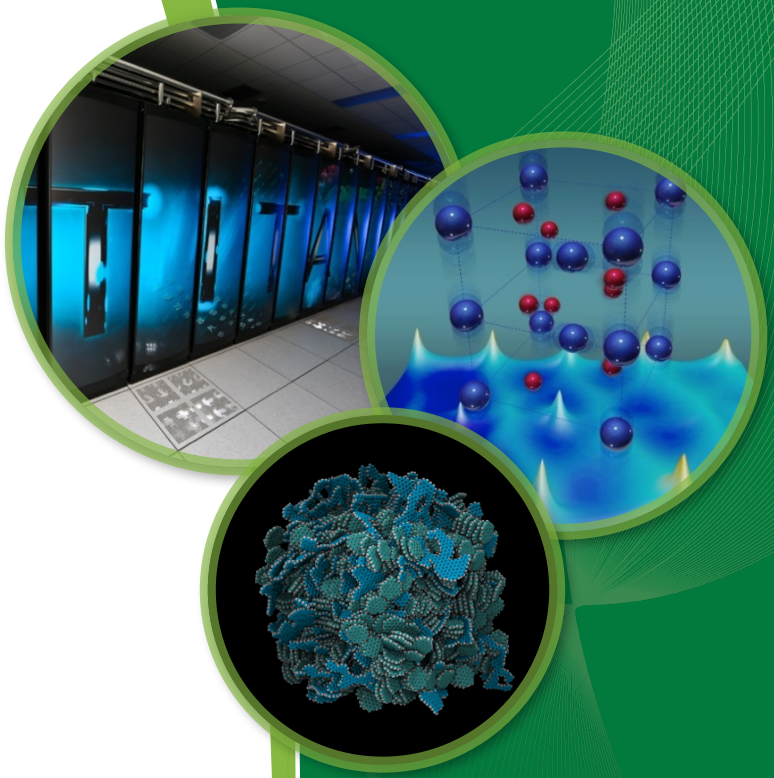
Characterizing Faults, Errors & Failures in Extreme-Scale Computing Systems



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, under contract number DE-AC05-00OR22725.

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?

Objectives

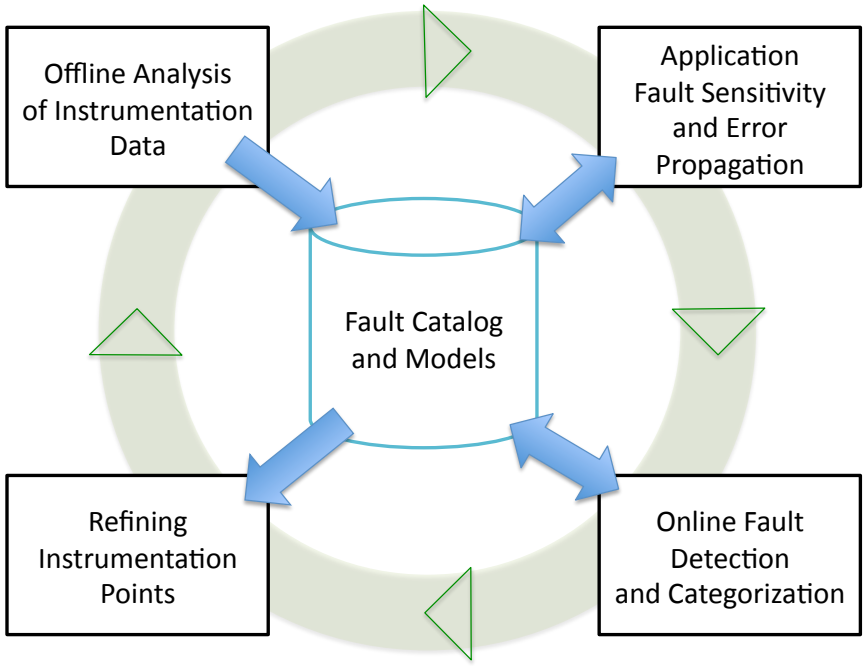
- This project identifies, categorizes and models the fault, error and failure properties of US Department of Energy (DOE) systems.
- It develops a fault taxonomy, catalog and models that capture the observed and inferred conditions in current systems and extrapolates this knowledge to exascale systems.
- This project provides a clear picture of the fault characteristics in the DOE computing environments.
- It 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)*
- *Martin Schulz, Lawrence Livermore National Laboratory (LLNL)*

Approach

- Create an HPC fault taxonomy
- Employ in-breadth offline data gathering and analysis techniques
- Create an initial HPC fault catalog and models of DOE systems
- Employ realistic in-depth fault vulnerability and error propagation studies with applications
- Employ in-breadth online data gathering and analysis techniques
- Update the HPC fault catalog and models of DOE systems
- Refine instrumentation points for improved fault detection



