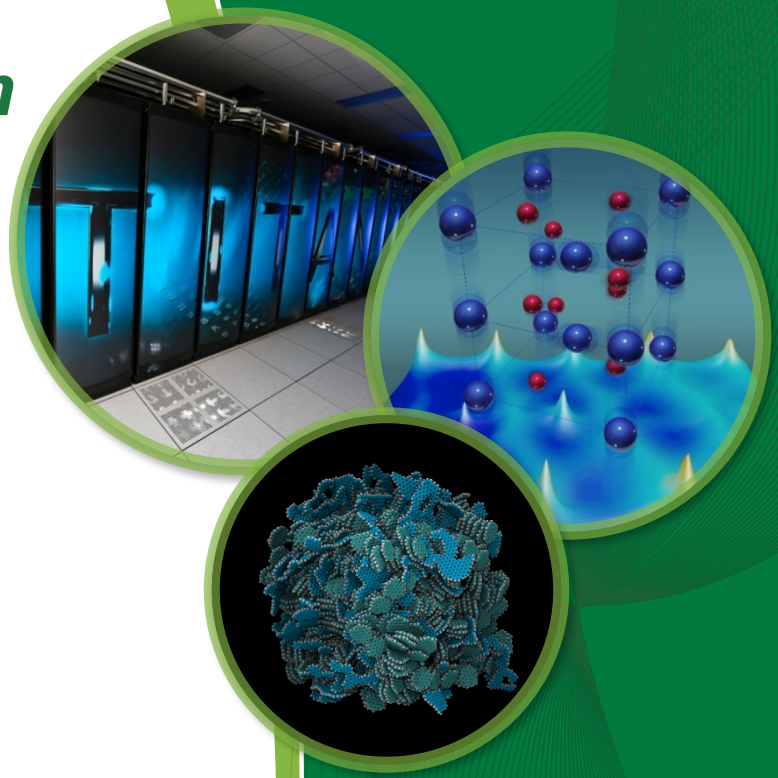
Resilience for Extreme Scale Systems: Understanding the Problem



Christian Engelmann,
Oak Ridge National Laboratory

Work supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, program manager Lucy Nowell and Robinson Pino, under contract number DE-AC05-00OR22725.

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

Motivation

- Resilience in extreme-scale systems is a optimization problem between the key system design and deployment cost factors:
  - Performance, resilience, and power consumption
- The challenge is to build a reliable system within a given cost budget that achieves the expected performance.
- This requires fully understanding the resilience problem and offering efficient resilience mitigation technologies.
  - What is the fault model of such systems?
  - What is the impact of faults on applications?
  - How can mitigation in hard-/software help and at what cost?

Catalog: Characterizing Faults, Errors, and Failures in Extreme-Scale Systems

- Identifies, categorizes and models the fault, error and failure properties of US Department of Energy (DOE) systems
- Develops a fault taxonomy, catalog and models that capture the observed and inferred conditions in current systems and extrapolates this knowledge to exascale systems
- Provides a clear picture of the fault characteristics in the DOE computing environments.
- Improves resilience through reliable fault detection at an early stage and actionable information for efficient mitigation

Principal Investigators:

- *Christian Engelmann, Oak Ridge National Laboratory (ORNL) - Lead*
- *Franck Cappello, Argonne National Laboratory (ANL)*
- *Ignacio Laguna, Lawrence Livermore National Laboratory (LLNL)*

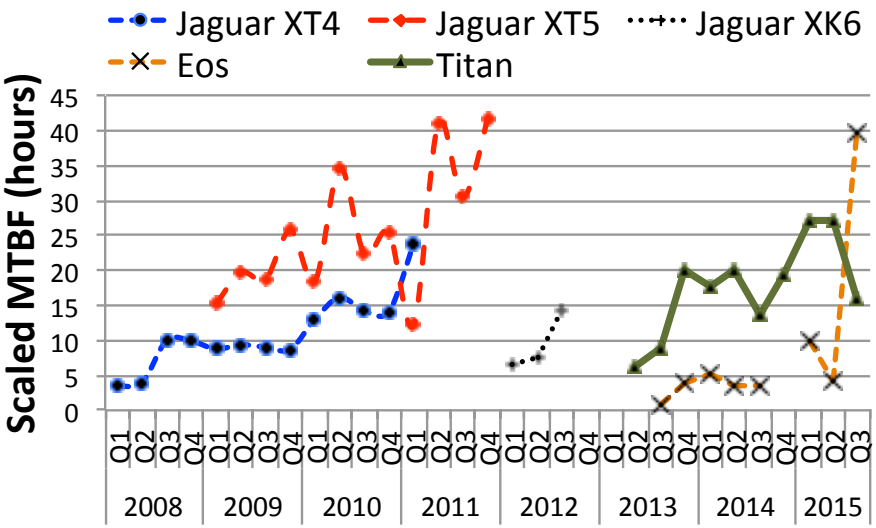
Argonne NATIONAL LABORATORY

Lawrence Livermore National Laboratory

OAK RIDGE National Laboratory

Reliability of HPC systems: Large-term Measurement, Analysis, and Implications (1/4)

