MAY 8, 2025

ANALYZING A LIFETIME OF FAILURES ON A CRAY XC40 SUPERCOMPUTER

KEVIN A. BROWN*
TANWI MALLICK
ROBERT ROSS
Argonne National Laboratory
**kabrown@anl.gov*

ZHILING LAN
University of Illinois Chicago

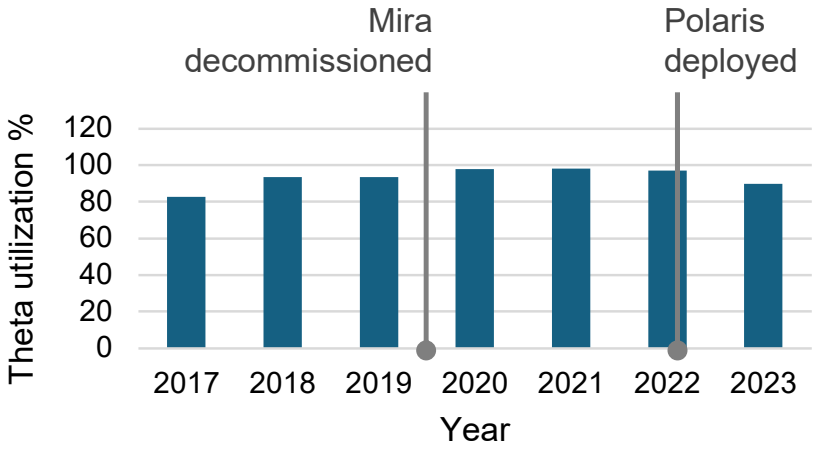
CHRISTOPHER D. CAROTHERS
Rensselaer Polytechnic Institute

U.S. DEPARTMENT of ENERGY
Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC.

Argonne NATIONAL LABORATORY

THETA SUPERCOMPUTER

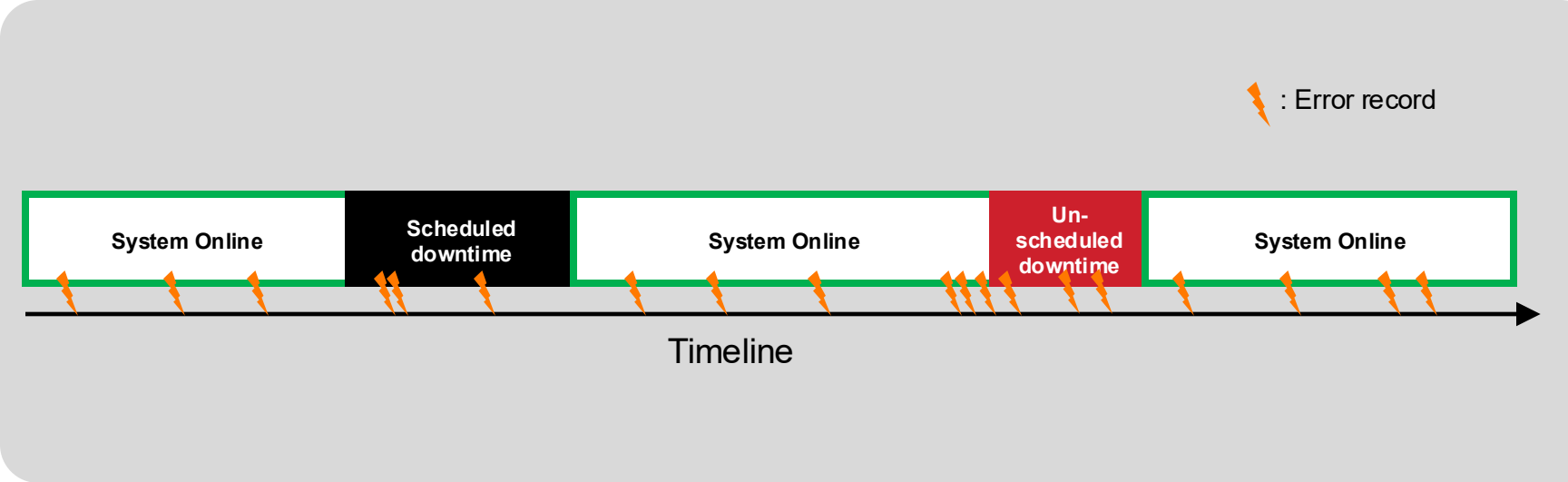
We investigate the failure patterns during normal system usage



Theta was a 4,392 Cray XC40 System deployed at the Argonne Leadership Computing Facility (ALCF) and heavily used throughout its lifetime