Iterative approach of developing the catalog & models

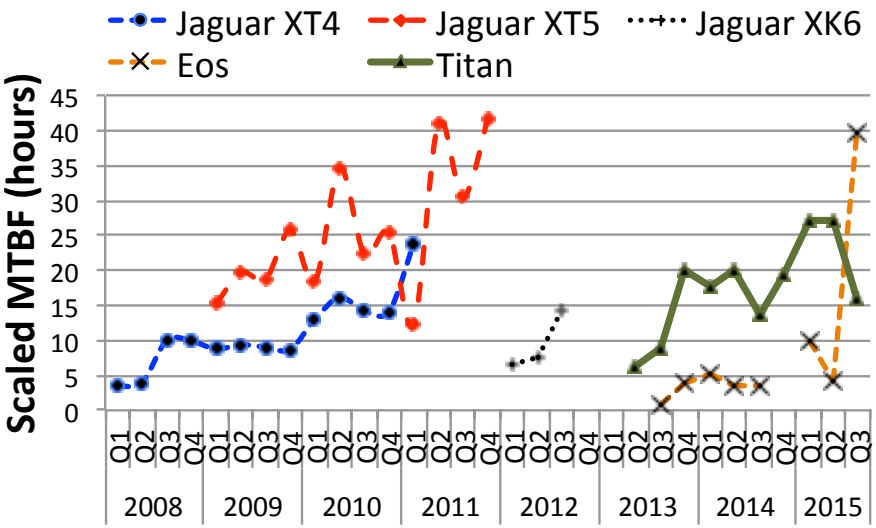
HPC Fault Taxonomy

- Clarified common terms, metrics and methods, such as:
  - {benign,dormant,active}
    {permanent,transient,intermittent}
    {hard,soft} fault
  - {undetected,detected}
    {unmasked,masked}
    {hard,soft} error
  - {undetected,detected}
    {permanent,transient,intermittent}
    {complete,partial,Byzantine} failure
  - {true,false} {positive,negative} error/failure detection
  - Error propagation vs. failure cascade
  - Reliability vs. availability

Principal Investigator: ***Christian Engelmann – ORNL***

Catalog: Oak Ridge Leadership Computing Facility (OLCF)

- 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 faults
- Analyzed the evolution of temporal and spatial fault behavior over the years
- Analyzed the correlation of different fault types
- Compared the mean-time between faults of the 5 systems

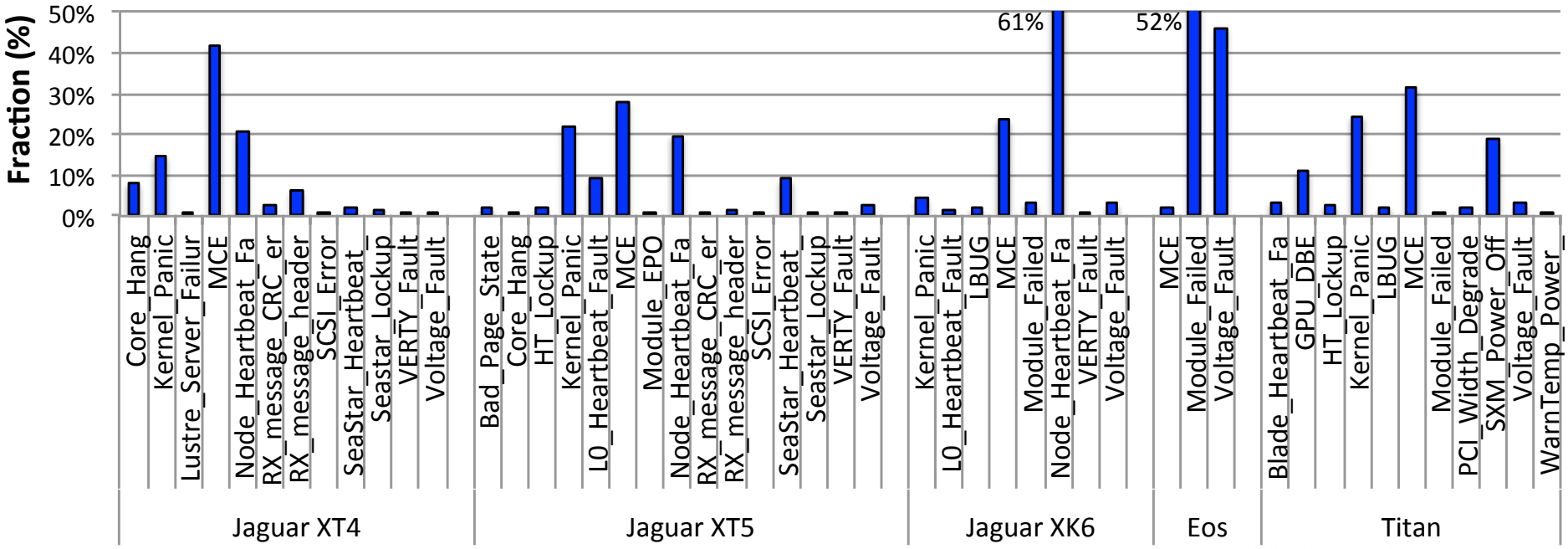


Scale-normalized MTBF of each system over time (averaged quarterly)

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

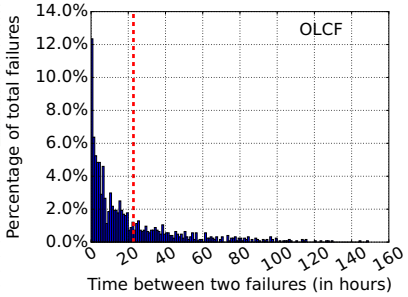
Saurabh Gupta, Devesh Tiwari, Tirthak Patel, and Christian Engelmann. **Reliability of HPC systems: Large-term Measurement, Analysis, and Implications.** SC'17.

