Failures in Large Scale Systems: Long-term Measurement, Analysis, and Implications

Saurabh Gupta
Intel Labs

Tirthak Patel
Northeastern University

Christian Engelmann
Oak Ridge Nat'l Lab

Devesh Tiwari
Northeastern University

Oak Ridge National Laboratory
Northeastern University
U.S. Department of Energy | Office of Science

Large-scale scientific applications are going to face severe resilience challenges at exascale!

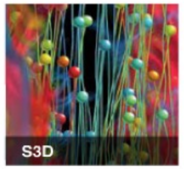
- "Top Ten Exascale Research Challenges", DOE ASCAC Subcommittee Report, Feb. 2014

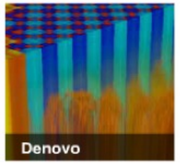
Long-running, large-scale scientific applications are interrupted by failures on HPC systems.

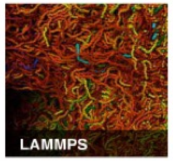
At exascale, an application is expected to be interrupted every couple of hours.

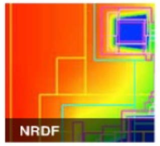
Why investigate the reliability characteristics of large-scale systems?

Reduce Checkpoint I/O Overhead on Large-scale Systems









Astrophysics, climate modeling, combustion and fusion applications periodically write checkpoints to permanent storage system, and recover from the last checkpoint in case of a failure.

Compute System

Excessive I/O overheads due to checkpoints

Domain	Application	Checkpoint data size
Astrophysics	CHIMERA	160 TB
Astrophysics	VULCUN/2D	0.83 GB
Climate	POP	26 GB
Combustion	S3D	5 TB
Fusion	GTC	20 TB
Fusion	GYRO	50 GB

Permanent Storage System

At exascale, applications may spend up to 60% of execution time in checkpointing and lost work!

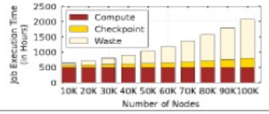


Figure 1. Impact of checkpoint/restart mechanism on a large-scale application checkpoint taken every hour (top), and every four hours (bottom). Checkpoint and restart time 30 mins and 15 mins, respectively. MTBF of each node is taken as 25 years and scaled according to the system size. Large-scale scientific applications are weak-scaling, i.e., the compute time per node remains a constant.

Expedite Scientific Discovery

Save Energy – A Positive Impact Beyond the Computing Facility

1 hour of lost work on the Titan supercomputer is roughly 5-9 MWhr

Systems: 5 Supercomputer Generations at ORNL

System	Number of Nodes	Period
Jaguar XT4 (31328 cores, quad-core AMD Opteron processor per node, SeaStar2)	7,832	Jan'08-Mar'11
Jaguar XT5 (149504 cores, four dual-core AMD Opteron processor per node, SeaStar2+)	18,688	Jan'09-Dec'11
Jaguar XK6 (298,592 cores, 16-core Opteron-6274 processor per node, Gemini)	18,688	Jan'12-Oct'12
Eos XC 30 (23,553 cores, 2 sockets of 16-core Intel Xeon E5-2670 (with hyperthreading) per node, Aries)	736	Sep'13-Sep'15
Titan XK7 (560,640 cores, one 16-core Opteron-6274 and one K20x Nvidia GPU per node, Gemini)	18,688	May'13-Sep'15

Failures in Over 1 Billion Compute Node Hours

Failure Event	Type	Component Affected
Bad Page State	Software	–
Blade Heartbeat Fault	Hardware	Module
Core Hang	Hardware	Node/CPU
GPU Double Bit Error (DBE)	Hardware	GPU
HT Lockup	Hardware	CPU
Kernel Panic	Software	OS
L0 Heartbeat Fault	Hardware	Module
Lustre Bug (LBUG)	Software	File System
Lustre Server Failure	Software	File System
Machine Check Exception (MCE)	Hardware	CPU/Memory
Module Emergency Power Off (EPO)	Hardware	Module
Module Failed	Hardware	Module
Node Heartbeat Fault	Hardware	Module/Node
PCI Width Degrade	Hardware	GPU
RX message CRC error	Hardware	Interconnect
RX message header CRC error	Hardware	Interconnect
SCSI Error	Hardware	–
SeaStar Heartbeat Fault	Hardware	Interconnect
Seastar Lockup	Hardware	Interconnect
SXM Power Off	Hardware	GPU
VERTY Fault	Hardware	Module
Voltage Fault	Hardware	Module
WarnTemp Power Off	Hardware	CPU