- 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

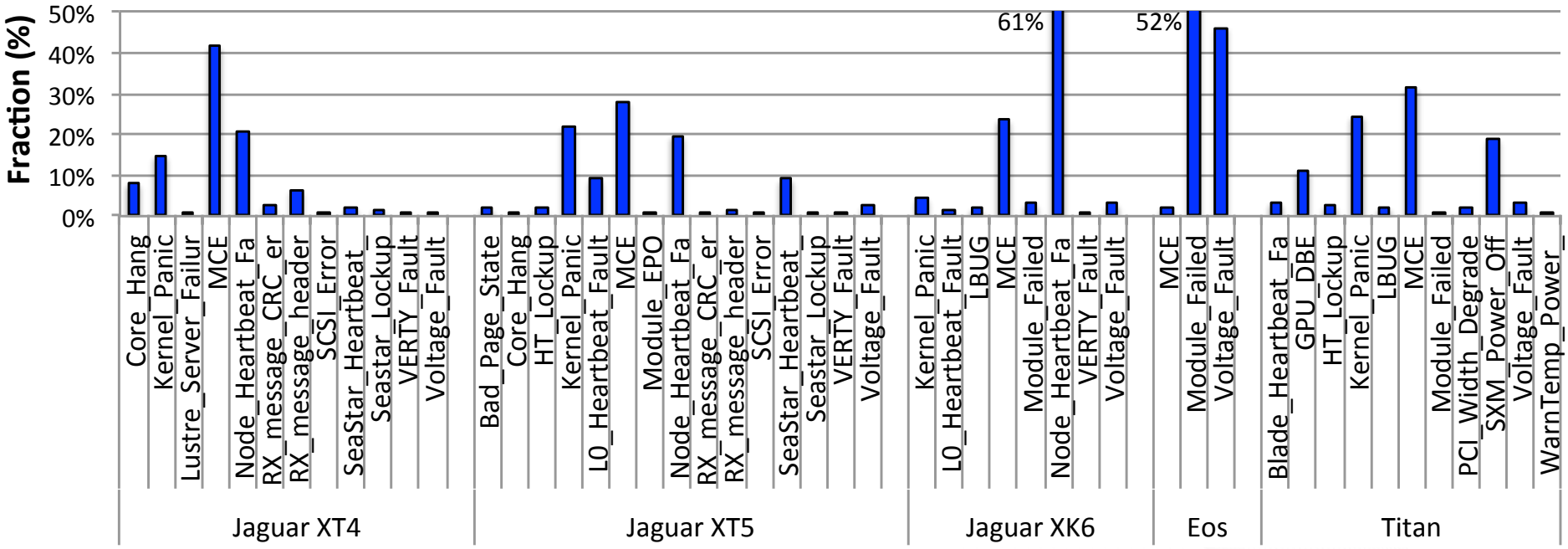


Scale-normalized MTBF of each system over time (averaged quarterly)

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

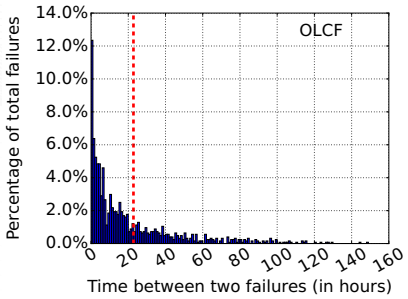
Saurabh Gupta, Devesh Tiwari, Tirthak Patel, and Christian Engelmann. **Reliability of HPC systems: Large-term Measurement, Analysis, and Implications.** SC'17.

Reliability of HPC systems: Large-term Measurement, Analysis, and Implications (2/4)

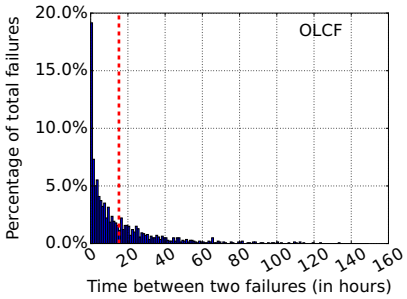


Fraction of each failure type on the studied systems

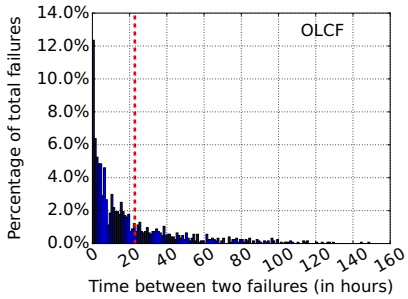
Reliability of HPC systems: Large-term Measurement, Analysis, and Implications (3/4)



