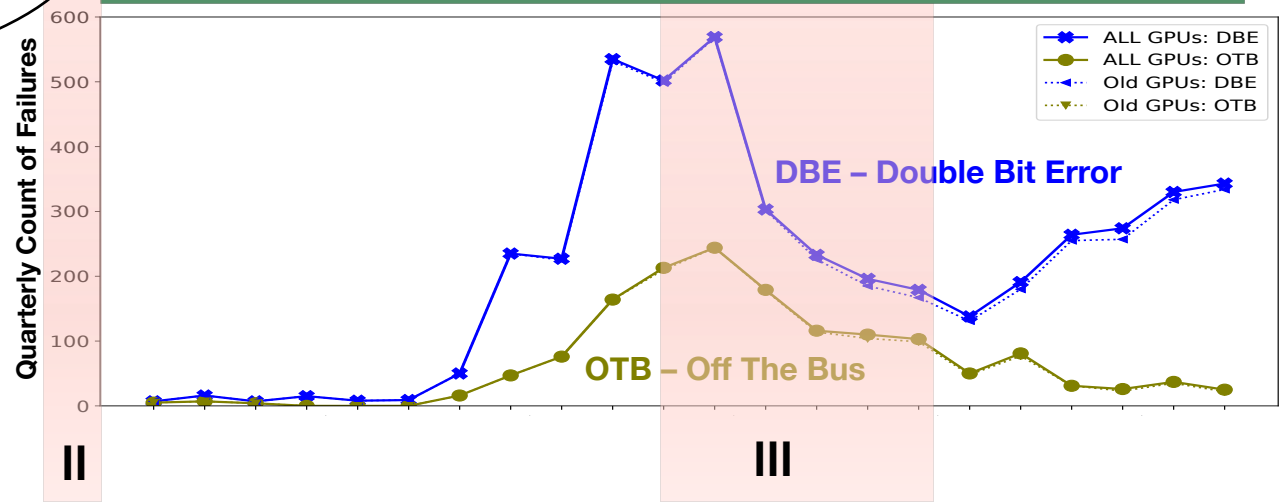
Three Rework Cycles and Years of Stable Operation

GPU swaps detected at inventories (narrow blue) and yearly sum totals for 2014 and later (wide gray)

