Devesh Tiwari

Assistant Professor

Northeastern University

tiwari@northeastern.edu

Research Talk at IFIP WG 10.4 Meeting

Longmont, CO June 2017

In collaboration with Saurabh Gupta, Evgenia Smirni, Christian Engelmann, Sudharshan Vazhkudai, Franck Cappello, Jim Rogers et al.

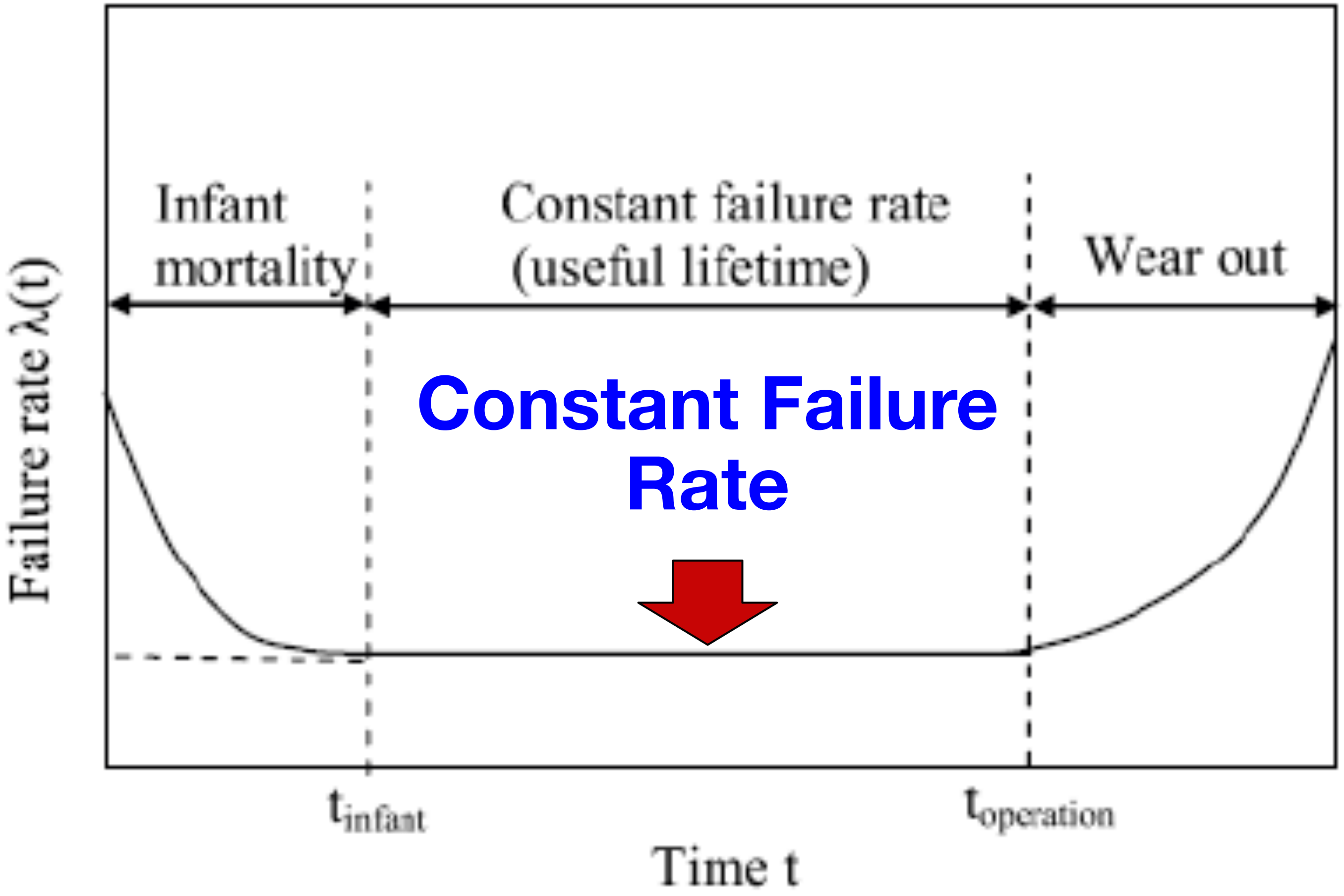
Useful Fake News

Devesh Tiwari
Assistant Professor
Northeastern University
tiwari@northeastern.edu

Research Talk at IFIP WG 10.4 Meeting
Longmont, CO June 2017

In collaboration with Saurabh Gupta, Evgenia Smirni, Christian Engelmann, Sudharshan Vazhkudai, Franck Cappello, Jim Rogers et al.