Scope and Limitations

Failures that cause application aborts

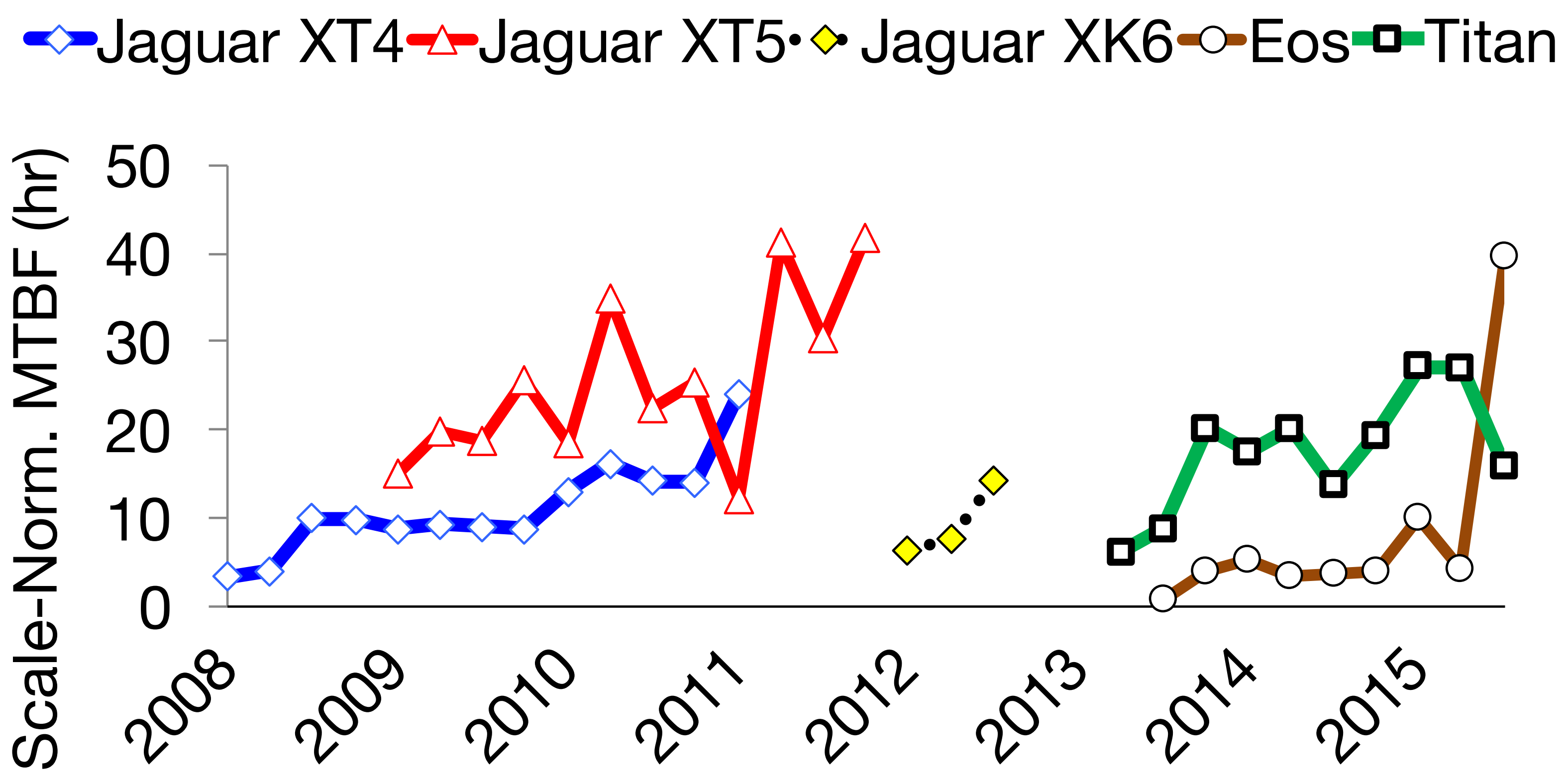
Difficult to isolate effects of multiple factors (300 second filter)

Dynamic operating environment

Root-cause analysis is not the goal
Easy to do (inaccurately)!

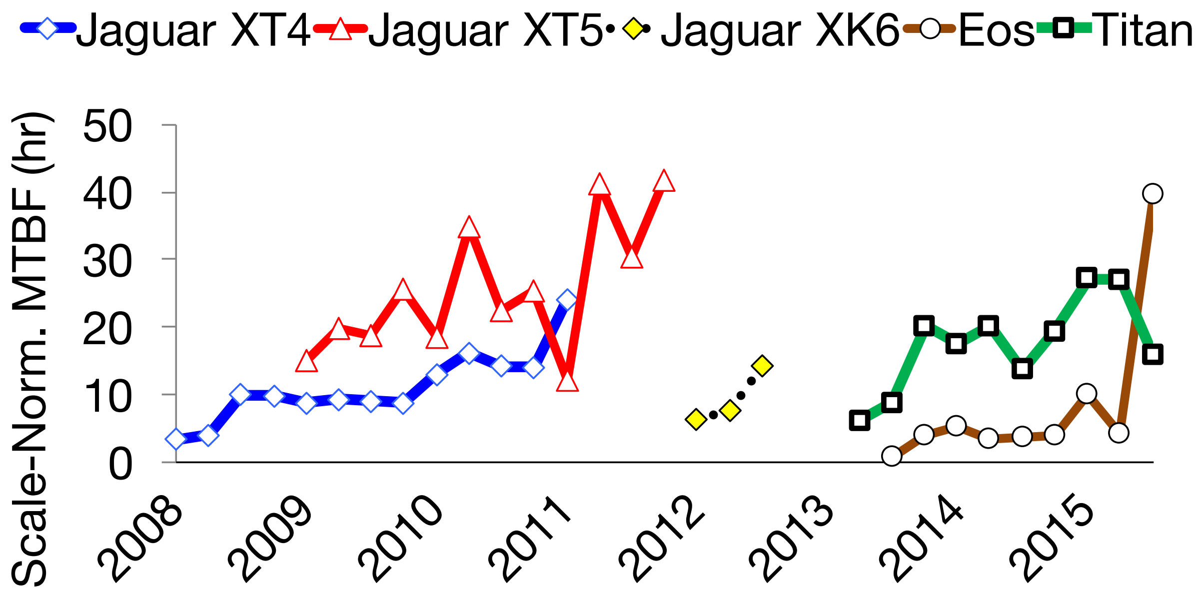
Are newer generations of HPC systems becoming less reliable?

