LA-UR-15-29236

Title: Data-Driven Decision Making in Resilience

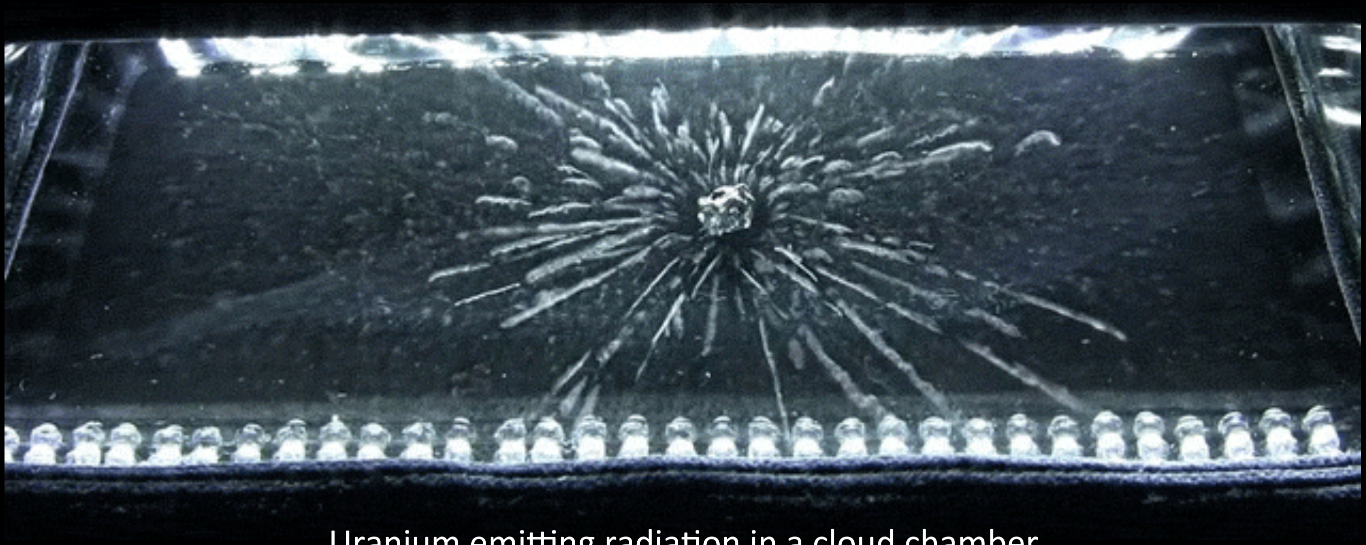
Author(s): Debardeleben, Nathan A.

Intended for: SC'15, 2015-11-18 (Austin, Texas, United States)

Issued: 2015-12-01 (Draft)

Disclaimer:
Los Alamos National Laboratory, an affirmative action/equal opportunity employer,is operated by the Los Alamos National Security, LLC for the National NuclearSecurity Administration of the U.S. Department of Energy under contract DE-AC52-06NA25396. By approving this article, the publisher recognizes that the U.S. Government retains nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Departmentof Energy. Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

http://sploid.gizmodo.com/watching-uranium-emit-radiation-inside-a-cloud-chamber-1689997373/