SYSTEM LIFETIME AND ERRORS

Errors are generated when the system is online and offline to users



DATASETS AND PRE-PROCESSING

DATASETS

We need to consider state of the system when errors are reported

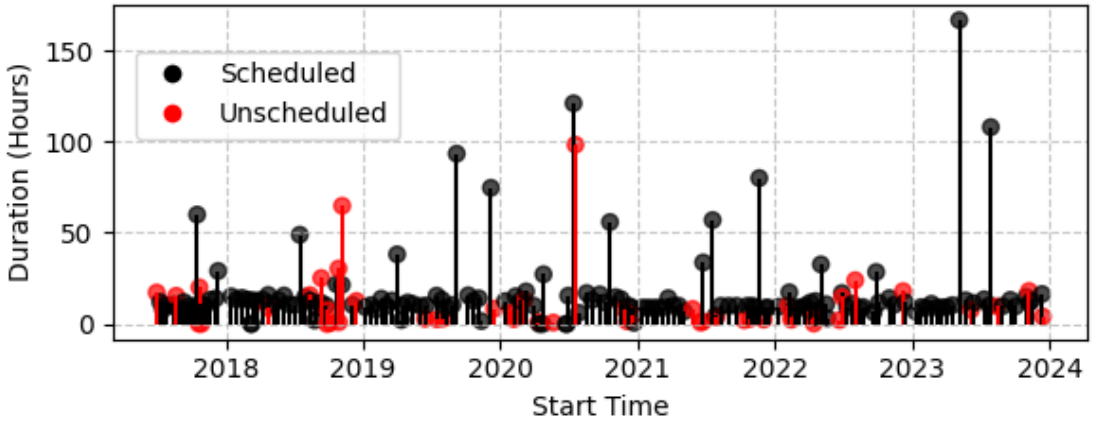
System Downtime Logs

- Logs the time periods when the system was down and unavailable for normal use.
  - Maintained by the systems administrators.
- Records indicate if a downtime was scheduled or unscheduled.
- Some records include an explanation of the cause of the downtime.

Hardware Error Logs

- Combines hardware errors logs generated from different sources.
  - Most logs are from **Cray** and **Intel** hardware monitoring mechanism.
- Records are timestamped and component-specific (NIC etc.).
- Contains duplicates – multiple records correspond to the same error.

THETA DOWNTIMES



Total Online Duration:	2231.04 Days
Total Downtime Duration:	125.69 Days
Total Online:	94.67%
Scheduled Downtime:	4.41%
Unscheduled Downtime:	0.92%

THETA HARDWARE ERRORS

- Large size of the error logs: approximately 430 GB
- High level of redundancy in error records
  - — Same fault has multiple entries with the same timestamp
  - — Same fault has multiple entries with different timestamps

FAILED_COMPONENT	COMPONENT_NAME	COMPONENT_TYPE	COMPONENT_TYPE_STRING	ERROR_MAP_JSON	INFO_MMR_JSON
50683276972656272	c11-1c0s11a0n2	72	rt_aries_nic	{}	[8, 0, 0, 0, 0, 0, 0, 0]
68697503683315340	c1-1c0s15a0	70	rt_aries	{"NL_ERR_INFO_RSP_P2F": {"MBE_SYNDROME": 0, "M...	[35184372089856, 25615097, 0, 0, 0, 0, 0, 0]
10696186697090816	c8-0c2s2n0	0	rt_node	{"ADDRV": 1, "VAL": 1, "BANK_DESCRIPTION": "In...	[22631252129480704, 360287978796351488, 951160...
10696186697090816	c8-0c2s2n0	0	rt_node	{"ADDRV": 1, "VAL": 1, "BANK_DESCRIPTION": "In...	[22631252129480704, 360287983074541568, 951160...

HARDWARE ERROR REDUNDANCY

1 Identical error records with the **same timestamp** – Same error reported by simultaneously by multiple threads/processes.

	ERROR_CODE	LOG_TIMESTAMP	COMPONENT_NAME
2	64784	2020-01-01 00:00:29	c10-1c2s0n3
3	64784	2020-01-01 00:00:29	c10-1c2s0n3
4	64784	2020-01-01 00:00:29	c10-1c2s0n3
5	64784	2020-01-01 00:00:29	c10-1c2s0n3
6	64784	2020-01-01 00:00:29	c10-1c2s0n3
7	64784	2020-01-01 00:00:29	c10-1c2s0n3
8	64784	2020-01-01 00:00:29	c10-1c2s0n3
9	64784	2020-01-01 00:00:29	c10-1c2s0n3
10	64784	2020-01-01 00:00:29	c10-1c2s0n3
11	64784	2020-01-01 00:00:29	c10-1c2s0n3
12	64784	2020-01-01 00:00:29	c10-1c2s0n3
13	64784	2020-01-01 00:00:29	c10-1c2s0n3
14	64784	2020-01-01 00:00:29	c10-1c2s0n3
15	64784	2020-01-01 00:00:29	c10-1c2s0n3
16	64784	2020-01-01 00:00:29	c10-1c2s0n3
17	64784	2020-01-01 00:00:29	c10-1c2s0n3
18	64784	2020-01-01 00:00:29	c10-1c2s0n3
19	64784	2020-01-01 00:00:29	c10-1c2s0n3
20	64784	2020-01-01 00:00:29	c10-1c2s0n3
21	64784	2020-01-01 00:00:29	c10-1c2s0n3

2 Identical error records with **sequential timestamps** Continuous monitoring of component component in a failed/error state.

	ERROR_CODE	LOG_TIMESTAMP	COMPONENT_NAME
0	23299	2020-01-01 00:00:15	c5-0c1s12a0n0
1	23299	2020-01-01 00:00:25	c5-0c1s12a0n0
166	23299	2020-01-01 00:01:15	c5-0c1s12a0n0
167	23299	2020-01-01 00:01:25	c5-0c1s12a0n0
168	23299	2020-01-01 00:01:35	c5-0c1s12a0n0
174	23299	2020-01-01 00:01:45	c5-0c1s12a0n0
208	23299	2020-01-01 00:02:05	c5-0c1s12a0n0
209	23299	2020-01-01 00:02:15	c5-0c1s12a0n0
210	23299	2020-01-01 00:02:25	c5-0c1s12a0n0
211	23299	2020-01-01 00:02:35	c5-0c1s12a0n0
254	23299	2020-01-01 00:02:55	c5-0c1s12a0n0
255	23299	2020-01-01 00:03:15	c5-0c1s12a0n0
256	23299	2020-01-01 00:03:25	c5-0c1s12a0n0
304	23299	2020-01-01 00:04:05	c5-0c1s12a0n0
305	23299	2020-01-01 00:04:15	c5-0c1s12a0n0
306	23299	2020-01-01 00:04:25	c5-0c1s12a0n0
307	23299	2020-01-01 00:04:35	c5-0c1s12a0n0
311	23299	2020-01-01 00:04:45	c5-0c1s12a0n0
343	23299	2020-01-01 00:05:15	c5-0c1s12a0n0
344	23299	2020-01-01 00:05:35	c5-0c1s12a0n0