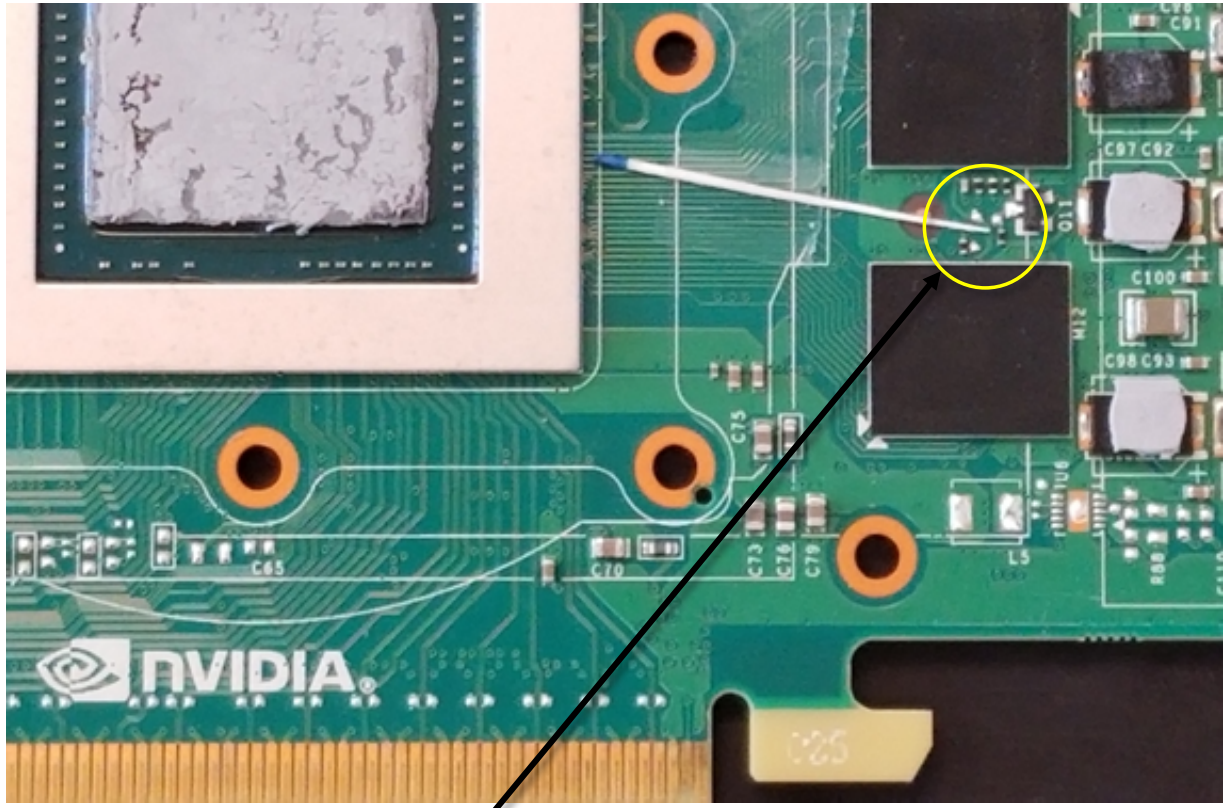


"Old" GPUs

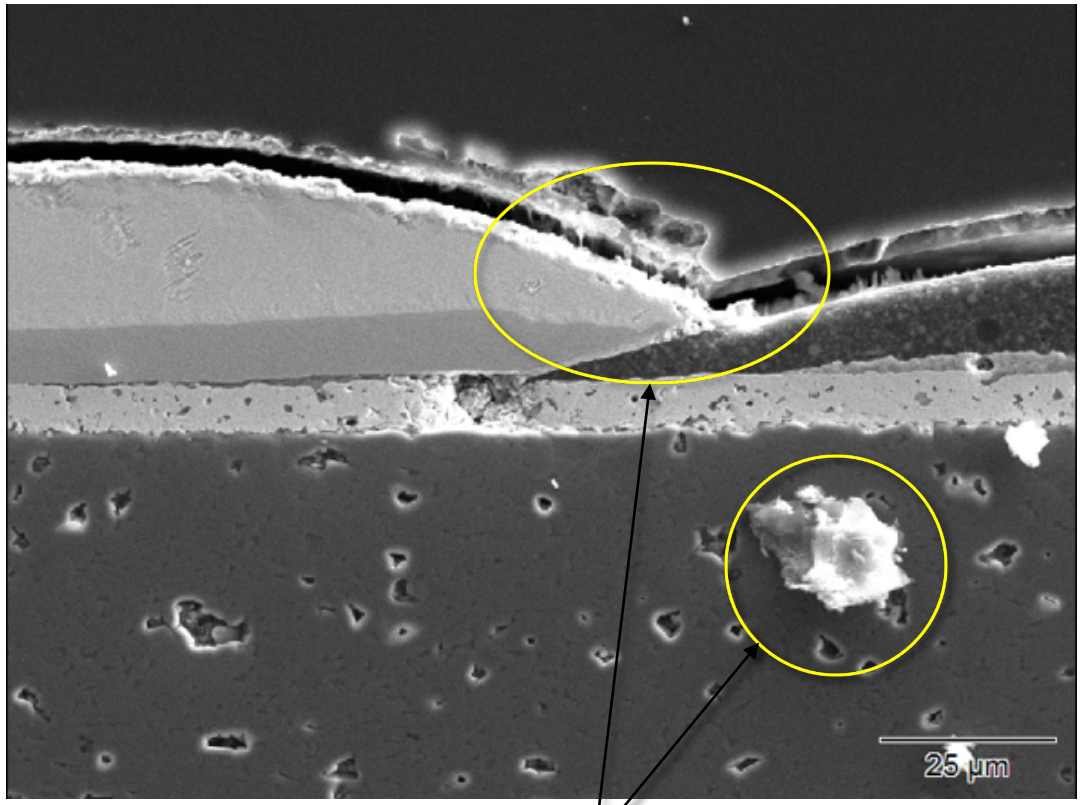
"New" GPUs



Root Cause: Non-ASR Components on SXM GPU



NVIDIA SXM – Location of a non-ASR

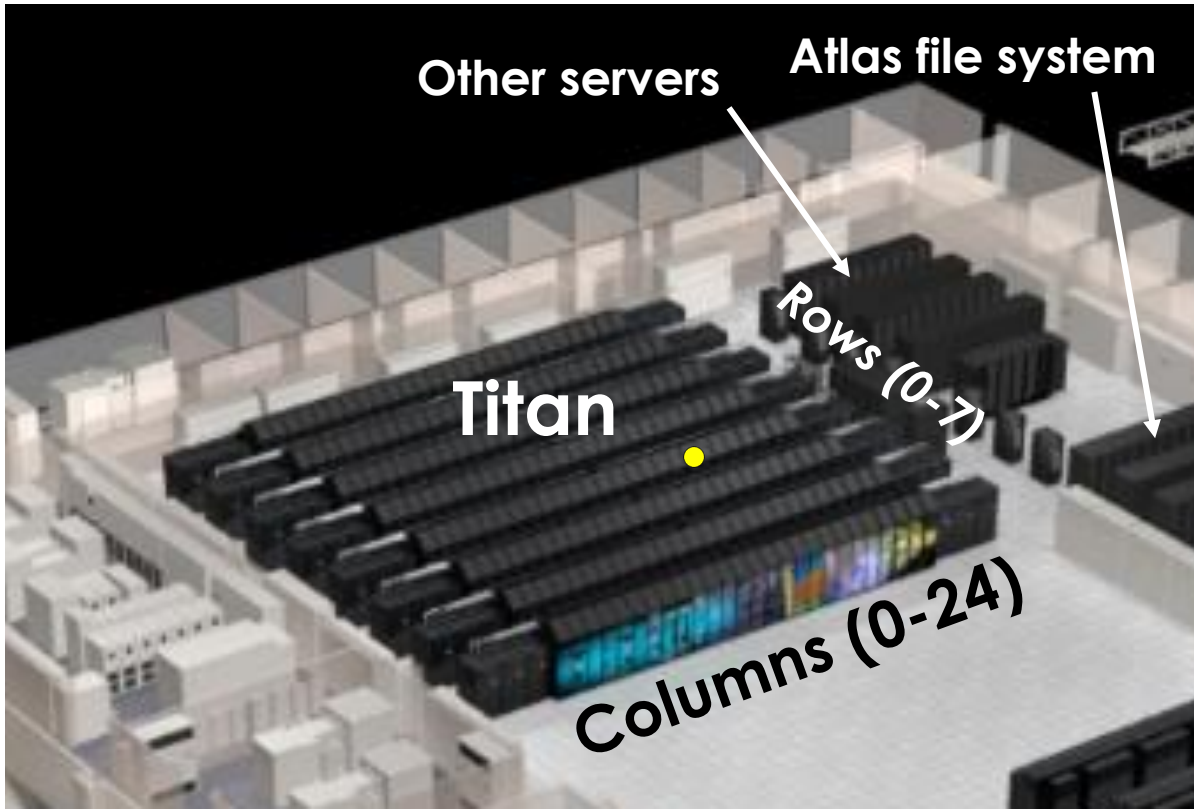


Silver-sulfide corrosion
"Flowers-of-Sulfur"

