OAK RIDGE National Laboratory

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

Rizwan A. Ashraf and Christian Engelmann,
Computer Science and Mathematics Div.,
Oak Ridge National Laboratory (ORNL), USA.

The 8th Workshop on Fault Tolerance for HPC at eXtreme Scale (FTXS) 2018 @ SC18,

Dallas, TX, 16th November, 2018.

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

U.S. DEPARTMENT OF ENERGY

Introduction: Significance of this Study

- Extreme-scale computing machines, aka, *supercomputers*, are complex systems used to simulate, understand and solve real-world problems.
  - Heterogeneous hardware and software components
  - Large number of independent components
  - Multiple vendors
  - *User applications*
- Reliability, Availability, Serviceability (RAS) log data analyzed in conjunction with application placement and scheduling log data.
  - Understand the relationship between user applications and commonly occurring RAS events in extreme-scale machines.
  - Impact of RAS events on application performance is measured by comparing with cases where no events are recorded.

The Study

- What are RAS events?
  - E.g., errors in memory, errors in parallel file system, errors in the interconnect, etc.
  - Both hardware and software induced events are recorded.
  - In some cases, users also cause RAS events to be recorded (segmentation fault, out-of-memory error, etc.)
- Previous work has shown that memory errors even when correctable through ECC cause severe performance degradation [1].
  - Characterization of resiliency of HPC applications on Blue Waters supercomputer [2].
  - Multiple studies of HPC systems to understand failure mechanisms, error patterns, and to quantify system time-to-failure, e.g., [3], [4].

Complexity of the Study

- Titan supercomputer at Oak Ridge National Laboratory.
  - 18,688-node Cray machine with heterogeneous architecture
  - Each computer node has 16-core AMD Opteron CPU and 32 GB RAM,
  - Each node also has NVIDIA Kepler GPU with 6 GB DDR5 memory connected via PCIe.
  - 3D torus interconnect, Lustre File System with 32 PB of disk space.
  - Cray Linux Environment.
- Complex mix of workloads from a variety of scientific domains including biology, chemistry, computer science, earth science, materials science, fusion, etc. are executed on Titan.
- We analyzed logs for a 13-month period. Most user/project allocations last for one year.

Questions Addressed

- *Is my application running slow* due to reliability issues in the system?
- What are the odds that my application will have events recorded during its execution?
- What components of the system mostly impact application performance?
- What is the expected slowdown in case a RAS event coincides with the execution of an application?
- **Benefits:** Increased system efficiency; Better job scheduling; Mitigation of overheads; Mechanisms in runtime systems and system software to incorporate RAS events;