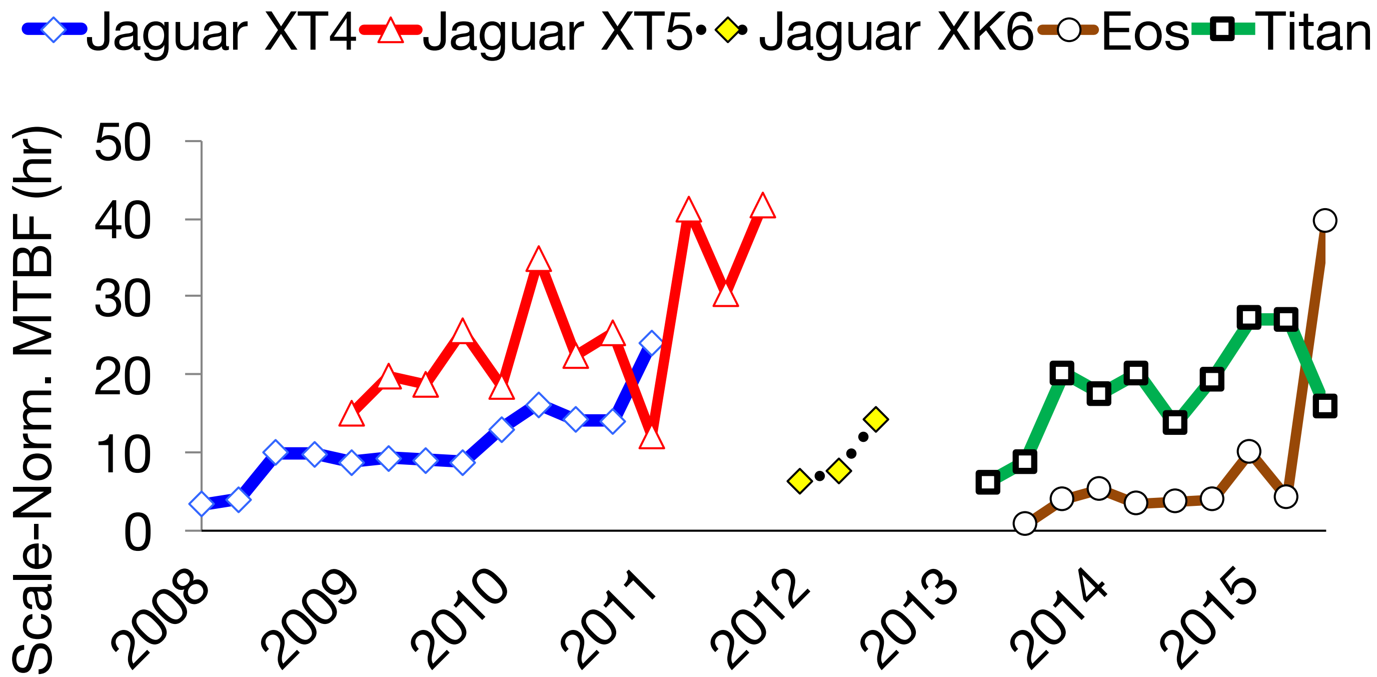
Bathtub Curve



Constant Failure Rate

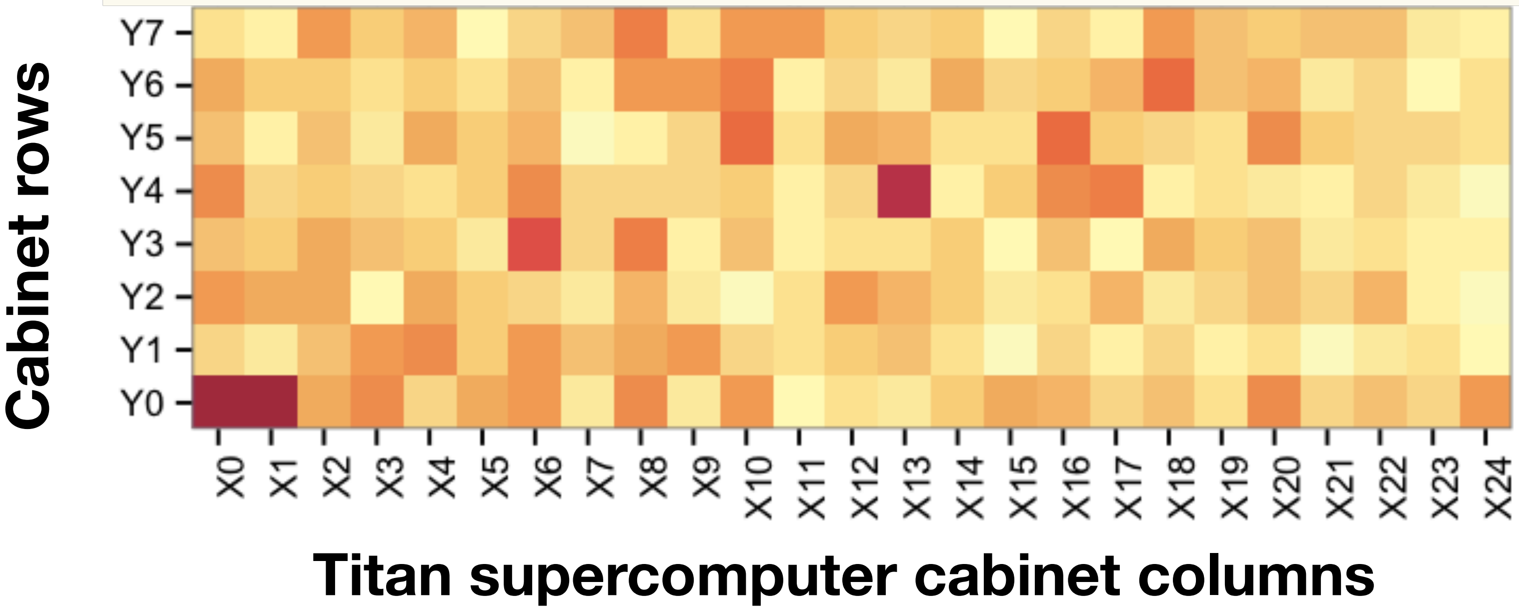
Uniformly random temporal and spatial distribution