Catalog: Oak Ridge Leadership Computing Facility

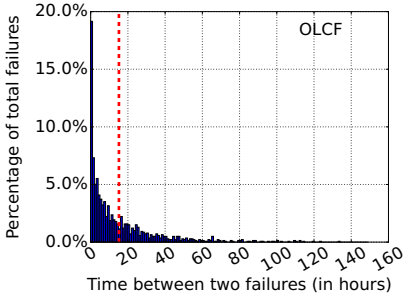


Fraction of each failure type on the studies systems

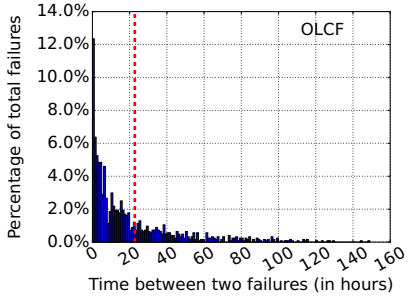
Catalog: Oak Ridge Leadership Computing Facility



(a) Jaguar XT4

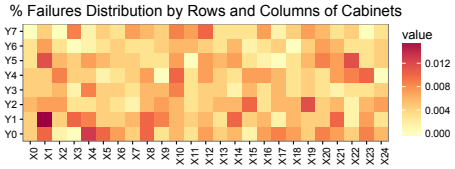


(b) Titan

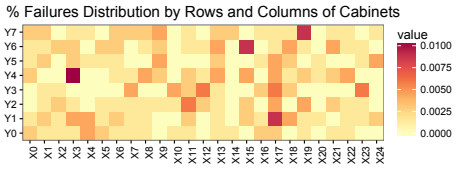


(c) Eos

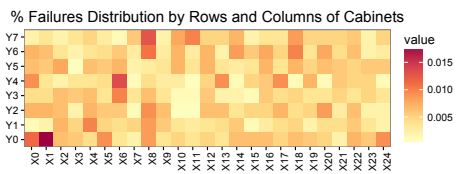
Failure inter-arrival time for 3 studied systems (MTBF as red vertical line)



(a) Jaguar XT5

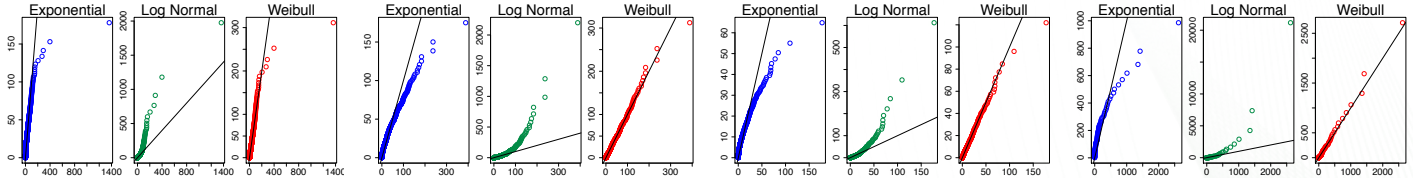


(b) Jaguar XK6



(c) Titan

Spatial distribution of failures among cabinets for 3 studied systems



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

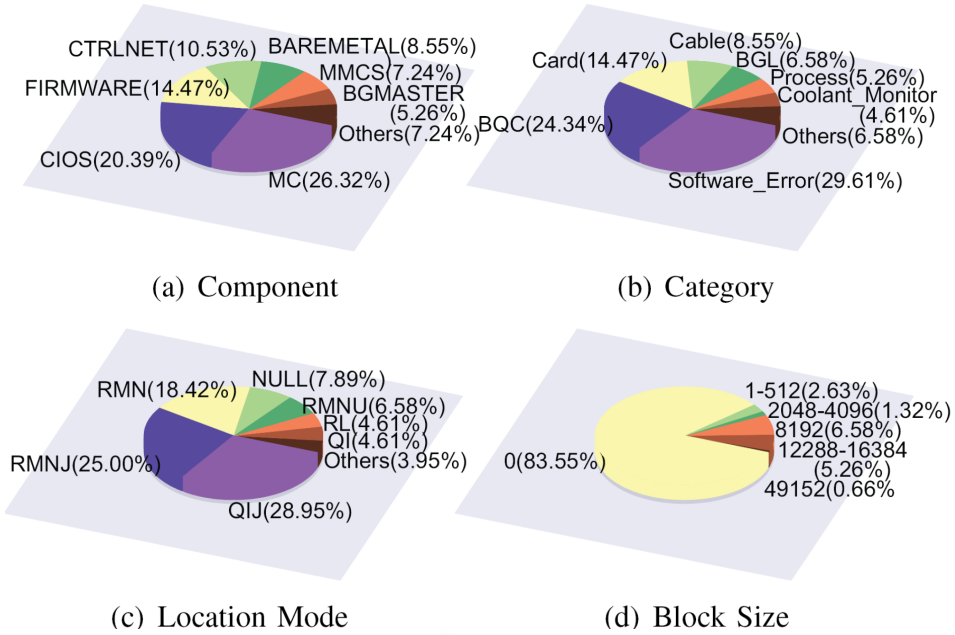
QQ-plots showing goodness of fit for the failure inter-arrival times for 4 studied systems with different failure probability density functions (Weibull fits best)