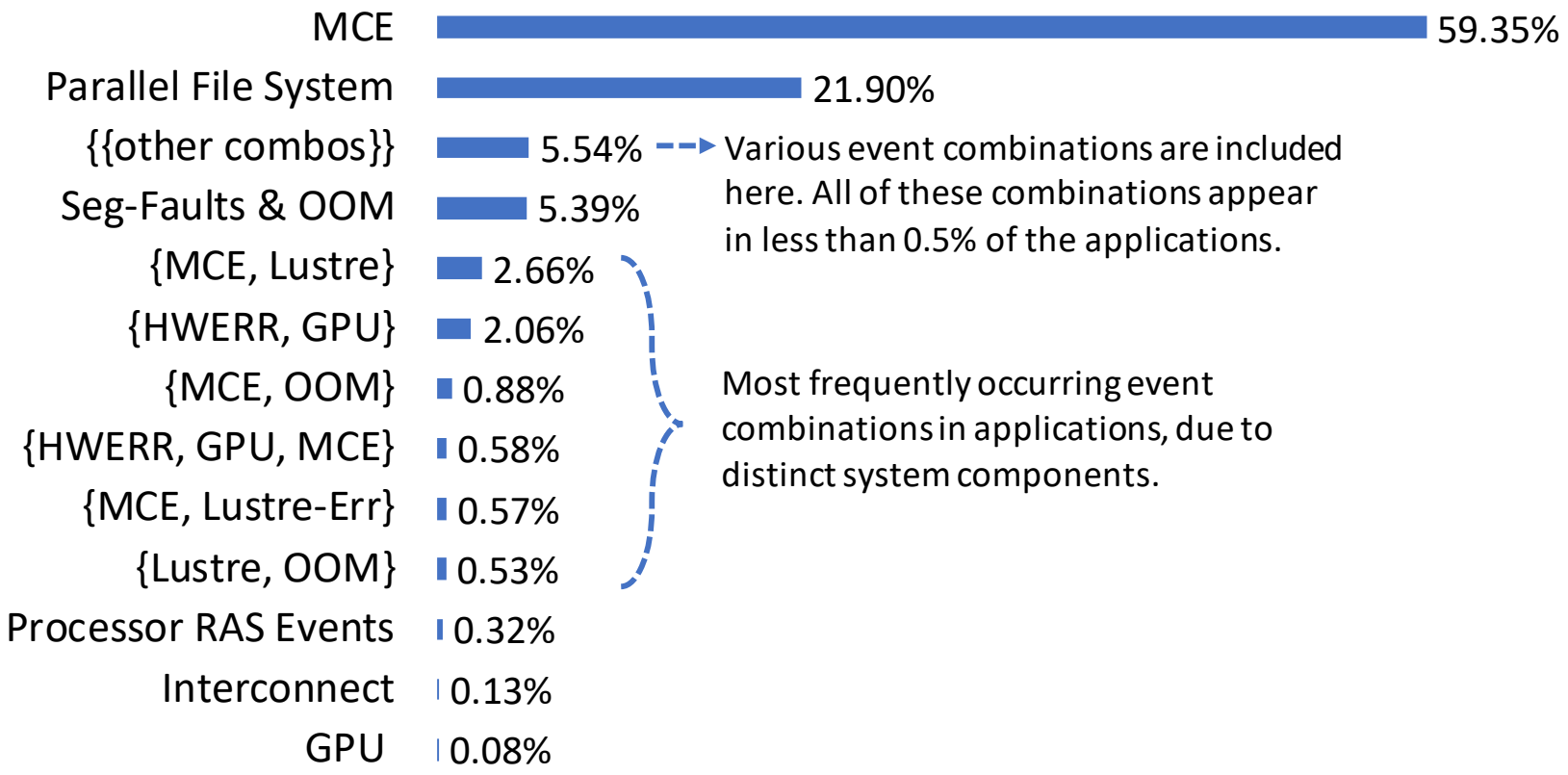
The Data and its Limitations

- **RAS Event log:** type of event, time of the event, identity of the node.
- **Application level placement scheduler (ALPS)**: name and identity of the executable, start time, end time, identities of allocated nodes.
- Data ingested in distributed database (Cassandra DB); Parallel analytics is performed using Spark [5].
- Exit codes of application runs are NOT available. This prohibits us from distinguishing between fatal and non-fatal runs and/or events.
- Resource utilization information is NOT available. E.g., memory utilization, I/O bandwidth, Network utilization, GPU utilization.

Categorization of RAS Events

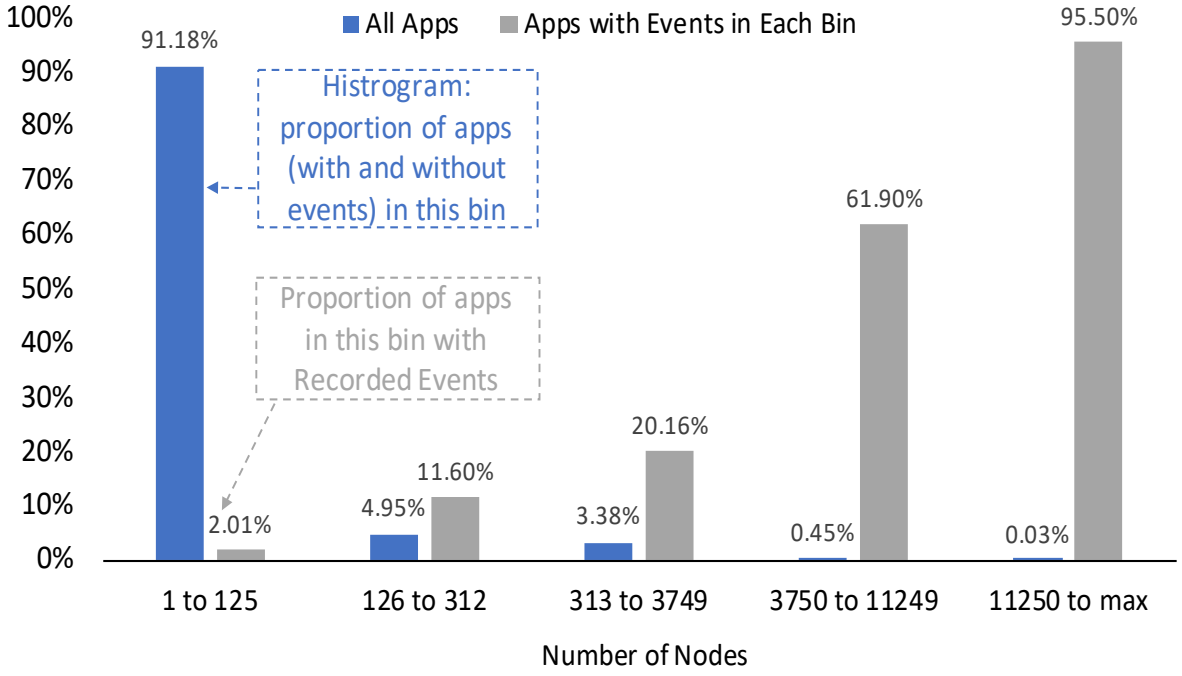
Event Class	Category	Percentage
Parallel File System: system software and Lustre network issues	Hardware/Software	73.7%
Processor: HWERR, Kernel Panic, Graphics Engine Error	Hardware/Software	15.7%
Machine Check Exceptions: mostly CPU memory errors	Hardware	6.5%
Seg-Faults & Out-of-Memory (OOM)	Software	1.9%
GPUs: errors due to user kernels, GPU driver, thermal issues, etc.	Hardware/Software	1.5%
Interconnect: Lane(s) inside a network link can go down	Hardware	0.8%