ASR = Anti-Sulfur Resistor

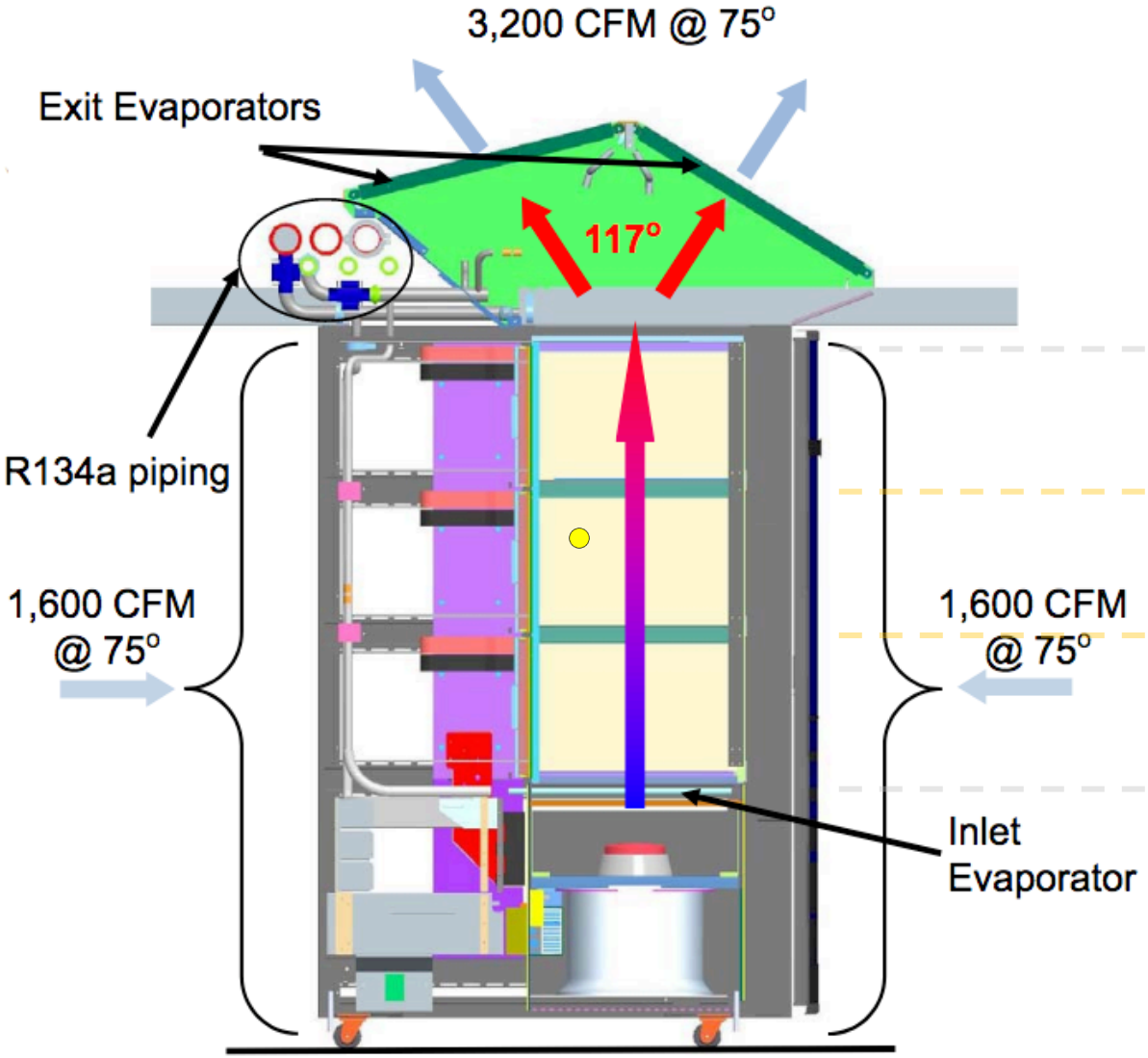
Machine Room Layout: GPU Locators

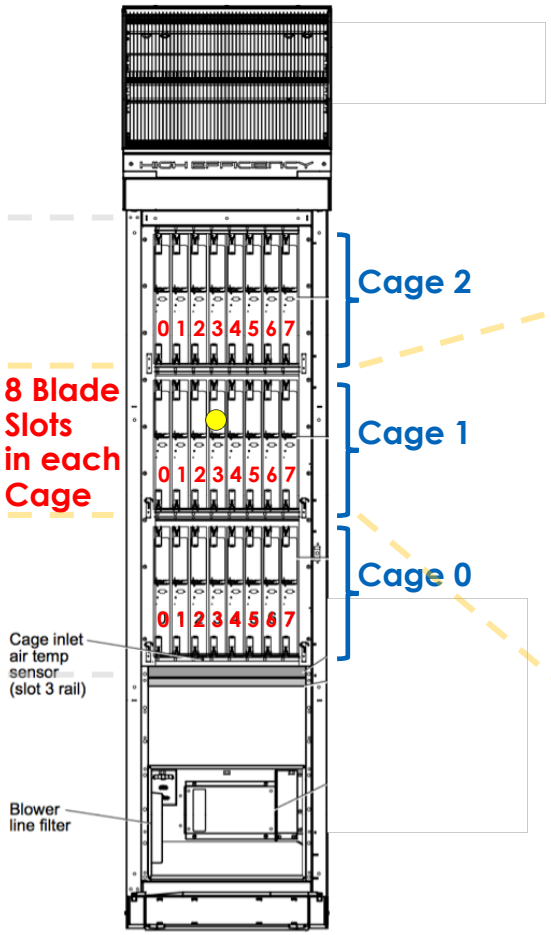
```
c##-#c#s#n#
a    a l o
b    g o d
i    e t e
n
e
t
```