Catalog: Oak Ridge Leadership Computing Facility - Lessons learned

- 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.

Catalog: Argonne Leadership Computing Facility (ALCF)

- Analyzed 1 year of of system logs from ALCF's Mira supercomputer
- Identified the frequency of fatal events based on different components and categories
- Identified the probability distribution of events
- Created daily and monthly statistics on fatal events and spatial correlations
- Performed across-field correlation of events
- Performed spatial correlation of events based on location in the torus network
- Performed temporal correlation of events based on event similarity



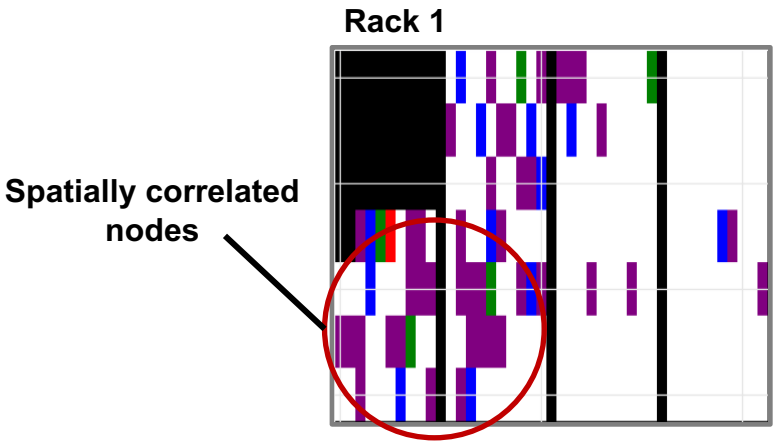
Daily count of the filtered fatal events (4min filtering interval)

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

Catalog: Livermore Computing

Lawrence Livermore National Laboratory

- Analyzed 1+ year of of memory error logs from LLNL's Linux cluster computing environments
- Focused on correctable single bit errors as a symptom
- Found spatial correlation of memory errors in the form of areas with significantly higher error rates
- Ongoing work focuses on correlation of room temperature data with spatial memory error locality

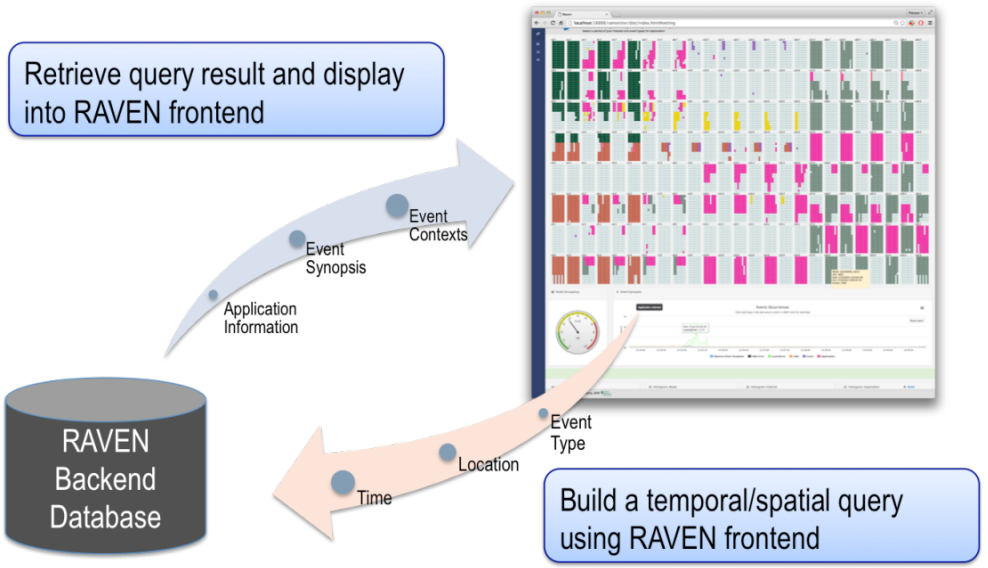


Spatial correlation of memory errors in a set of racks

Ayush Patwari, Ignacio Laguna, Martin Schulz, Saurabh Bagchi. **Understanding the Spatial Characteristics of DRAM Memory Errors in HPC Clusters.** FTXS@HPDC'17.