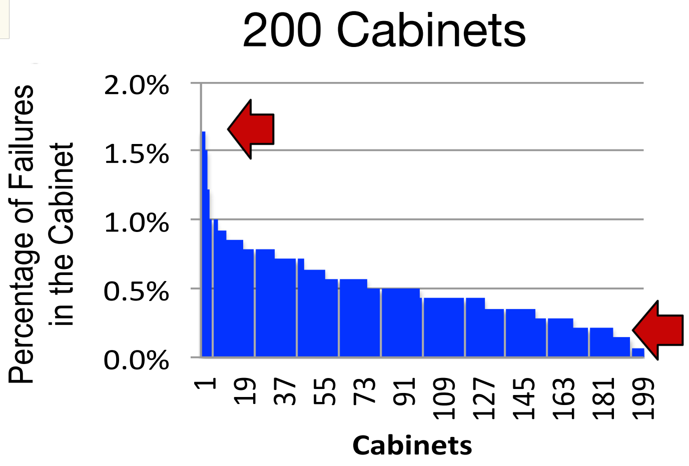
Bathtub Curve

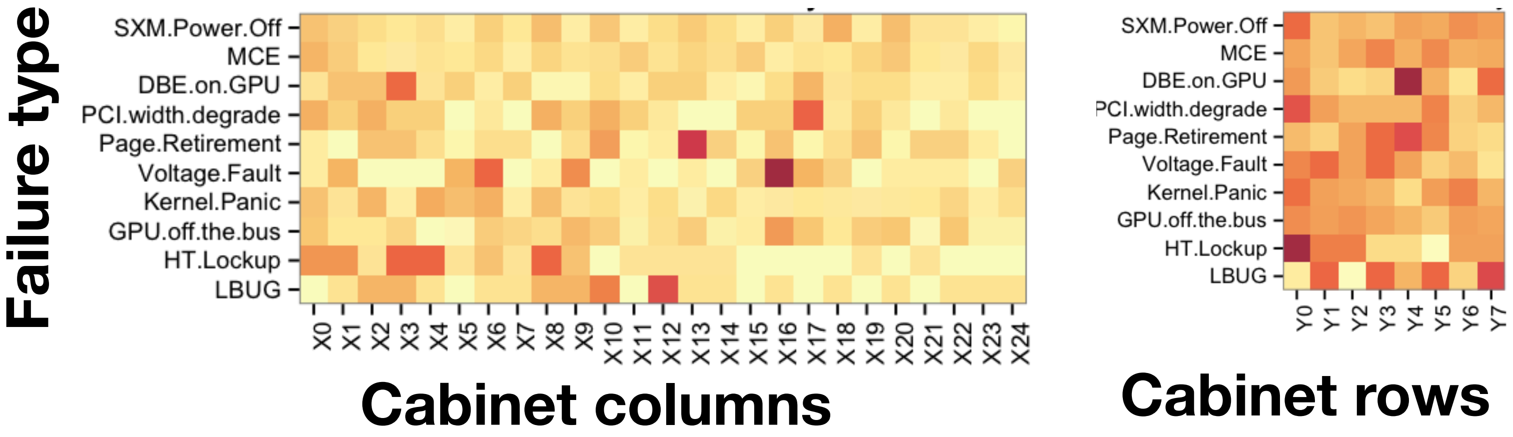


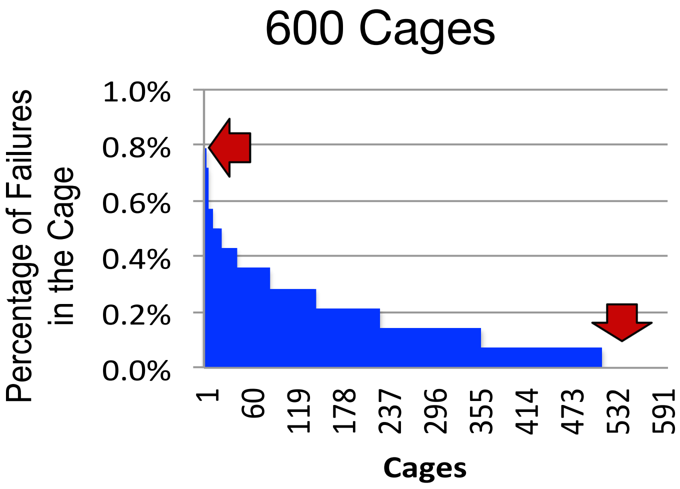
Up to 4x variation in MTBF during useful lifetime