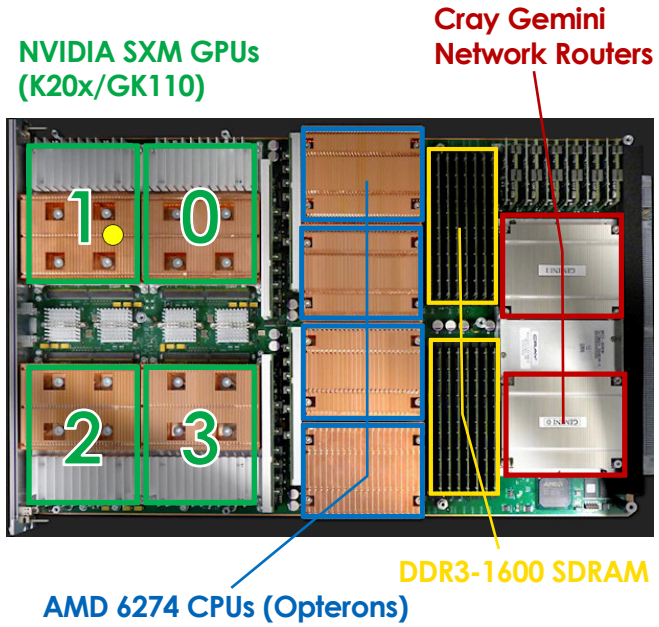


- **c17-4c1s3n1** = cabinet in column **17** – row **4**

Cabinet Mechanical Packaging: Locating a GPU

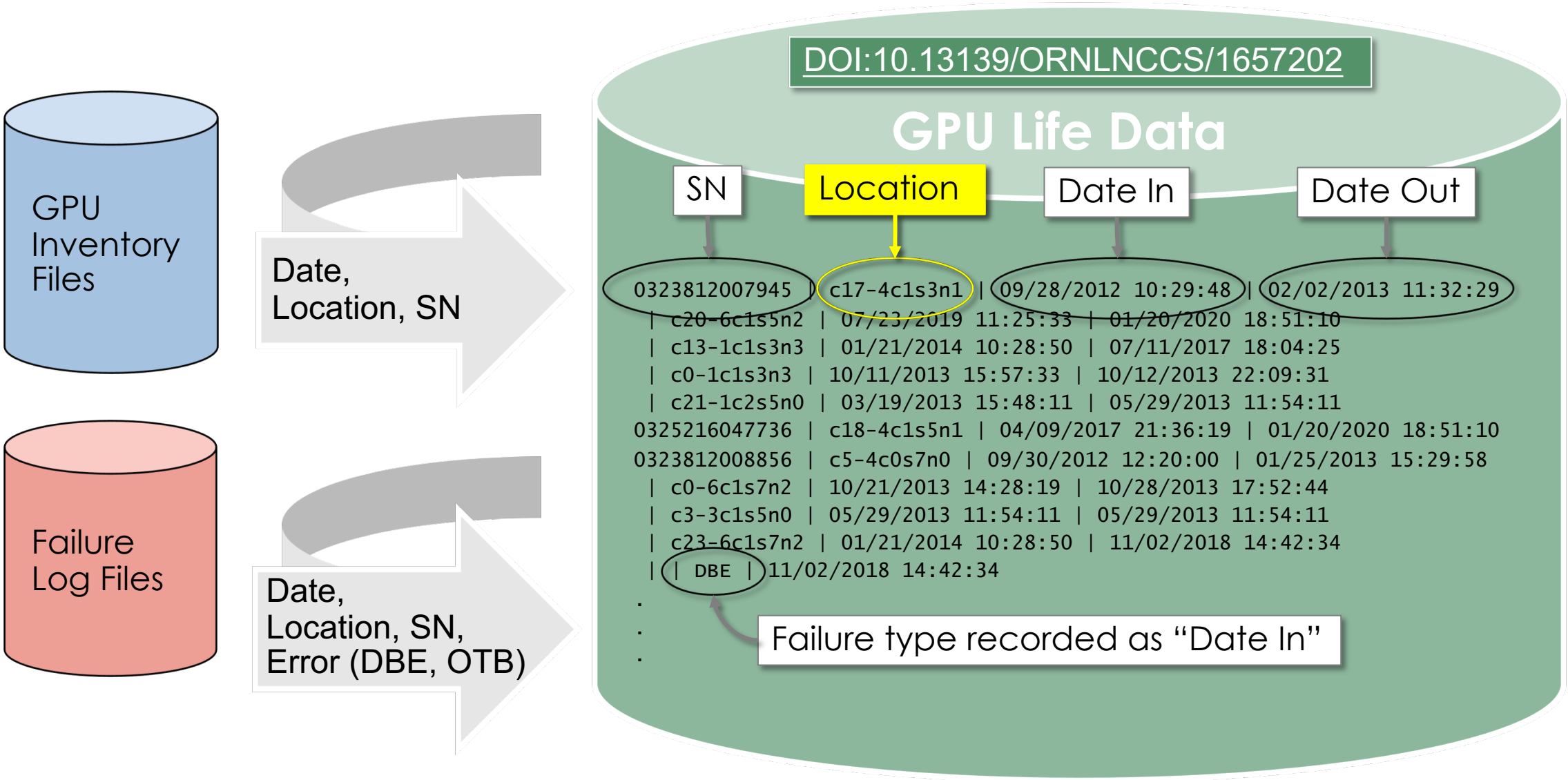






- c17-4c1s3n1 = cage **1**, slot **3**, node **1**

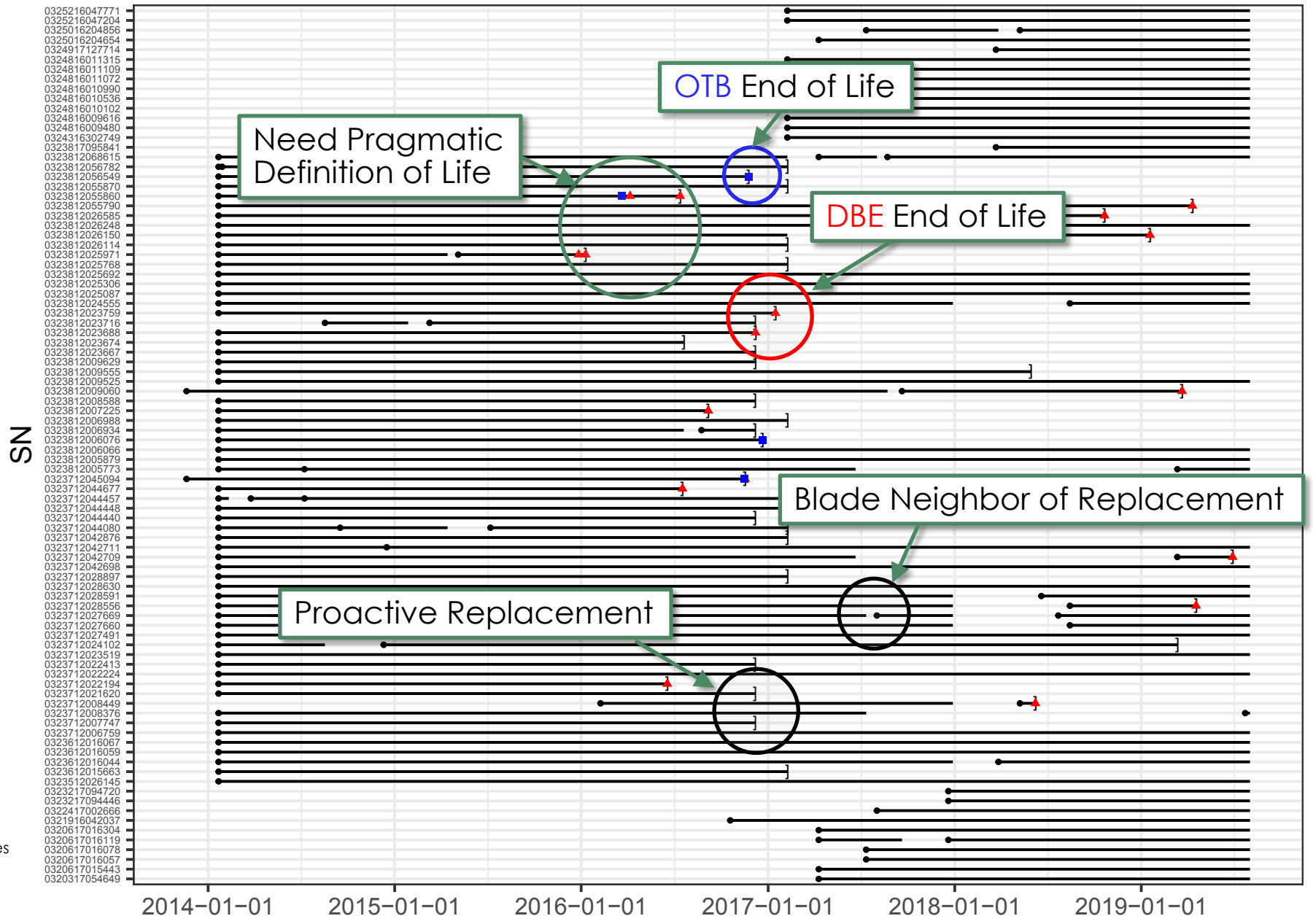
GPU Life Data Built Incrementally from Two Sources



GPU Life Visualization: Serial Number View

Critical for:

- Understanding data
- Defining GPU Life
- Data processing verification

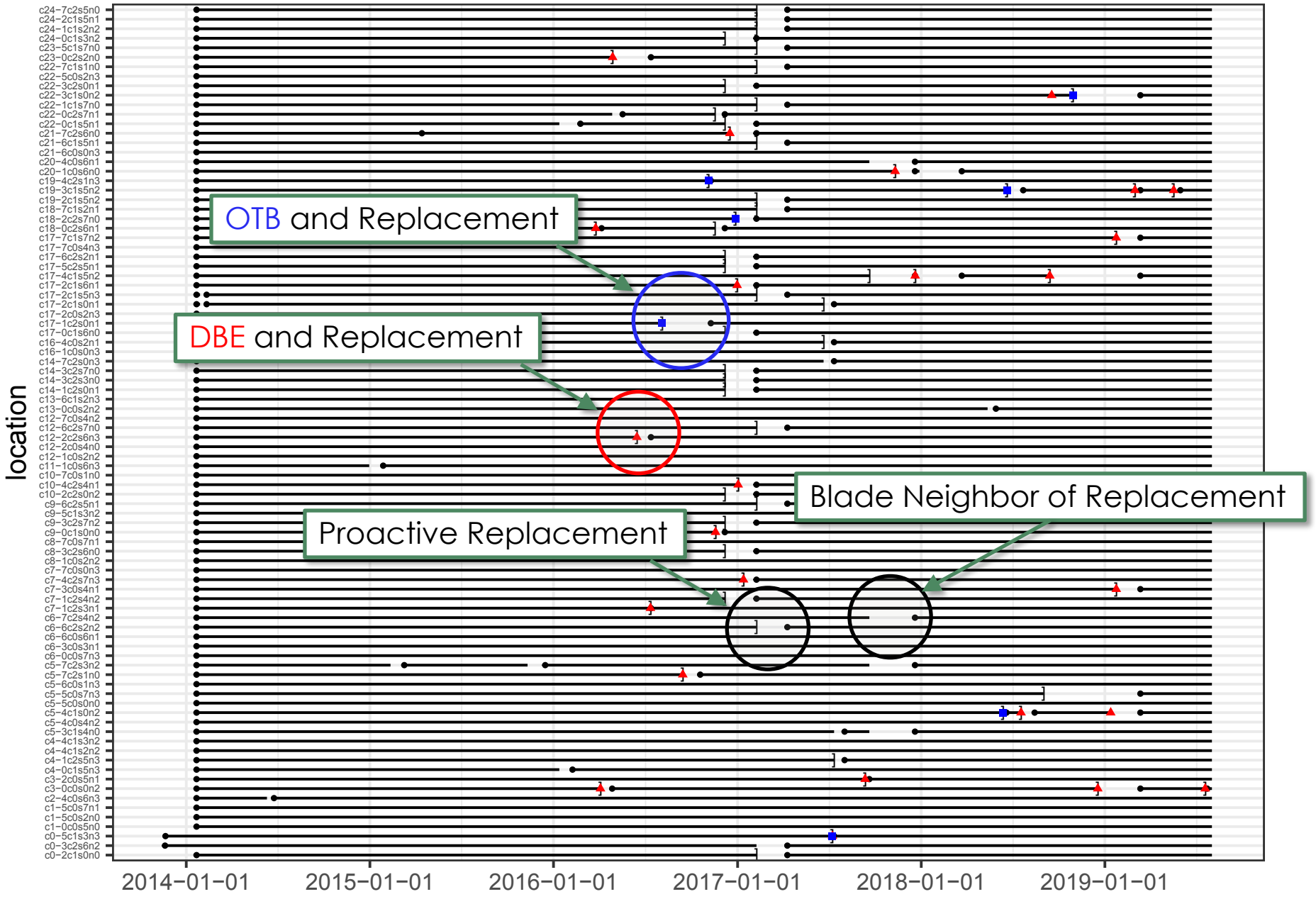


Produced in R via `ggplot2` and `lubridate` packages

GPU Life Visualization: Location View

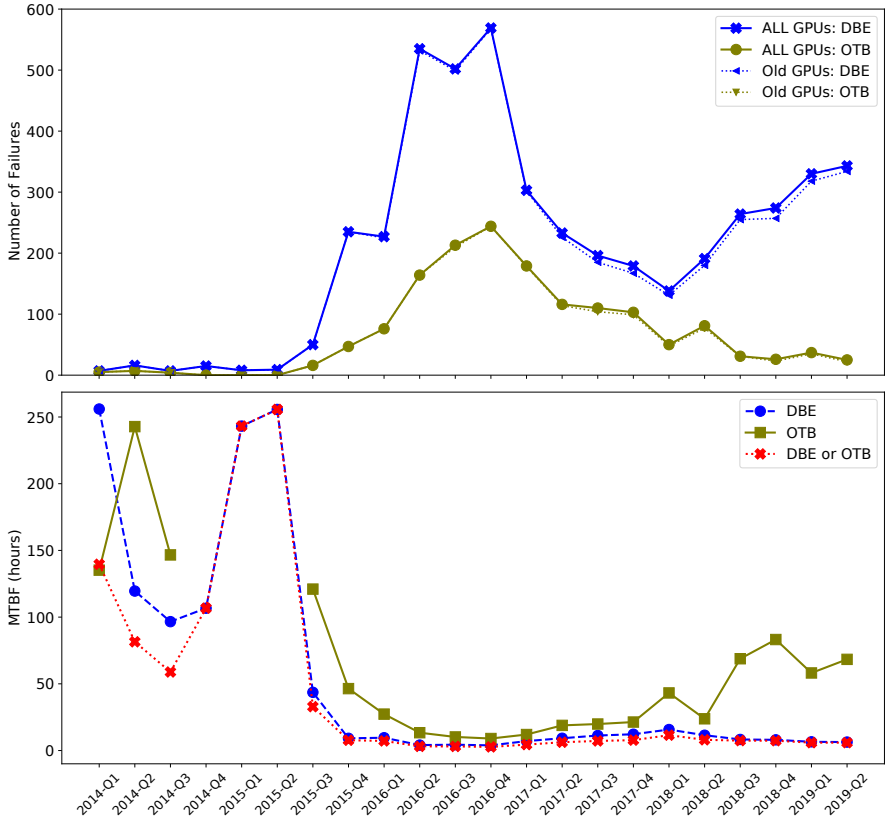
Critical for:

- Understanding data
- Defining GPU Life
- Data processing verification

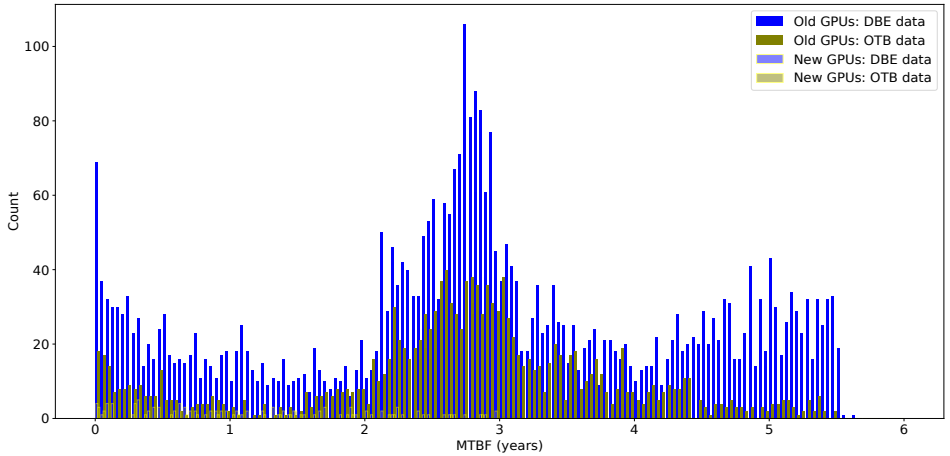


Produced in R via `ggplot2` and `lubridate` packages

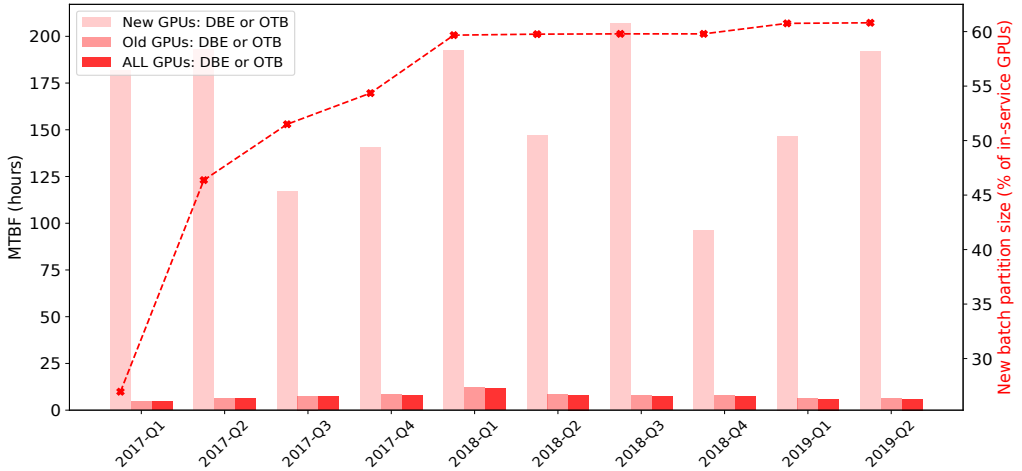
Traditional Reliability in HPC is Focused on MTBF



System-wide Reliability: Quarterly number of failures (top) and MTBF (bottom).



Individual GPU Reliability: MTBF histogram for units that had at least one failure. Interpret carefully: lacks information from units with no failures!



Old–New as Two Partitions: MTBF differs by 12x factor!