HARDWARE ERROR CATEGORIES

Different error classifications from the different vendors

Five categories from Cray

- **GHAL_ARIES_CRITICAL_ERRORS**
- **GHAL_ARIES_TRANSACTION_ERRORS**
- **GHAL_ARIES_TRANSIENT_ERRORS**
- **GHAL_ARIES_CORRECTABLE_MEMORY_ERRORS**
- **GHAL_ARIES_INFORMATIONAL**

One category from Intel

- **MCE_ERROR**
  - Five classification of MCE errors
    - UC: *Uncorrected error*
    - SRAR: *Software recoverable action required*
    - SROR: *Software recoverable action optional*
    - UCNA: *Uncorrected no action required*
    - CE: *Corrected Error*

Systemwide analysis requires a uniform classification

UNIFORM ERROR CLASSIFICATION

Guided by vendor manuals

CRITICAL: Requires hardware reset/replacement
MAJOR: Require application/service restart/reconfiguration

INTERMEDIATE: Require operation to be retried
MINOR: Automatically corrected errors

Error Classification	ERROR_CATEGORY_STRING	ERROR_MAP_JSON: UC	ERROR_MAP_JSON: PCC	ERROR_MAP_JSON: VAL
Critical	GHAL_ARIES_CRITICAL_ERRORS			
	MCE_ERROR	1	1	1
Major	GHAL_ARIES_TRANSACTION_ERRORS			
	MCE_ERROR	1	0	1
Intermediate	GHAL_ARIES_TRANSIENT_ERRORS			
Minor	GHAL_ARIES_CORRECTABLE_MEMORY_ERRORS			
	MCE_ERROR	0		1
No-considered	GHAL_ARIES_INFORMATIONAL			
	MCE_ERROR			0

HARDWARE ERROR PRE-PROCESSING

Filtering and labeling

Redundancy Removal

- Raw dataset:
  - 479 million records
  - 320 GB
- Unique error dataset:
  - 7.92 million records
  - 5 GB

Pruning Non-hardware and Downtime Errors

- Non-hardware (environmental) errors:
  - 59 records
- Errors recorded during downtimes:
  - 457,500 (5.78% of unique records)

Uniform Error Classification (Labeling)

- Uniform classification for labeling all Intel and Cray errors:
  - **Critical**
  - **Major**
  - **Intermediate**
  - **Minor**
  - Informational/Debug
    - 1275 records

RESULTS AND ANALYSIS

FOCUS OF OUR ANALYSIS

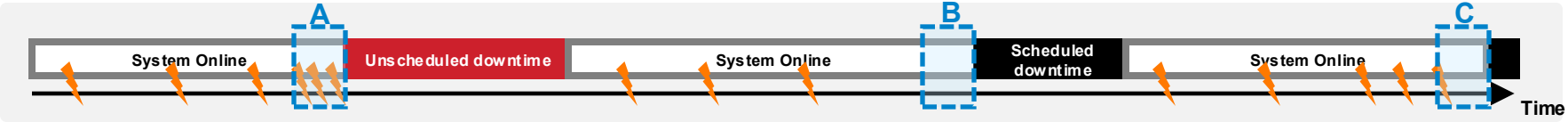
Demonstrate the utility of our data processing approach

Distribution of errors

- How do errors vary across time?
  - Daily, monthly, yearly
- How do errors vary across component?
  - Nodes, NICs, and switches

Correlation with system downtimes

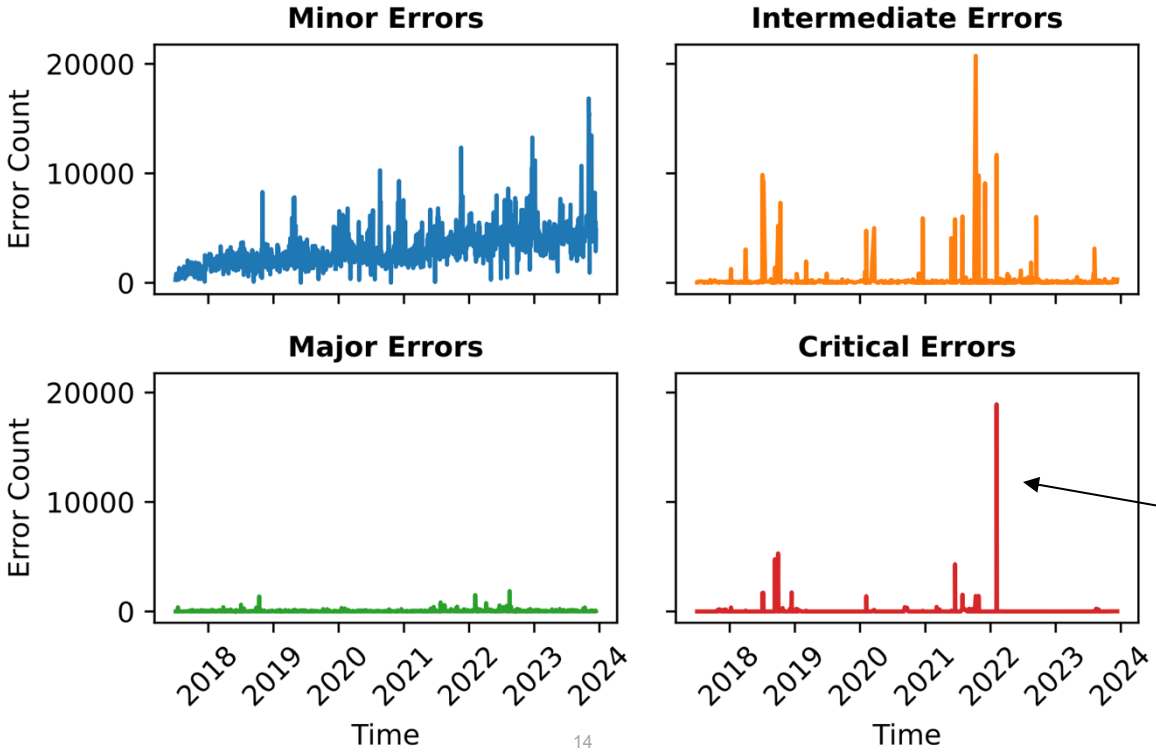
- How do error patterns change with schedule and unscheduled downtimes?
- What error messages are correlated with unscheduled downtimes?