During the stable operational period, does the reliability of the system change significantly? If so, by how much?



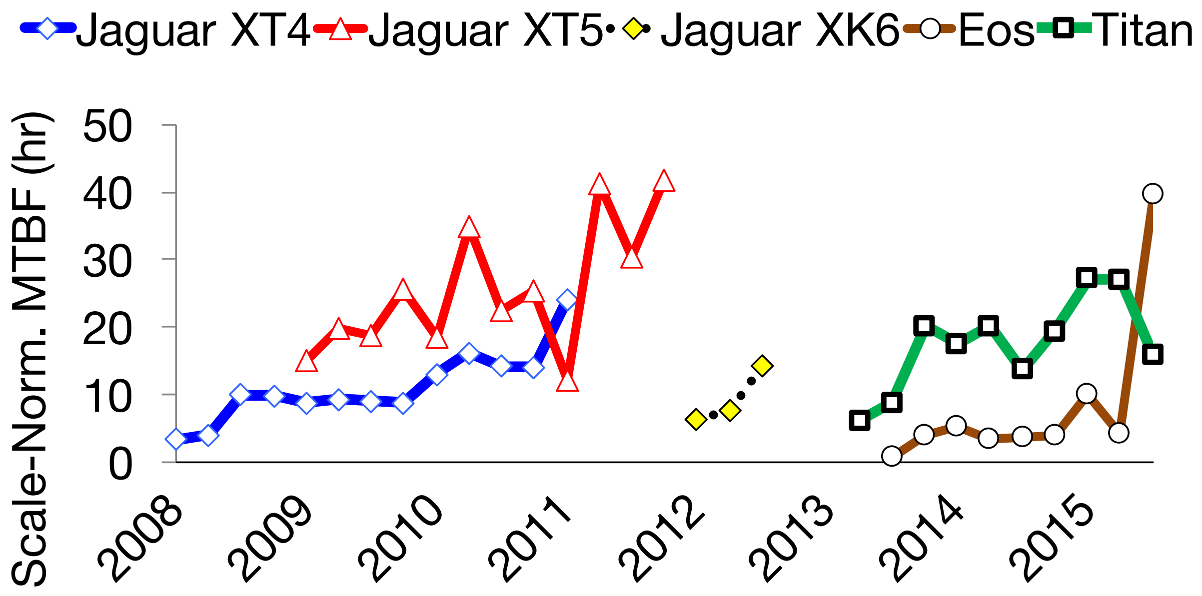
Scale normalized MTBF of each system

$$\text{Scale-Normalized MTBF} = \frac{\text{MTBF} \times \text{Num of Nodes in the System}}{\text{Max Number of Nodes across all Systems}}$$



Scale normalized MTBF of each system

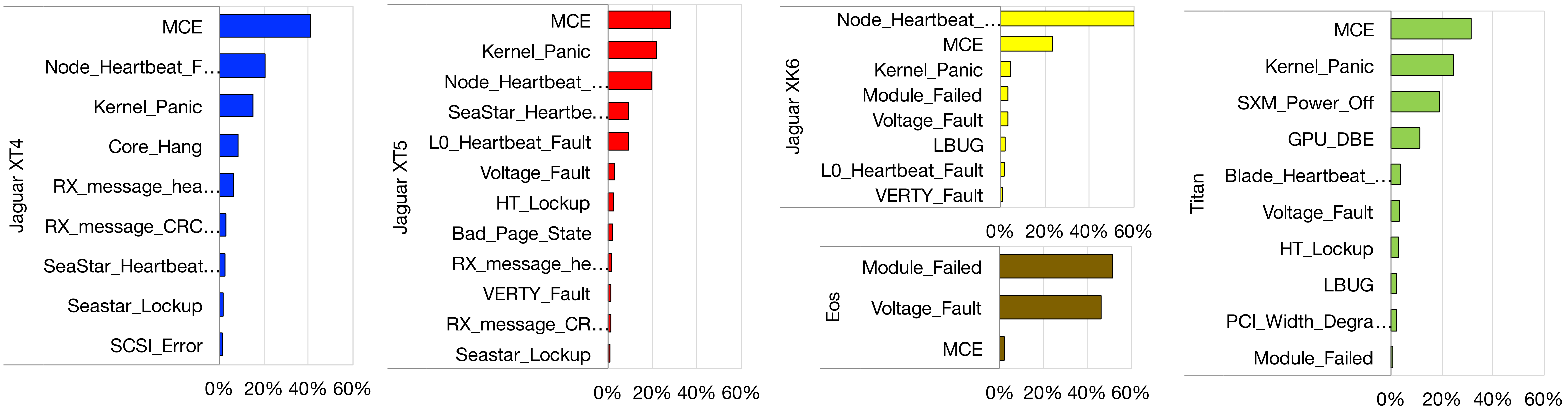
Newer generation of HPC systems are not necessarily consistently less reliable than previous generation systems.