Survival Analysis Requires a Definition of GPU Lifetime

Lifetime with observed failure:

- Unit experiences at least one DBE or OTB
- Unit is removed from the system
- Lifetime is the cumulative operation time over all its locations

Censored lifetime:

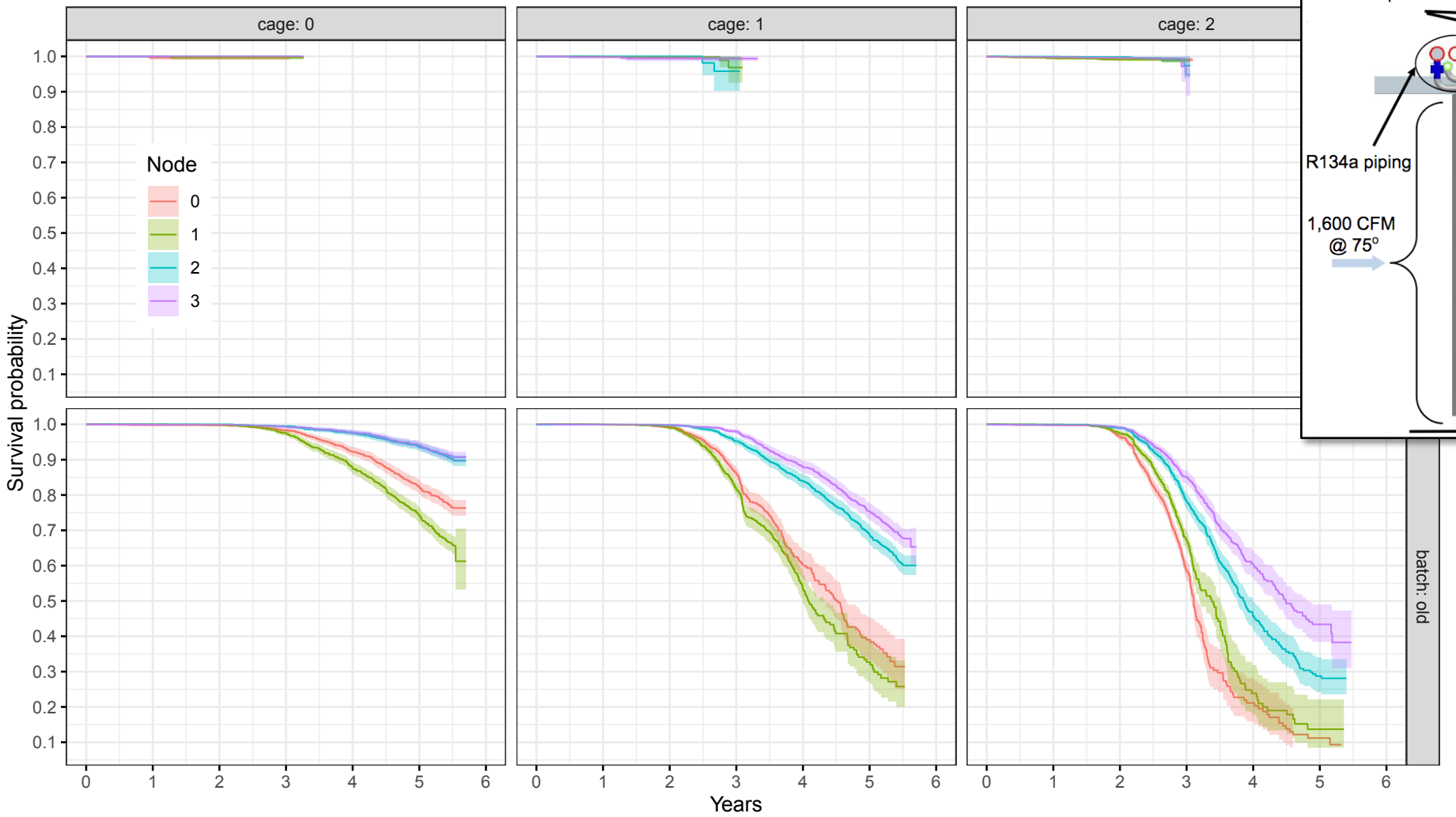
- Unit is removed or is present when system turned off
- If removed, no DBE or OTB events
- Lifetime is the cumulative operation time over all its locations
- Ignore censoring due to observation only at inventories

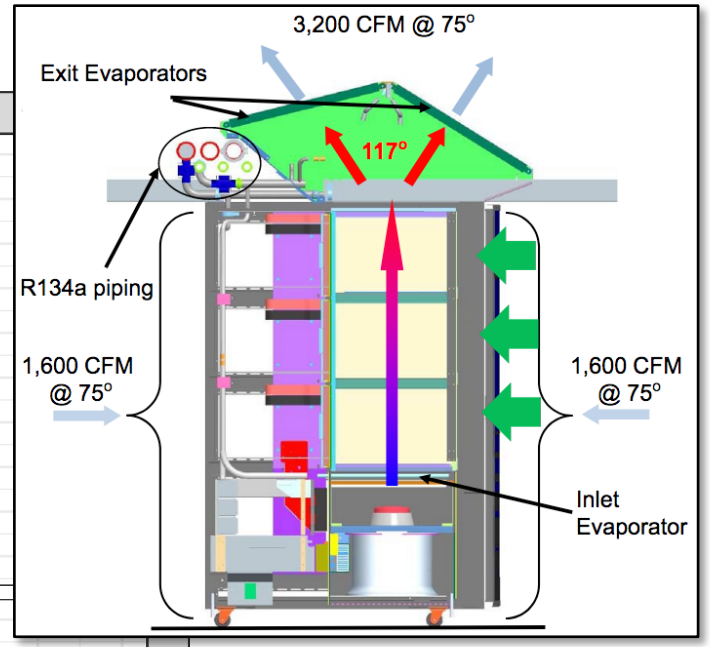
Kaplan-Meyer Survival Analysis

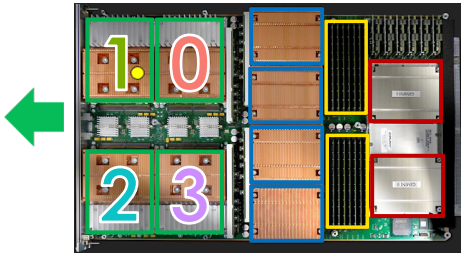
- Commonly used in Biostatistics and Biomedical research*
- Nonparametric
  - If T is failure time and $F(t) = \Pr\{T < t\}$ is the cumulative failure distribution function
  - Then the survival probability, $S(t) = \Pr\{T \geq t\} = 1 - F(t)$, is its complement
  - Recursive computation $S(t_2) = \Pr\{\text{survive from } t_1 \text{ to } t_2\}\, S(t_1)$
- Able to incorporate censoring
- Split population into groups
- Available uncertainty estimate

*E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American Statistical Association*, vol. 53, no. 282, pp. 457–481, 1958.

Cage and Node Effect Explainable by Airflow in Cabinet







Cox Proportional Hazards Regression Model

- Commonly used in Biostatistics and Biomedical research*
- Able to adjust for covariate effects
- Each GPU is like a patient, affected by its location (treatment)
- The hazard for patient k is $H_k(t) = H_0(t)e^{\sum_1^n \beta_i x_i}$
  - Base hazard rate, $H_0(t)$, multiplied by a function of covariates (hazard coefficient)
- Semiparametric model
  - Baseline hazard is nonparametric (no functional shape assumption)
  - Hazard coefficient is a parametric function of covariates
- Assumes hazards are proportional

*D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 34, no. 2, pp. 187– 220, 1972.
We use R packages `survival` and `survminer`.