(a) Jaguar XT4

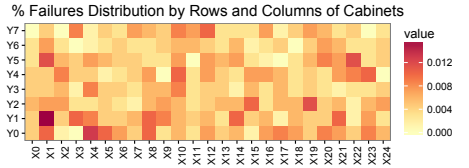


(b) Titan

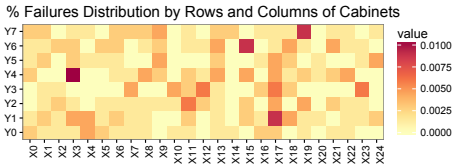


(c) Eos

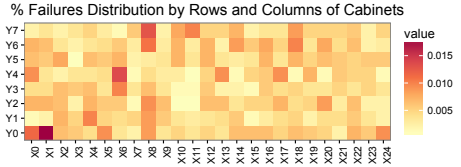
Failure inter-arrival time for 3 studied systems (MTBF as red vertical line)



(a) Jaguar XT5

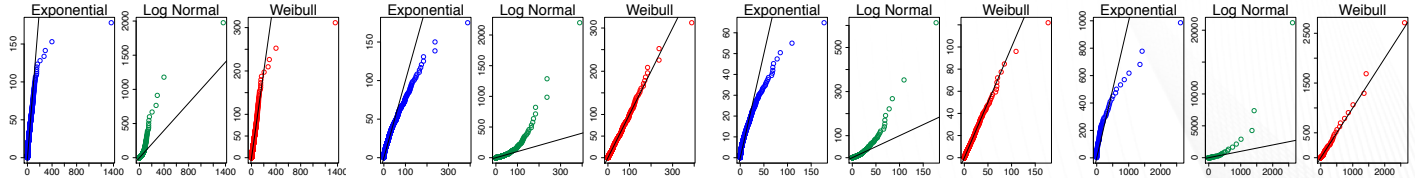


(b) Jaguar XK6



(c) Titan

Spatial distribution of failures among cabinets for 3 studied systems



(a) Jaguar XT4 (b) Jaguar XT5 (c) Jaguar XK6 (d) Eos

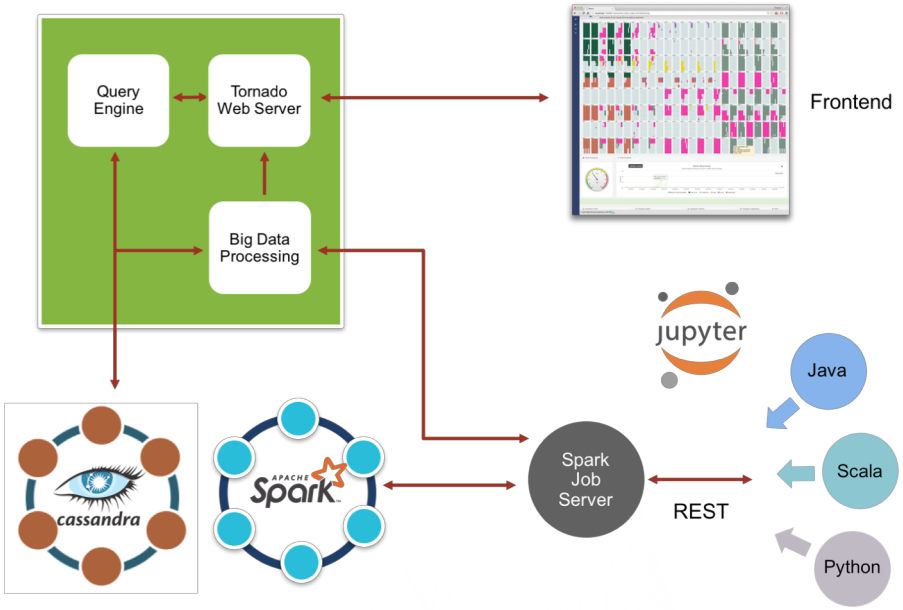
QQ-plots showing goodness of fit for the failure inter-arrival times for 4 studied systems with different failure probability density functions (Weibull fits best)

Reliability of HPC systems: Large-term Measurement, Analysis, and Implications (4/4)

- MTBF can change significantly over time, with often a non-monotonic trend, which makes it averaged over lifetime an unattractive choice as metric.
- A set of dominant failure types is common across systems. Only very few types contribute most of the failures for each system.
- The degree of temporal locality, which is very high in all studied systems, captures temporal characteristics better than auto-correlation.
- Several failure types are likely to reoccur within a short amount of time (e.g., one hour) on all systems.
- Spatial locality exists in all systems at all granularities.
- 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.
- The studied systems best fit the Weibull distribution.

LogSCAN: A Big Data Analytics Framework for HPC Log Data

- Improved the analysis of supercomputer reliability, availability and serviceability (RAS) logs using modern Big Data analytics tools to understand resilience issues at scale
- Assessing system status, identifying reliability event patterns and correlating events with application performance improves supercomputer efficiency by identifying error and failure modes
- Created a multi-user Big Data analytics framework –Log processing by Spark and Cassandra-based ANalytics (LogSCAN) – in ORNL's private cloud of Compute and Data Environment for Science (CADES)

