OAK RIDGE National Laboratory

Understanding the Interplay between Hardware Errors and User Job Characteristics on the Titan Supercomputer

Seung-Hwan Lim (lims1@ornl.gov),
Ross G. Miller,
Sudharshan S. Vazhkudai

Oak Ridge National Laboratory

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

U.S. DEPARTMENT OF ENERGY

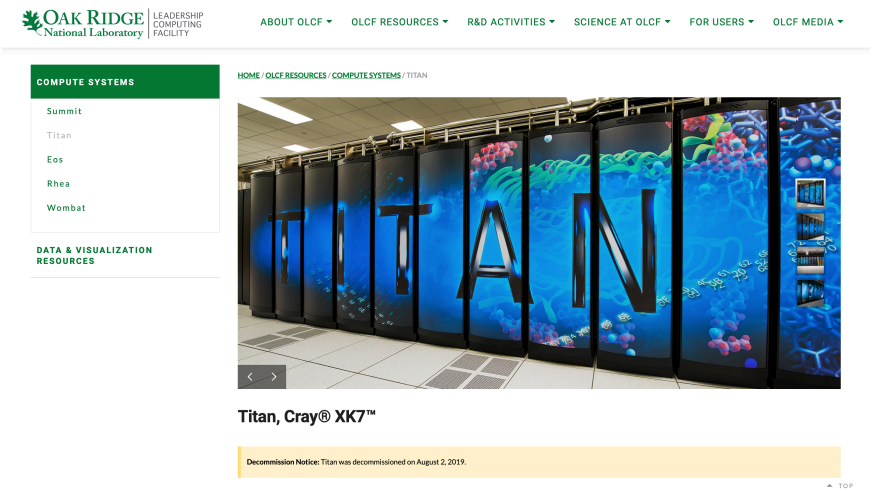
Outline

- Introduction :
  - The understanding of errors is the foundation of designing dependable systems
- System overview: Titan supercomputer at ORNL
- Log collection
- Overview of collected errors
- Data analysis
  - Error characteristics
  - Node characteristics
  - Job characteristics
  - User characteristics
- Conclusions

The understanding of errors is the foundation of designing dependable systems

- An error is the discrepancy between the intended behavior of a system and its actual behavior at runtime.
- Error may not cause faults, if handled by fault tolerance mechanisms.
- Errors happen at runtime; thus factors are the software/hardware system fault, the system operation (e.g. job scheduler), and user's usage (e.g. application)
- Given hardware and software components, the understanding of error patterns under **various** applications and users are of historical importance.

System overview: Titan supercomputer at ORNL



The best environments to record statistics of errors from various applications and users at runtime.

- Titan
  - Operation began in 2012
  - Top1 in Top500 list in 2012
  - Top12 in Top500 list in June 2019 just two months before the commission (Aug, 2019)
  - Peak performance: 17.59PF
  - The first large scale GPU cluster, 18K NVIDIA K20 GPUs on 18K nodes.
  - **Diverse applications** from 30 science domains, including traditional simulations and complex data analysis tasks like deep learning.

Log collection mechanism

OPTERON PROCESSOR AMD

Machine Check Exception Log

Job submission

Job submission Log

splunk>

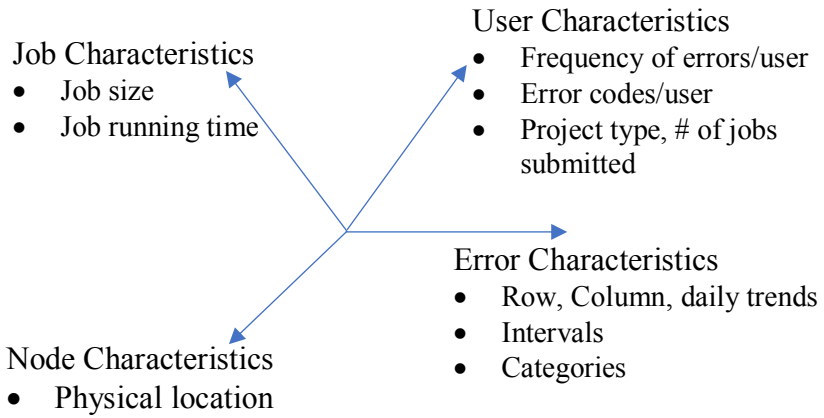
- Machine Check Exception is supported by hardware (Machine Check Architecture), which reports hardware and processor errors to system software
- Thus, our log collection obtains error logs and job submission logs with minimal interruption of the normal job execution.