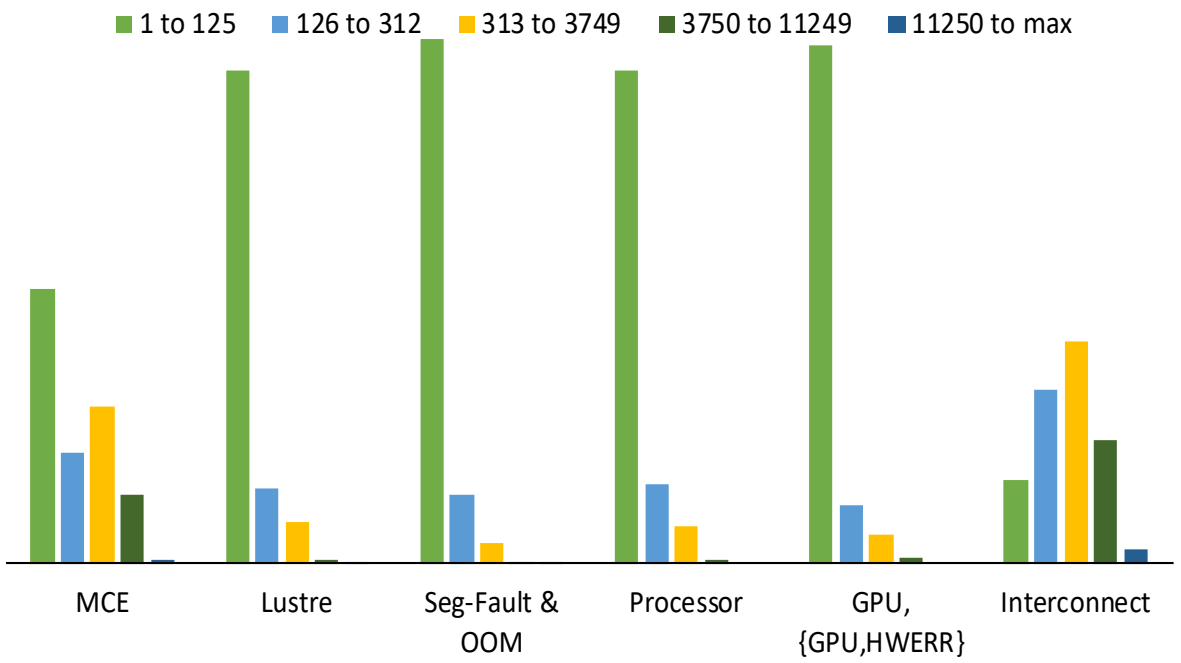
Categorization of Applications based on Recorded RAS Events



Characteristics of Applications with RAS Events

- Majority of applications use only a fraction of the total nodes.
- Likelihood of seeing an event increases for larger sized applications.
- Lustre: contention of resources; Seg-Fault, OOM: testing of codes;