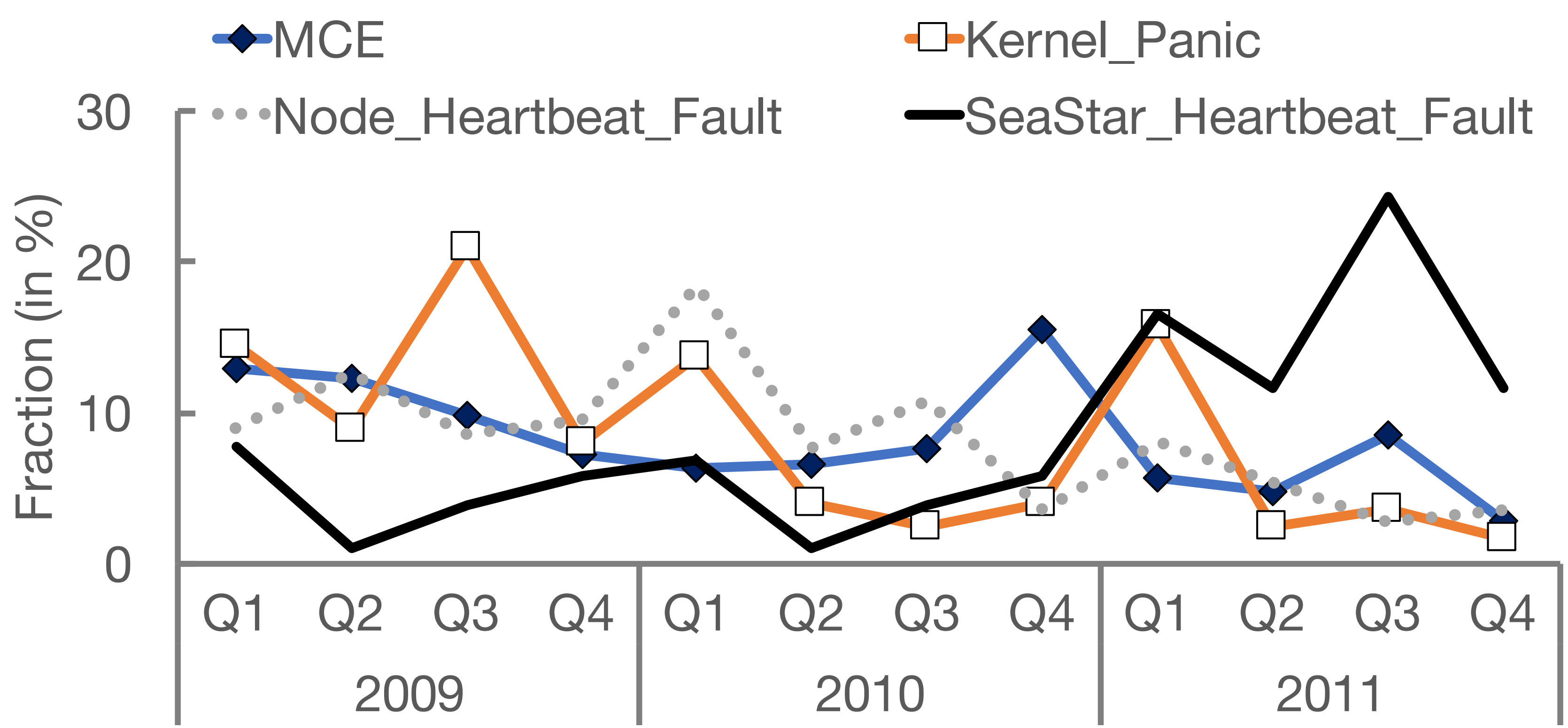
Scale normalized MTBF of each system

The MTBF of HPC systems doesn't necessarily decrease monotonically over different generations.
Even during the stable operational period, the MTBF may change by up to 4x!

What is the impact and temporal behavior of different failure types?



Contribution of different failure types across systems

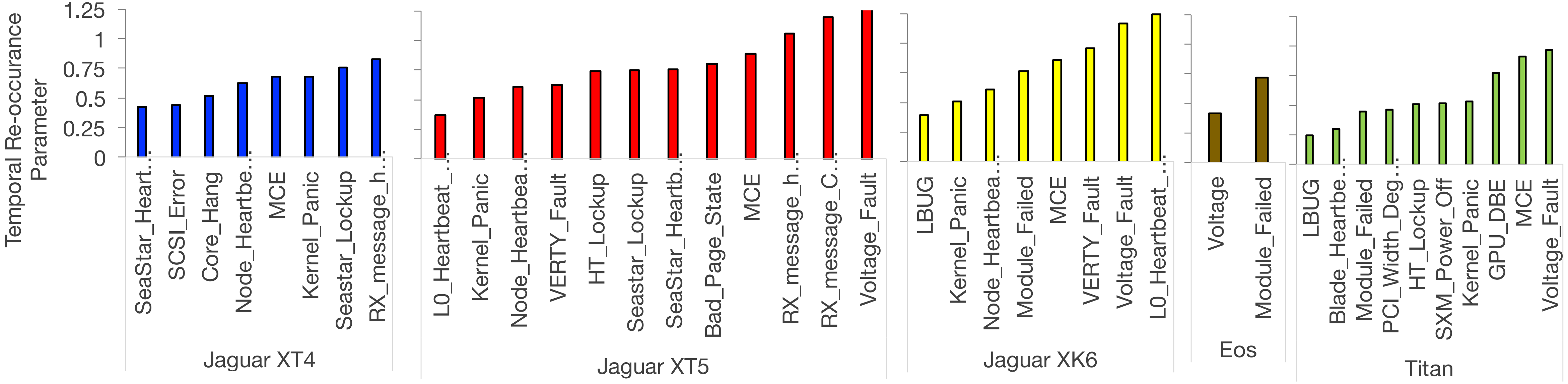


Contribution of different failure types over time (for Jaguar XT5)