Spatial Distribution of System Failures

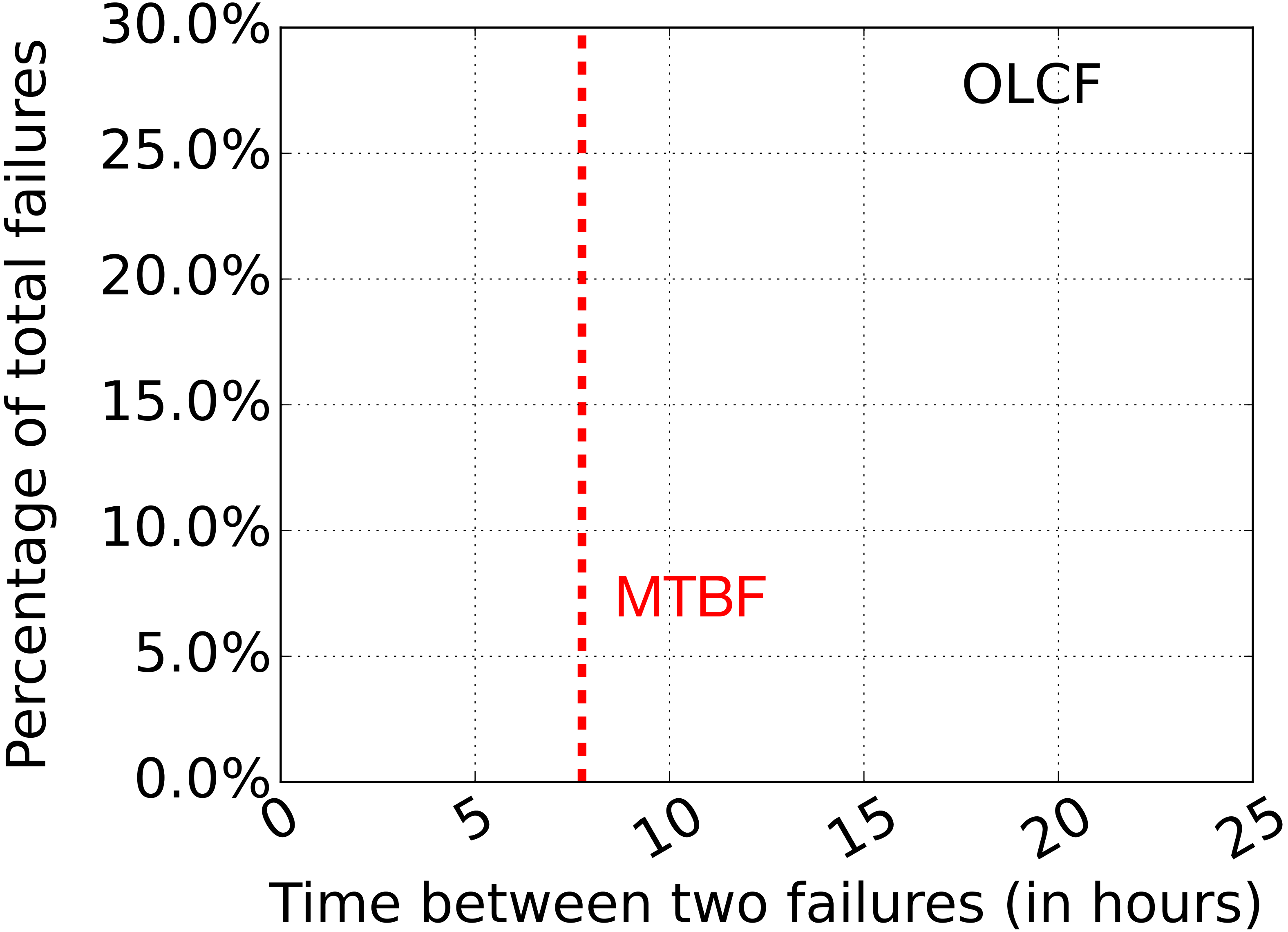


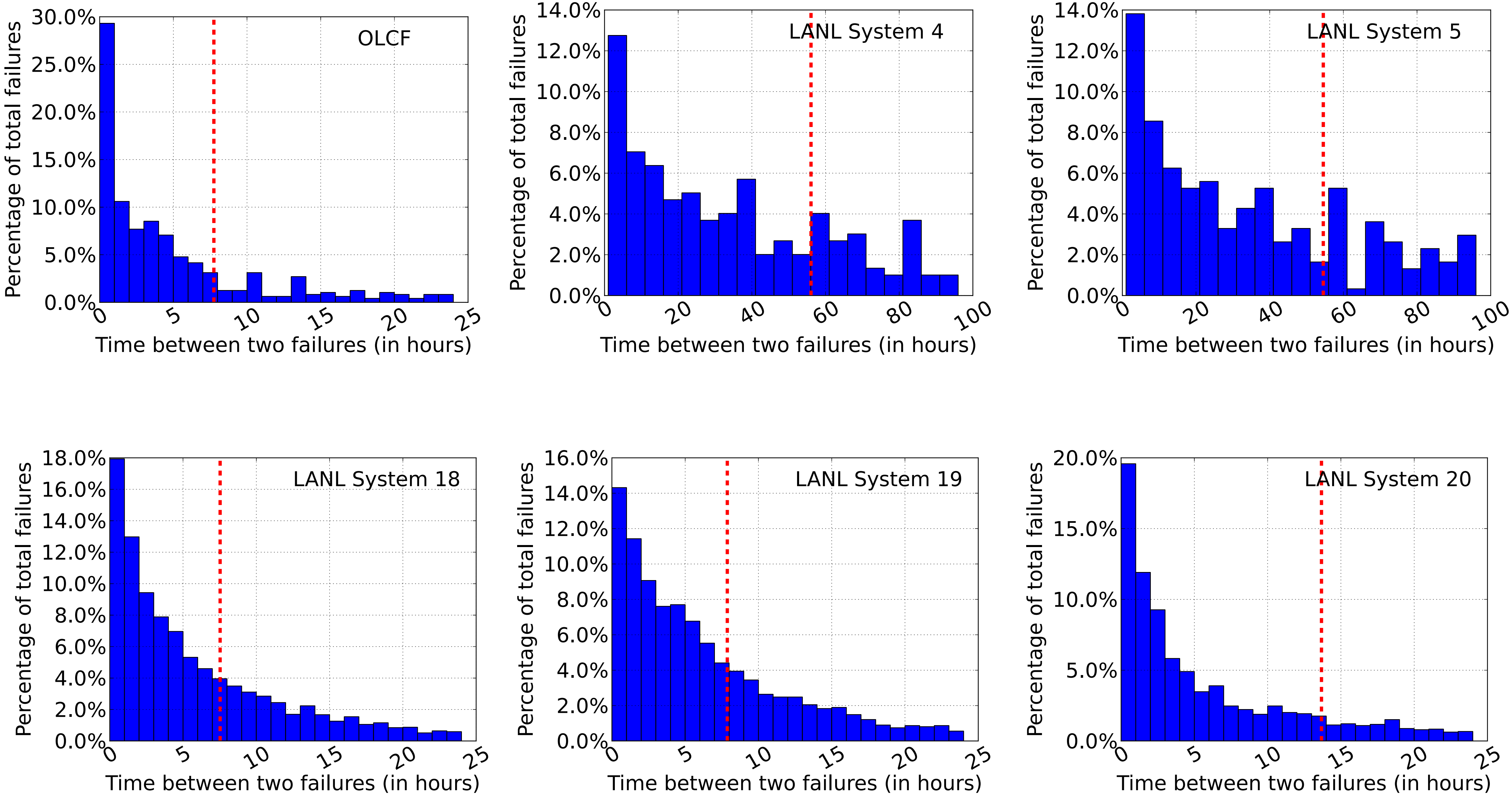


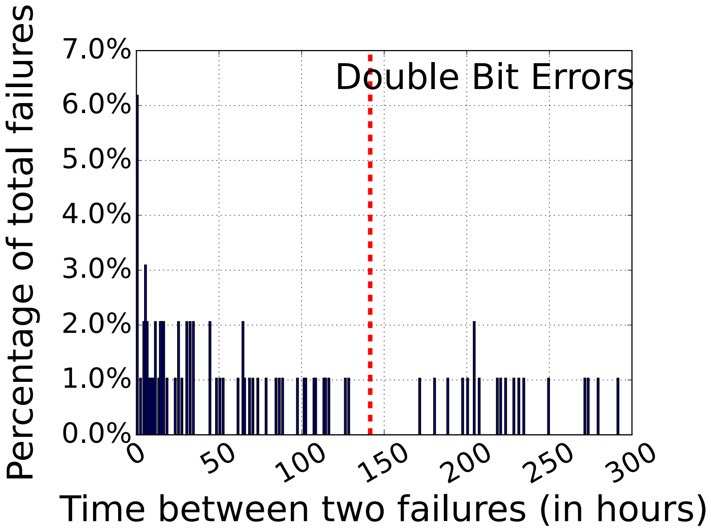


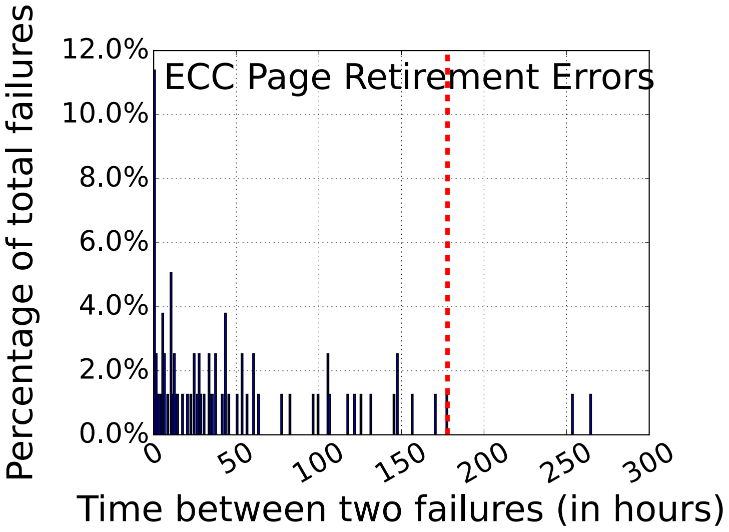


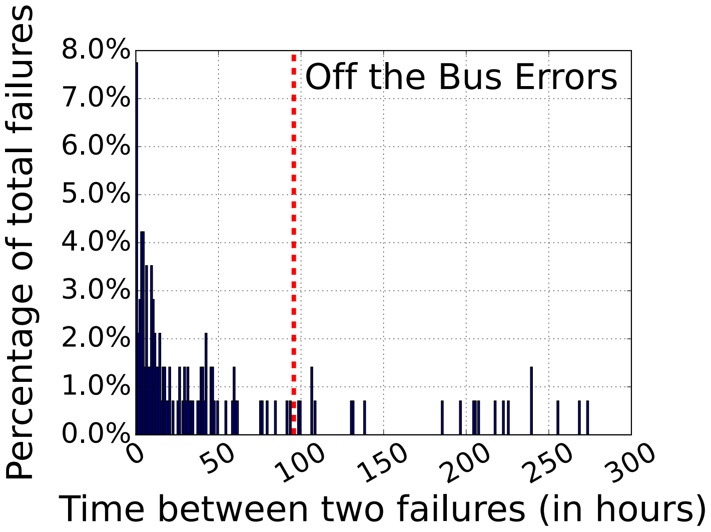
System failures are not uniformly randomly distributed in space.
This holds true for individual failure types, different time windows, spatial granularity.


30.0%
25.0%
20.0%
15.0%
10.0%
5.0%
0.0%
Percentage of total failures
0
5
10
15
20
25
Time between two failures (in hours)
OLCF
MTBF