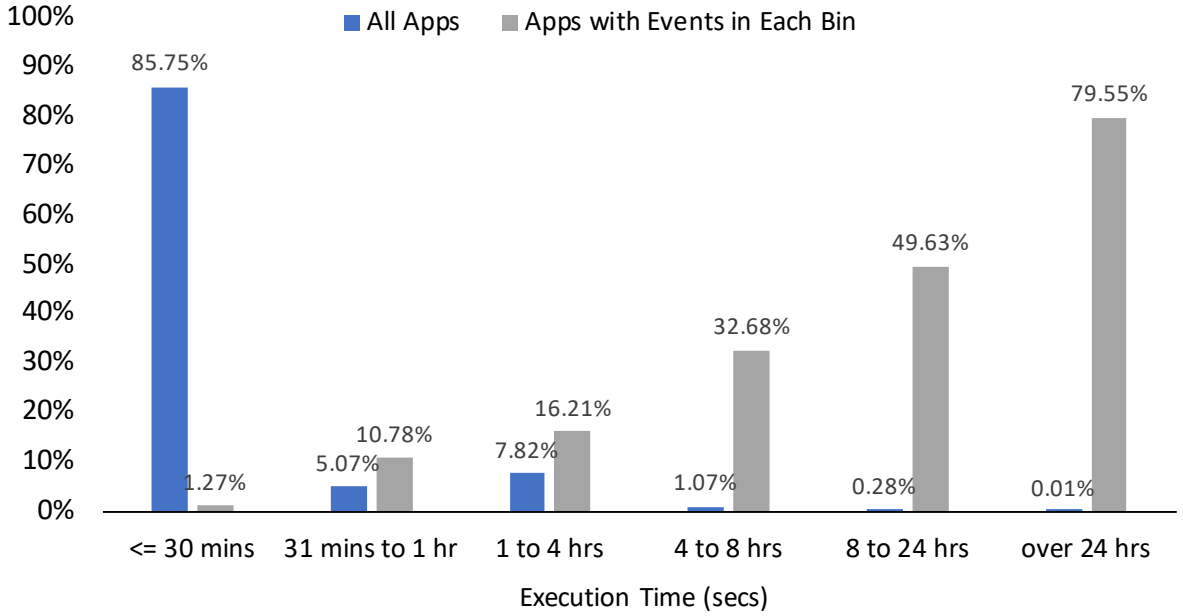


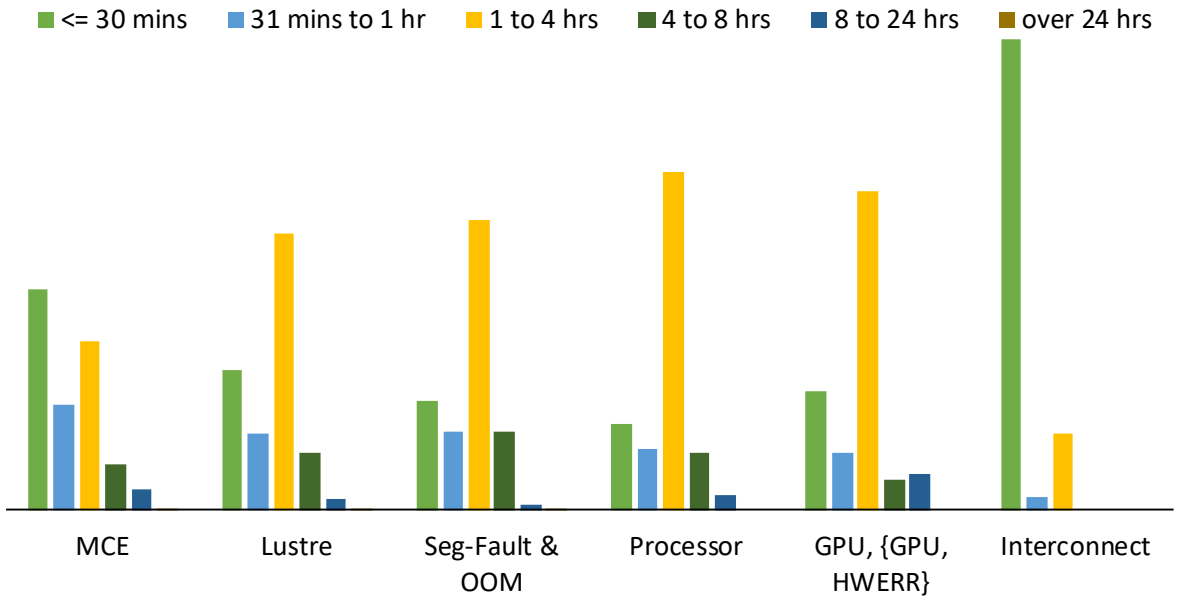


OAK RIDGE National Laboratory

Characteristics of Applications with RAS Events

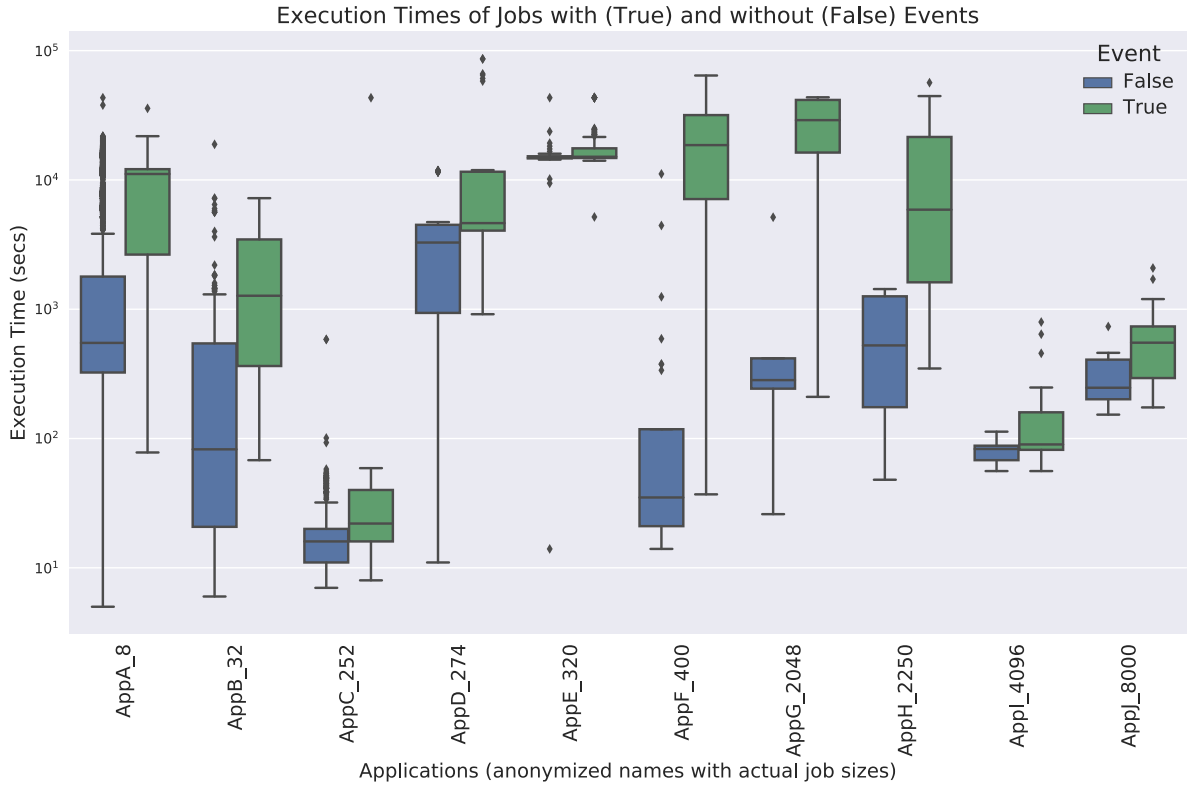
- Majority of applications are of short duration.
- Likelihood of seeing an event increases for longer running applications.
- Lustre, Seg-Fault, OOM, Processor, GPU: 1 to 4 hour time window?
- Interconnect: medium to large-scale and short-lived applications.