A few failure types constitute a major fraction of all failures. Hardware related errors (e.g., uncorrectable memory errors) are dominant across systems over the whole period of time – implicating the importance of better provisioning and replication of CPU and GPU memory against such errors.

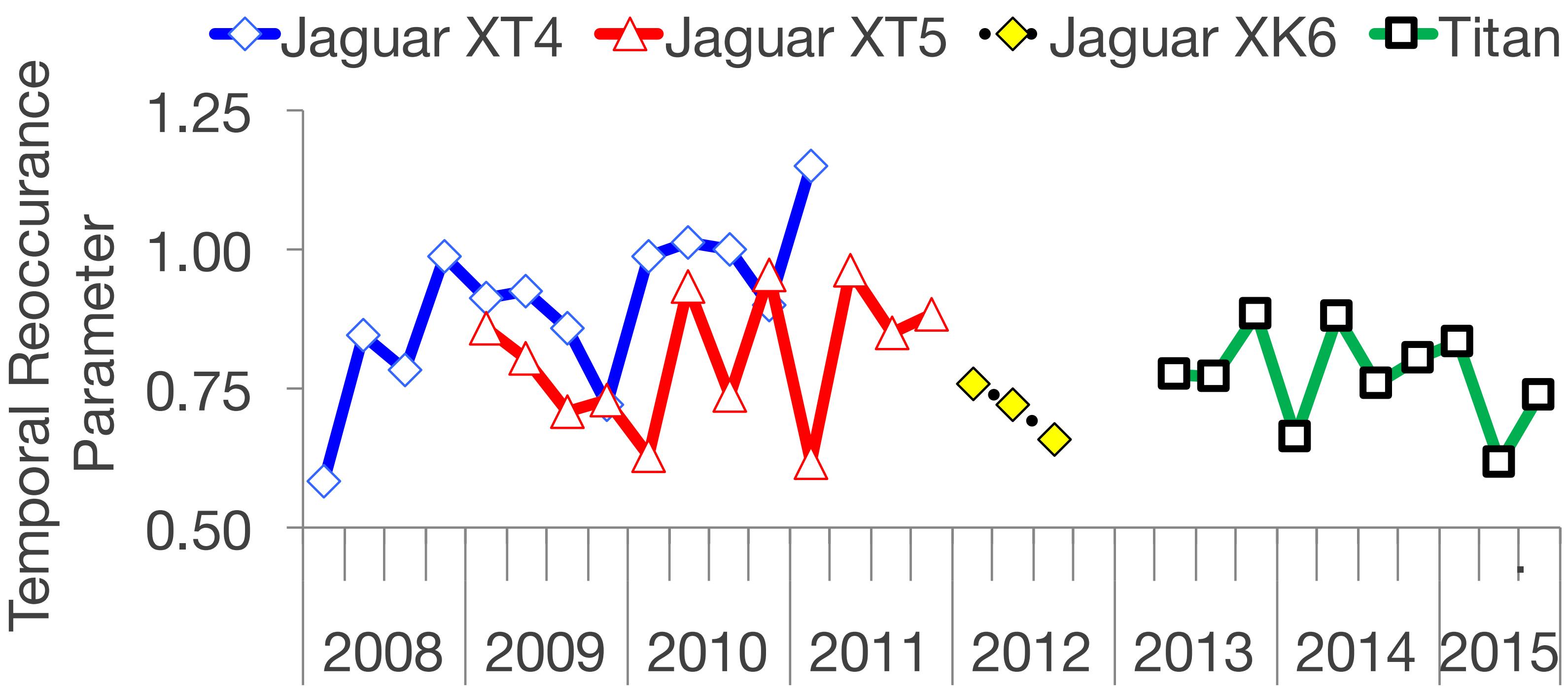
Given the significant variance in MTBF among different failure types, HPC system acquisition teams should also consider adding MTBF bounds for different failure types as a key metric in the request for proposals and contracts.

Temporal locality in failures: Does it vary across failure types and over time?