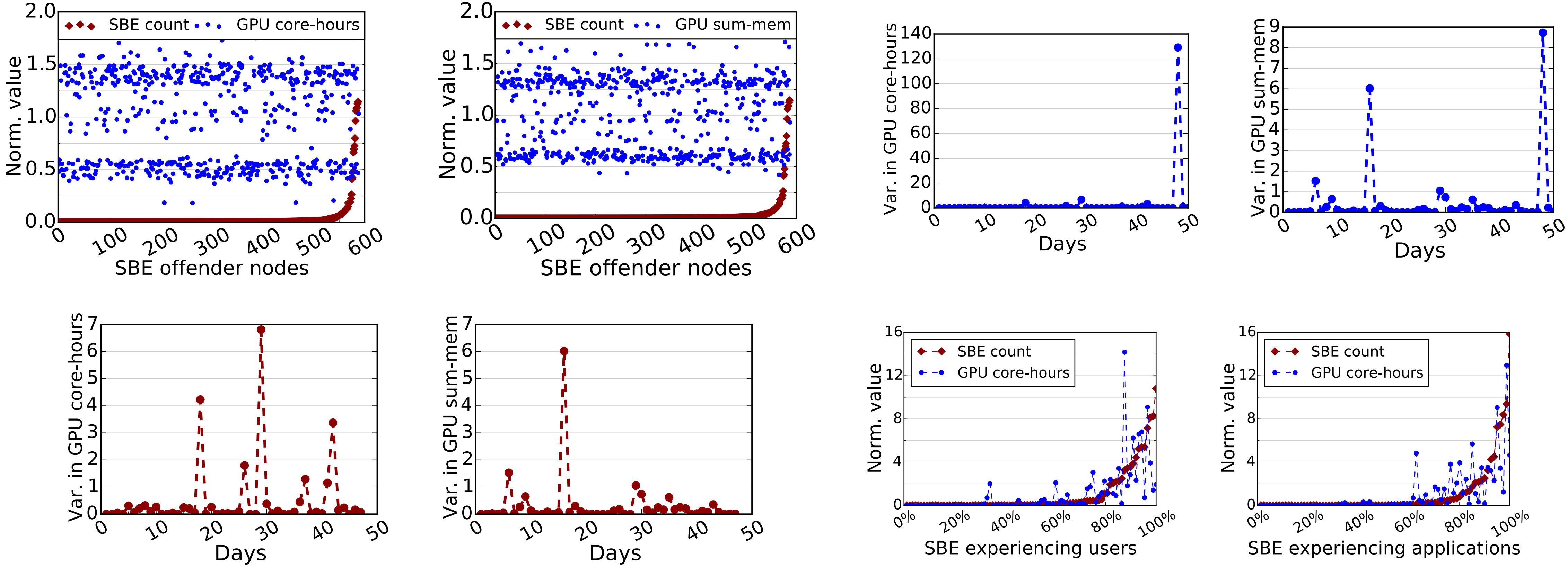

OLCF
LANL System 4
LANL System 5
LANL System 18
LANL System 19
LANL System 20
Percentage of total failures
Time between two failures (in hours)
0.0%
5.0%
10.0%
15.0%
20.0%
25.0%
30.0%
2.0%
4.0%
6.0%
8.0%
10.0%
12.0%
14.0%
16.0%
18.0%
0
5
10
15
20
25
40
60
80
100







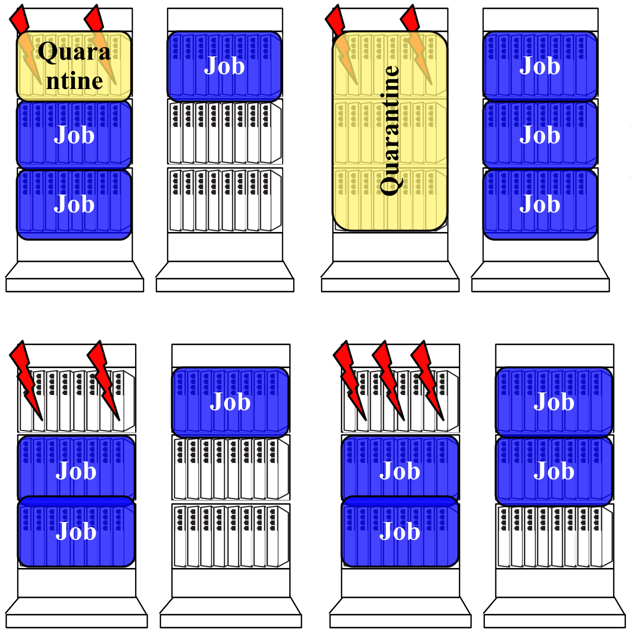
System failures exhibit strong temporal locality (not same as the traditional early bath-tub curve).

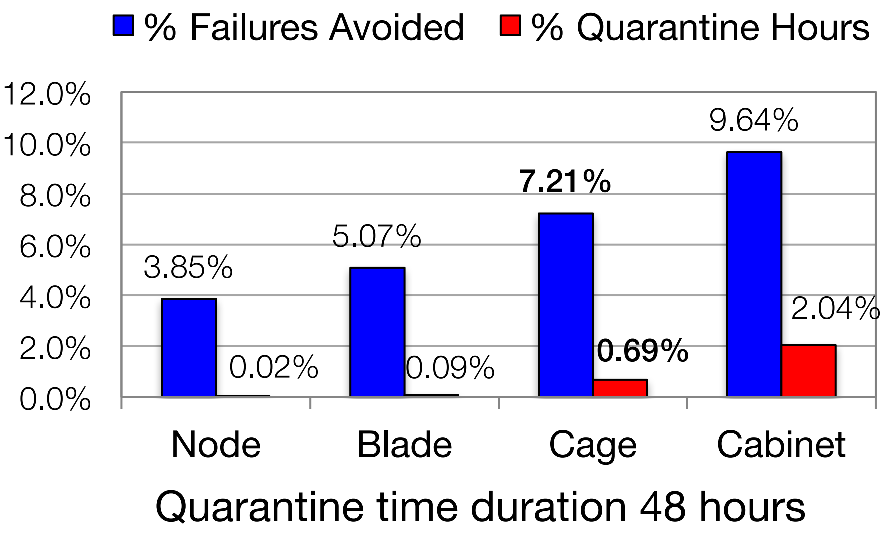


GPU core utilization or variance in utilization does not necessarily correlate with soft-errors, application and users do!

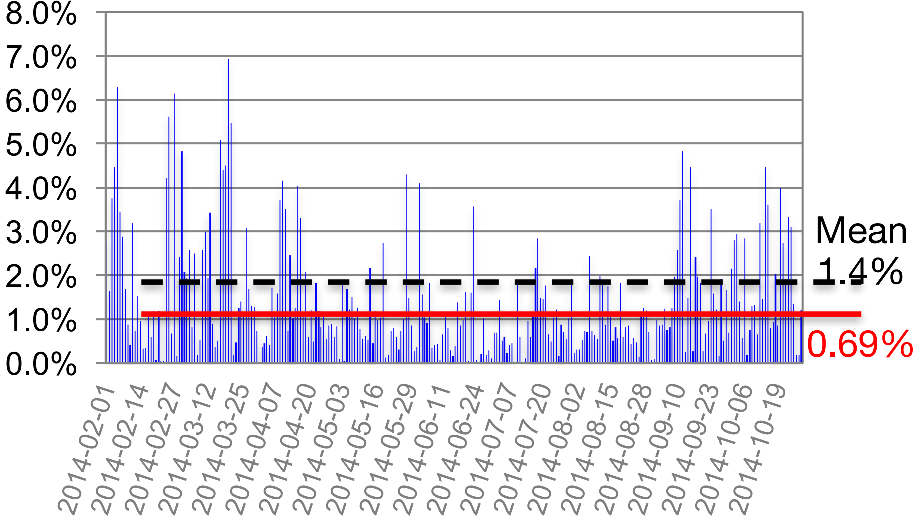
[HPCA 2016] A Large-Scale Study of Soft-Errors on GPUs in the Field