Comparative Analysis

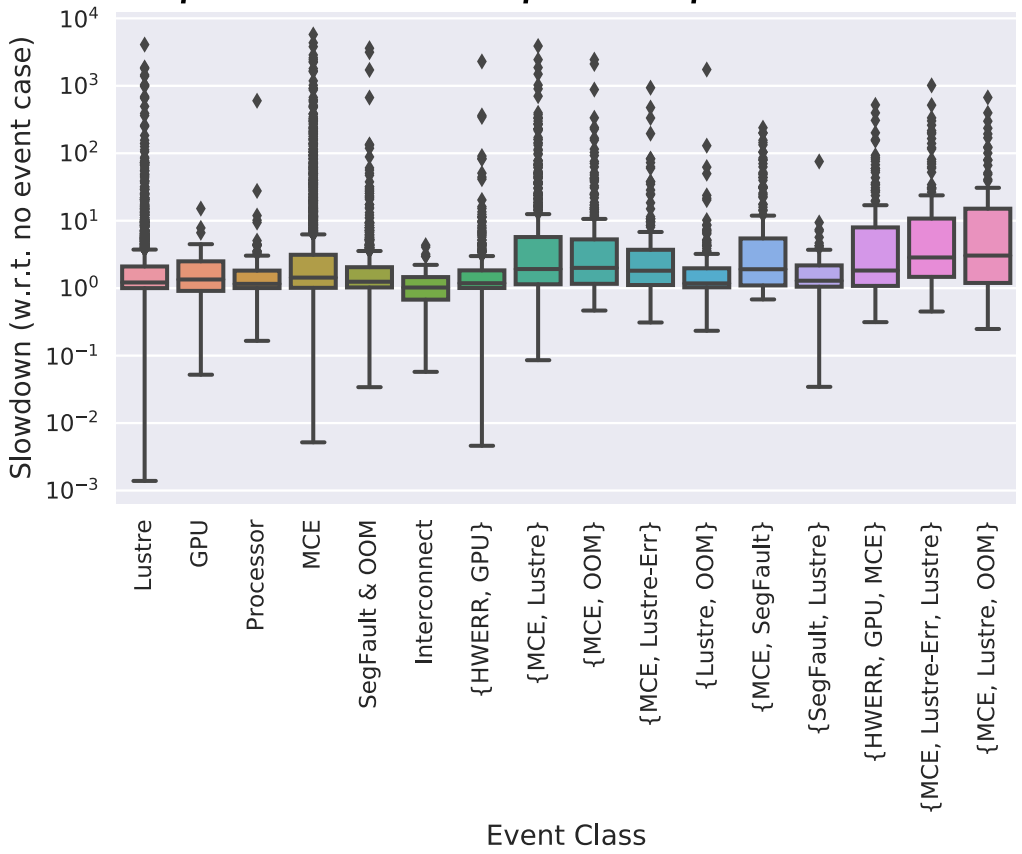
- Null Hypothesis: "*Execution times of applications with and without events are same.*"
- Two sample t-tests reveal that the null hypothesis can be rejected with a p-value of less than 0.1 in 48.1% cases.



Comparison between runtimes of applications with same binary names and using same number of nodes.

Slowdown Analysis: Event type (1/2)

- Average runtimes of different runs for each application normalized w.r.t. average runtimes of corresponding cases with no events.
- *Do different system components impact performance differently?*



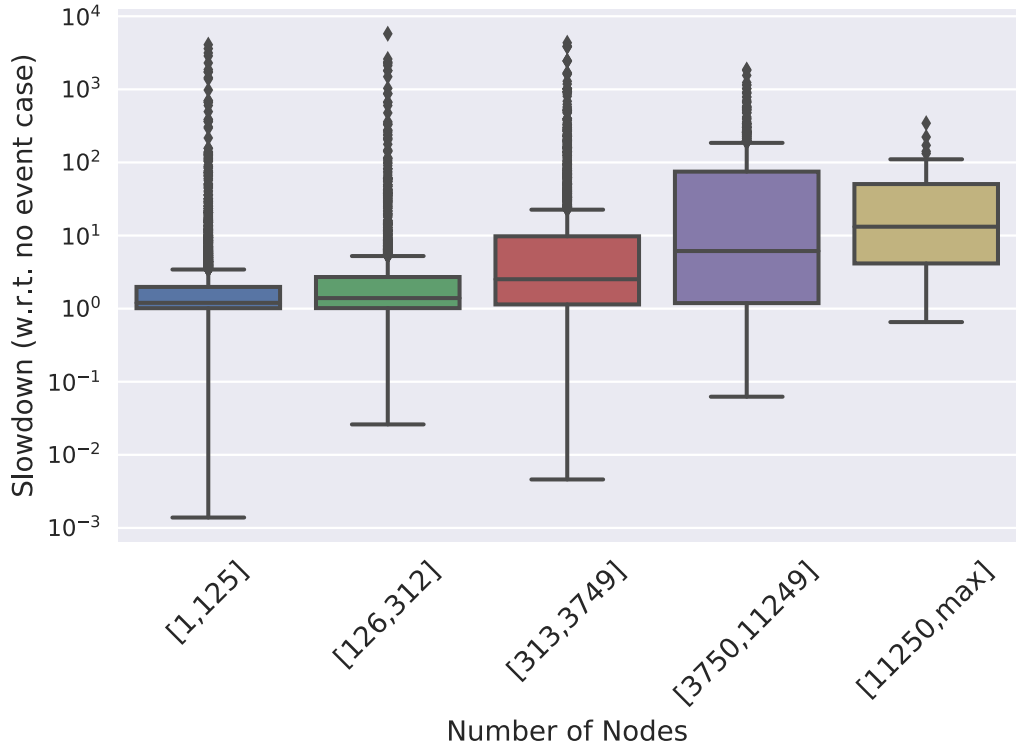
Slowdown Analysis: Event type (2/2)