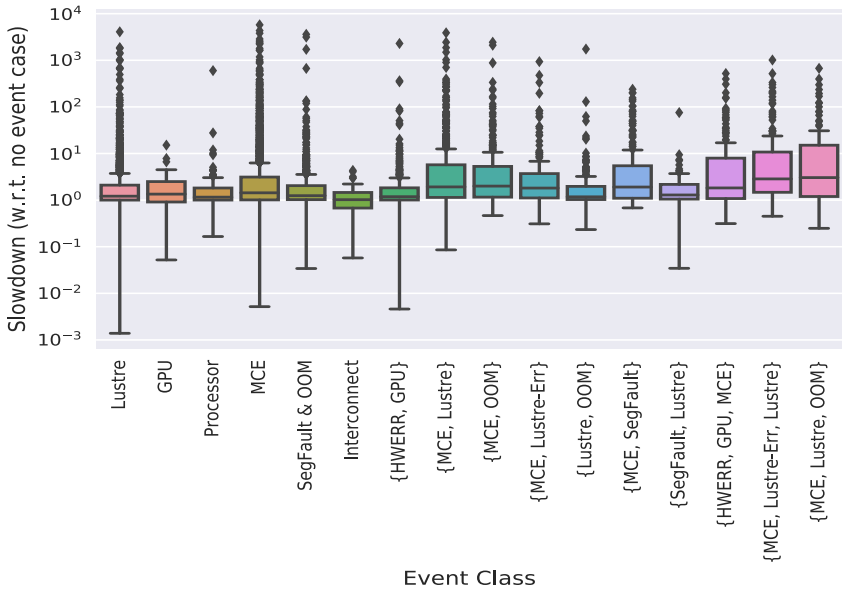


LogSCAN uses Apache Spark and Cassandra to analyze logs and health data in a combined offline/online fashion

B. H. Park, Y. Hui, S. Boehm, R. A. Ashraf, C. Layton, and C. Engelmann. **A Big Data Analytics Framework for HPC Log Data: Three Case Studies Using the Titan Supercomputer Log**. HPCMASPA'18.

Analyzing the Impact of System Reliability Events on Applications in the Titan Supercomputer

- Created an understanding of the impact of non-fatal reliability events on scientific application performance.
- Co-analyzed 13 months of application scheduling and reliability event data from ORNL's Titan supercomputer
- Studied the performance characteristics of scientific applications which are most affected by RAS events
- Identified system components that are most likely to impact the performance of scientific applications
- Quantified the slowdown of scientific application jobs due to RAS events from different components



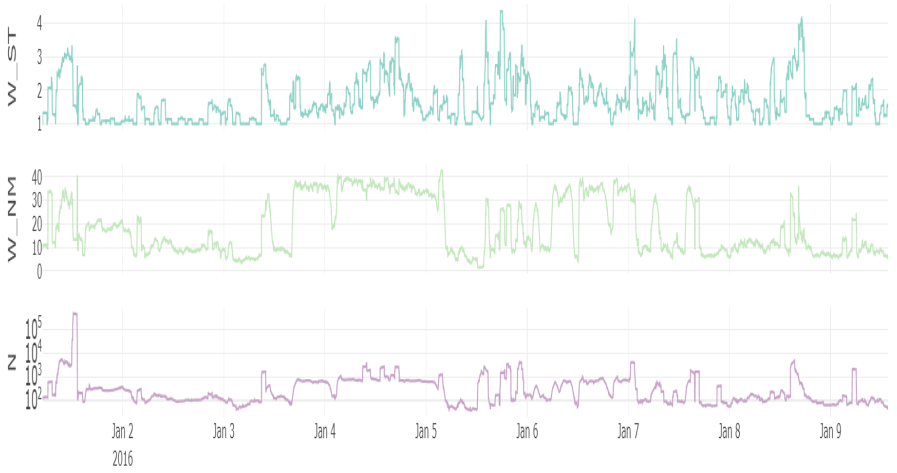
Slowdown assessment of applications executed on ORNL's Titan supercomputer due to reliability issues in various system components:

R. A. Ashraf and C. Engelmann. **Analyzing the Impact of System Reliability Events on Applications in the Titan Supercomputer**. FTXS'18.

OAK RIDGE National Laboratory

System Information Entropy: A Comprehensive Informative Metric for Analyzing HPC System Status (1/2)