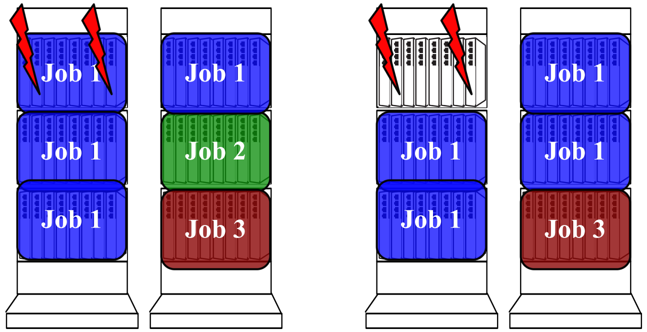
Quarantine Job Scheduling





Percentage of Node-hours used by debug jobs



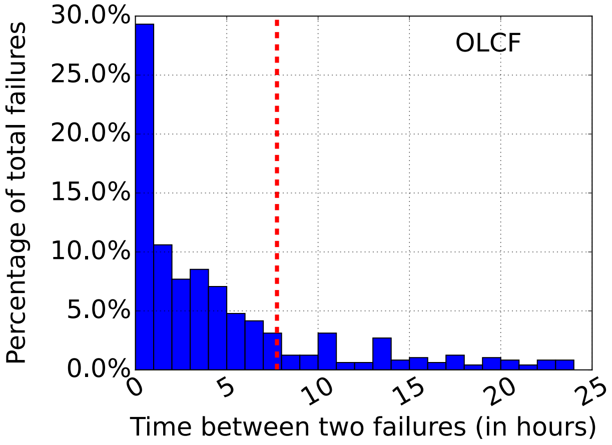


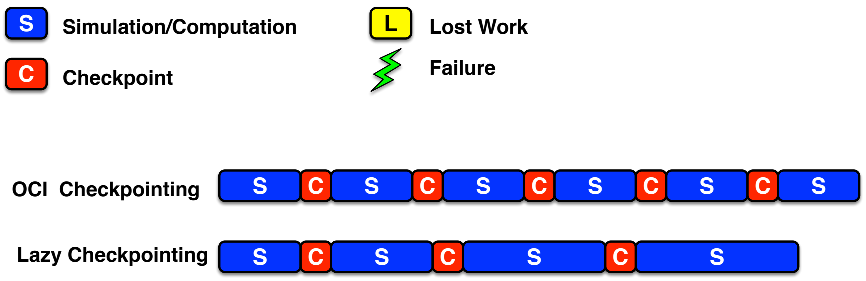
Significant fraction of failures can be avoided from interrupting production applications

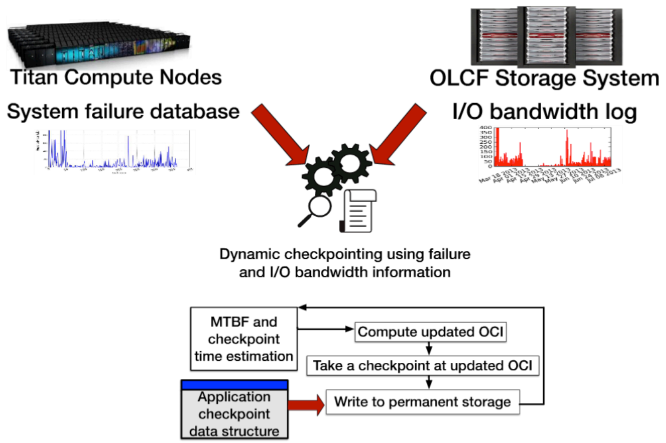
Debug or non-production jobs can be scheduled on quarantine nodes

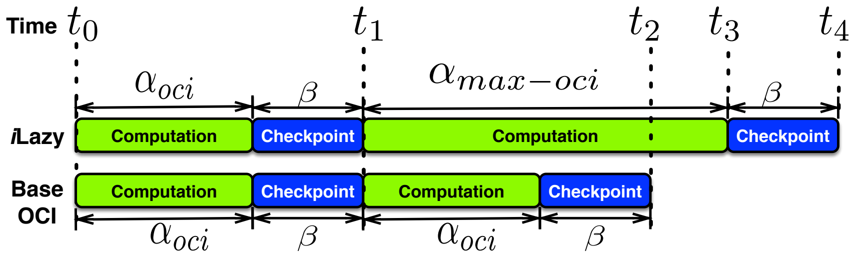
[DSN 2015] Understanding and Exploiting Spatial Properties of System Failures on Extreme-Scale HPC Systems

Lazy Checkpointing






Titan Compute Nodes
System failure database
OLCF Storage System
I/O bandwidth log
Dynamic checkpointing using failure and I/O bandwidth information
MTBF and checkpoint time estimation
Compute updated OCI
Take a checkpoint at updated OCI
Application checkpoint data structure
Write to permanent storage