Overview of collected logs

Category		Description
Hardware error log (MCE log)	Corrected Error	Corrected by hardware. No software action is required
	Uncorrected Error	Correction by hardware is not possible. Software action is required.
	Deferred Error	Hardware correction is not possible, the impact can be contained. Do not require software action.
Job scheduler log	Job end/start	Start, end time, and user id (UID)

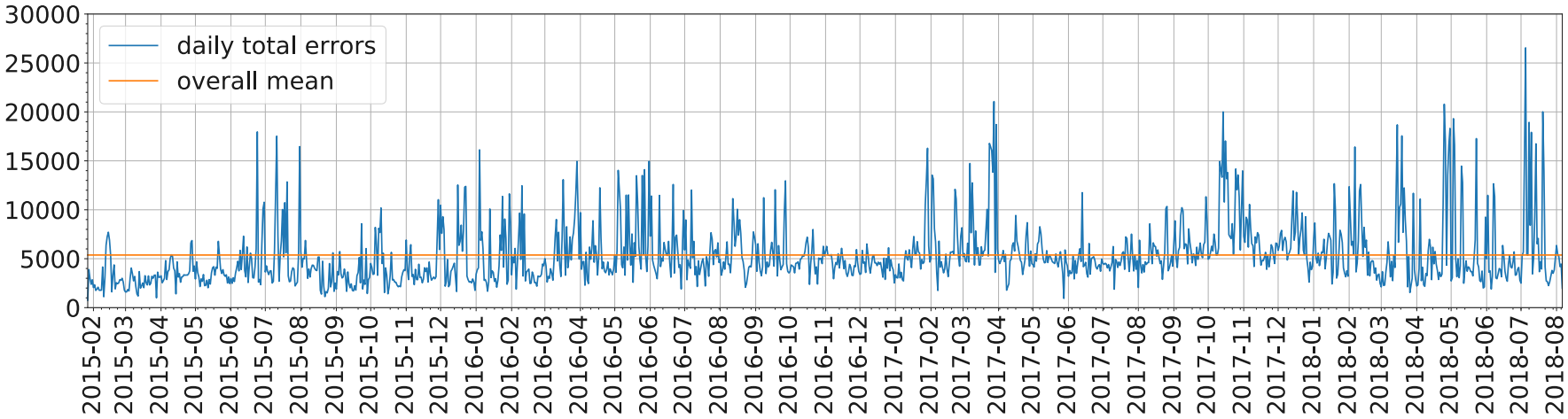
Job monitoring period	Feb,9, 2018 to Aug 6, 2018
Error monitoring period	Jan 27, 2015 to Aug 6, 2018
Job entries	312,215
Error entries	6,908,297
Jobs with errors	24,337
Errors during the job monitoring period	1,096,666
# of users	342

Overview of analysis



- What kinds of errors happen?
- Which factors are the most influential to the error trend?
  - Hardware
    - Component failure
    - Location
  - Job
    - Allocation size
    - Running time
  - User
    - Application (and parameters)

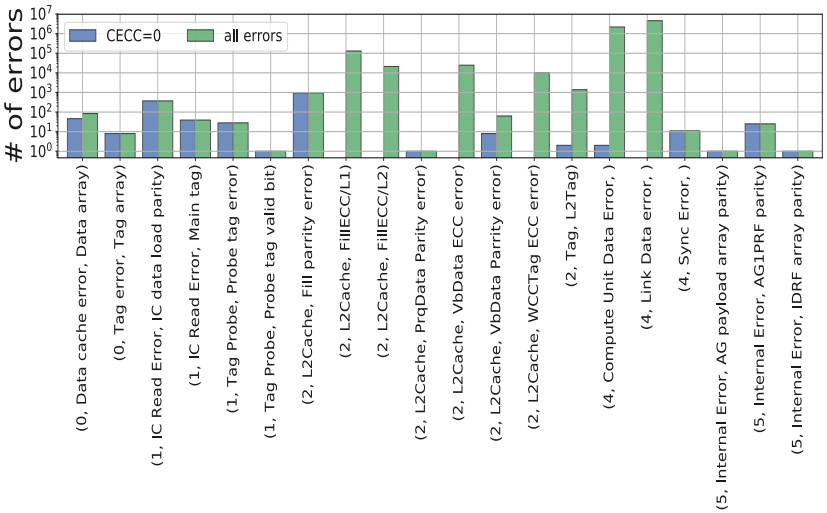
Error characteristics: (1) daily trend



- The average number of errors in a day: 5,376 errors
- Max number of errors in a day: 26,550 errors.