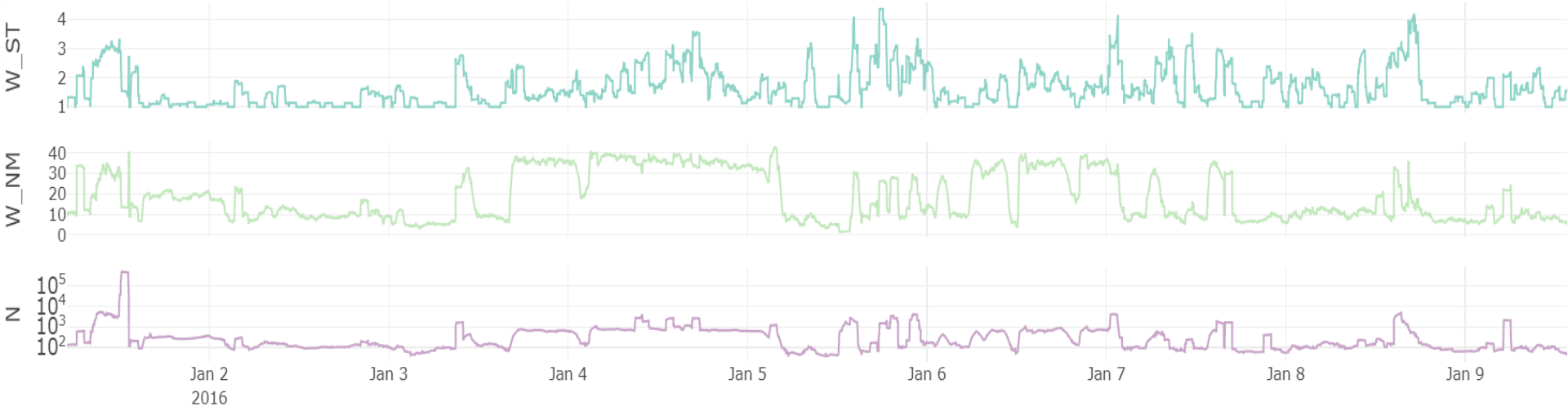
- Created the System Information Entropy (SIE) metric to concisely represent health status in a time series
- This metric aids operators in assessing system health status by easily and quickly identifying its changes
- Used ORNL's multi-user Big Data analytics framework (LogSCAN)
- Analyzed 3+ years of log data from ORNL's Titan (Jan. 2015 – Mar. 2018)
- Applied Principal Component Analysis and Shannon Entropy Theory to calculate SIEs based on different record vs. feature views of the data



SIE with Source Type layout (top), SIE with Nodal Map layout (middle), and Total event count (bottom)

Y. Hui, B. Park, and C. Engelmann. **A Comprehensive Informative Metric for Analyzing HPC System Status using the LogSCAN Platform**. FTXS'18.

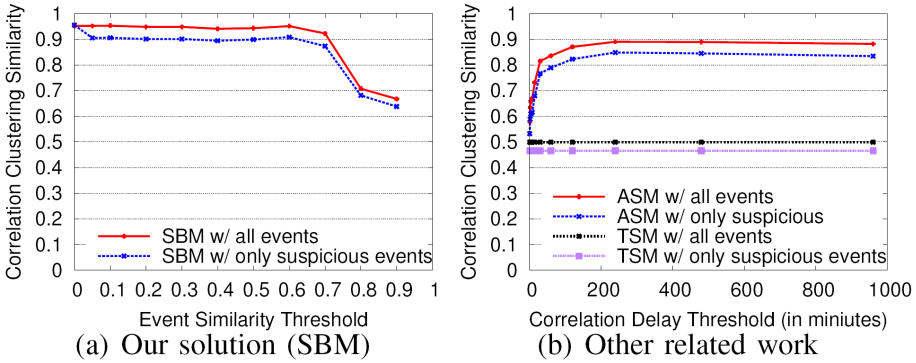
System Information Entropy: A Comprehensive Informative Metric for Analyzing HPC System Status (2/2)



SIE with Source Type layout (top),
SIE with Nodal Map layout (middle), and
Total event count (bottom)

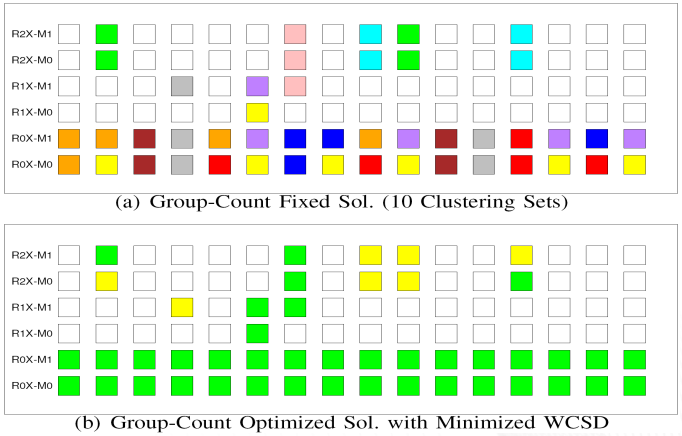
LogAider: Offline Mining of Correlations in Supercomputer System Logs

- A tool that for mining fault, error and failure correlations in supercomputer system logs
- Explores correlation across fields
- Permits spatial correlation analysis
- Enables temporal correlation analysis
- Helps in identifying propagation chains



(a) Our solution (SBM) (b) Other related work

Accuracy of temporal correlation analysis



(a) Group-Count Fixed Sol. (10 Clustering Sets)