Temporal reoccurrence parameter across systems and failure types

See the paper for the formal mathematical formulation of the temporal reoccurrence parameter



Temporal reoccurrence parameter over time and across systems

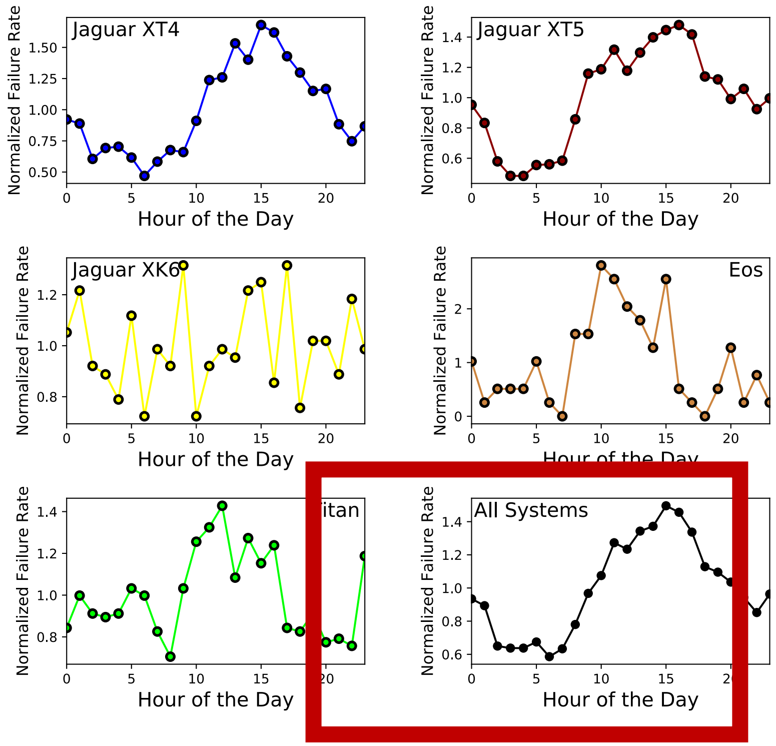
The temporal reoccurrence property varies significantly over time for a given system.

The temporal reoccurrence property for different failure types is significantly different, but similar across systems.

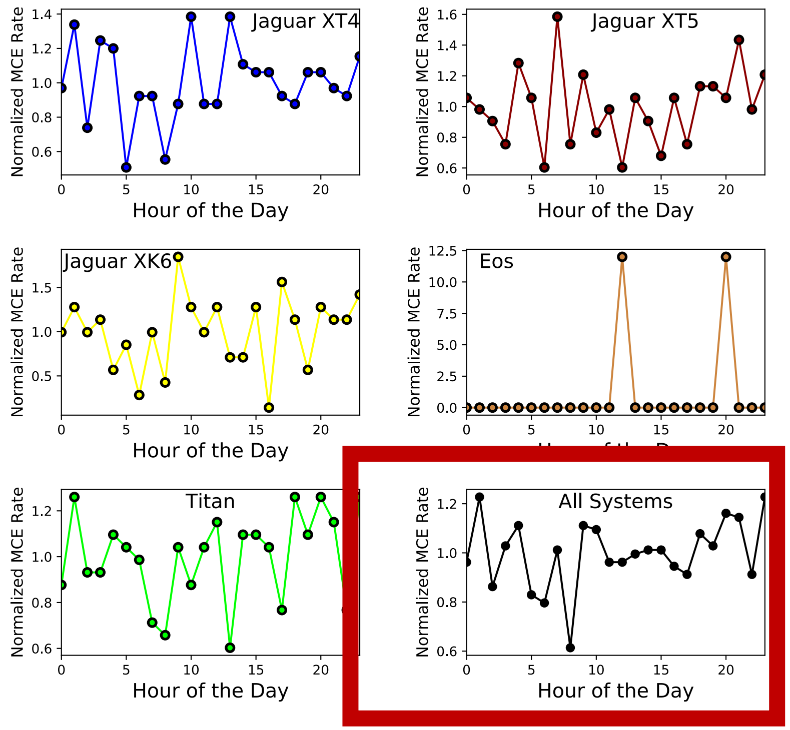
Implications for failure prediction.

The MTBF and the temporal reoccurrence parameter capture two different aspects of system reliability – any one alone is not sufficient.

Is there periodicity or are there temporal trends in failures?