Uranium emitting radiation in a cloud chamber

Data-Driven Decision Making in Resilience

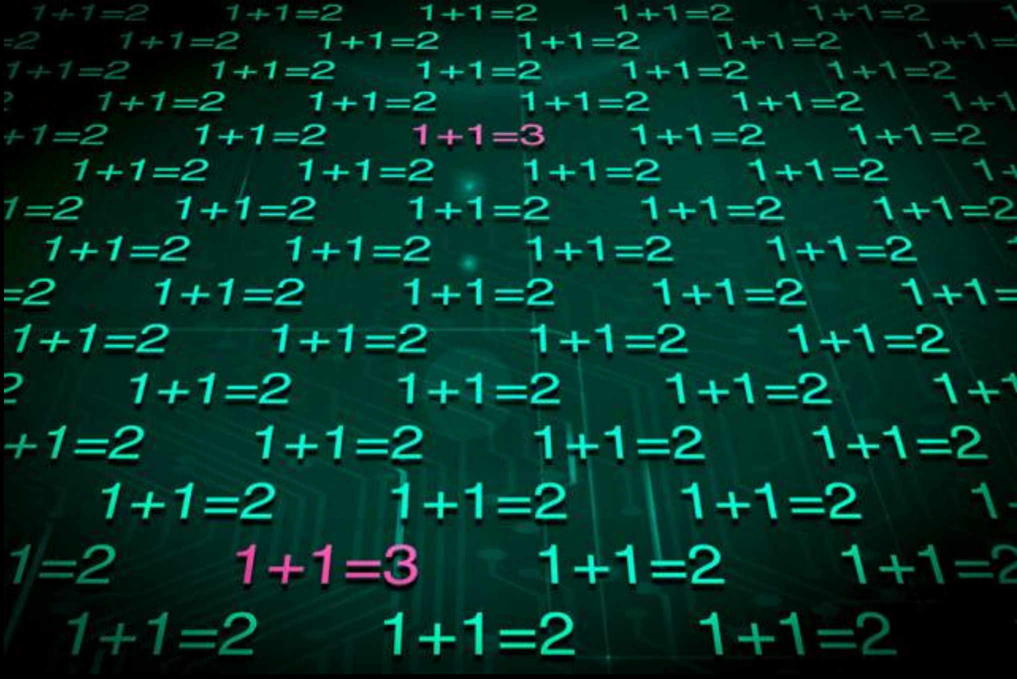
Nathan DeBardeleben, Ph.D.
Los Alamos National Laboratory

High Performance Computing
Ultrascale Systems Research Center Lead

LA-UR-15-23012

Los Alamos
NATIONAL LABORATORY
EST.1943

LANL System Data Analytics (Reliability Focus)



Jose-Luis Olivares/MIT

- What do we have?
- What do we do with it?
- What are we sharing?

LANL Supercomputers

- Over a dozen production supercomputers:

Capacity Technology Systems

Advanced Technology Systems

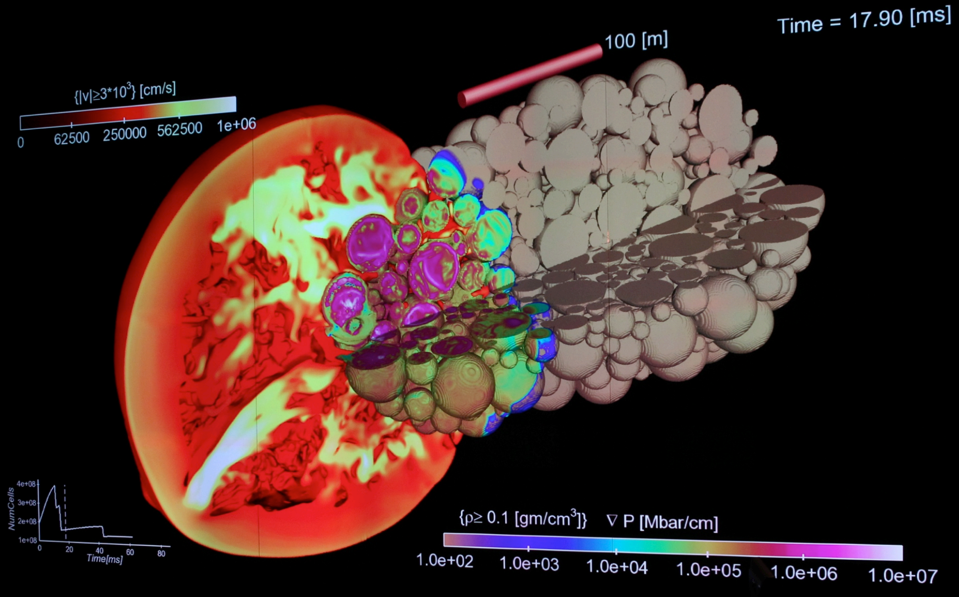
LANL Supercomputers

- Systems of this scale crash at times

Capacity Technology Systems

Advanced Technology Systems

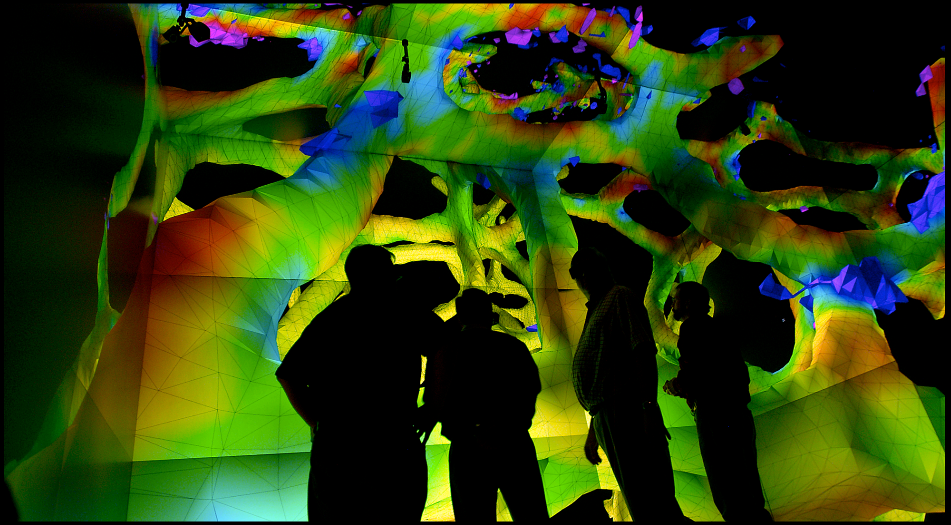
What Can We Learn?



Extracted from LA-UR-13-27862

- Logs – memory, CPU, disk, network, scheduler, resource manager, hardware replacements, as well as full syslog
- Lots of sensitive data
- All data has to be curated before sharing externally
- Very time consuming process
  - Potentially beneficial

How Can I Get Some LANL System Data?



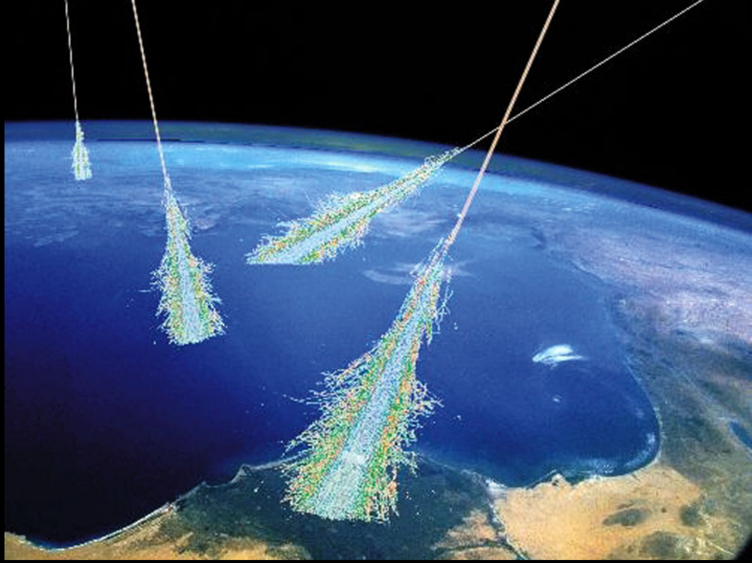
Extracted from LA-UR-13-27862

- Collaboration
- U.S. citizenship almost a requirement
- Send us your students to process the data (they can't take the data home)
- Have us run your tools on our data internally
- All of these are challenging as analytical results are not guaranteed to be releasable
- Almost a guarantee that new raw field data will not be released
- This is where the open sciences communities need to be contributing more – NSF, Office of Science, universities, etc.

What Can You Do With This Data?

- Let’s look at a sampling of results!

