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

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.

Excessive I/O overheads due to checkpoints

At exascale, applications may spend up to 60% of execution time in checkpointing and lost work!
Save Energy – A Positive Impact Beyond the Computing Facility

1 hour of lost work on the Titan supercomputer is roughly 5-9 MWhr
Table 2: This section collected from these HPC systems is regularly parsed on software.

The data is from January 2008 to September 2015. In total, this is 11.26 systems for this study using different instrumentation and logging methods, and key component analysis of failures is not possible given the complexity and granularity of data measurement, collection, and dynamic operational updates doesn’t skew our findings. We only consider the actual parent events in our analysis. We note that only statistically significant results are expected from our analysis to avoid side-effects of the actual failure event. We effect such modifications on the validity of our results.

Our data processing step revealed that some failure events are recorded multiple times in the system logs because of multiple locations reporting the side-effects. Other works such as [12] have effectively dropped such events from our analysis to avoid side-effects in system monitoring and diagnostics, to carefully drop such events.
Failures in Over 1 Billion Compute Node Hours

### 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)!**

#### Table 1: HPC Systems analyzed in this study.

<table>
<thead>
<tr>
<th>System</th>
<th>Type</th>
<th>Node/CPU</th>
<th>18,688</th>
<th>7,832</th>
</tr>
</thead>
<tbody>
<tr>
<td>Titan XK7</td>
<td>Hardware</td>
<td>560,640</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Eos XC 30</td>
<td>Hardware</td>
<td>23,553</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Jaguar XK6</td>
<td>Hardware</td>
<td>298,592</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>Jaguar XT4</td>
<td>Hardware</td>
<td>31328</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

#### Table 2: List of failure types observed on the systems in this study.

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

This study analyzes system failure data from different production HPC systems. These systems are Jaguar XT4, Jaguar XK6, Titan XK7, Eos XC 30, and Jaguar XT5 supercomputer, Jaguar XK6 supercomputer, Jaguar XT4 supercomputer, and Titan XK7 supercomputer. The data measured and collected is also used to report problems to vendors and build a knowledge tract record of failures, problems, and potential issues. This data is from January 2008 to September 2015. In total, this is 11.26 systems base for system administration. The data collected from these HPC systems is regularly parsed on software.
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?
ways where both parties always have rewards and risks associated. The overall number of failures and what can we learn from their dramatic below a certain threshold due to hardware issue, the support available should the reliability drops below a certain threshold.

Upper bound on the variance in MTBF as a key metric in the request acquisition teams at HPC centers can use this. Outage still counts as one failure and hence, does not account for upgrades. As discussed later in detail, different generations of the HPC systems. Even during the stable operational periods previous generation of systems are more reliable – systems have higher scale-normalized MTBF, while during some time period the reliability doesn't necessarily decrease monotonically over different granularities.

Summary

We use the mean time between failure (MTBF) as the impact. This is a simple, direct way to study a system’s reliability. System MTBF has been a commonly

We also considered that a small subset of failure types constitute a significant fraction of hardware related errors, such as SeaStar errors (interconnect related errors) interrupted application.

Machine Check Exception (MCE), Kernel Panic, and Node failure could be a result of all failures of that particular type occurring in a short period of time. These events have resulted in a system failure only when these events have resulted in system failure. This also indicates that improved system maintenance and operational cost in the future.

According to studies [11, 37, 39], that dominant failure types occur throughout the period instead of at different times. Fig. 2 shows the fraction of each failure type with respect to the total number of failures on the system.

Table 3: Scale-Normalized MTBF of each system

<table>
<thead>
<tr>
<th>System</th>
<th>Scale-Normalized MTBF (hr)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jaguar XT5</td>
<td>42.67</td>
</tr>
<tr>
<td>Jaguar XT4</td>
<td>22.67</td>
</tr>
<tr>
<td>Eos</td>
<td>7.45</td>
</tr>
<tr>
<td>Titan</td>
<td>8.93</td>
</tr>
<tr>
<td>Jaguar XK6</td>
<td>18.68</td>
</tr>
</tbody>
</table>

From Table 1, we note that Jaguar XT5 has the highest scale-normalized MTBF over time. The plot shows that a small subset of failure types constitute a significant fraction of hardware related errors, such as SeaStar errors (interconnect related errors) interrupted application.