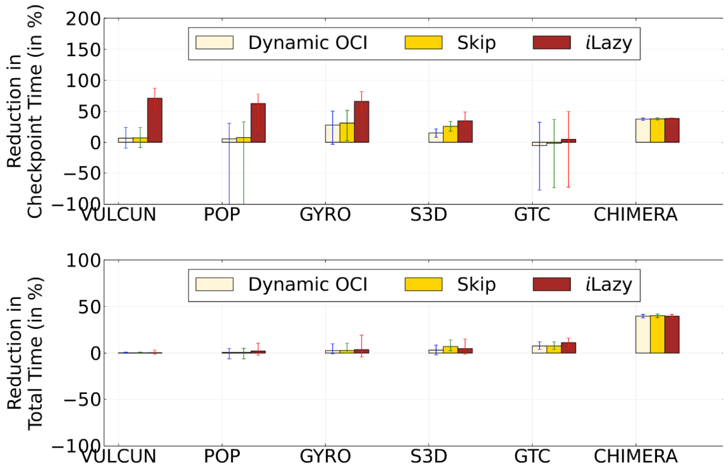


$$\text{performance gain} = \beta e^{-(\frac{t_3}{\lambda})^k}$$

$$\text{performance loss} = (\alpha_{max-oci} - \alpha_{oci})(e^{-(\frac{t_2}{\lambda})^k} - e^{-(\frac{t_4}{\lambda})^k})$$

$$= (\alpha_{max-oci} - \alpha_{oci})(e^{-(\frac{2(\alpha_{oci}+\beta)}{\lambda})^k} - e^{-(\frac{\alpha_{max-oci}+\alpha_{oci}+2\beta}{\lambda})^k})$$

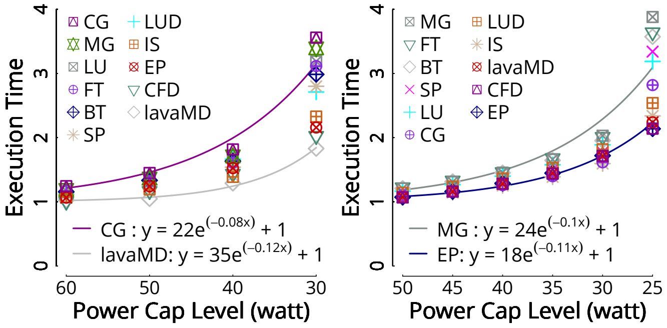
$$\beta e^{-(\frac{\alpha_{max-oci}+\alpha_{oci}+\beta}{\lambda})^k} = (\alpha_{max-oci} - \alpha_{oci})e^{-(\frac{2(\alpha_{oci}+\beta)}{\lambda})^k}$$
$$-(\alpha_{max-oci} - \alpha_{oci})e^{-(\frac{\alpha_{max-oci}+\alpha_{oci}+2\beta}{\lambda})^k}$$



Being Lazy is good (but with some bounds!)

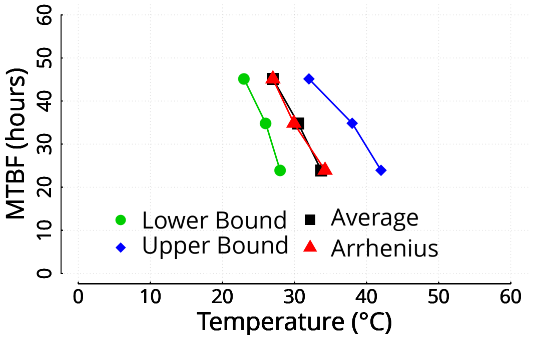
[DSN 2014] Lazy Checkpointing: Exploiting Temporal Locality in Failures to Mitigate Checkpointing Overheads on Extreme-Scale Systems

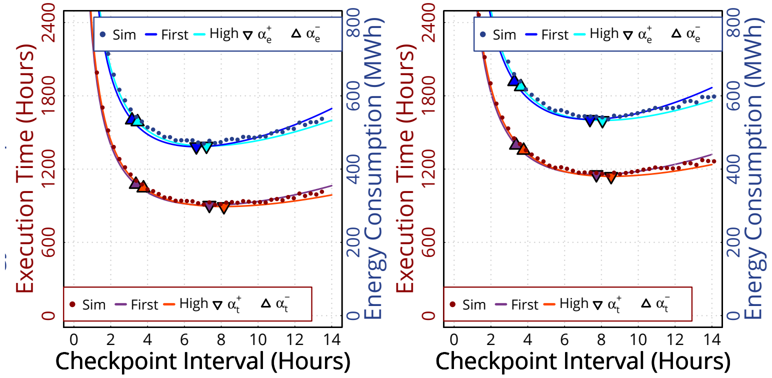
Optimal Checkpointing under Power-Constraints



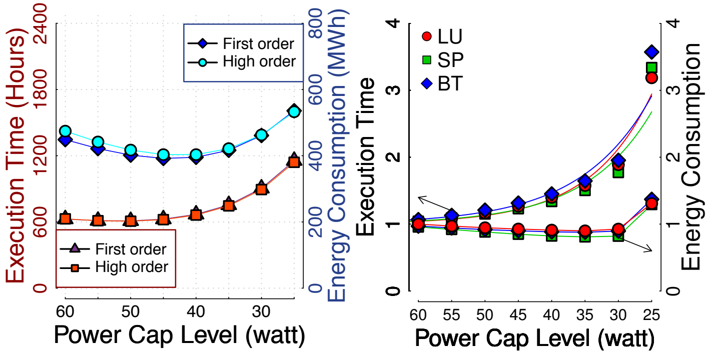
(a) Xeon E5-2670

(b) Xeon E5-2630





Under power-capping optimal checkpointing interval is quite different than traditional optimal checkpointing interval



(a) Optimal checkpointing interval

(b) No checkpointing

Optimal power capping level is different than no-checkpointing case.

[DSN 2016] Power-capping Aware Checkpointing: On the Interplay among Power-capping, Temperature, Reliability, Performance, and Energy

Thanks!

Devesh Tiwari
tiwari@northeastern.edu

Research Talk at IFIP WG 10.4 Meeting
Longmont, CO June 2017