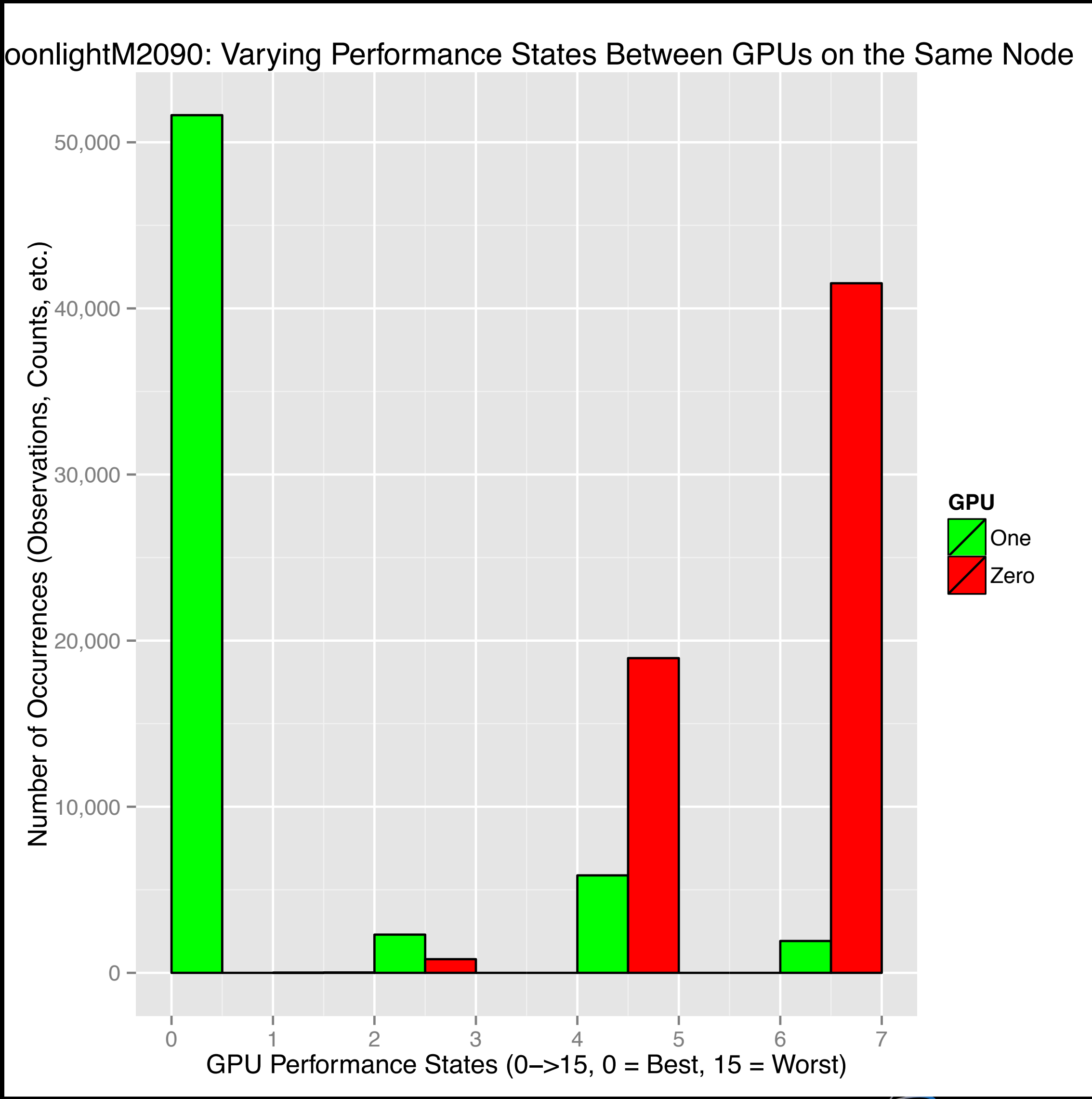
Our Systems are Aggressively Maintained



- Cielo – ~0.3 correctable errors / min ← LANL
- Hopper – ~1 correctable error / min
- Titan – ~1.4 correctable errors / min
- BlueWaters (DSN2014) – ~4.2 correctable errors / min
- This is an artifact of hardware log monitoring and aggressive replacement of failing hardware