DAILY ERRORS

Different class of errors show different trends

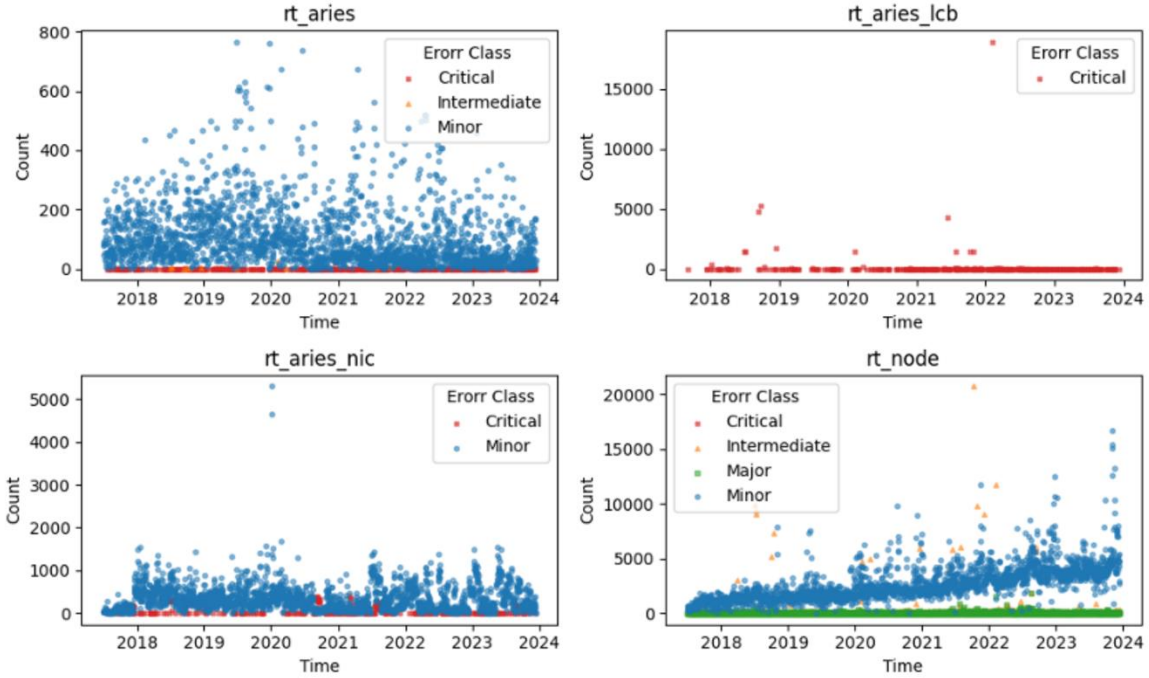
Minor errors (soft memory errors etc.) increase as the system ages