- Average runtimes of different runs for each application normalized w.r.t. average runtimes of corresponding cases with no events.
- *Do different system components impact performance differently?*

Event Class	Median	Average	% Applications
Parallel File System	1.21	9.08	75.7%
Processor Events	1.16	6.41	76.5%
Machine Check Exceptions	1.43	22.14	78.3%
Graphics Processing Unit	1.34	2.16	72.5%
Seg-Faults & Out-of-Mem.	1.24	12.30	82.2%
Interconnect	1.02	1.23	52.4%
Event Groups – selected, sorted by # of cases			
(MCE, Lustre)	1.92	36.45	88.7%
(HWERR, GPU)	1.18	10.39	76.8%
(MCE, Out-of-Mem.)	2.00	32.77	89.3%
(Lustre, Out-of-Mem.)	1.18	13.08	84.0%
(MCE, Lustre-Err)	1.81	16.33	84.4%
(MCE, Seg-Fault)	1.91	35.96	90.2%
(HWERR, GPU, MCE)	1.82	18.04	84.6%
(MCE, Lustre-Err, Lustre)	2.85	31.87	89.9%
(Seg-Fault, Lustre)	1.29	2.34	85.9%
(MCE, Lustre, Out-of-Mem.)	3.04	28.06	87.8%

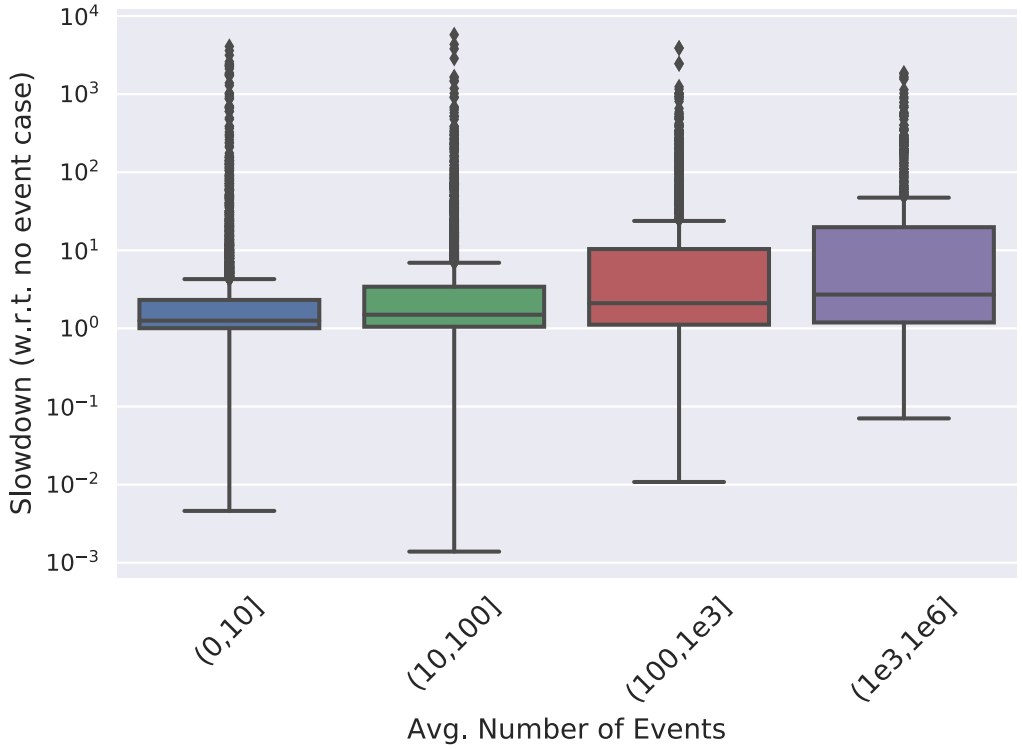
Slowdown Analysis: Application Size

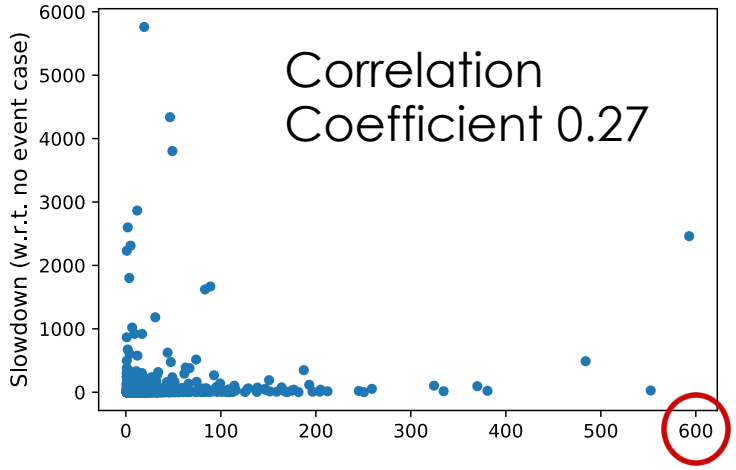
- *Do larger-scale applications see a greater impact on performance as compared to medium- or small-scale applications?*



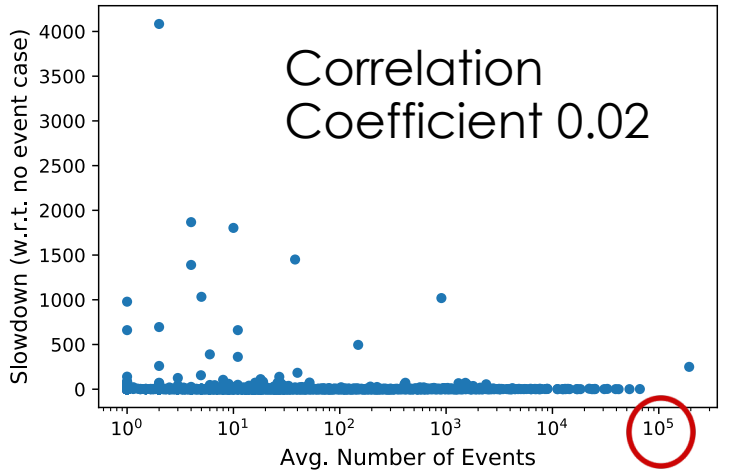
Slowdown Analysis: Number of Events

- *Do high number of recorded events during an application run cause a greater impact on application performance?*