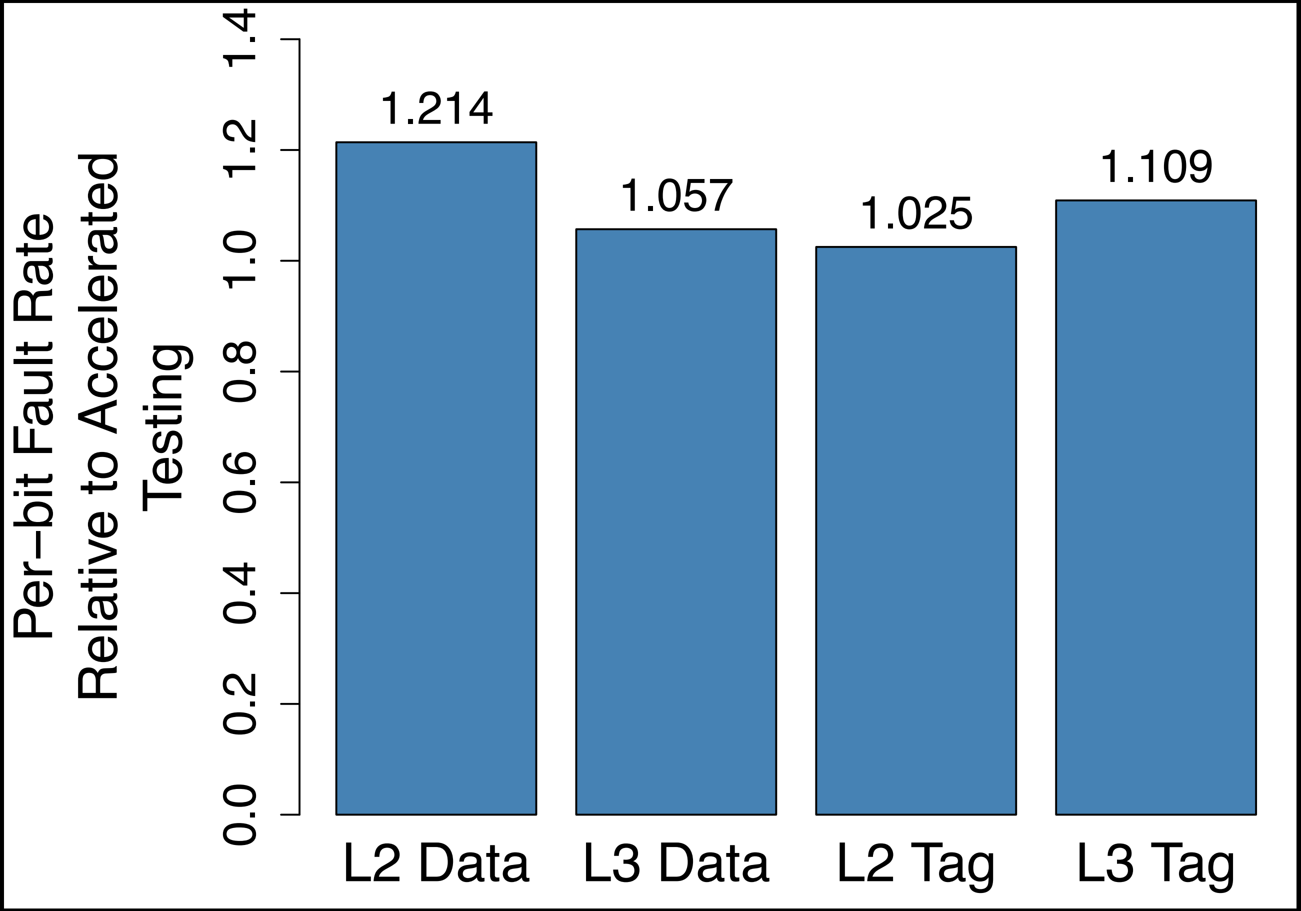
Tightly Coupled GPUs

- Tightly coupled numerical codes run at the speed of the slowest component
- 2 GPUs on a node are running in different pStates (throttling)