February 2022 multiday filesystem failures

DAILY ERRORS – BY COMPONENT TYPE

Some components drive the trend for specific error classes

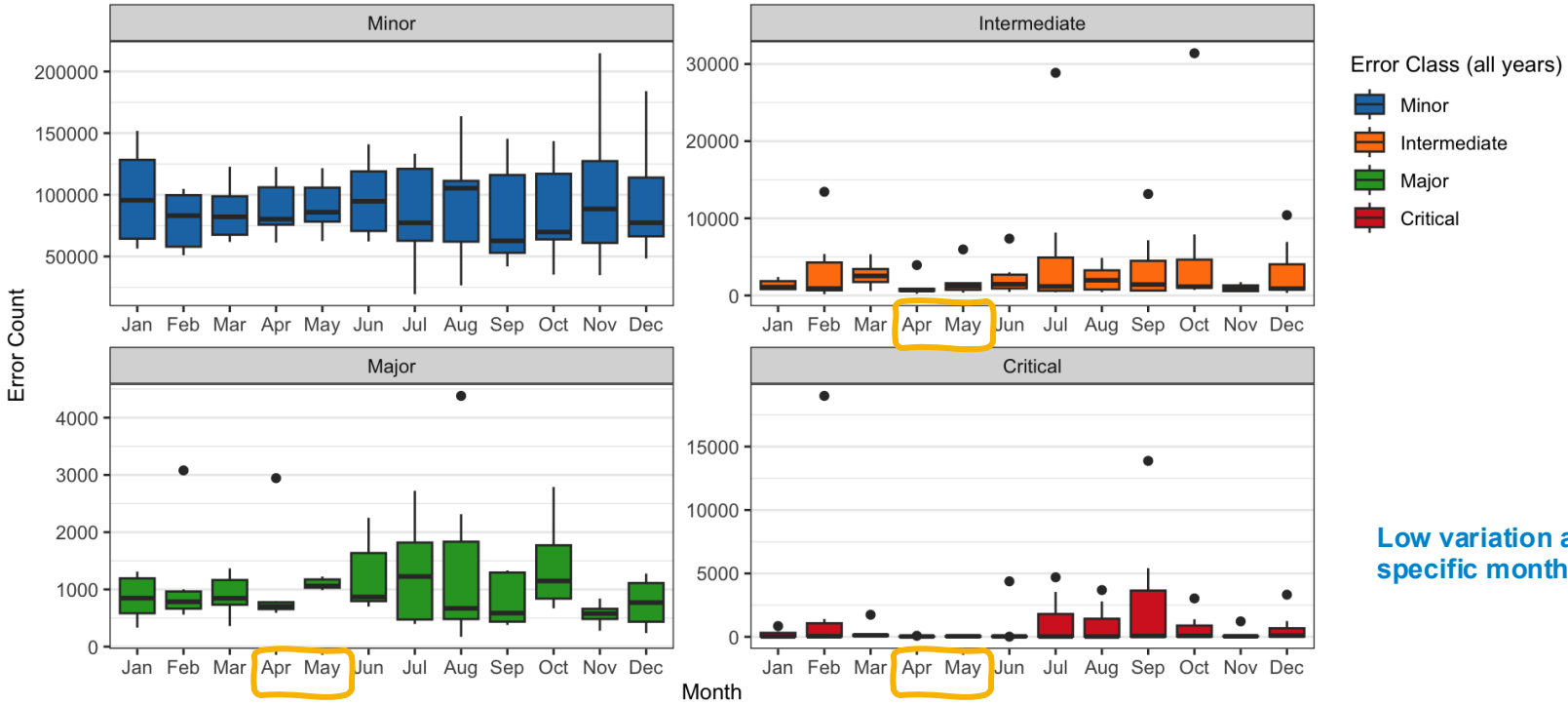


- LCBs report only critical errors and the most of all components

- Major errors are only reported on nodes
- Nodes drive minor & intermediate error trends