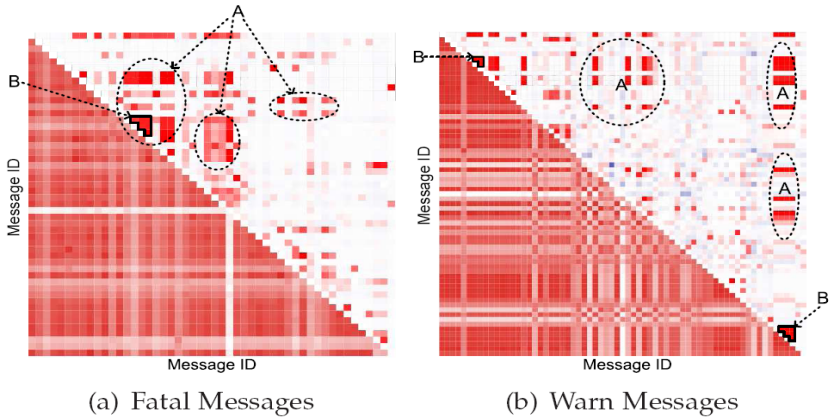
(b) Group-Count Optimized Sol. with Minimized WCSD

Spatial K-means clustering based on torus network

Sheng Di, Rinku Gupta, Marc Snir, Eric Pershey, and Franck Cappello. **LOGAIDER: A tool for mining potential correlations of HPC log events.** CCGrid'17.

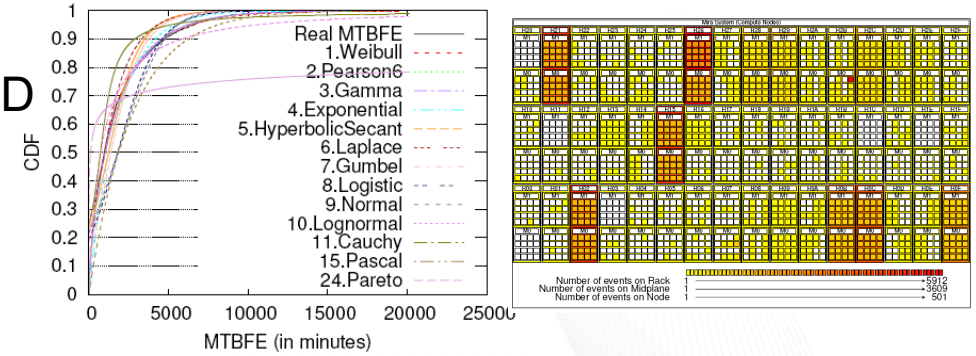
Exploring Properties and Correlations of Fatal Events in a Large-Scale HPC System

- Analyzed 5 years of of system logs from ALCF's Mira supercomputer using the LogAider tool for mining correlations
- Takeaways:
  - ~80% of the fatal events are covered by ~20% of attribute values
  - Strong correlations between different message IDs, especially between fatal and warn message ID
  - Mean time between fatal events (MTBFE) is 1.3 days
  - Weibull is best fit for most events



(a) Fatal Messages (b) Warn Messages

Fatal and warn messages both exhibit transitivity property

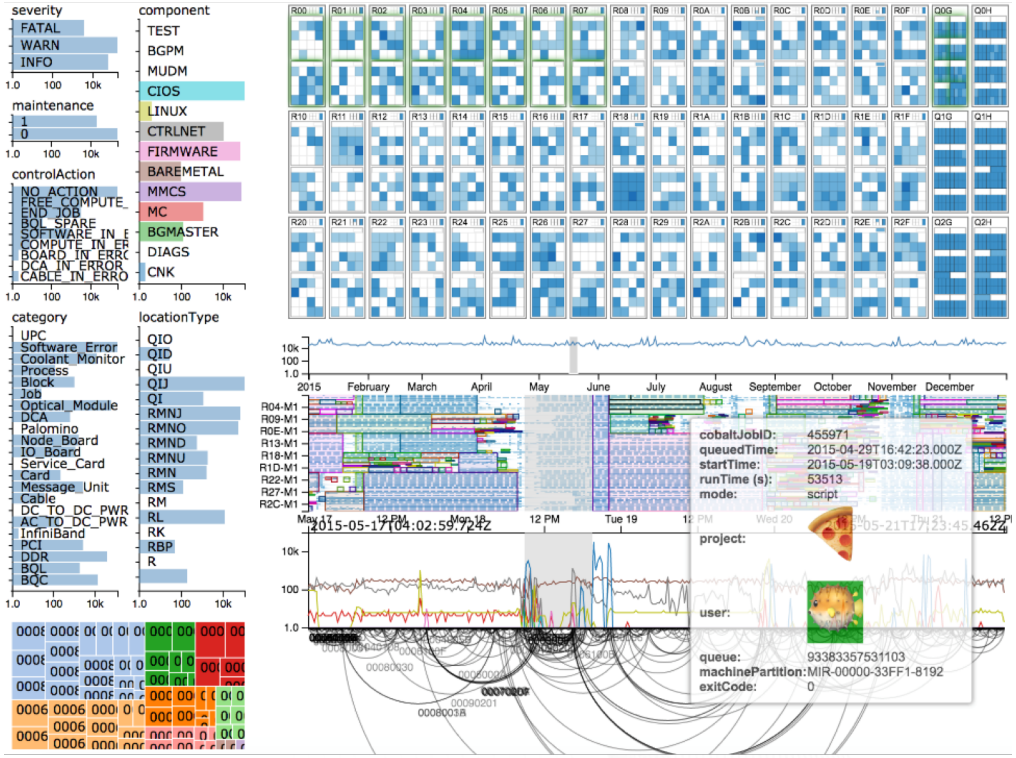


Explore best fit distribution of fatal event intervals (left) and spatial correlation of fatal events (right)