MCE

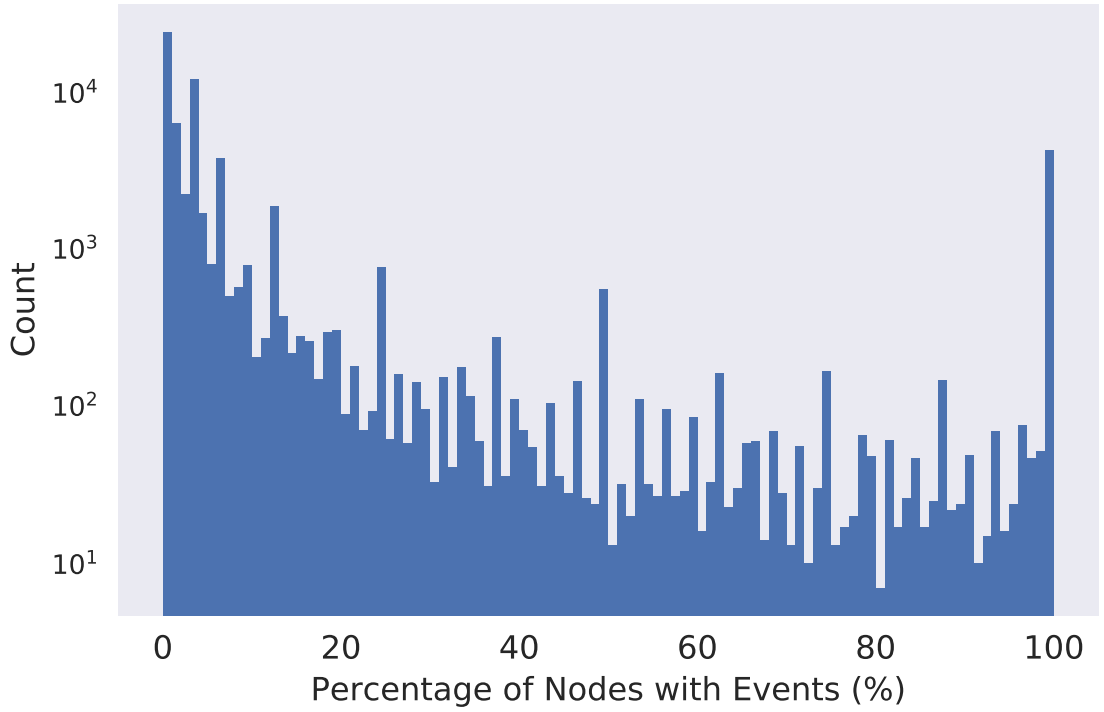


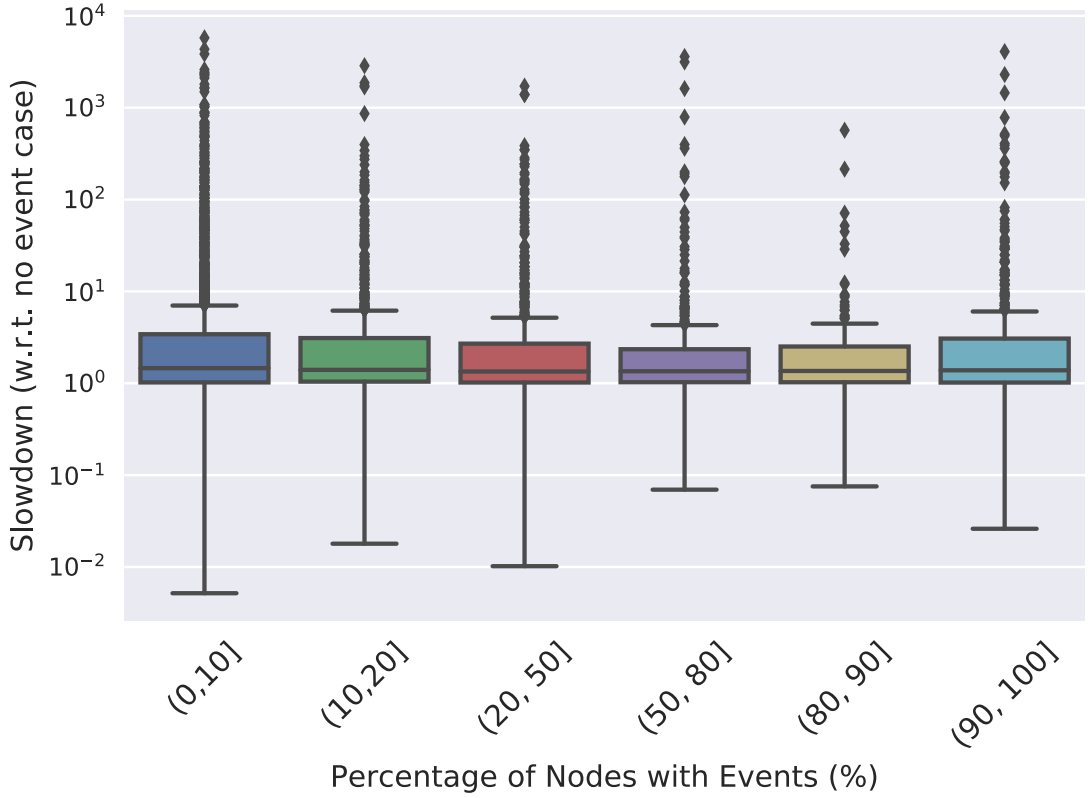
Lustre

OAK RIDGE National Laboratory

Slowdown Analysis: Proportion of Nodes with Events

- *Do occurrence of events across multiple nodes in an application cause a greater impact on performance?*





Conclusions and Future Work

- Study shows that RAS events do correspond to application slowdown in majority of cases.
- Large scale and longer duration applications most likely to be impacted.
- Slowdown analysis shows that some system components/situations cause more slowdown as compared to other components/scenarios.
- Implications for users and system operators, and a good starting point is to provide feedback to users on completion of their runs.
- Need to develop fault injection mechanisms to trigger RAS events in different system components, irrespective of failures.

References

1. M. Gottscho, M. Shoaib, S. Govindan, B. Sharma, D. Wang, and P. Gupta, "*Measuring the impact of memory errors on application performance*," IEEE Computer Architecture Letters (CAL), vol. 16, no. 1, pp. 51–55, Jan 2017.
2. C. D. Martino, W. Kramer, Z. Kalbarczyk, and R. Iyer, "*Measuring and understanding extreme-scale application resilience: A field study of 5,000,000 HPC application runs,*" in 2015 45th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN), June 2015, pp. 25–36.