MONTHLY ERRORS

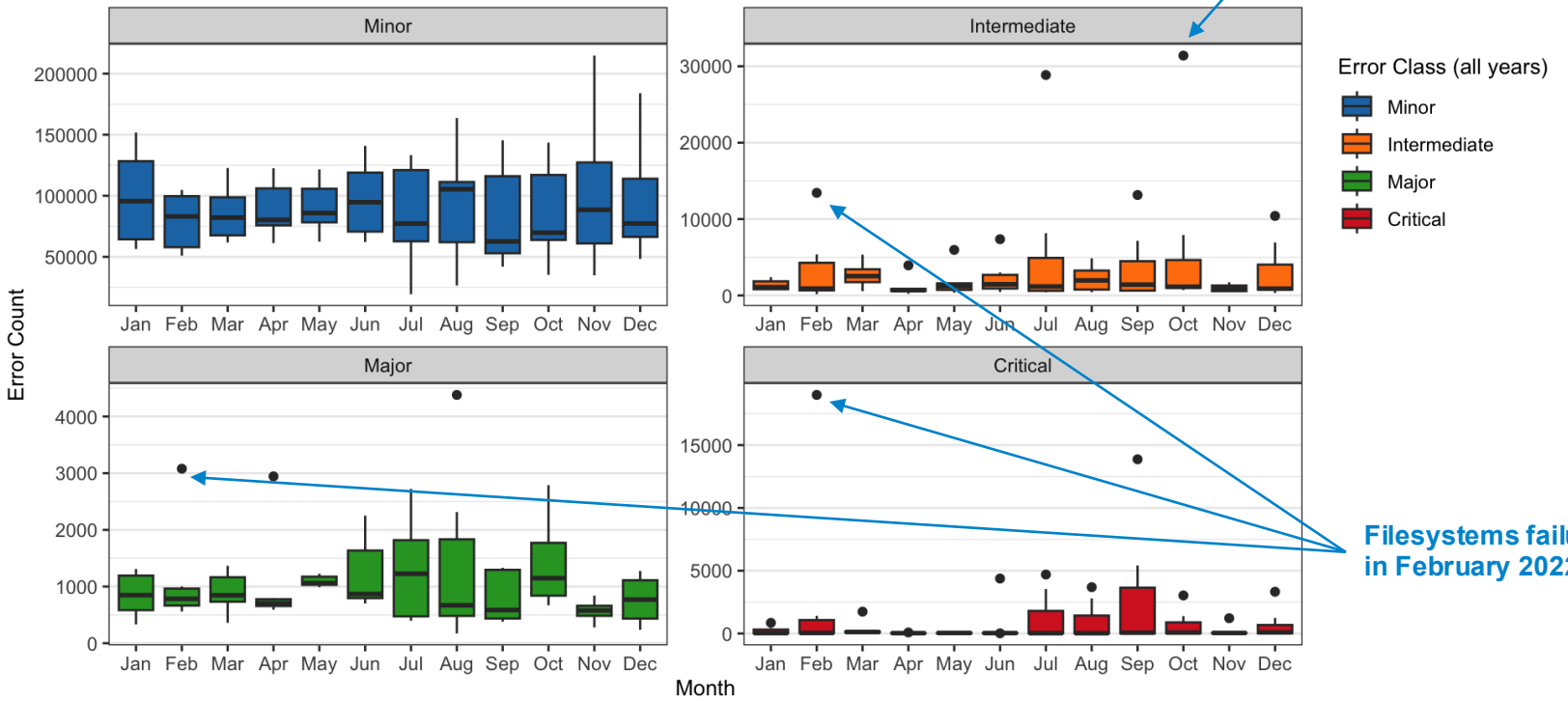
Observable monthly trends across error



Low variation across specific months

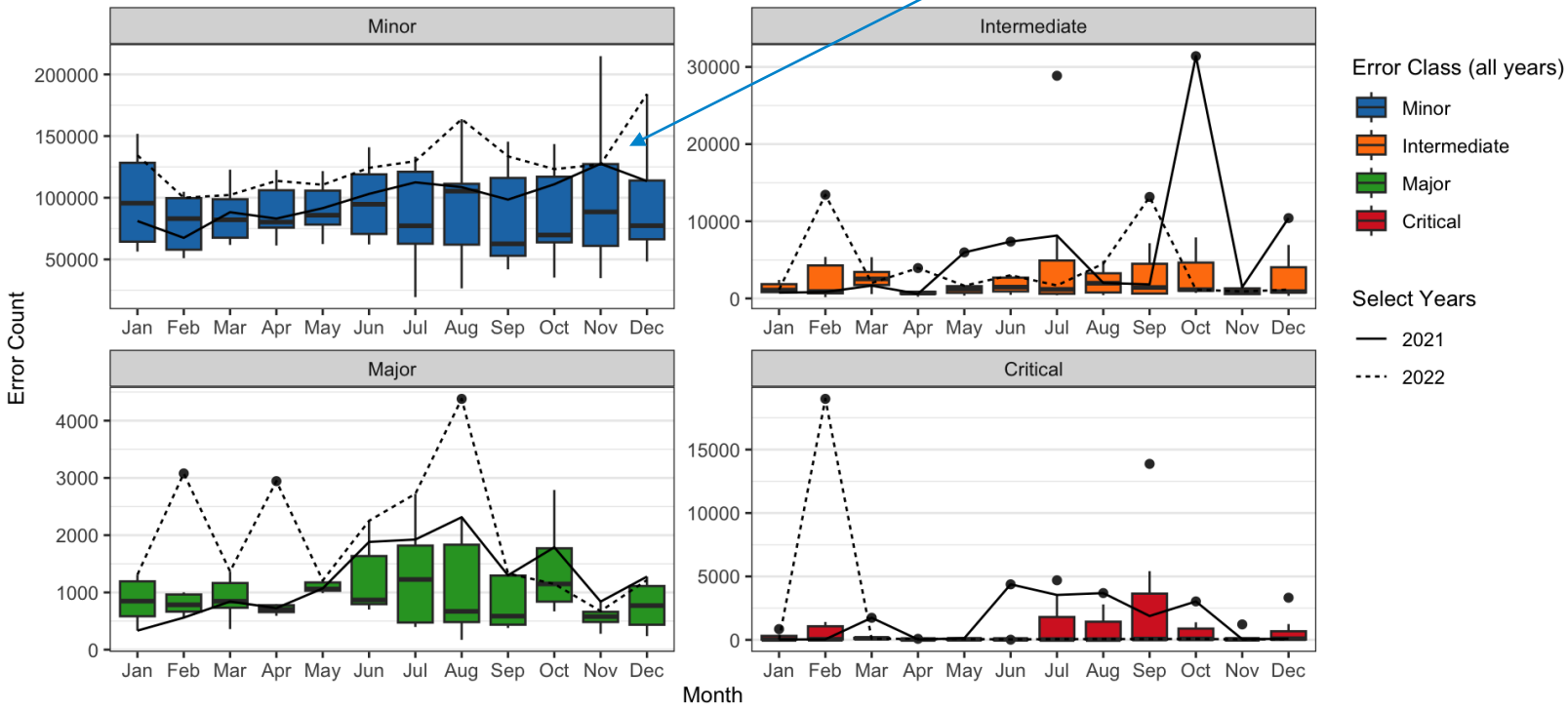
MONTHLY ERRORS

Some spikes attributed to major system outages



MONTHLY ERRORS

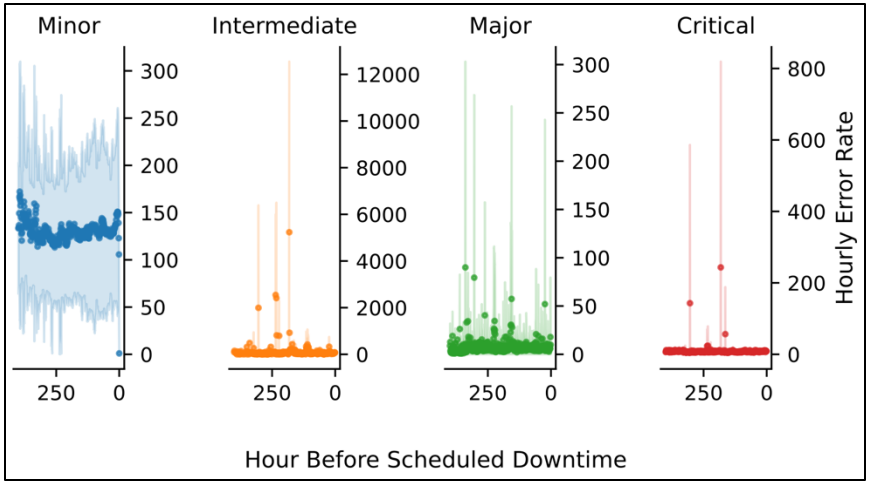
Some trends may vary across years



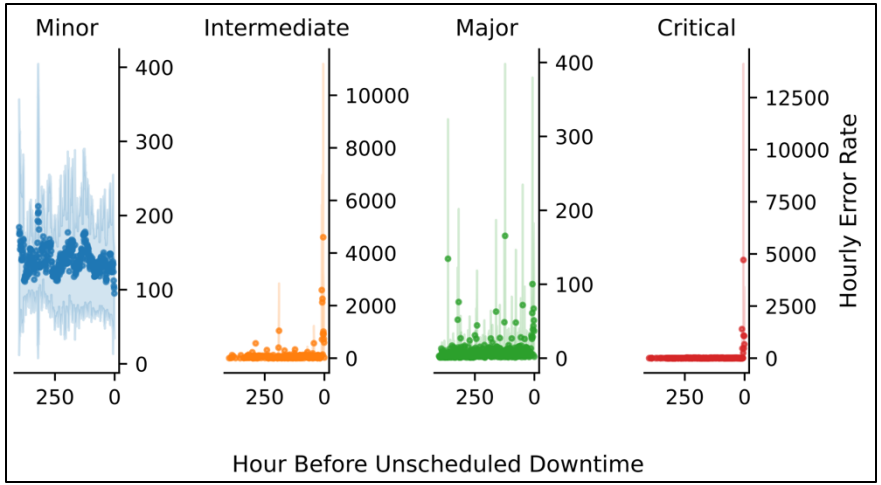
ERRORS PRECEDING DOWNTIMES

More critical, major, & intermediate errors occur a few hours before unscheduled downtimes compared to scheduled downtimes

Before Scheduled Downtimes

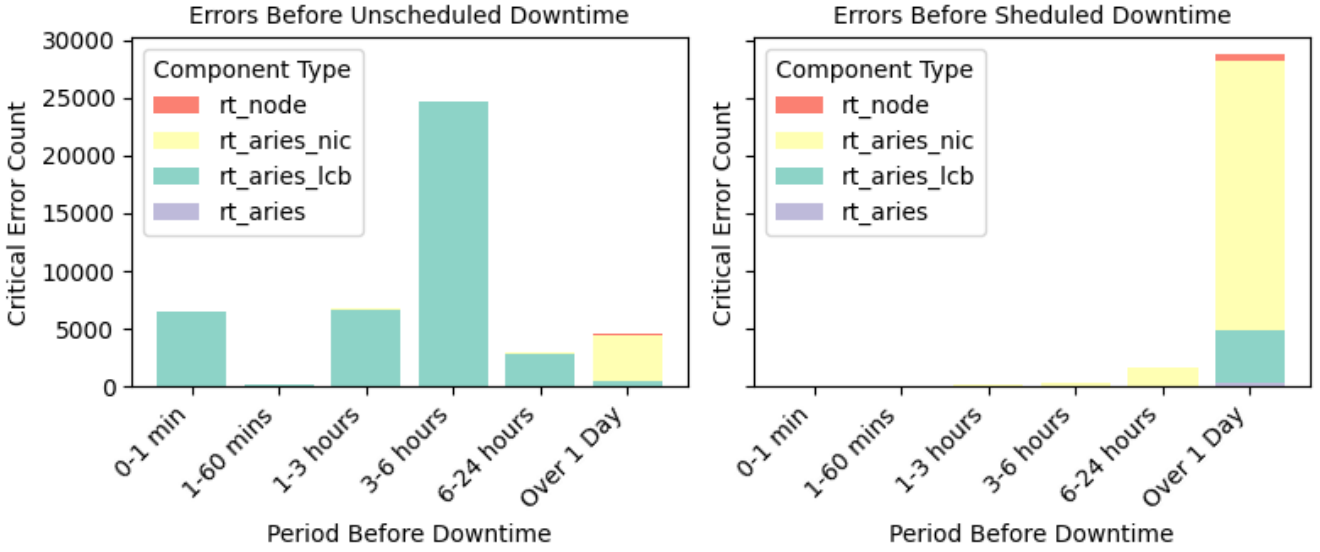


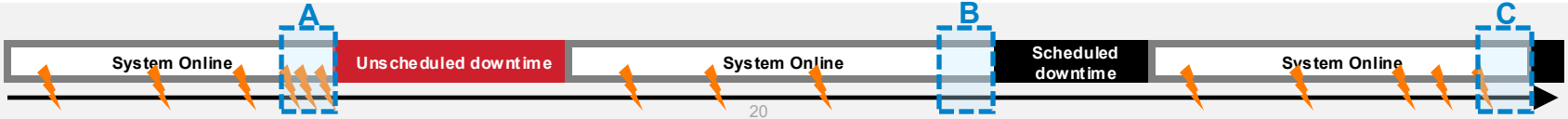
Before Unscheduled Downtimes



ERRORS PRECEDING DOWNTIMES

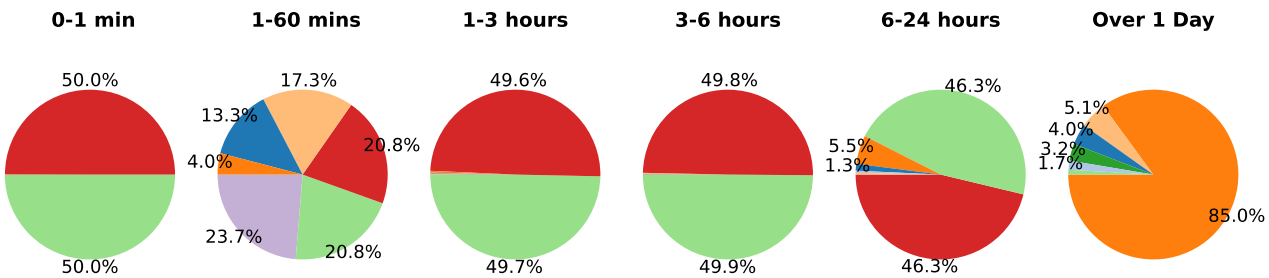
Count of error by component type for each period before downtimes





ERRORS PRECEDING DOWNTIMES

Top error messages reported during each period before downtimes

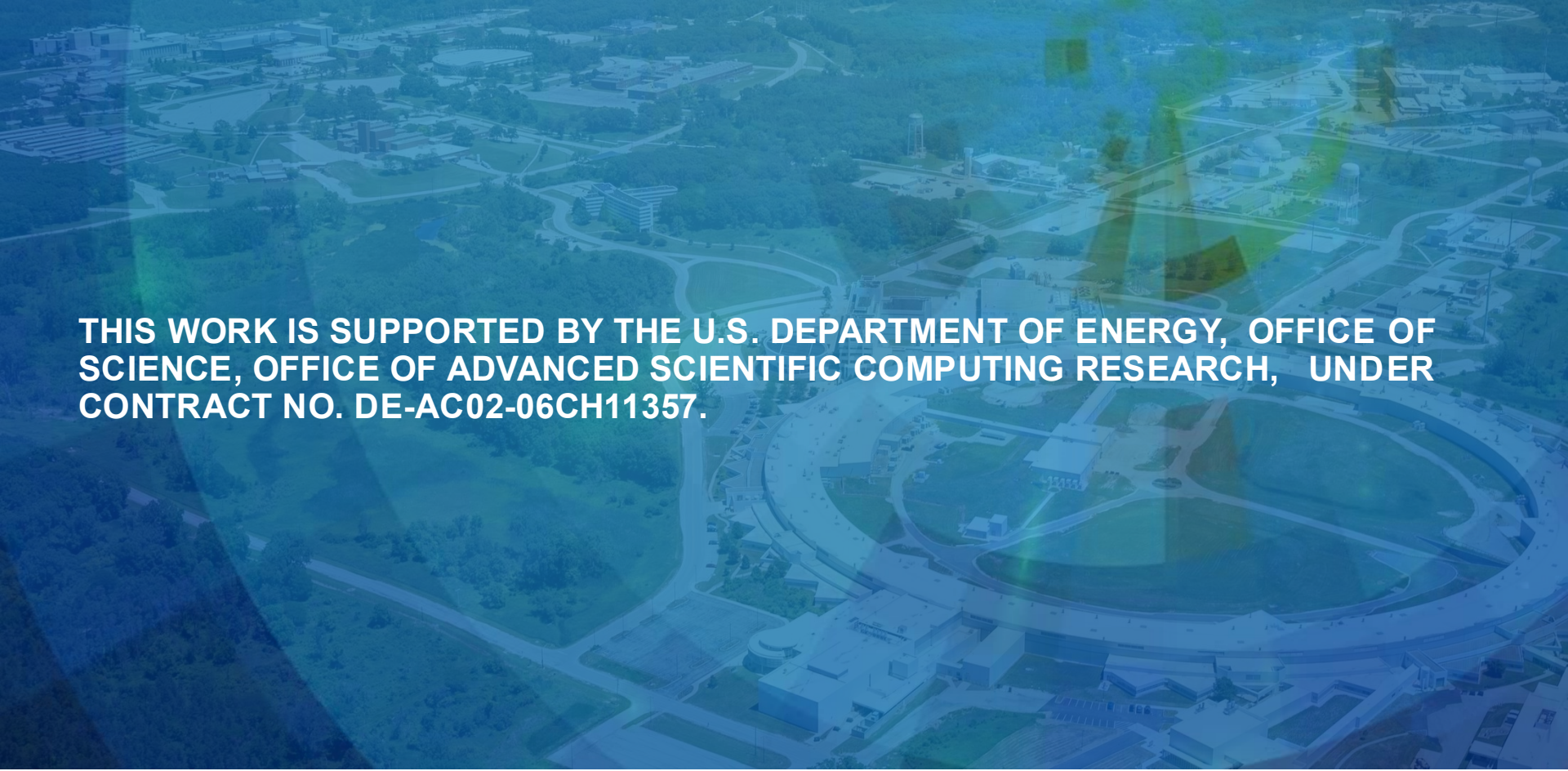


Error Messages

- Others
- BUS0 MCA Error
- Current state of the LTSSM
- Aries LCB operating badly and will be shut down
- Denali Core Fatal Error
- Will set with PAUSE_ON_REINIT and will clear if the pause is released
- The link went down
- This flag sets when the column buffer encounters a packet longer then max packet length Error
- The SSID received an unexpected response
- Set when a response packet comes into the Response FIFO for which there is no corresponding request in the ORD RAM

TAKEAWAYS

- Our data pre-processing approach is effective at reducing redundancy without loss of integrity
  - Reduces storage requirements by up to **60X**
  - Retains unique records & allows effective spatial and temporal analysis
- System failure analysis should incorporate system state information
  - Highlights the impact of regular utilization on error trends
  - Effectively exposes errors related to downtimes
- **Opportunities and next steps**
  - Improve facility's data collection pipelines with online redundancy removal
  - Utilize facility and operational information to develop novel hardware failure models



THIS WORK IS SUPPORTED BY THE U.S. DEPARTMENT OF ENERGY, OFFICE OF SCIENCE, OFFICE OF ADVANCED SCIENTIFIC COMPUTING RESEARCH, UNDER CONTRACT NO. DE-AC02-06CH11357.


Argonne
NATIONAL LABORATORY
U.S. DEPARTMENT of ENERGY