Sheng Di, Hanqi Guo, Rinku Gupta, Eric R. Pershey, Marc Snir, and Franck Cappello. **Exploring Properties and Correlations of Fatal Events in a Large-Scale HPC System**. in TPDS

La VALSE: Visual Analysis of Logs in Supercomputers (1/2)

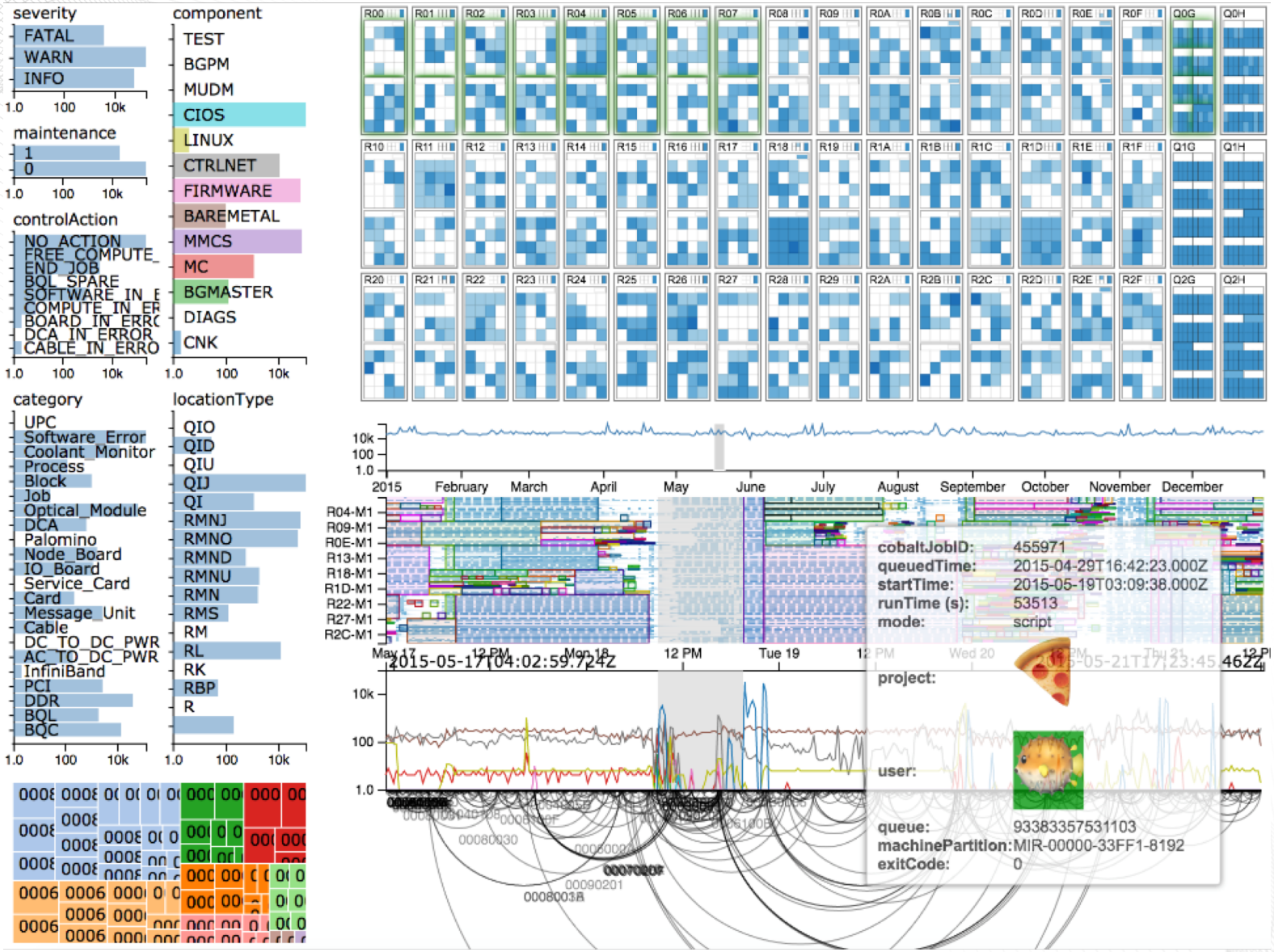
- A tool for interactive exploration of logs with automatic analysis
- For system admins and users
- Multidimensional view: Filtering 11M events with attributes
- Machine view: Visualizing and querying 100K components with levels-of-details rendering
- Timeline view: Visualizing trends and individual events with novel visual designs
- Automatic analysis: Spatiotemporal correlation with dynamic time warping and longest common subsequence