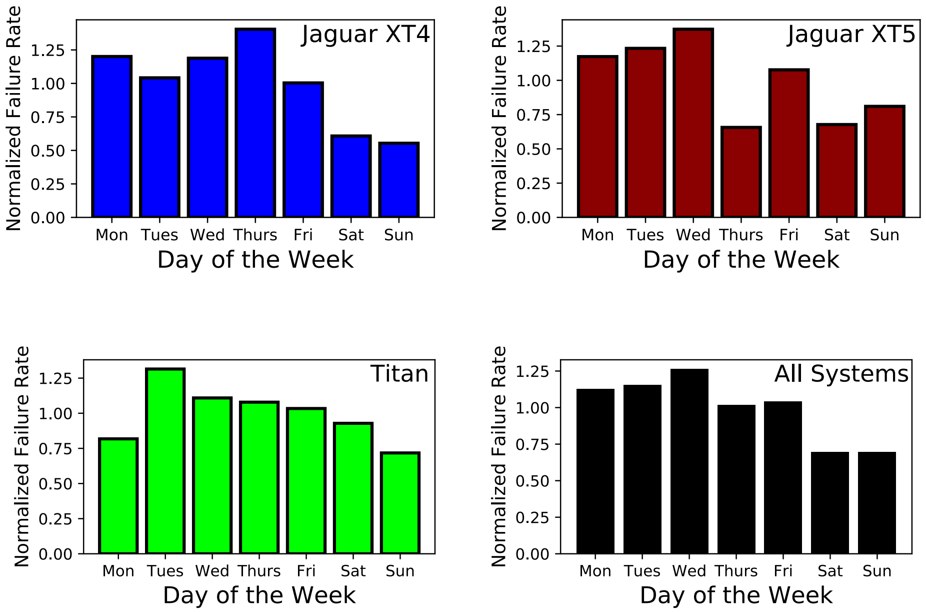


All failures over hour of the day

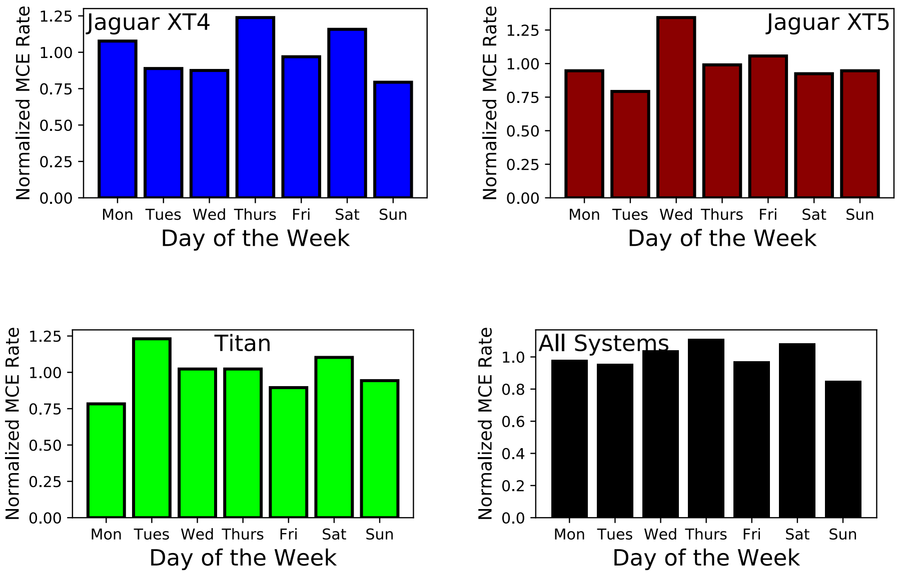


Memory errors over hour of the day

Failure rate increases during afternoon hours by up to 40%. However, this is not true for all failure types. Memory errors do not necessarily show increased failure rate during afternoon hours.



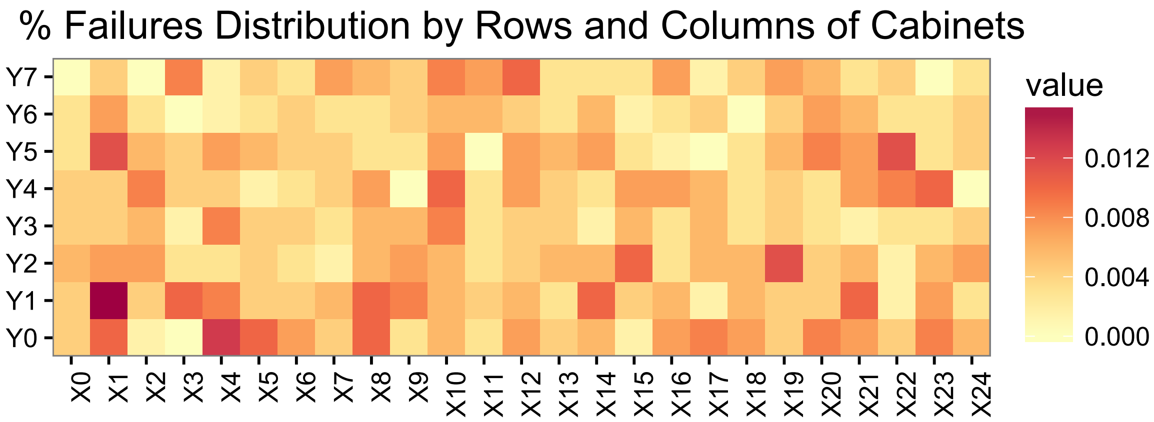
All failures over day of the week



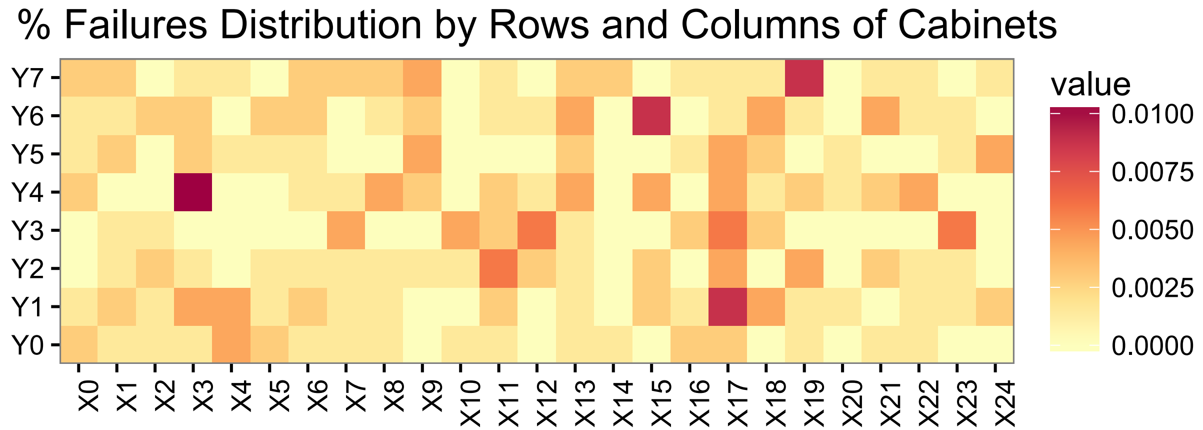
Memory errors over day of the week

Failure rate seem to decrease during the weekend. However, memory errors do not necessarily show this trend. Implications about utilization and error reporting.