Offline Log Analytics with RAS Data Analysis Through Visually Enhanced Navigation (RAVEN)

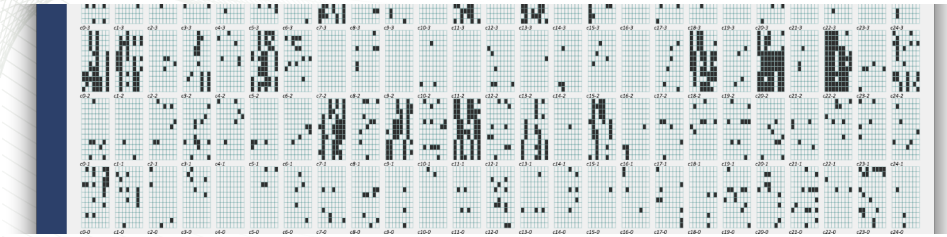
- A tool for studying faults in supercomputers using an interactive visual interface
- Performs context-preserving transformation of events in system logs
- Uses a database designed for scalable/flexible insertion/retrieval for events and query results
- Offers a spatial/temporal query engine that permits user-driven investigations
- Interfaces with Titan at the Oak Ridge Leadership Computing Facility, including its Lustre parallel file system



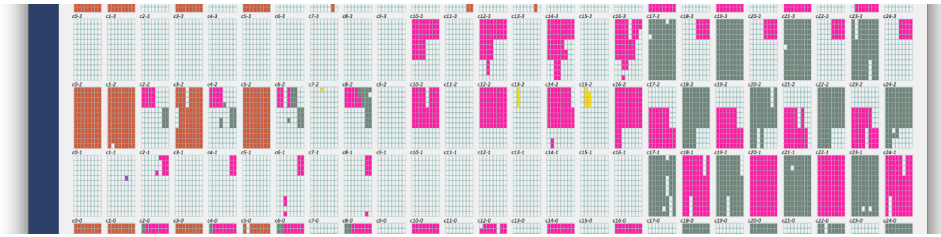
RAVEN framework architecture

Principal Investigators: ***Byung-Hoon (Hoony) Park and Christian Engelmann – ORNL***

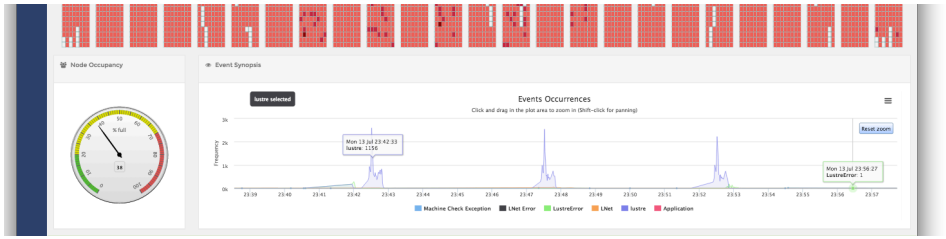
Offline Log Analytics with RAS Data Analysis Through Visually Enhanced Navigation (RAVEN)



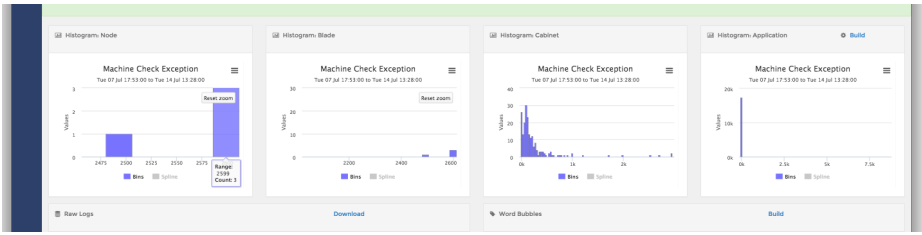
(a) Event occurrences on compute nodes



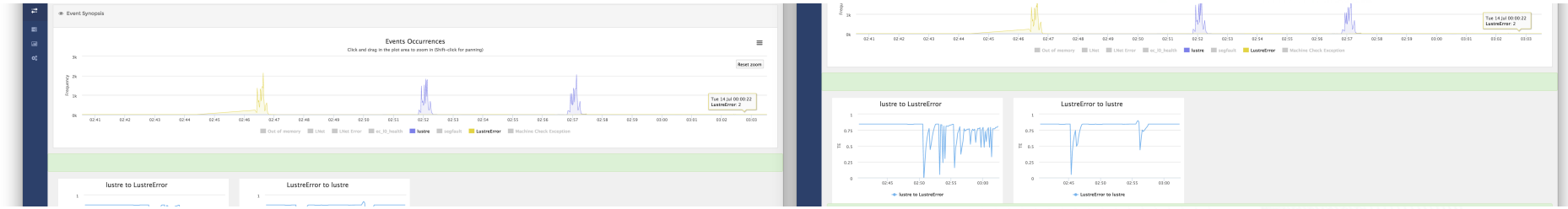
(b) Application displacement on compute nodes