Neutron Beam Testing is Good Approximation for Field Experiences

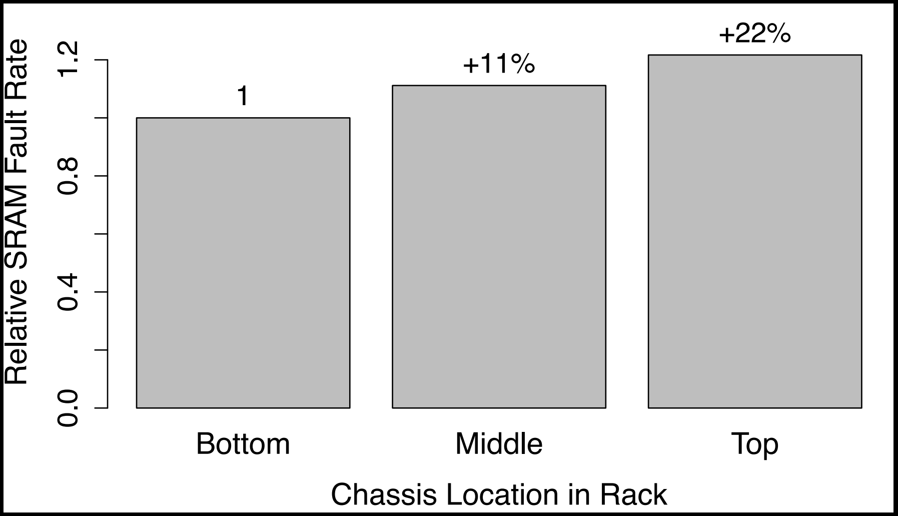
- Working with AMD we find that years of field data from supercomputers lines up reasonable well with neutron beam experiments

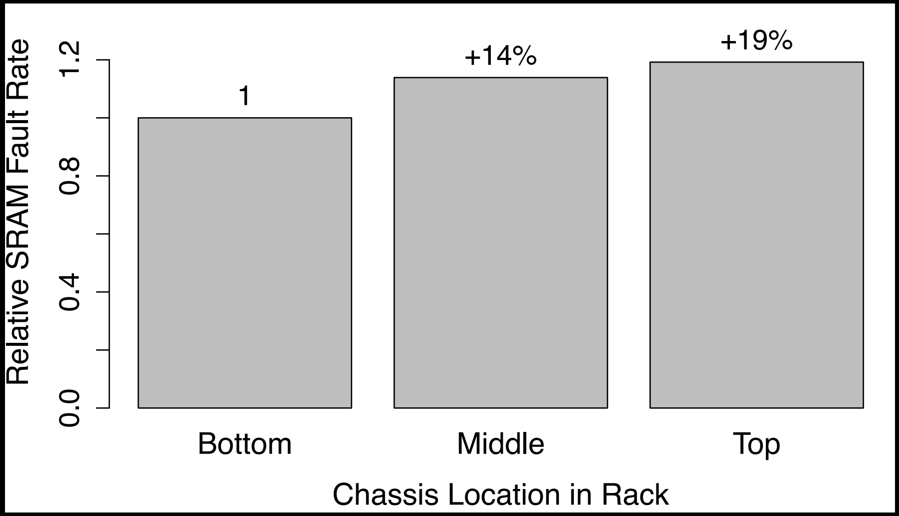


Memory Errors in Modern Systems: The Good, the Bad, and the Ugly,
Vilas Sridharan, et. al., ASPLOS 2015

More Faults Higher in the Rack

- ~10% increase in SRAM fault rates at each chassis level
- Temperature?
- Cosmic radiation shielding?