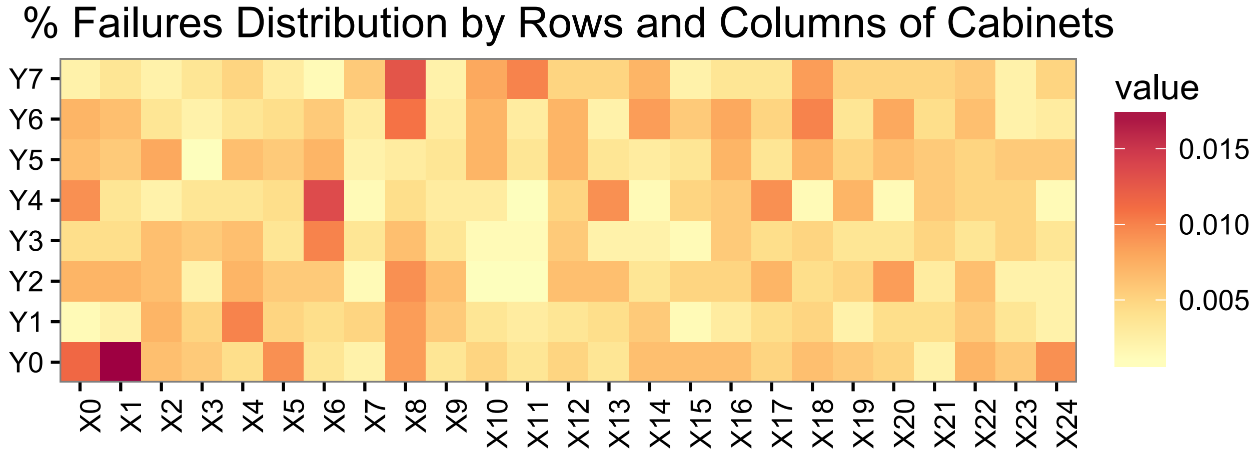
What about neighborhood effects in failures?



(a) Jaguar XT5

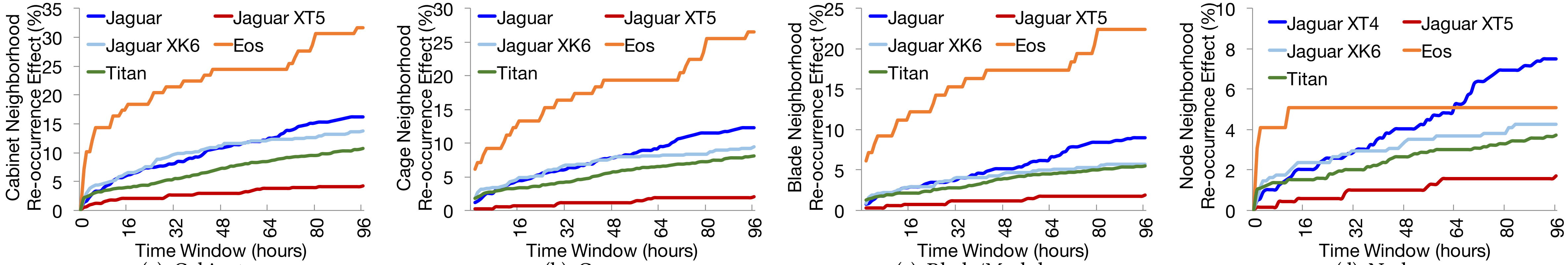


(b) Jaguar XK6



(c) Titan XK7

See the paper for the formal mathematical formulation of the neighborhood reoccurrence property



(a) Cabinet (b) Cage (c) Blade/Module (d) Node

Neighborhood reoccurrence property at different granularity across systems for a fixed time window

The spatial distribution of failures is not uniform at any compute granularity across systems.

Implications for job scheduler and users.

The neighborhood reoccurrence effect is not strongly correlated with the MTBF or the degree of temporal reoccurrence.

The neighborhood reoccurrence effect should be used as a separate reliability characteristic of a system.

It can not be subsumed by temporal characteristics, such as MTBF or temporal reoccurrence.

Conclusion

Systems show significant variations in reliability characteristics, even during the stable operational period.

Metrics beyond MTBF are needed to capture system failure characteristics.

Spatial and temporal characteristics of failures are often left unexploited.

Implications for job scheduler, sys admins, and system acquisition team.

Acknowledgements

Global Resilience Institute at Northeastern University

US Department of Energy, Office of Science, Office of Advanced Scientific Computing Research - Program manager Lucy Nowell

Failures in Large Scale Systems: Long-term Measurement, Analysis, and Implications

Saurabh Gupta
Intel Labs

Tirthak Patel
Northeastern University

Christian Engelmann
Oak Ridge Nat'l Lab

Devesh Tiwari
Northeastern University