(c) Time series representations at different scales



(d) Histograms of nodes, blades, cabinets, and applications



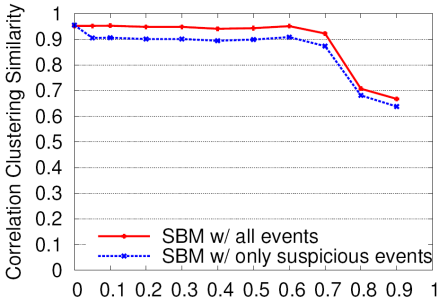
(f) Causality analysis: Occurrences of two event types (left) and their mutual influences measured as transfer entropy

Baseline functionalities (a) and (b), and the extended features (c)-(f) of RAVEN

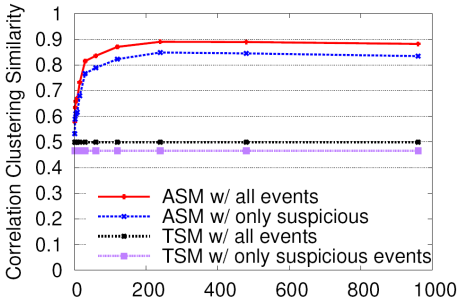
C. Engelmann. A Catalog of Faults, Errors, and Failures in Extreme-Scale Systems. ISC, Frankfurt, Germany, June 18-22, 2017.

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

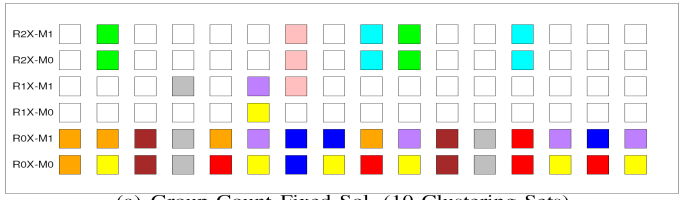


(a) Our solution (SBM)

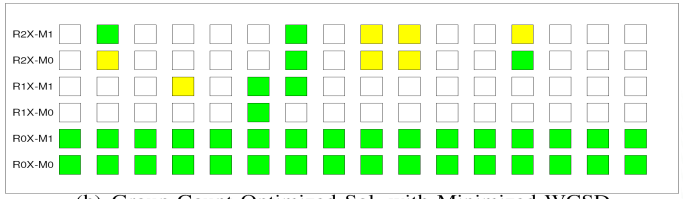


(b) Other related work

Accuracy of temporal correlation analysis



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



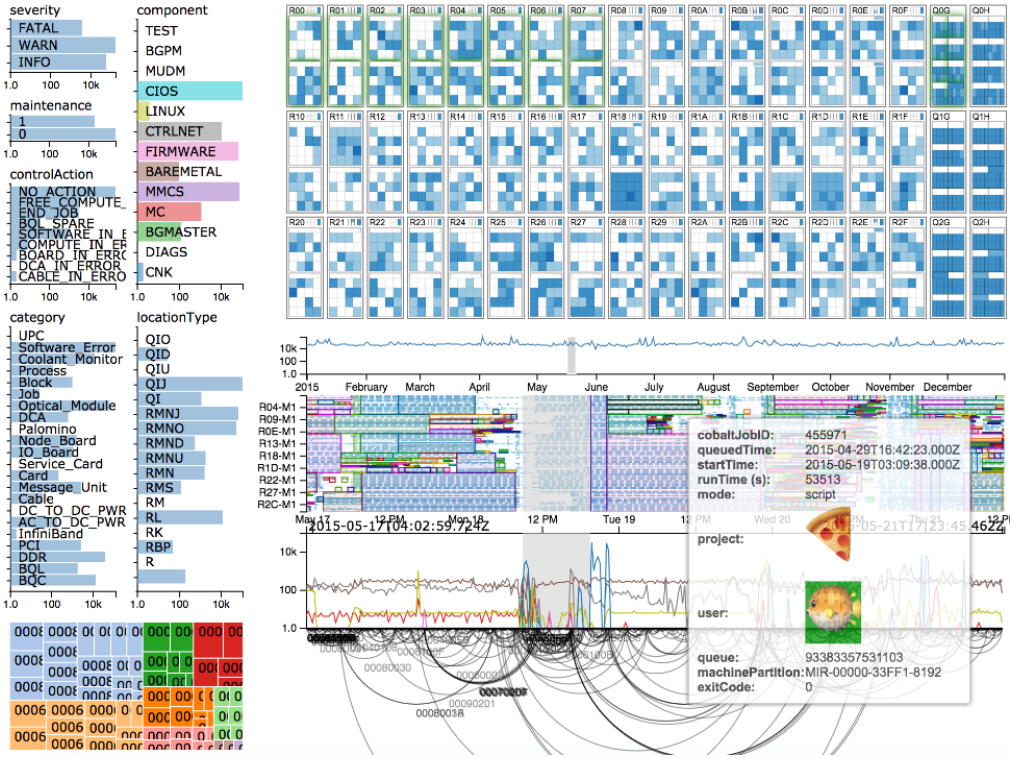
(b) Group-Count Optimized Sol. with Minimized WCSD

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.