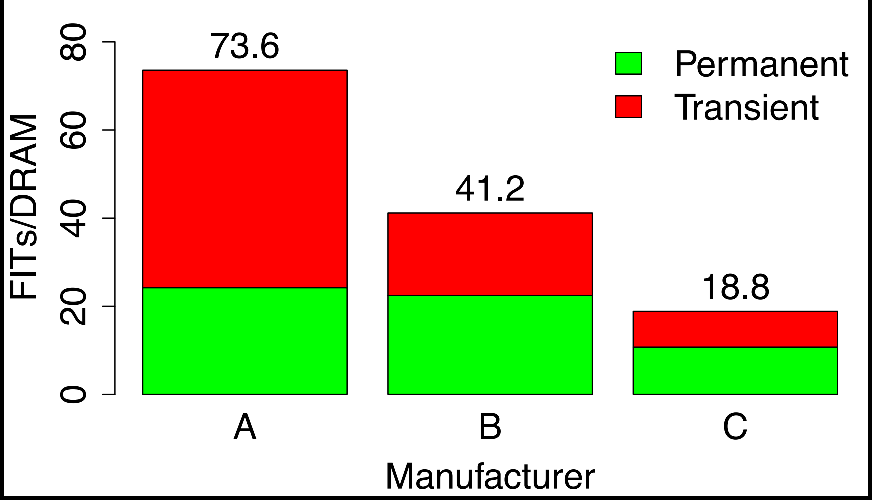




Feng Shui of Supercomputer Memory, Vilas Sridharan, et. al., SC 2013

Not all DRAM Vendors are Created Equal

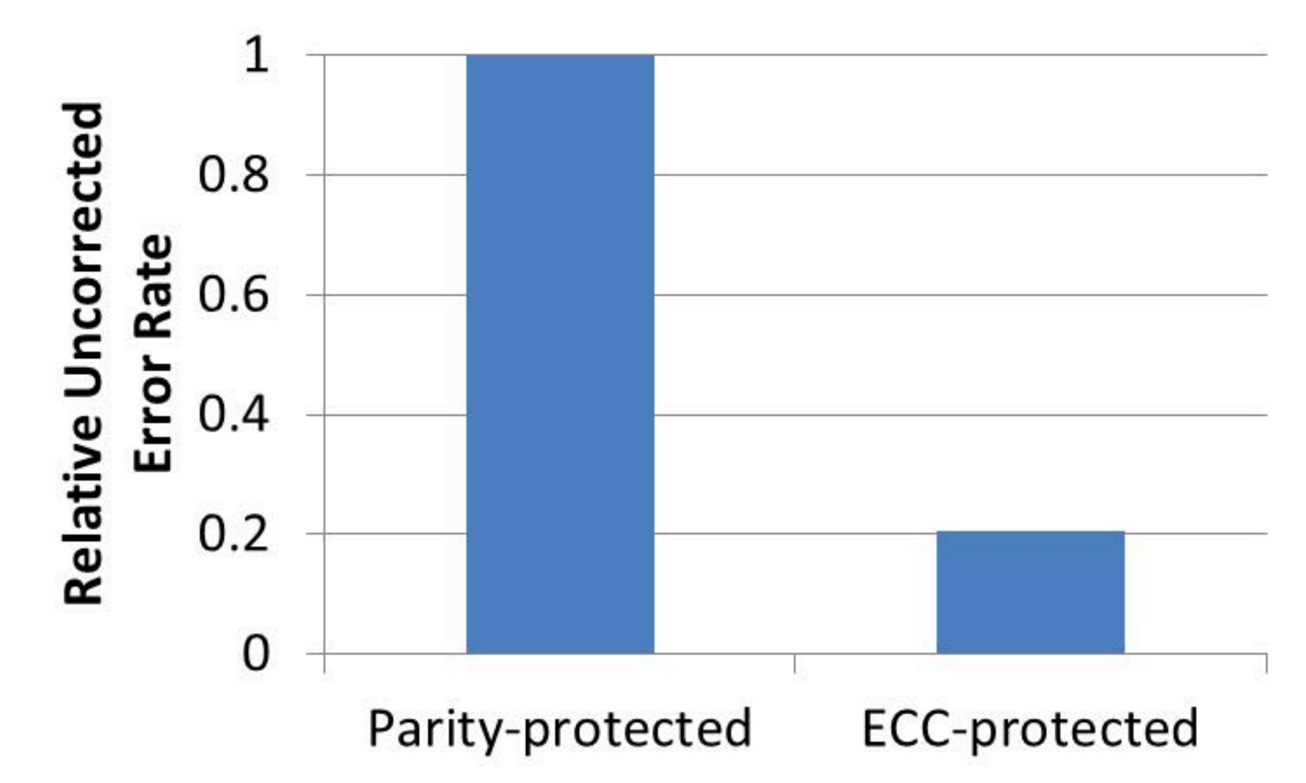
- Must study your DRAM by vendor, not just faults alone
- All 3 vendors about the same wrt permanent errors
- Vendor A has transient error problems