Table 3: Scale-Normalized MTBF of each system

<table>
<thead>
<tr>
<th>System</th>
<th>Scale-Normalized MTBF (hr)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jaguar XT5</td>
<td>42.67</td>
</tr>
<tr>
<td>Jaguar XT4</td>
<td>22.67</td>
</tr>
<tr>
<td>Eos</td>
<td>7.45</td>
</tr>
<tr>
<td>Titan</td>
<td>8.93</td>
</tr>
<tr>
<td>Jaguar XK6</td>
<td>18.68</td>
</tr>
</tbody>
</table>

We also considered that a small subset of failure types constitute a significant fraction of hardware related errors, such as SeaStar errors (interconnect related errors) interrupted application.
Newer generation of HPC systems are not necessarily consistently less reliable than previous generation systems.
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?
Summary

Failures in Large Scale Systems: Long-term Measurement, Analysis, and Implications SC17, November 12–17, 2017, Denver, CO, USA

- The key metric in the request for proposals and contracts.
- Consider adding MTBF bound for different failure types, HPC system acquisition teams should also.

Given the significant variance in MTBF among different failure types, we found that a few failure types constitute a major fraction of all failures. We also observed that the number of failure events were not enough to pass the test of statistical significance.

To the best of our knowledge, we are the first to investigate if the degree of temporal recurrence varies over time and across systems. For example, the temporal recurrence parameter for Jaguar XT5, Jaguar XK6, and Titan are 0.81, 0.71, -0.97, and -0.03.

We found that the number of failure events were not enough to pass the test of statistical significance. Eos XC30 system is not included in this analysis because systems. Fig. 5 shows the temporal recurrence parameter for Jaguar XT4 varies between 0.58 and 0.99.

Interestingly and counter-intuitively, in the same period of time, the MTBF of Jaguar XT4 increases significantly over time across different systems. We found the correlation coefficient between the sequence of MTBFs across systems. The MTBF and temporal recurrence parameter capture two different aspects of the system reliability and can not be used as a measure of overall system reliability.

This shows that a high likelihood of temporal recurrence doesn't necessarily mean lower MTBF. It varies significantly across systems. We found the correlation coefficient between the sequence of MTBFs across systems. The MTBF and temporal recurrence parameter capture two different aspects of the system reliability and can not be used as a measure of overall system reliability.

On the other hand, we computed the correlation between the sequence of MTBFs across systems. The closer the value of the correlation coefficient is to 1 (i.e., closer to exponential distribution), the higher is the degree of recurrence. Moreover, the lower is the degree of recurrence. Interestingly, the degree of recurrence changes significantly over time and across systems. For example, the temporal recurrence parameter for Jaguar XT4 varies between 0.58 and 0.99.

Prior studies have shown that system failures have temporal recurrence property. To the best of our knowledge, we are the first to investigate if the degree of temporal recurrence varies over time and across systems. For example, the temporal recurrence parameter for Jaguar XT5, Jaguar XK6, and Titan are 0.81, 0.71, -0.97, and -0.03.

We found that the number of failure events were not enough to pass the test of statistical significance. Eos XC30 system is not included in this analysis because systems. Fig. 5 shows the temporal recurrence parameter for Jaguar XT4 varies between 0.58 and 0.99.

Interestingly and counter-intuitively, in the same period of time, the MTBF of Jaguar XT4 increases significantly over time across different systems. We found the correlation coefficient between the sequence of MTBFs across systems. The MTBF and temporal recurrence parameter capture two different aspects of the system reliability and can not be used as a measure of overall system reliability.