La VALSE: Visual Analysis of Logs in Supercomputers

- 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

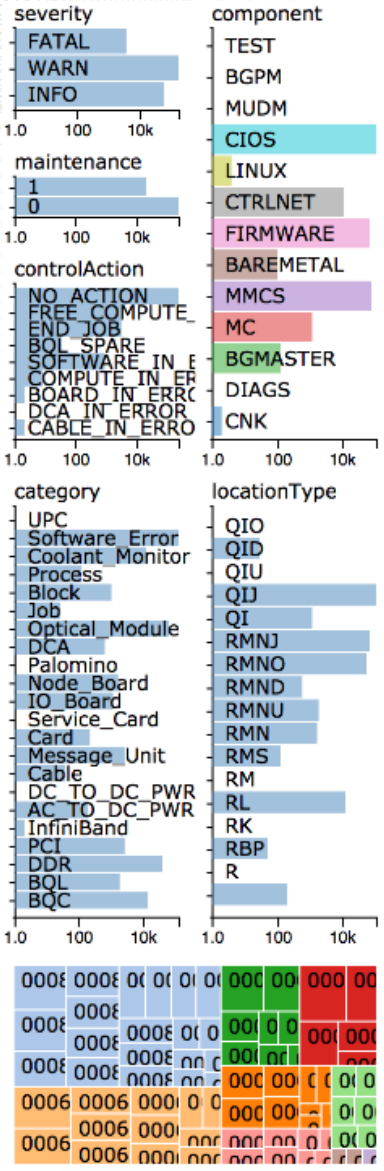


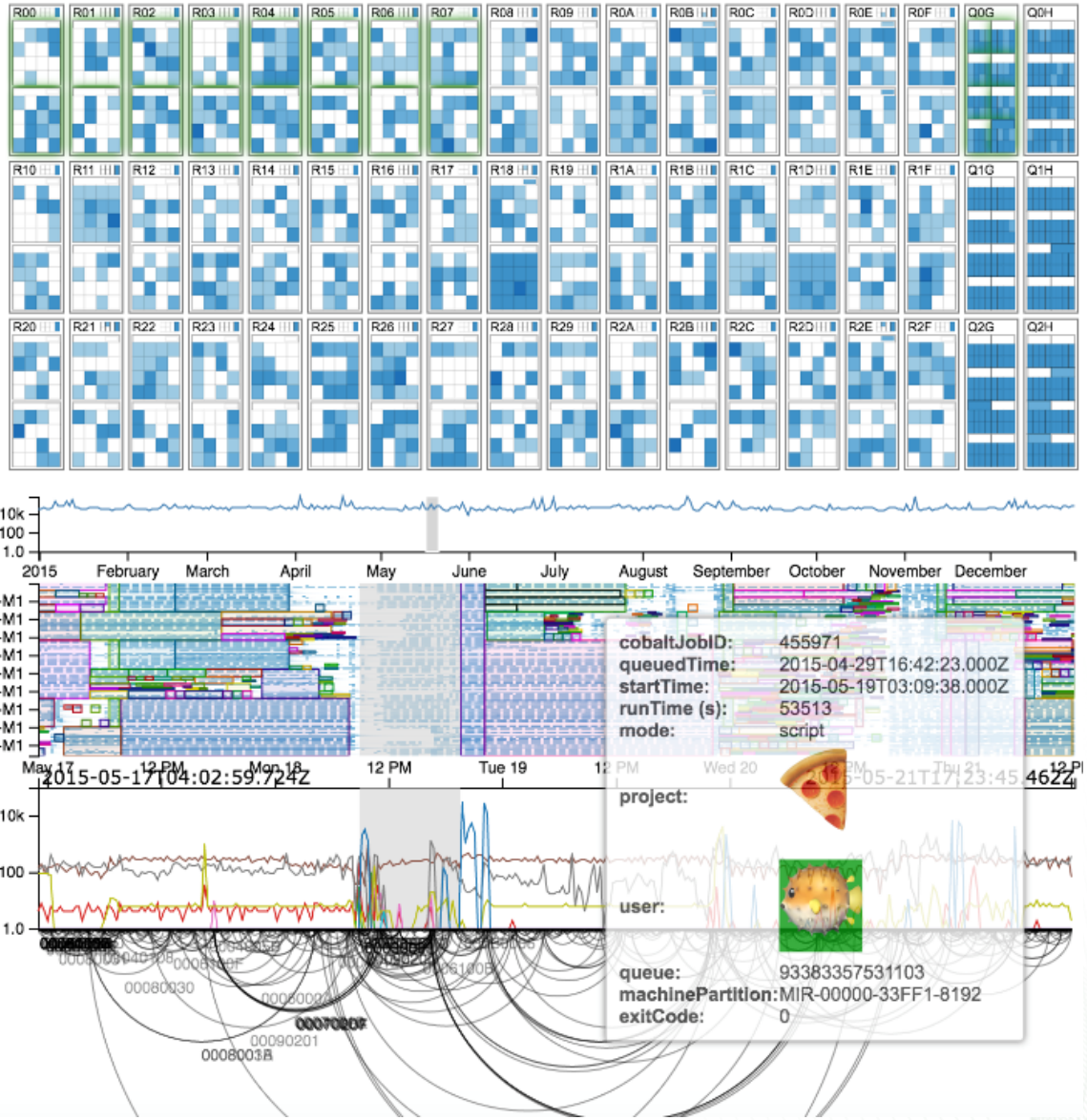
La VALSE's graphical user interface

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

C. Engelmann. A Catalog of Faults, Errors, and Failures in Extreme-Scale Systems. ISC, Frankfurt, Germany, June 18-22, 2017.