Feng Shui of Supercomputer Memory, Vilas Sridharan, et. al., SC 2013

SRAM Uncorrected Error Rates

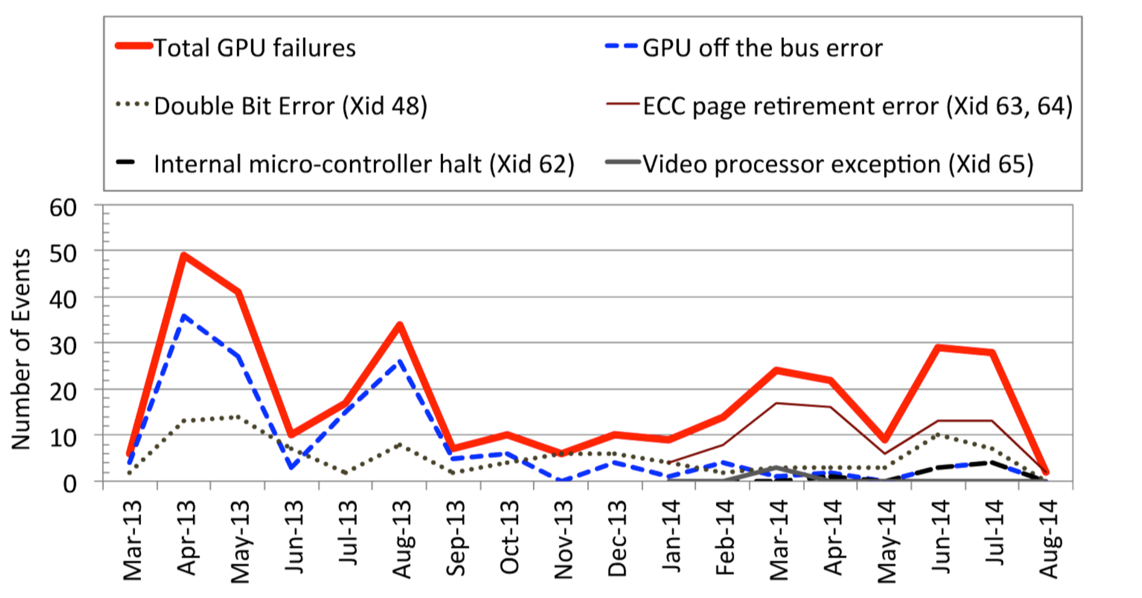
- Studying the rates of ECC protected ECC compared to parity-protected can provide insights on required error protection levels



Extra Bits on SRAM and DRAM Errors – More Data From the Field, Nathan DeBardeleben, et. al., SELSE 2014

GPU Failure Rates on Titan

- ~1 GPU failure per day on Titan
- Better than previous generation but not good enough for exascale
- DOE needs 1 failure / day across the *SYSTEM* not 1 component



Understanding GPU Errors on Large-scale HPC Systems and the Implications for System Design and Operations, Devesh Tiwari, et. al., HPCA 2015

Conclusions

- Data is valuable
- We work *very* closely with hardware vendors and supercomputer integrators to:
  - Understand how the systems are behaving
  - Work to improve current systems through configuration
  - Work to improve next generation systems through insights from trends
- Working with us is hard, but we would welcome more collaboration
- Strongly encourage more open groups to share their data openly