This shows that a high likelihood of temporal recurrence doesn't necessarily mean lower MTBF. It varies significantly across systems. We found the correlation coefficient between the sequence of MTBFs across systems. The MTBF and temporal recurrence parameter capture two different aspects of the system reliability and can not be used as a measure of overall system reliability.
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?
A system could also be proactively stress tested for some particular characteristics amongst different failure types after the first occurrence of the failure – to avoid recurrence parameter. This observation can help in the planning for proactive action and repair mechanism for a given failure type.

Next, we investigate the temporal recurrence characteristics of each failure type across different systems. Fig. 4 shows the temporal recurrence parameter for different failure types is presented as heatmaps. We explored deeper and found that Seastar lockup may cause interesting temporal recurrence relationship between a Seastar lockup and other failures. Temporal reoccurrence parameter across different 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?
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.
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?
As shown in above equations, the neighborhood recurrence $e_{t}$ can be both positively and negatively correlated with the temporal recurrence in failures, where subsequent failures occur in the vicinity of previous failure events. This further demonstrates that certain locations may experience higher failures temporarily after an event. There are two different phenomena that are responsive to the neighborhood recurrence parameter for each quarter. The correlation coefficient $r$ is quite large. Second, other than Jaguar XT4, which has relatively smaller scale system (i.e., Eos) at each granularity (in different time windows), other systems also see an increase in neighborhood recurrence as the time window is changed. For Jaguar XT5, the neighborhood recurrence is significantly different compared to other systems. Jaguar XT4 and Eos have higher neighborhood recurrence $e_{1}$, whereas Jaguar XK6 and Titan show similar neighborhood recurrence trends. For Jaguar XT5, the neighborhood recurrence $e_{1}$ for relatively smaller scale system (i.e., Eos) at each granularity (in different systems). Jaguar XT5 and Eos have higher neighborhood recurrence $e_{1}$, whereas Jaguar XK6 and Titan show similar neighborhood recurrence trends. For Jaguar XT5, the neighborhood recurrence $e_{1}$ for relatively smaller scale system (i.e., Eos) at each granularity (in different systems).

See the paper for the formal mathematical formulation of the neighborhood reoccurrence property.
Granularity across systems for a fixed time window indicates that the autocorrelation of system failure events represents some periodicity in the signal, while the correlation dies down with increasing lag [18]. The intensity of system failures is monotonically increasing with time, and includes reliability growth models to understand if the inter-arrival times are memoryless, i.e., the failures are independent of each other; the best distribution is used to infer the system's behavior. For example, if distributions used to represent empirical data and the best fitting distribution is Exponential, then the system exhibits temporal recurrence. Neighborhood recurrence at different granularities is not strongly correlated with MTBF or degree of spatial correlation in system failures. We compare and contrast the data from Argonne National Laboratory, which have independently verified results for their systems as well [30].

System failures are complex to allow system administrators and users to compare reliability characteristics to improve the reliability of HPC systems. For example, Liang et al. have provided a thorough understanding of the impact of failure types, inter-arrival patterns and different event types is explored by previous works [8]. The analysis and reliability metrics derived from the data are system-specific and can not be subsumed by temporal characteristics such as MTBF. The development of such analysis methodologies, tools and predictive models can lead to deeper insights about the underlying system behavior but it remains too complex to make parts of the data available for researchers.

System RAS logs for multiple HPC systems investigated system RAS logs for multiple HPC systems include reliability growth models to characterize temporal and spatial properties of failures and impact on DRAM reliability [20]. The analysis and reliability metrics derived from the data are system-specific and can not be subsumed by temporal characteristics such as MTBF. DRAM-focused insights presented here focus on different systems, and different systems. This work has introduced two new metrics to characterize temporal and spatial properties of failures and impact on DRAM reliability. The analysis and reliability metrics derived from the data are system-specific and can not be subsumed by temporal characteristics such as MTBF. The analysis and reliability metrics derived from the data are system-specific and can not be subsumed by temporal characteristics such as MTBF.
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