La VALSE's graphical user interface

Principal Investigators: ***Hanqi Guo, Sheng Di, Rinku Gupta, and Franck Cappello – ANL***

La VALSE: Visual Analysis of Logs in Supercomputers (2/2)

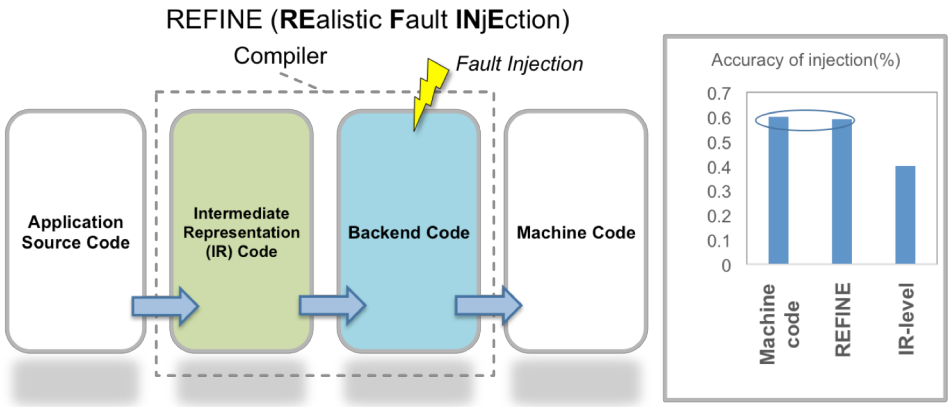


C. Engelmann. Resilience for Extreme Scale Systems: Understanding the Problem. SIAM CSE'19, Spokane, WA, USA.

REFINE: Compiler-level Fault Injection with High Accuracy and Flexibility

Lawrence Livermore National Laboratory

- Allows error propagation studies, while maintaining a high accuracy when injecting faults
- REFINE is a fault injector that instruments applications at the compiler backend code
- Because instrumentation occurs at the backend, it is closer to machine code and as a result can consider a larger set of instructions when injecting faults
- It provides more accurate and more efficient fault injection than state-of-the-art IR or application-level fault injection tools



Instrumentation method and accuracy of REFINE

Giorgis Georgakoudis, Ignacio Laguna, Dimitrios S. Nikolopoulos, Martin Schulz. **REFINE: Realistic Fault Injection via Compiler-based Instrumentation for Accuracy, Portability and Speed.** SC'17.

OAK RIDGE National Laboratory

FlipTracker: Understanding Natural Error Resilience in HPC Applications

Lawrence Livermore National Laboratory

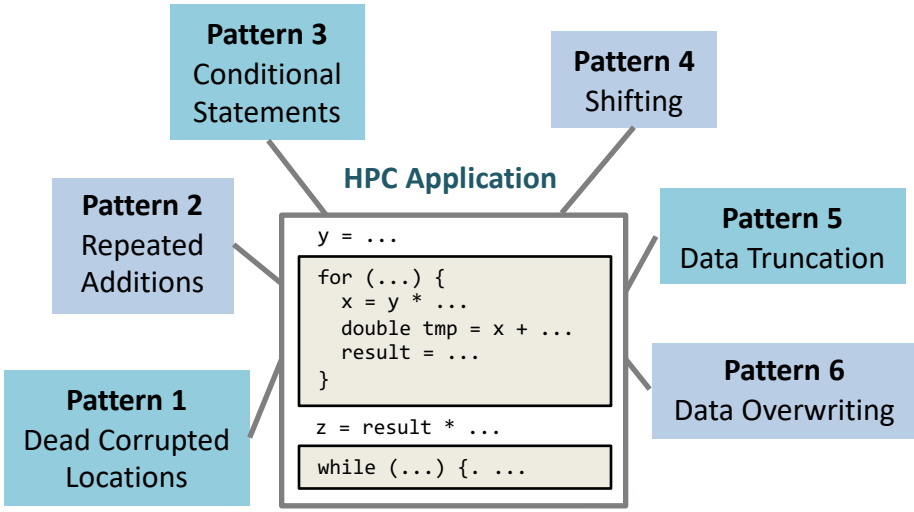
Technical University of Munich

University of California UC Merced

- Identified computational patterns that explain why some science applications are naturally resilient to errors
- FlipTracker is a framework that identifies naturally resilient code patterns

> A resilient pattern example is ***repeated additions***, where an erroneous value is amortized by many correct addition operations

- We found six computational resilience patterns in ten science programs:
  - LULESH, CG, MG, LU, BT, IS, DC, SP, FT, KMEANS



HPCwire

This article was featured in HPCwire's "***What's New in HPC Research***" article on Sep/18/18.

Luanzheng Guo, Dong Li, Ignacio Laguna, Martin Schulz. FlipTracker: **Understanding Natural Error Resilience in HPC Applications.** SC'18.

OAK RIDGE National Laboratory

Questions?

- Web site: https://ornlwiki.atlassian.net/wiki/display/CFEFIES
- Contact:
  - Christian Engelmann, engelmannc@ornl.gov





