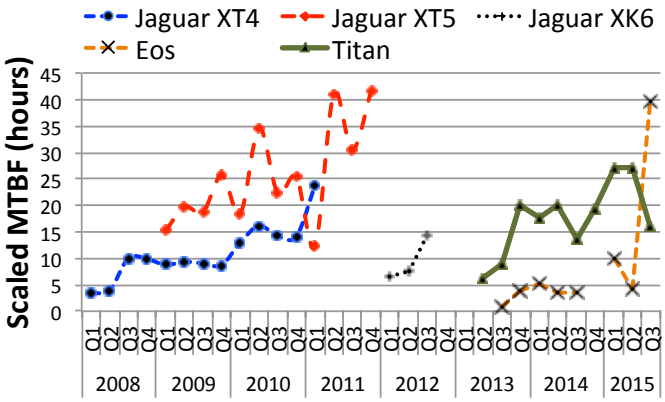
Failures in Large Scale Systems: Long-term Measurement, Analysis, and Implications

Achievement: Analyzed 1.2 billion node hours of system logs from the Jaguar, Titan, and Eos supercomputers at the Oak Ridge Leadership Computing Facility (OLCF).

Significance and Impact: This work offers an understanding of failures in today's supercomputers with a catalog of faults, errors and failures.

Research Details:

- Analyzed 1.2 billion node hours of logs from 5 different OLCF supercomputers.
- Combined information from different logs and created a consistent log format for analysis.
- Used standard and created new methods to model the temporal and spatial behavior of failures.
- Analyzed the evolution of temporal and spatial behavior over the years.
- Analyzed the correlation of different failure types.
- Compared the mean-time between failures of the 5 systems.



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

Scale-normalized MTBF of each system over time (averaged quarterly): MTBF can change significantly over time, with often a non-monotonic trend, which makes it averaged over lifetime an unattractive choice as metric.

Sponsor/Facility: This work was performed at ORNL in collaboration with Northeastern University. The work was sponsored by DOE ASCR (Program manager: Lucy Nowell).

PI and affiliation: Christian Engelmann – Oak Ridge National Laboratory

Team: S. Gupta, T. Patel, D. Tiwari, and C. Engelmann.

Publication: S. Gupta, T. Patel, D. Tiwari, and C. Engelmann. **Failures in Large Scale Systems: Long-term Measurement, Analysis, and Implications**. In *Proceedings of the 30th IEEE/ACM International Conference on High Performance Computing, Networking, Storage and Analysis (SC) 2017*, Denver, CO, USA, November 12-17, 2017.

Overview:

Resilience is one of the key challenges in maintaining high efficiency of future extreme scale supercomputers. Researchers and system practitioners rely on field-data studies to understand reliability characteristics and plan for future HPC systems. In this work, we compare and contrast the reliability characteristics of multiple large-scale HPC production systems. Our study covers more than one billion compute node hours across five different systems over a period of 8 years. We confirm previous findings which continue to be valid, discover new findings, and discuss their implications. The lessons learned are:

- System MTBF: The MTBF can change significantly over time, with often a non-monotonic trend, which makes it averaged over lifetime an unattractive choice as metric. As the optimal checkpointing interval depends on the MTBF, it needs to be adjusted accordingly.
- Dominant failure types: A set of dominant failure types is common across systems. Only very few types contribute most of the failures for each system. The set of major contributors can change significantly over time, which makes monitoring, diagnostics and mitigation a continuously evolving effort.

- Temporal failure locality:
  - The degree of temporal locality, which is very high in all studied systems, captures temporal characteristics better than auto-correlation, which is a measure of periodicity.
  - Different failure types can have a significantly different degree of temporal locality. Some types common between different systems also show a similar degree, such as voltage faults, operating system kernel panic, and Lustre parallel file system error.
  - While the MTBF may vary across systems, the degree of temporal locality may be very similar. The degree of temporal locality varies significantly over time for different systems, while a higher degree is not indicative of a lower MTBF.
- Spatial failure locality:
  - Spatial locality exists in all systems at all granularities (node, blade, cage, and cabinet). Jaguar XT4, Jaguar XK6 and Titan show similar trends, while Eos and Jaguar XT5 are significantly different.
  - Titan is the only system where spatial locality may be an artifact of the power/cooling infrastructure, i.e., hotter parts of the system experience more failures.
  - Spatial locality varies significantly over time and between systems. Some failure types have more locality than others. Though, the same types can have significantly different locality in different systems.
- Capturing system reliability better: The failure probability density function can be fitted into a parametric model to define the distribution under standard probability density models. The studied systems best fit the Weibull distribution. This methodology characterizes a system's spatio-temporal behavior and permits comparison.