Error characteristics: (2) error categories

Register Bank	Error type	Error subtype	Description	Counts
Bank0	Data cache error	Data array	load-store unit (LS), including data	84
	Tag error	Tag array	cache	8
Bank1	IC read error	IC data load parity	instruction fetch unit (IF), including instruction cache	370
	IC read error	Main tag		39
	Tag probe	Probe tag error		28
	Tag probe	Probe tag valid bit		1
Bank2	L2 Cache	Fill parity error	combined unit (CU), including L2 cache	951
	L2 Cache	FillECC/L1		130,084
	L2 Cache	FillECC/L2		21,130
	L2 Cache	PrqData parity error		1
	L2 Cache	VbData ECC error		24,545
	L2 Cache	VbData Parity error		63
	L2 Cache	WCCTag ECC error		10,165
	Tag	L2Tag		1379
Bank4	Compute unit data error	-	northbridge (NB)	2,171,095
	Link data error	-		4,548,314
	Sync error	-		11
Bank5	Internal error	AG payload array parity	execution unit (EX)	1
	Internal Error	AG1PRF parity		25
	Internal Error	IDRF array parity		1



- All errors fall into 29 different error codes.
- Majority of errors are related to Bank 4 (NorthBridge)
  - DRAM access across cores
  - I/O access
  - Most of these errors are also ECC errors
- ECC errors in L2 Cache are the third common errors.

Error characteristics: (3) critical errors

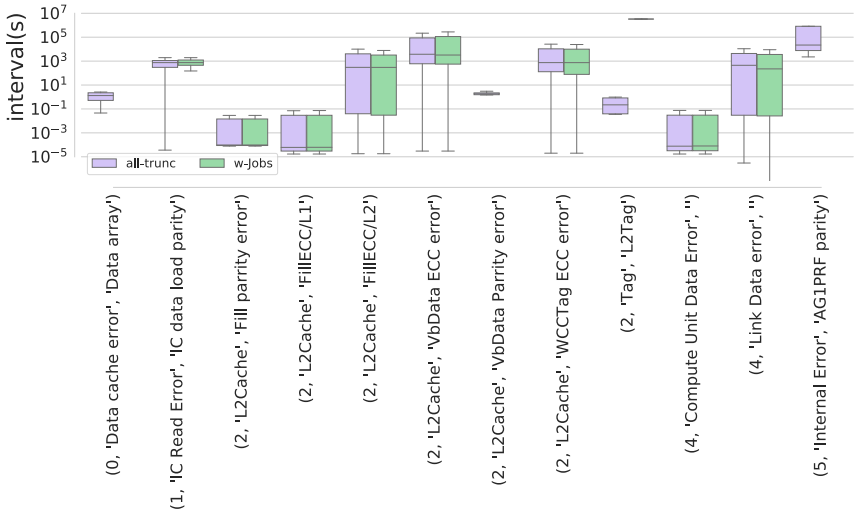
- Most errors are related to ECC
  - Correctable ECC errors (99.97%)
  - Uncorrectable ECC errors
- Hardware always does not request software (processor) to correct the error
  - Deferred errors
    - If double errors happen in the memory location where the current process may not access.
- Program context corrupt
  - Kernel panic, system shutdown

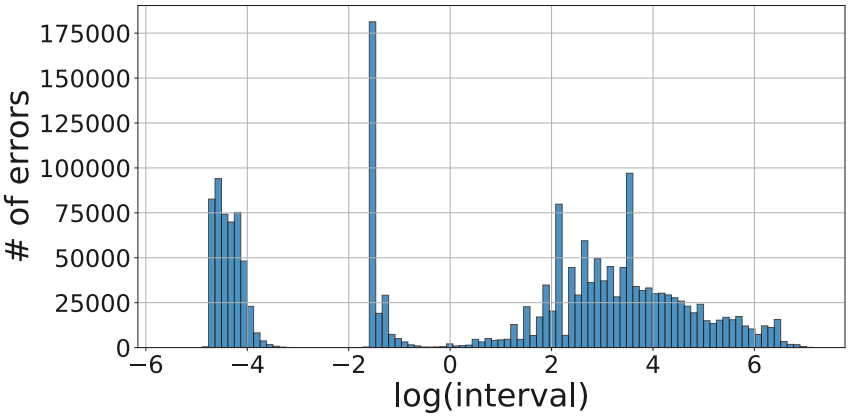
CECC	UECC	PCC	UC	counts
0	0	0	0	1,464
			1	7
		1	1	19
	1	1	1	2
1	0	0	0	6,906,805