La VALSE: Visual Analysis of Logs in Supercomputers



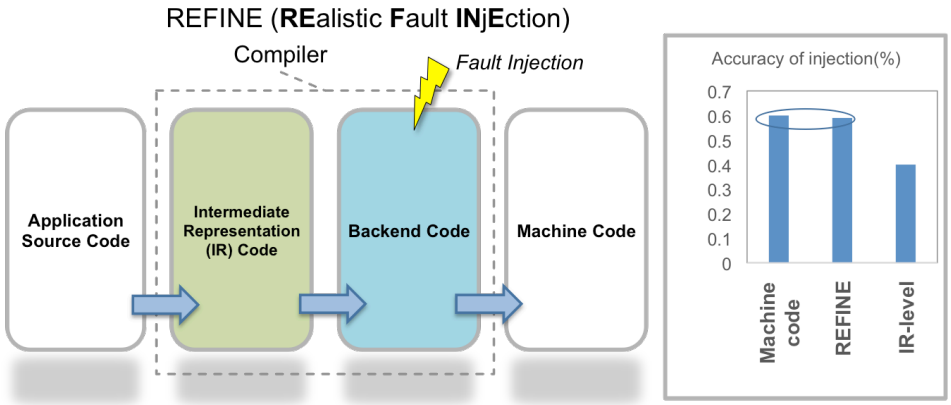


C. Engelmann. A Catalog of Faults, Errors, and Failures in Extreme-Scale Systems. ISC, Frankfurt, Germany, June 18-22, 2017.

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



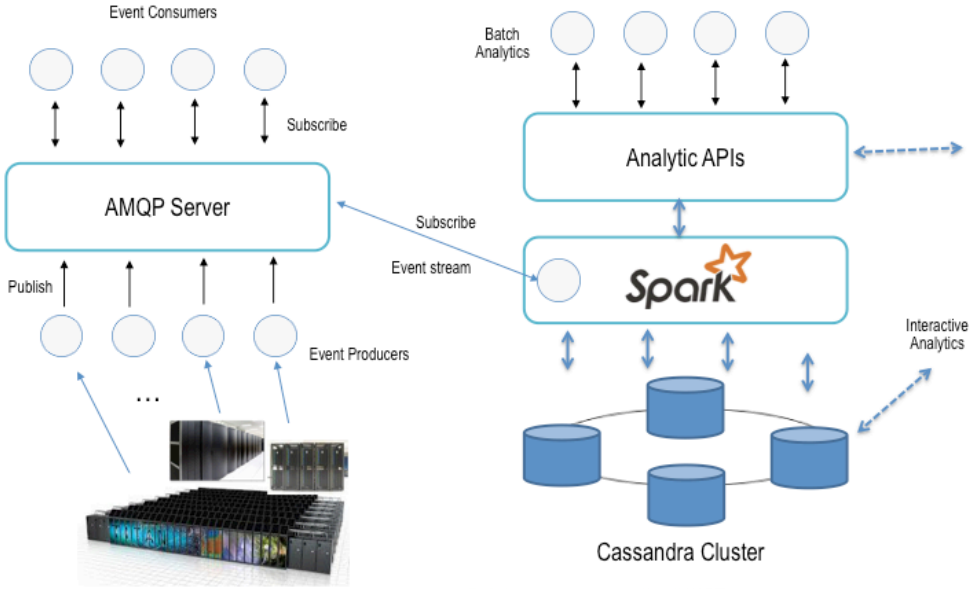
Instrumentation method and accuracy of REFINE

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

An Online Analytics Framework for the "Big Data" Approach to Studying Supercomputer Reliability

- Use Apache Spark and Cassandra to analyze logs and health data in a combined offline/online fashion
- Interface with Advanced Message Queueing Protocol server for real-time data from operational systems
- Utilize local Cloud infrastructure for flexible and on demand computing and storage
- Offer real-time & post-mortem analytics services:
  - Notification about an ongoing system health crisis
  - Notification about root causes of application aborts



Online analytics framework architecture

Principal Investigators: ***Byung-Hoon (Hoony) Park, Saurabh Hukerikar, Ryan Adamson, and Christian Engelmann – ORNL***

Current Status & Ongoing Work

- We have identified, categorized and modeled the fault, error and failure properties of DOE supercomputers.
- We have created a fault taxonomy, catalog and models of the conditions in current systems.
- We have developed a number of tools and frameworks.
- The results are published at SC'17, CCGrid'17 and FTXS@HPDC'17.
- We are still in the process of publishing more results.
- Ongoing work focused on:
  - Error propagation studies with applications
  - Online analysis framework and models
  - Refinement of instrumentation points

Questions?

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