Strong Signal in old Batch, Pattern Similar to K-M Analysis

Variable		N	Events
col			
	0 (X− 1)	744	173
	1 (X− 2)	759	169
	2 (X−25)	772	97
	3 (X− 3)	739	217
	4 (X−24)	755	88
	5 (X− 4)	759	266
	6 (X−23)	752	125
	7 (X− 5)	744	225
	8 (X−22)	758	131
	9 (X− 6)	758	259
	10 (X−21)	739	180
	11 (X− 7)	760	233
	12 (X−20)	756	228
	13 (X− 8)	759	242
	14 (X−19)	743	228
	15 (X− 9)	753	257
	16 (X−18)	754	224
	17 (X−10)	758	207
	18 (X−17)	758	194
	19 (X−11)	751	192
	20 (X−16)	766	217
	21 (X−12)	759	218
	22 (X−15)	755	244
	23 (X−13)	743	220
	24 (X−14)	743	193
row			
	0	2363	795
	1	2352	639
	2	2351	565
	3	2363	596
	4	2371	687
	5	2334	577
	6	2339	472
	7	2364	696
cage			
	0	6240	1107
	1	6275	1746
	2	6322	2174
slot			
	0	2355	637
	1	2353	592
	2	2349	575
	3	2352	553
	4	2364	606
	5	2349	633
	6	2357	694
	7	2358	737
node			
	0	4712	1388
	1	4722	1494
	2	4697	1195
	3	4706	950

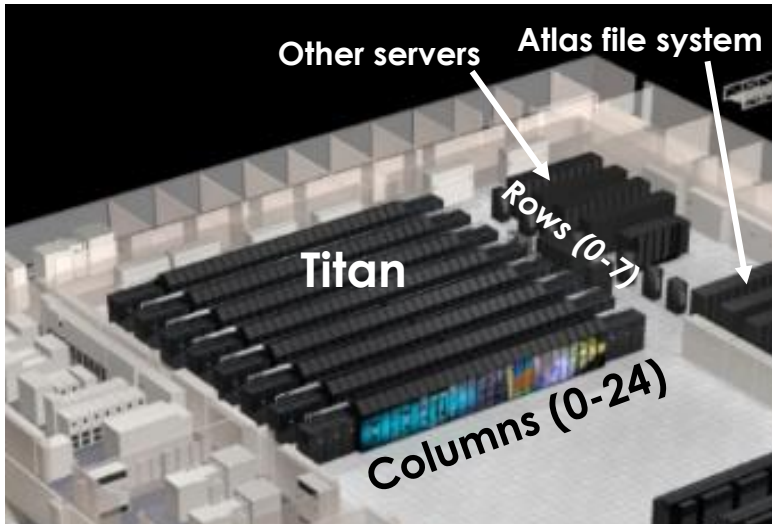
Hazard ratio

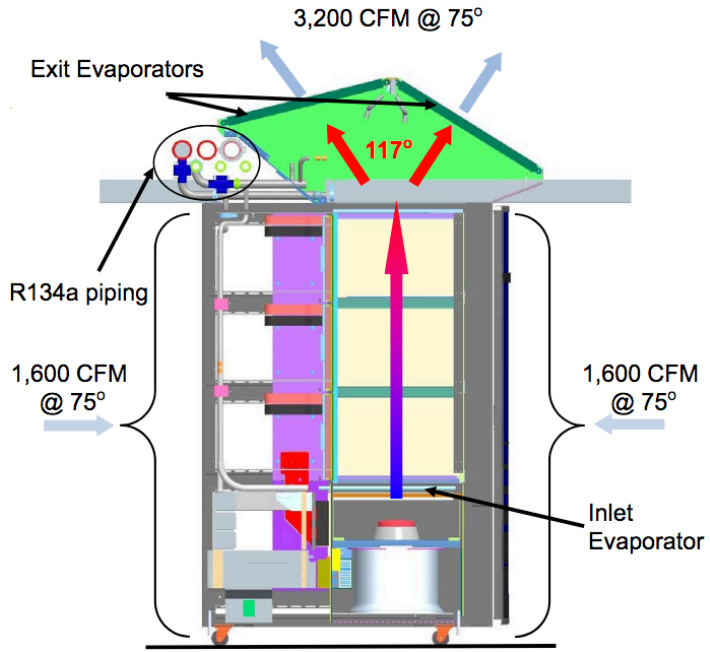
0.5 1 2 5 10 20

Variable		N	Events
col			
	0 (X− 1)	400	1
	1 (X− 2)	389	1
	2 (X−25)	384	1
	3 (X− 3)	383	2
	4 (X−24)	380	1
	5 (X− 4)	399	2
	6 (X−23)	378	0
	7 (X− 5)	451	2
	8 (X−22)	443	2
	9 (X− 6)	473	0
	10 (X−21)	457	2
	11 (X− 7)	471	5
	12 (X−20)	461	2
	13 (X− 8)	462	3
	14 (X−19)	469	5
	15 (X− 9)	478	0
	16 (X−18)	473	5
	17 (X−10)	494	4
	18 (X−17)	507	4
	19 (X−11)	494	3
	20 (X−16)	518	1
	21 (X−12)	508	4
	22 (X−15)	526	5
	23 (X−13)	488	2
	24 (X−14)	484	5
row			
	0	1875	15
	1	1662	6
	2	1638	6
	3	1277	9
	4	1285	11
	5	1256	4
	6	1241	5
	7	1136	6
cage			
	0	777	2
	1	4123	16
	2	6470	44
slot			
	0	1434	13
	1	1389	6
	2	1410	5
	3	1427	5
	4	1414	3
	5	1424	10
	6	1434	9
	7	1438	11
node			
	0	3495	19
	1	3513	20
	2	2209	10
	3	2153	13