3. S. Gupta, T. Patel, C. Engelmann, and D. Tiwari, "*Failures in large scale systems: Long-term measurement, analysis, and implications*," in Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC17), 2017, pp. 44:1–44:12.
4. V. Sridharan, N. DeBardeleben, S. Blanchard, K. B. Ferreira, J. Stearley, J. Shalf, and S. Gurumurthi, "*Memory errors in modern systems: The good, the bad, and the ugly*," in Proceedings of the Twentieth International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS' 15), 2015, pp. 297–310.
5. B. H. Park, Y. Hui, S. Boehm, R. A. Ashraf, C. Engelmann, and C. Layton, "*A Big Data analytics framework for HPC log data: Three case studies using the Titan supercomputer log*," in Proceedings of the 19th IEEE International Conference on Cluster Computing (CLUSTER) 2018: 5th Workshop on Monitoring and Analysis for High Performance Systems Plus Applications, Belfast, UK, Sep. 10, 2018

Contact & Acknowledgements

- Project website https://ornlwiki.atlassian.net/wiki/spaces/CFEFIES/overview
- Rizwan Ashraf ashrafra@ornl.gov Christian Engelmann engelmannc@ornl.gov
- Work supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Resilience for Extreme Scale Supercomputing Systems Program, with program manager Lucy Nowell, under contract number DE-AC05-00OR22725.
- Work was also supported by the Compute and Data Environment for Science (CADES) facility and the Oak Ridge Leadership Computing Facility (OLCF) at the Oak Ridge National Laboratory, which is managed by UT-Battelle, LLC for the U.S. Department of Energy (under contract number DE-AC05-00OR22725)