CECC=1: Correcctable ECC errors.
UECC=1: Uncorrectable ECC error, a deferred error if converted to poison data.
UC=1: actually corrected by processor
PCC=1: processor context corrupt, leading to node shutdown.

ECC protection must be considered at every level of system components

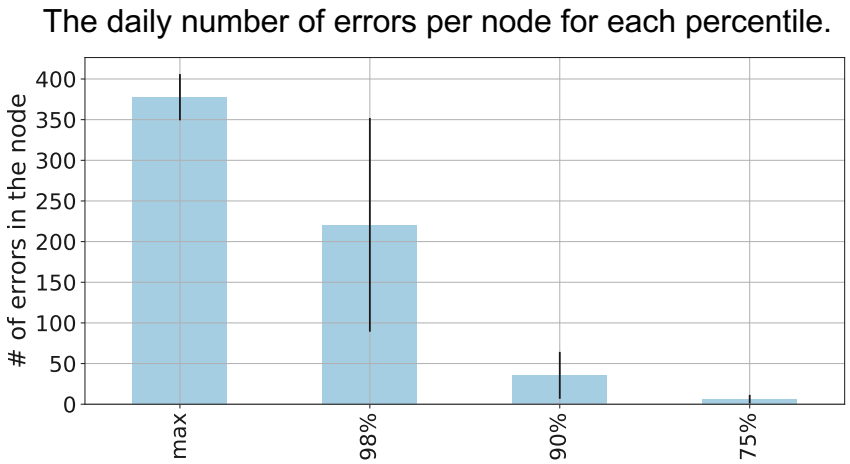
Error characteristics: (4) error intervals

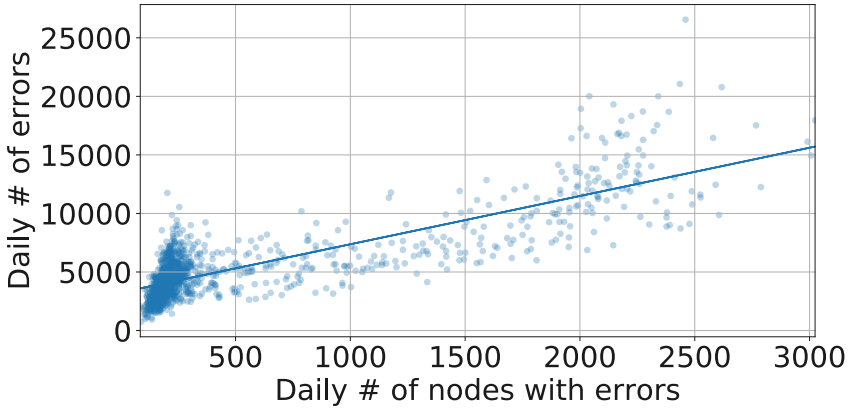




- Most of errors happen within the range of a job
  - Max job execution time is 24hr ($8*10^5$ sec)
- Each error have different range of intervals.
  - Bank 4 (NorthBridge) errors that consist the majority of errors have two distinctive error intervals.
- Interval of errors show a mixture of three different distributions according to histogram.