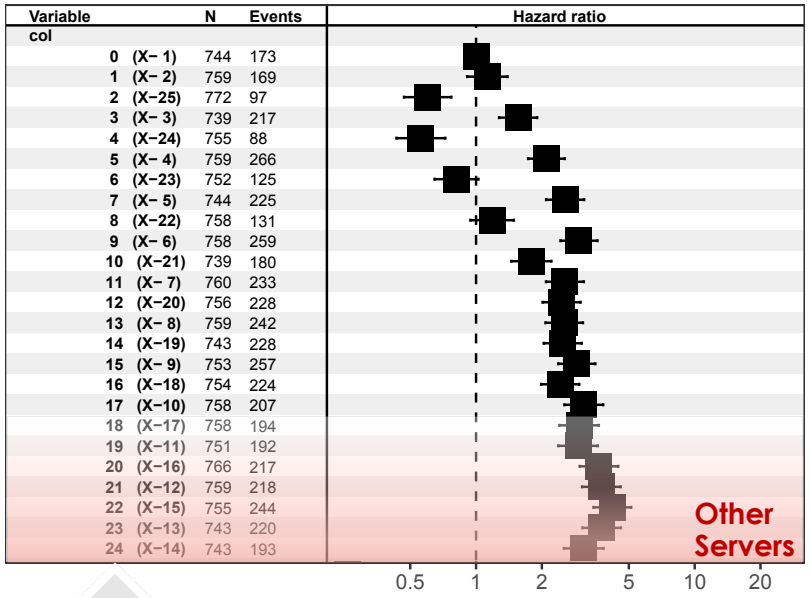
Hazard ratio

0.05 0.1 0.2 0.5 1 2 5 10 20

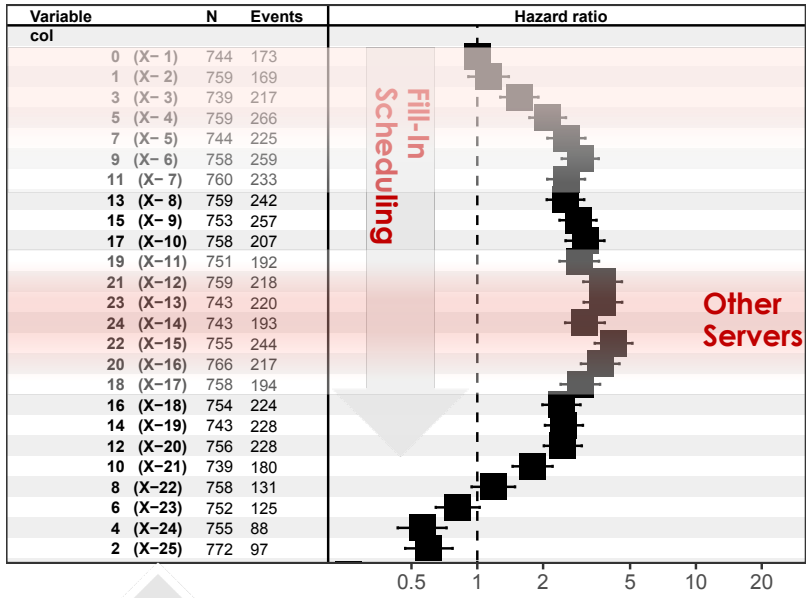




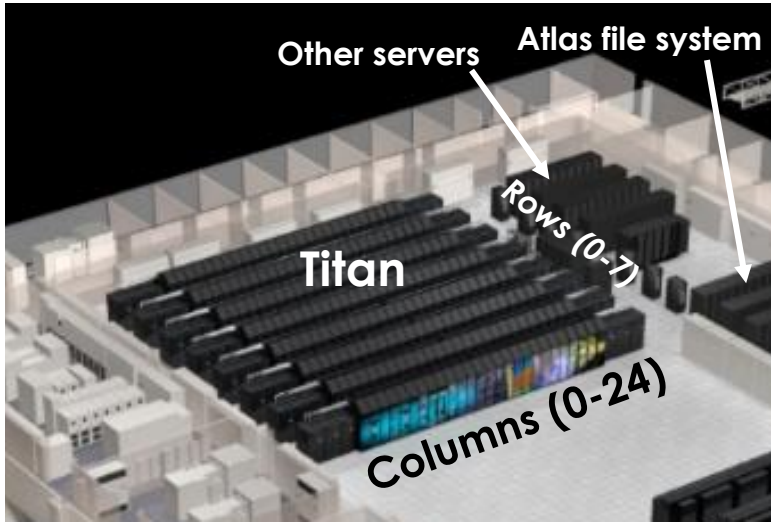
Fill-in Scheduling Effect Explainable via Torus Coordinate



Column Order



Torus Coordinate Order



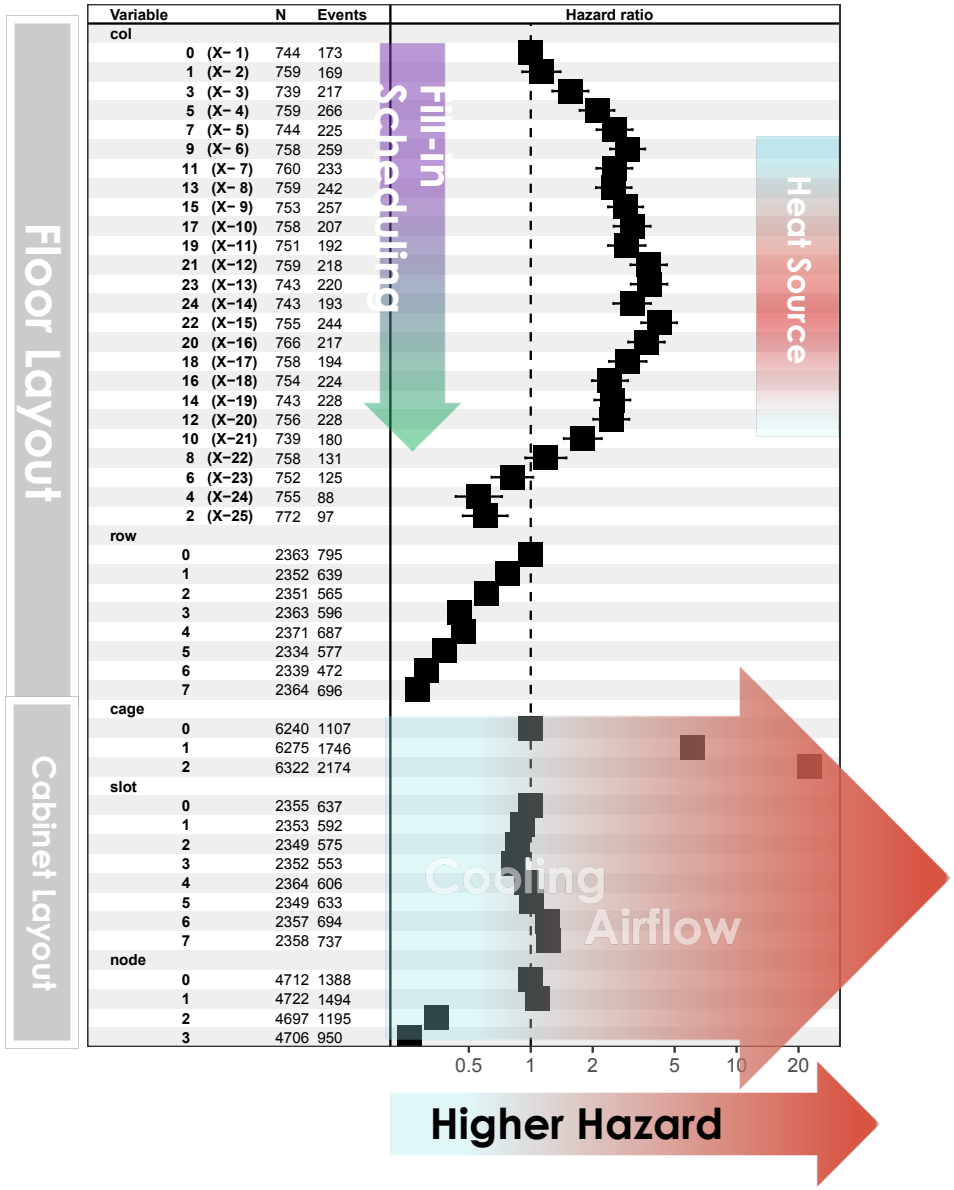
Cooling Architecture and Scheduling Affect Reliability

Conclusions for Future Systems

- Expect the unexpected with available resources
- Collect more detailed sensor data
- Use advanced statistics for configurable causal analysis
  - Provide early warning of reliability issues
  - Inform mitigation strategies

The Science

- 100,000 collective years of GPU life
- Advanced statistical methods typical in biomedicine
- Novel component lifetime visualizations
- Data and analysis codes made publicly available



Thank You!

Acknowledgements:

This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

This work was supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Resilience for Extreme Scale Supercomputing Sys- tems Program, with program managers Robinson Pino and Lucy Nowell.