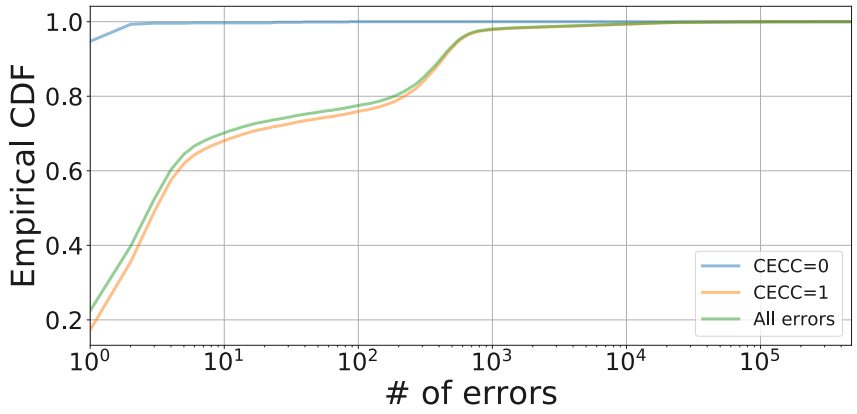
Node characteristics: (1) daily trend





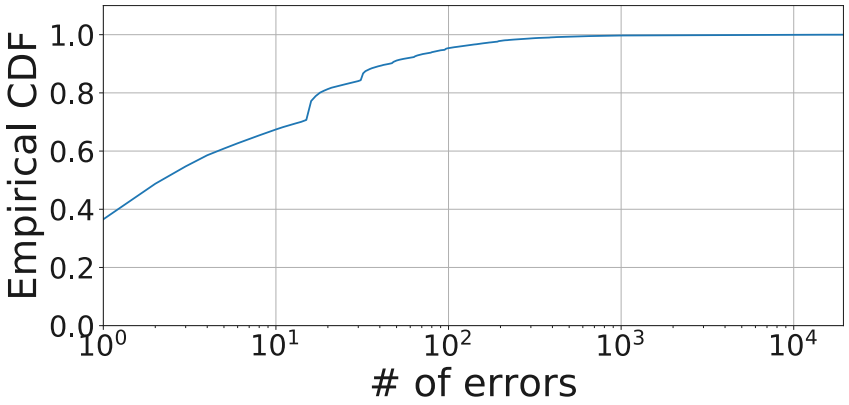
- We find a small set of nodes have more errors than others in a day.
  - The worst node of the day typically records around 370 errors.
  - 75% of nodes records less than 6 errors in a day.
- For each day, daily # of errors has a positive linear relationship ($R^2 = 0.68$) with daily # of nodes with error.

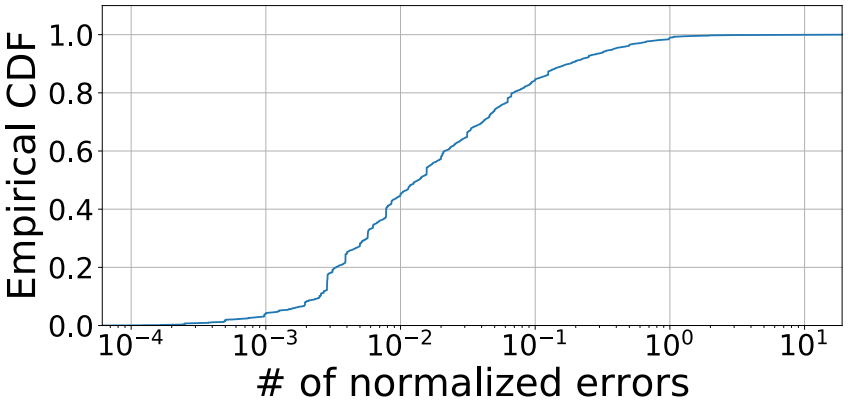
Node characteristics: (2) cumulative distribution



- A small set of nodes have generated errors
  - 23.9% of nodes have more than 100 errors during the monitoring period, accounting 98% of the entire errors.
  - 2% of nodes (373 nodes) have more than 1,000 errors, accounting 77.5% of errors.
  - However, no indication that those error prone nodes have more critical non-correctable ECC errors.

Job characteristics: (1) Job vs. Errors





- s

Job characteristics: (2) Job vs. Node

User characteristics: (1) user vs. error

User characteristics: (2) user vs. job

Conclusions

- This work provides statistical information on the first generation of large-scale CPU-GPU heterogeneous environments, which becomes extremely popular in both simulation and machine learning.
- Summary of findings:
  - Errors are highly concentrated in a small number of nodes, jobs, and users
  - User behavior is the most dominant factor, rather than the impacts from user-independent job characteristics or specific hardware component defects.
  - Current proactive and agile system operation practices at HPC centers may have contributed to minimize the impact from hardware faults.
    - such as a rapid identification and replacement of defective hardware components (e.g., DIMM modules and GPUs.)
- Full slidedeck: