A Rare Look at Real World Data Analysis of Supercomputer Faults –

DRAM, SRAM, and GPGPUs

Nathan DeBardeleben, Ph.D. (LANL)

52nd HPC User Forum
Santa Fe, April 2014

Los Alamos NATIONAL LABORATORY EST. 1943

LA-UR-14-22329

Many Collaborators

- **HPC-5 (Sean Blanchard, Laura Monroe, Phil Romero, Craig Idler, Daryl Grunau, Cornell Wright, . . .)**
- **HPC-3 – lots of folks getting data for me (Kyle Lamb, Cory)**
- **Cray – data! Greg Hamilton**
- **AMD – Vilas Sridharan, Sudhanva Gurumurti**
- **Sandia – Jon Stearley**
- **UIUC – Rakesh Kumar, Xun Jian**
- **NVIDIA – Timothy Tsai**
- **. . . etc**
- **<u>This is a team effort</u>**

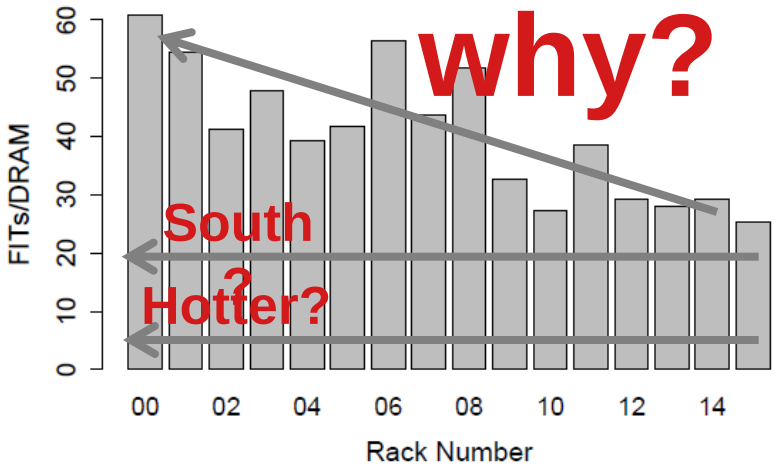
Topics Covered

- **Systems:**
  - Cray systems:
    - Cielo (@LANL)
    - Jaguar (@ORNL)
    - Hopper (@NERSC)
  - Moonlight GPGPU cluster (@LANL)
- **Topics:**
  - Corrected and uncorrected faults in DRAM and SRAM
  - Altitude effects
  - Positional effects (in a datacenter, separately we've looked at positional effects on a DRAM chip)
  - GPGPU variability in memory bandwidth, performance, and reliability

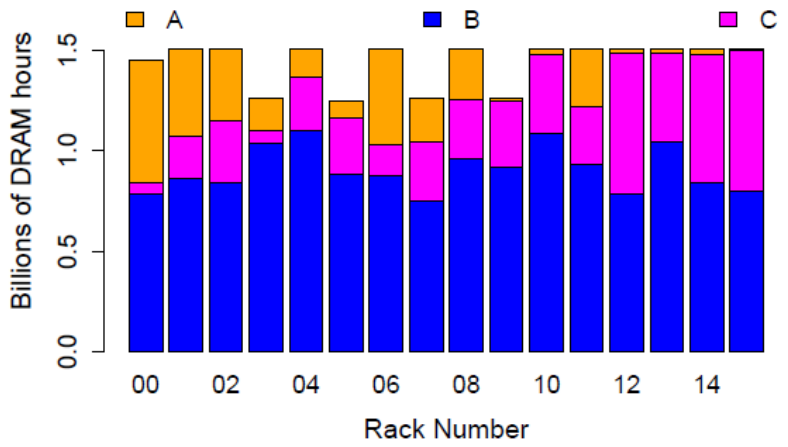
Cielo Data – A More Detailed Analysis Than Others Do

- **Cielo affords us more data sources than are often available**
- **Furthermore, there is *so very little* published reports or available data in the public**
  - Our report is one of the largest ever done – 24billion DRAM hours
  - Our report demonstrated the problems others have done in their analysis (fault vs. error) and how wrong conclusions have been drawn (see: *DRAM Errors in the Wild* and then promptly forget what you read, please).
- **It's very easy to jump to wrong conclusions with the kind of data usually available**
- **Such as . . .**

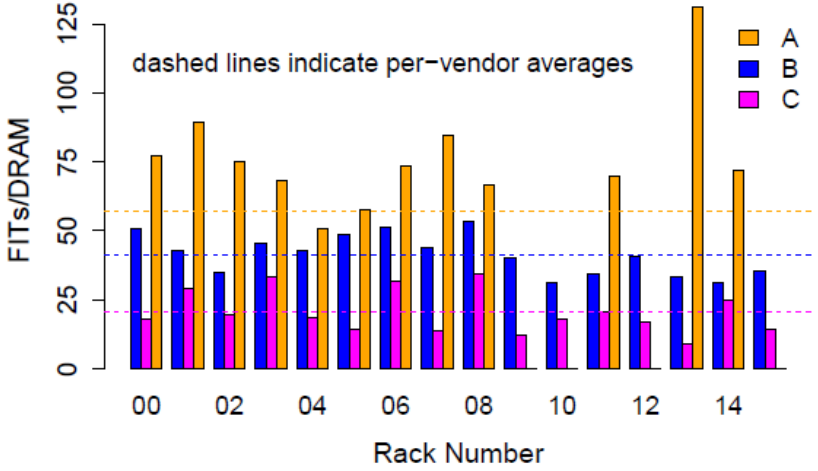
Sridharan et al., *Feng Shui of Supercomputer Memory*, SC 2013



- A correlation to physical location...



- ...is due to non-uniform distribution of vendor...



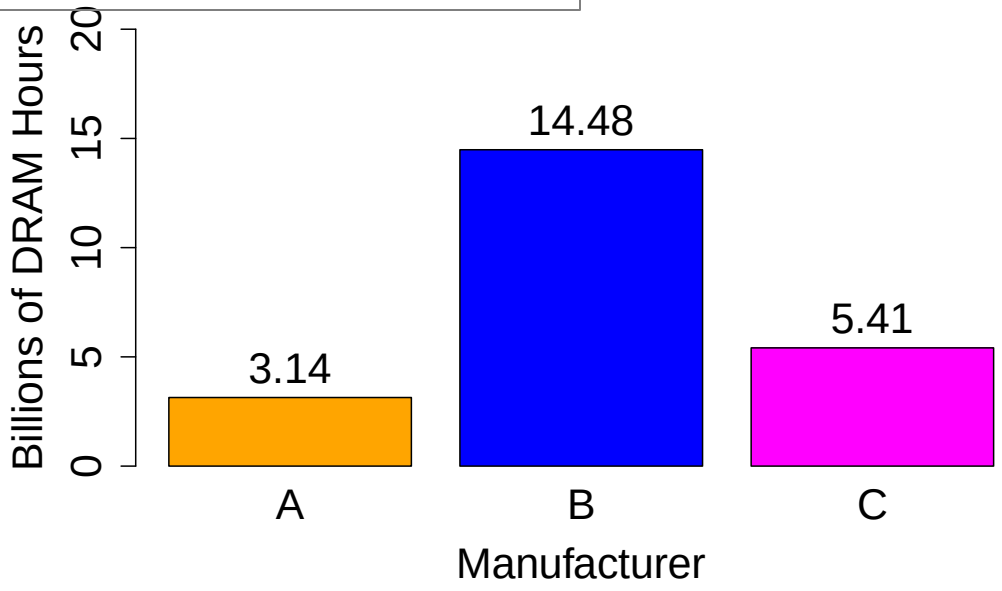
- ...and disappears when examined by vendor.

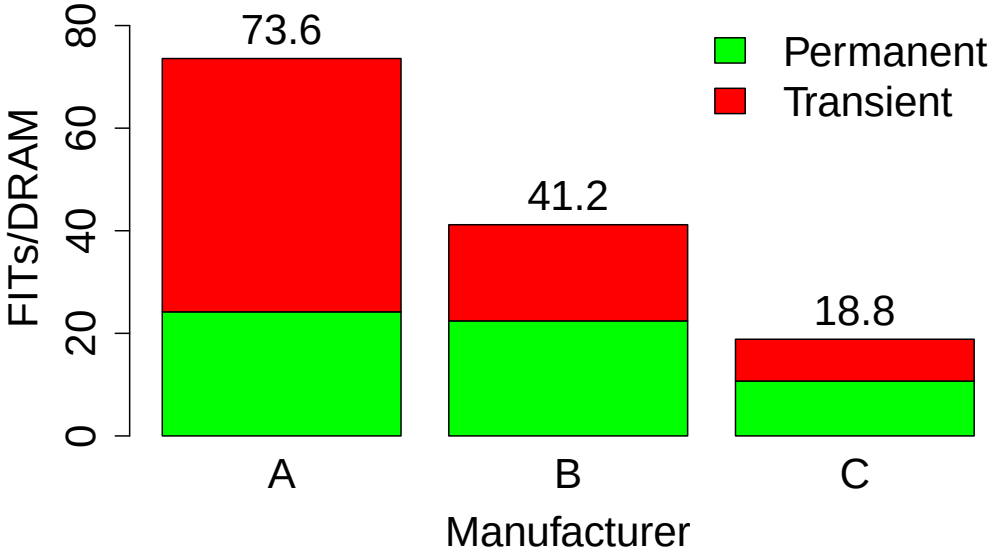
- DRAM reliability studies must account for DRAM vendor or risk inaccurate conclusions

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

Field Data – It Matters

- **Not all vendors are created equal**
- **As much as a 4x difference in FIT rate depending on DRAM vendor used in Cielo nodes**
- **While B and C are about 50/50 vulnerable to permanent/transient, vendor is is closer to 30/70 permanent/transient**

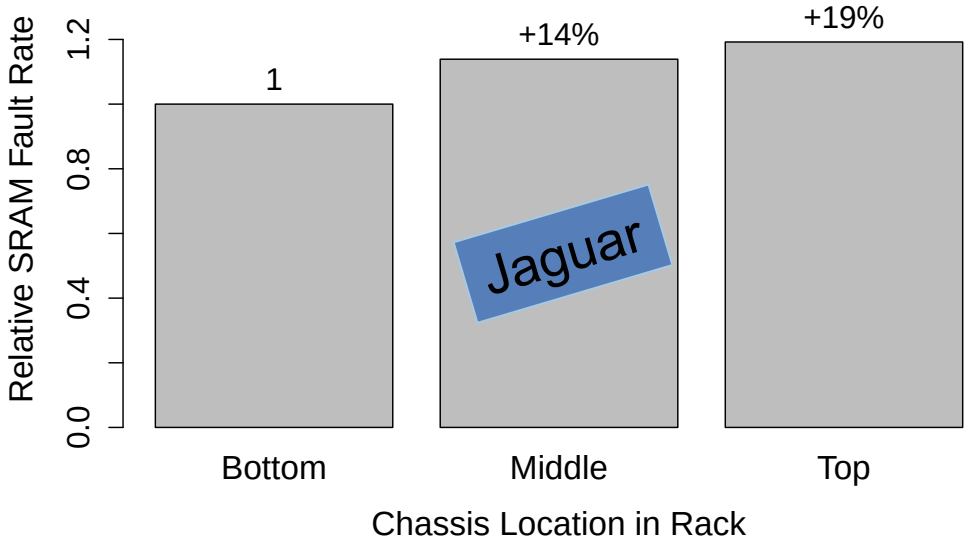




Sridharan et al., *Feng Shui of Supercomputer Memory*, SC 2013

Positional Effects

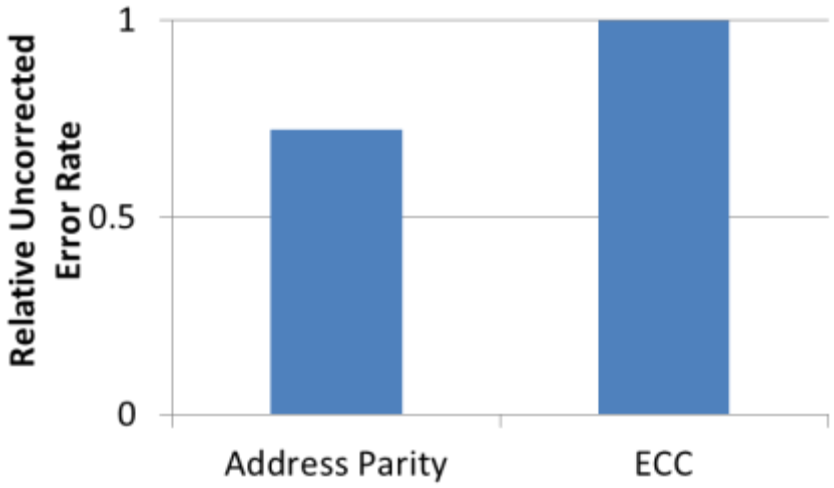
- **Fault rates increase as you go vertically in a rack**
- **Shielding? Temperature?**
- **We found a similar correlation in DRAM**
- **More studies are needed to explain why we see this**



Los Alamos NATIONAL LABORATORY EST. 1943

DDR Command and Address Parity – It's a Good Thing

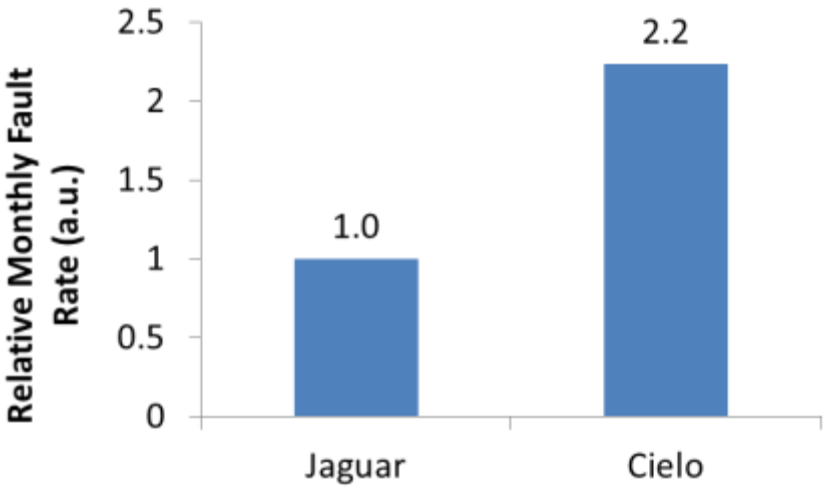
- **Key feature of DDR3 (and on) is the ability to add parity-check logic to the command and address bus.**
- **Can have a significant positive impact on DDR memory reliability**
  - Not previously shown empirically
- **DDR3 sub-system on Cielo includes command and address parity checking.**
- **Rate of command/address parity errors was 75% that of the rate of uncorrected ECC errors.**
- **Increasing DDR memory channel speeds may cause an increase in signaling-related errors.**



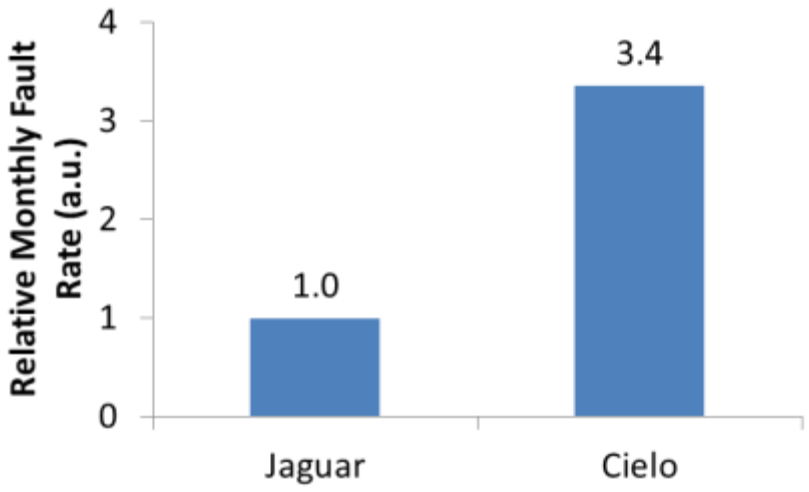
SRAM Altitude Effects

- **Higher altitude = increased fault rate due to increased cosmic ray-induced neutron strikes**
- **Whether this ultimately affects the application is correlated to the level of memory protection**
- **Cielo's AMD processors are *basically* the same as Jaguar's (@ORNL) AMD processors**
- **So what do we see in the data?**

SRAM Altitude Effects – Field Data



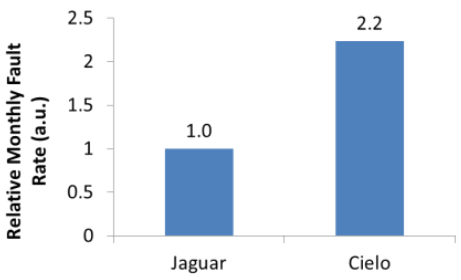
L2



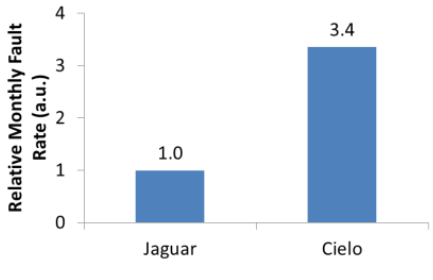
L3

SRAM Altitude Effects – Field Data

- **L2**
  - Cielo experiences 2.3x increase in SRAM transient fault rate relative to Jaguar in L2
- **L3**
  - Cielo experiences 3.4x increase in SRAM transient fault rate relative to Jaguar in L3



L2

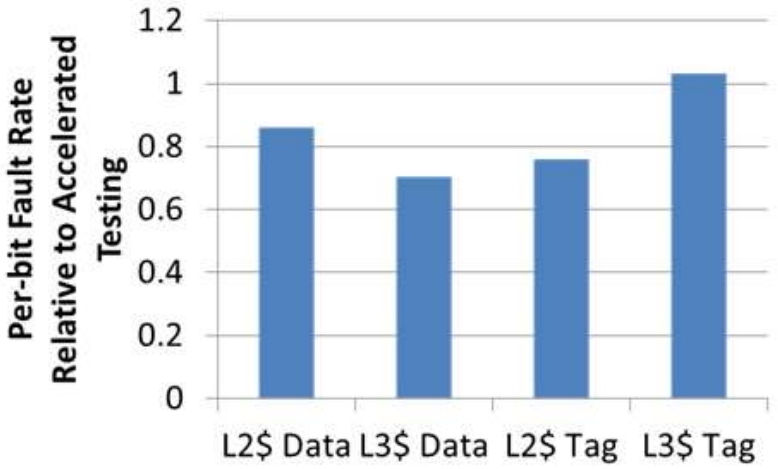


L3

- **Average flux ratio between LANL and ORNL (without accounting for solar modulation) is 4.39.**
- **Cielo fault rate increase is less than predicted due to altitude alone.**
- **This likely means that there are additional sources of SRAM faults, such as alpha particles**
- **How does this compare with accelerated testing (beam testing)?**

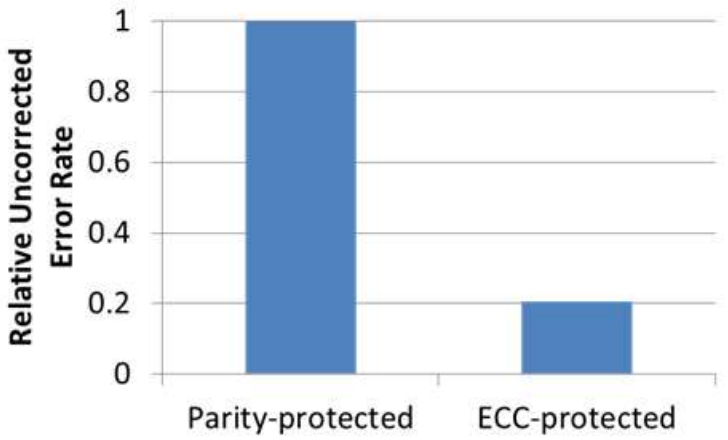
SRAM Altitude Effects – Accelerated Testing

- **Accelerated testing predicts a higher fault rate than seen in the field . . .**
- **. . . Except in the L3 tag array which was slightly higher in the field**
- **The per-bit fault rate seen in the field is expected to be somewhat lower than the rate seen in accelerated testing**
  - Many reasons for this, essentially related to the applications run during beam testing != applications used in the field
- **Conclusion is that the majority of SRAM faults in the field are caused by known high-energy particles**



SRAM Uncorrected Errors – Field Data

- **The majority of uncorrected errors in Cielo came from parity-protected structures**
- **Even though these structures are *far* smaller than the ECC-protected structures**
- **Conclusion: The best way to reduce SRC uncorrected error rates simply is to extend single-bit correction (e.g. ECC) through additional structures in the processor.**
- **Addressing multi-bit faults may be more challenging**



Predictions for Exascale – DRAM and SRAM

- **We have several papers on this subject**
  - Analytical modeling
  - Monte Carlo simulations
- **All show an expected increase in SRAM and DRAM faults at larger scales (not surprising)**
- **Highly dependent upon theoretical exascale system's DRAM and SRAM configuration**
- **For the configurations we studied, we expect between 5x and 90x increase in uncorrected errors in these structures**
  - I realize this is a large variation
  - It is, again, because of the high dependency on memory configuration choices
- **Takeaway:**
  - Uncorrected error rates in the future are going to be higher
  - We need to pay attention to memory system design, they can have a huge impact!

GPGPU Field Study

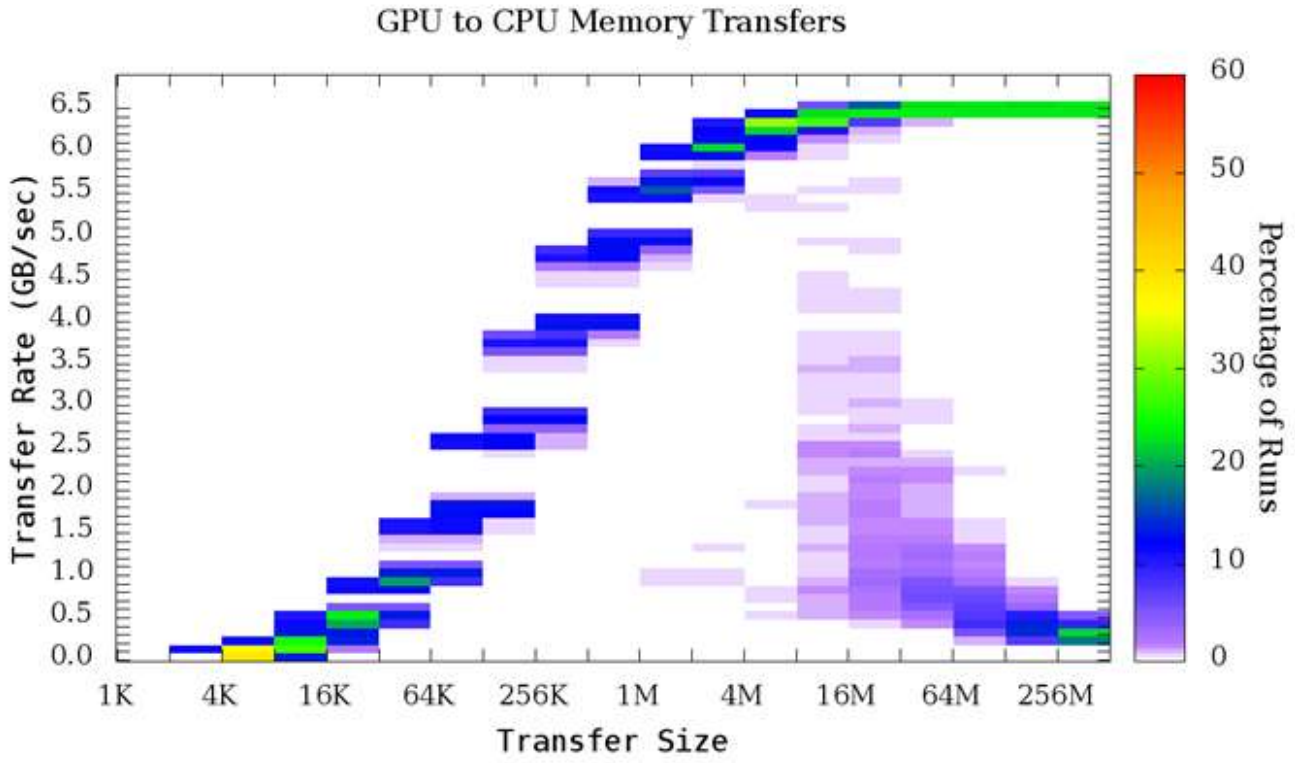
The Moonlight GPGPU Cluster

- **Deployed in 2011**
- **As high as 74 on top500 (not a large machine)**
- **308 nodes**
- **Dual socket**
- **8-core Intel Sandy Bridge**
- **32GB memory / node**
- **PCIe-3**
- **Two NVIDIA Tesla M2090 GPGPUs per node (Fermi generation)**
  - No large-scale Kepler deployment available at LANL at this time
- **Compute cores: 4,928 CPUs + 315,392 GPGPUs**

Overview of the Moonlight Saga

- An early foray into GPGPU computing at LANL
- A story of memory bandwidth problems, GPU performance variability, and correctness issues
- And also a story of an evolving software and driver stack, problems encountered and problems fixed
- The point is – the data I will show you were the problem data that lead to fixes
- Most importantly, the data shown here were acquired over 1.5 years and involved many CUDA driver upgrades
- At the end, we'll get to the question about whether things are better in the Kepler generation of cards

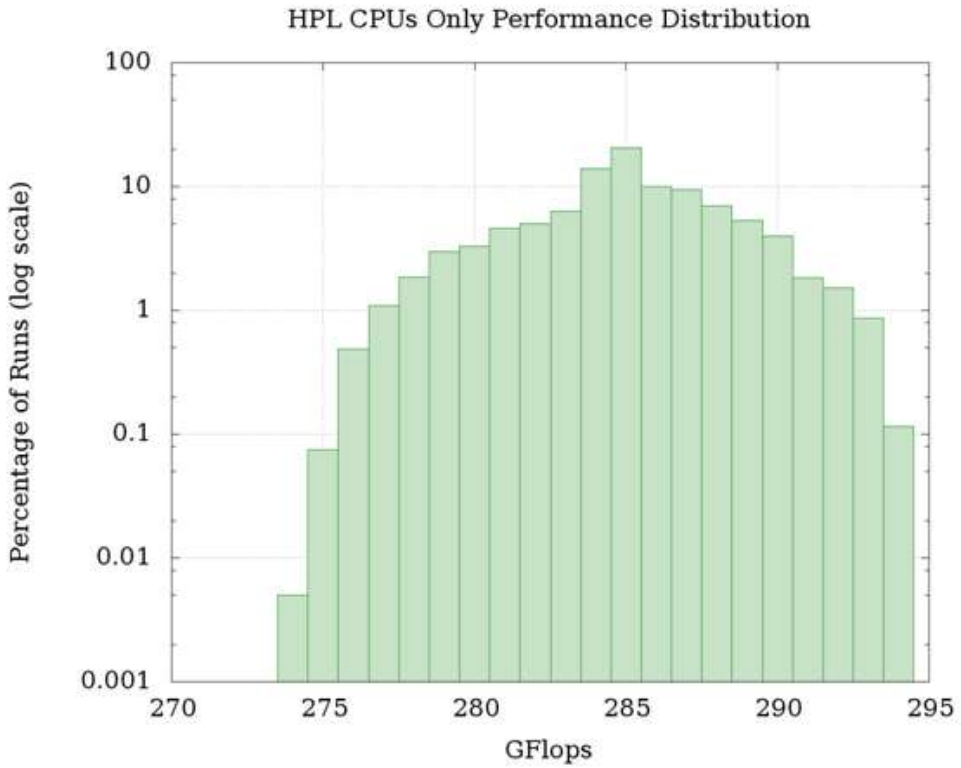
GPGPU Memory Bandwidth Woes



- **GPU to CPU transfer rate, not always consistent**
  <u>These bandwidth problems went away with driver upgrade</u>

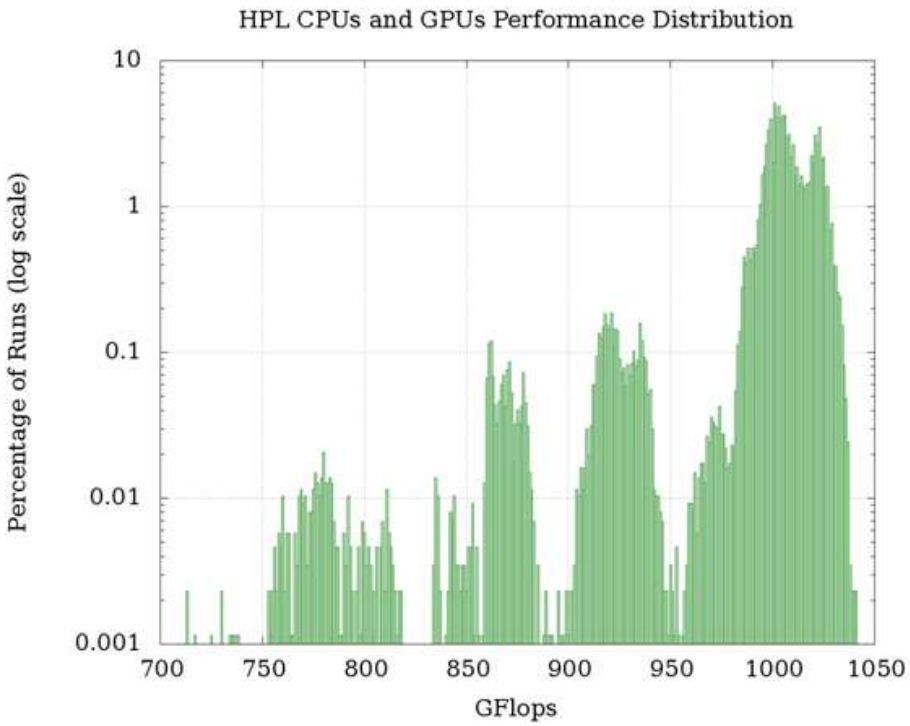
Moonlight HPL CPU Performance

- CPU performance is “normal”
- This is what we expect



Moonlight HPL CPU/GPU Performance

- **Performance is good . . .**
  - But irregular
- **Recall our tightly coupled numerical simulations run at the speed of the slowest component**

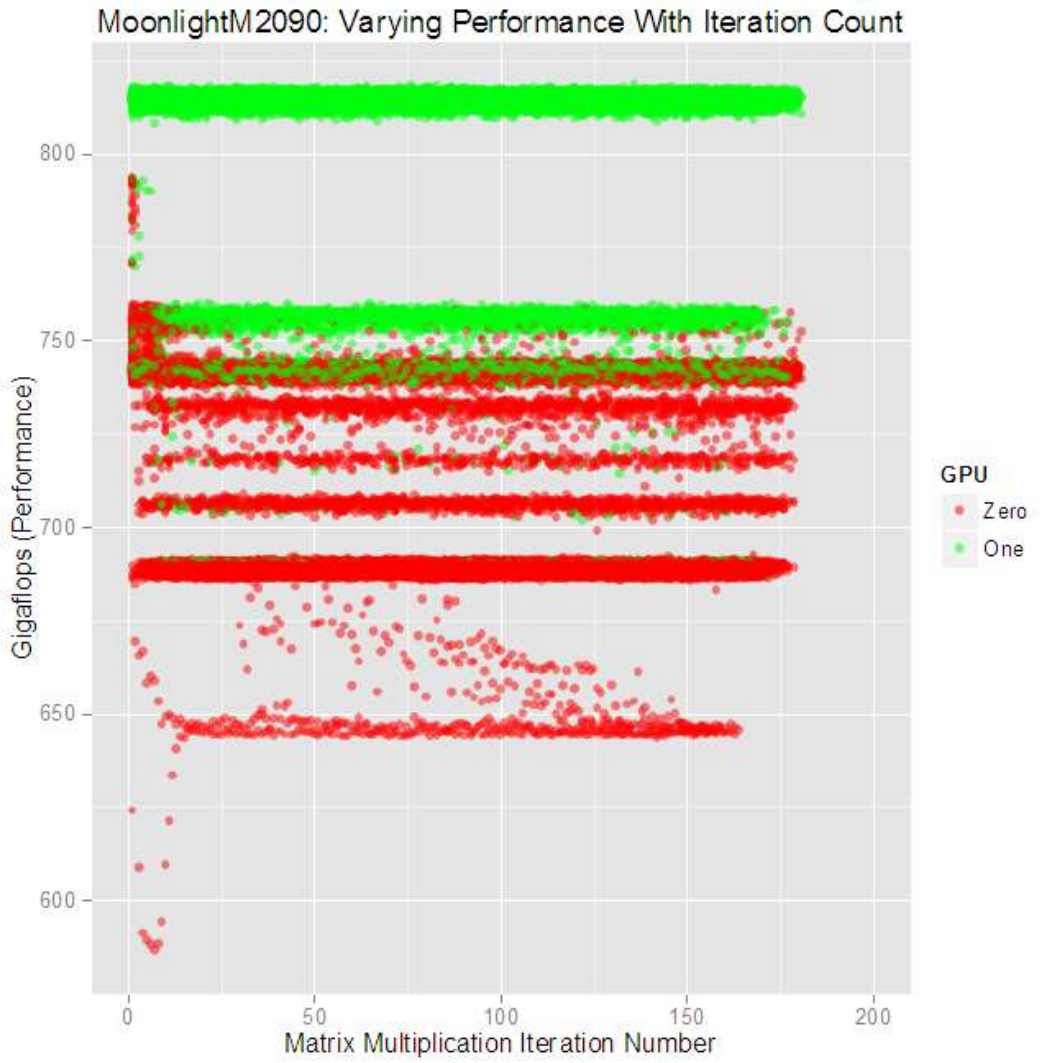


Matrix Multiplication Test Code

- **Is this problem unique to HPL?**
- **Perhaps, we built a simple matrix multiplication GPU code to explore that question**

ML M2090 Performance

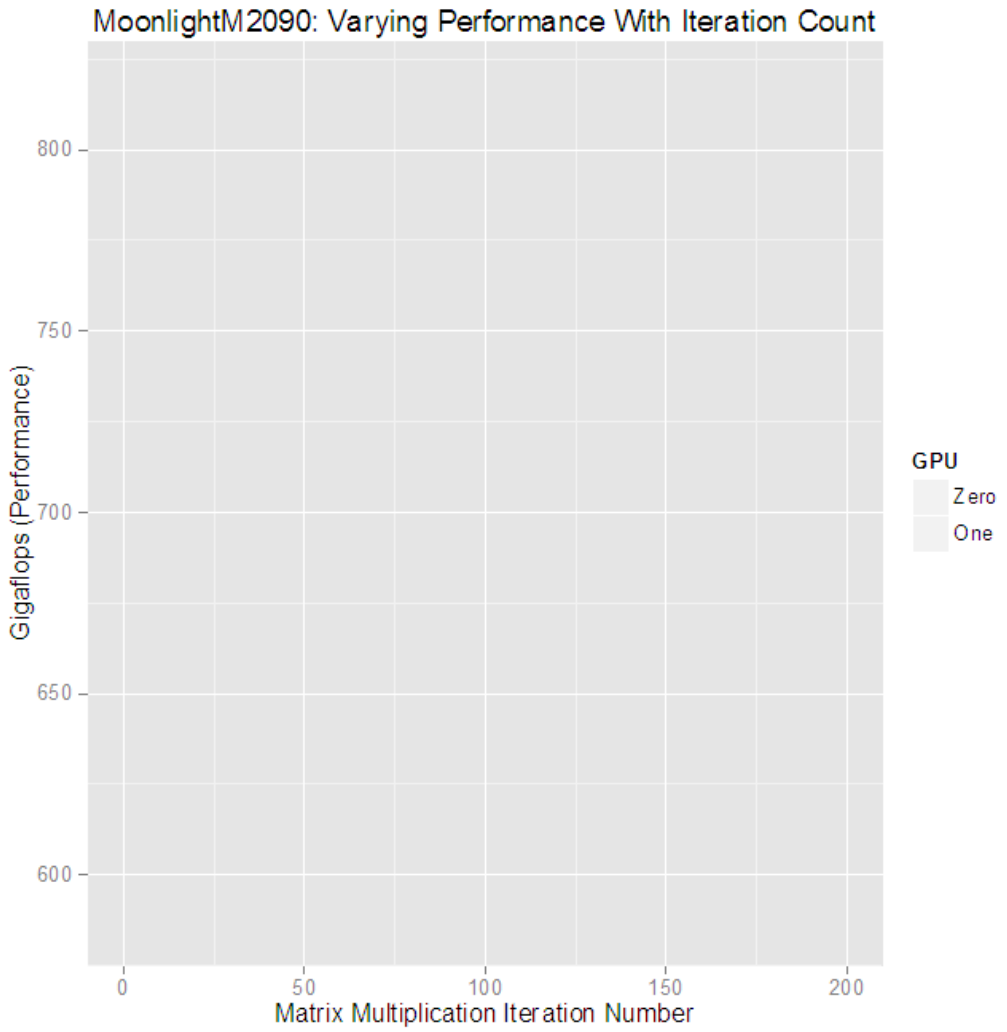
- Performance is variable on this test application as well
- Seem to be “bands” of performance
- GPU zero (red) is clearly under performing



LA-UR-14-22329

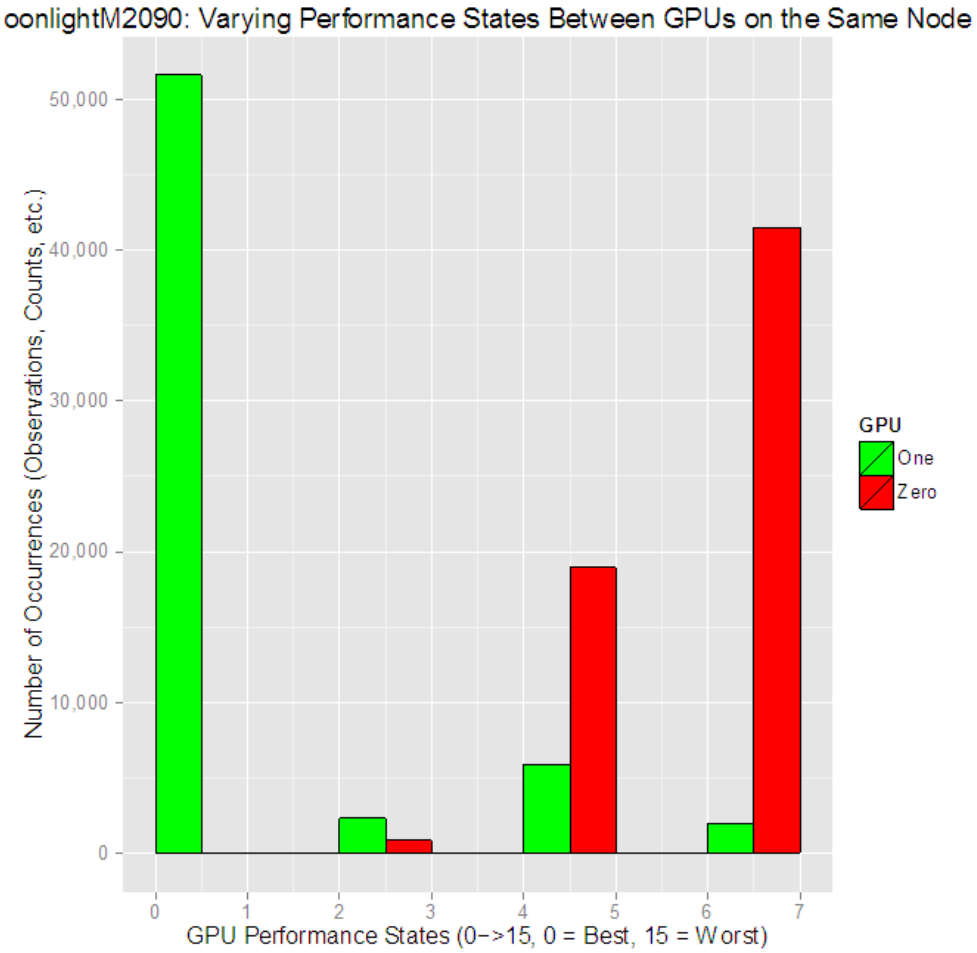
ML M2090 Performance

- Adding an alpha blur shows us that there are clear modes of performance
- GPU zero (red) is clearly under performing
- Why?



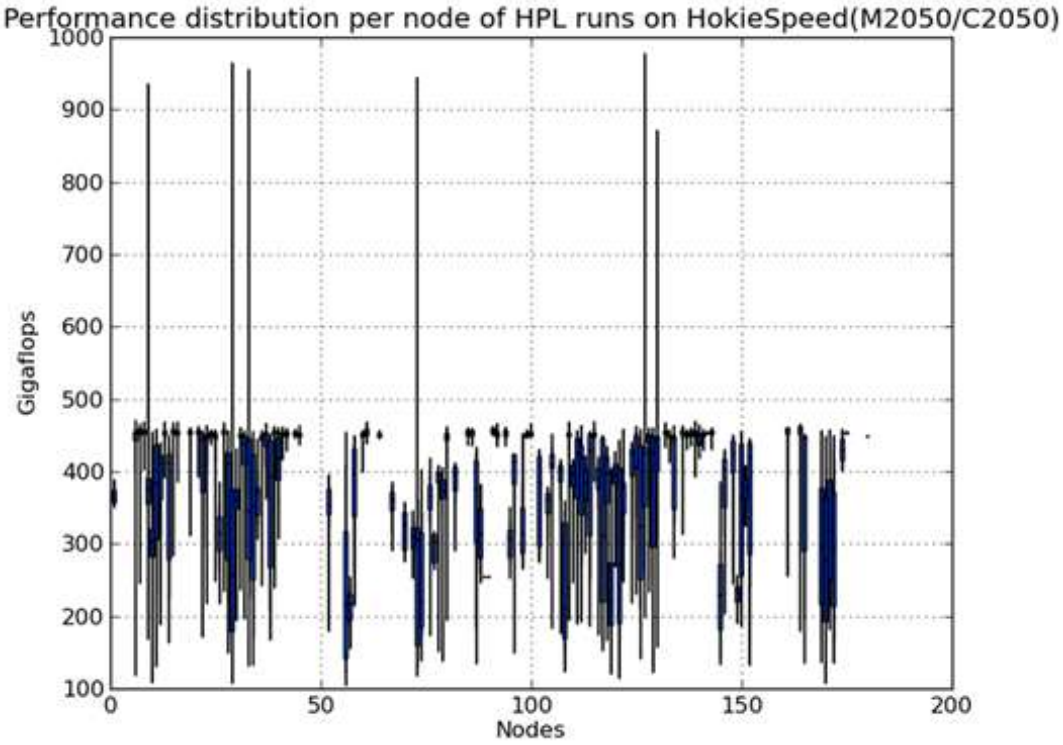
M2090 Performance Throttling?

- Something is causing the M2090s on this machine to throttle
- Sadly, the API calls to find out what is causing it don't appear until the Kepler generation



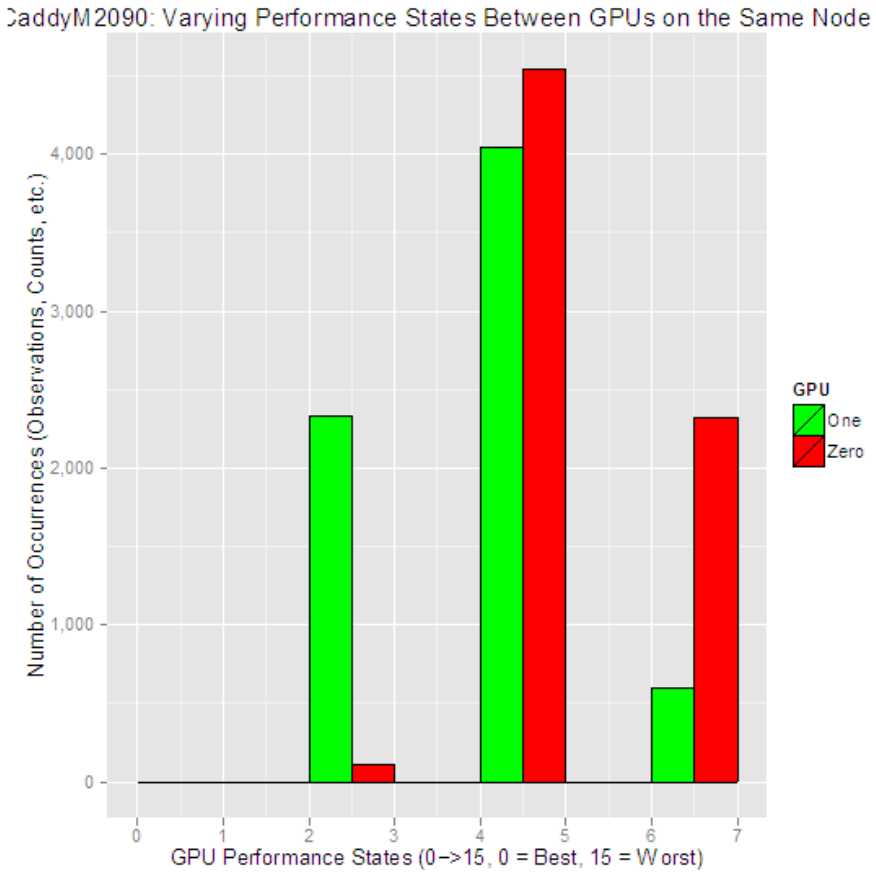
Unique to This One Machine?

- **M2050s tested by Argonne National Laboratory (Franck Cappello, Leonardo Bautista Gomez) on a Virginia Tech GPU cluster**



Unique to This One Machine?

- **M2090 small testbed at LANL also shows this performance variability**



Moonlight Bit Errors

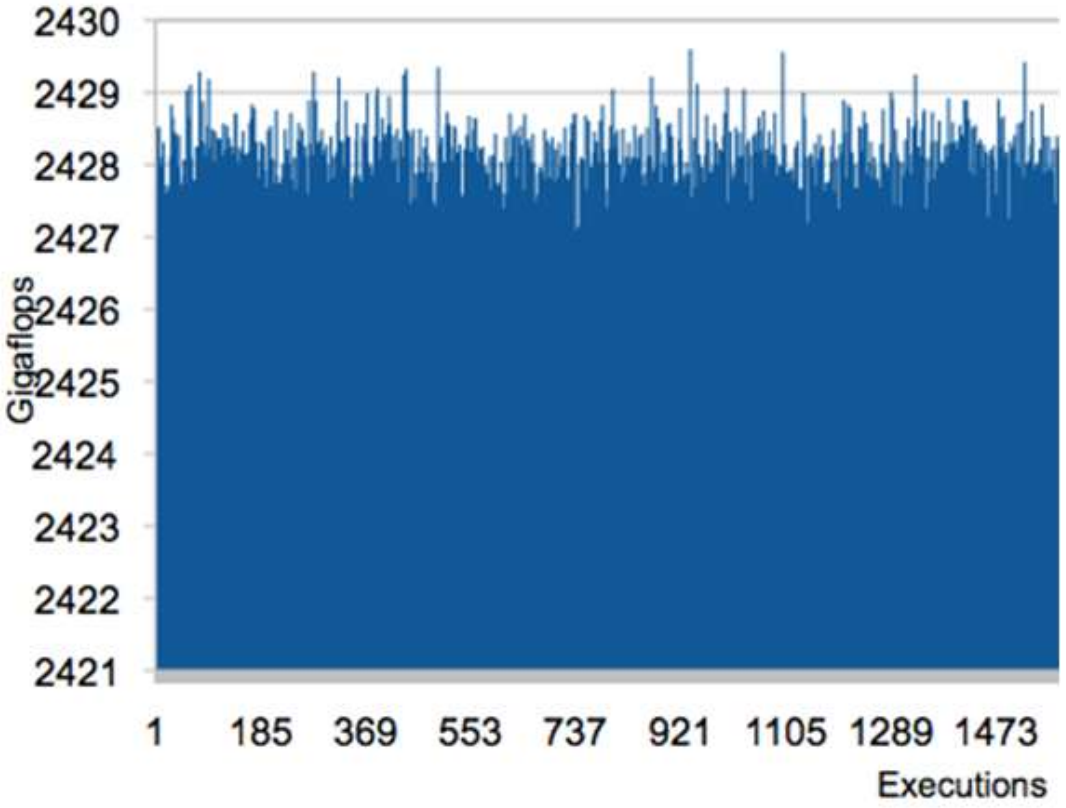
- **1.32 double bit errors / day reported in the logs**
- **~10,000x higher double bit error rate than LANL's Cielo supercomputer (DRAM only) when scaled to memory size**
  - Cielo is DDR3 with single chipkill correct, double chipkill detect
  - Moonlight is GDDR5 with SECDED
- **Inconsistent error logging**
  - This is a very long story, I will be happy to explain this offline to folks if they are interested
  - In short, we see every permutation of logs in the counters of the cards and the syslog leaving us with little confidence in either of them

Do Things Get Better in Kepler Generation?

- **Performance variation on our two (2) K20s at LANL are extremely regular!**
- **Look at that massive performance improvement too!**

Do Things Get Better in Kepler Generation?

- Argonne colleagues confirm similar results on a K20 system in Europe
- Performance variability seems very much improved in Kepler generation



GPGPU Takeaway

- **The cards have evolved a lot since deployed**
- **Sadly, our scientists were asked to evaluate the cards as deployed, not as they are now**
- **Upgrading drivers and studying systems is *extremely* challenging on a production HPC system**
- **Vendors need to be *extremely* cautious of shipping hardware that is not "ready for prime time" in the community they are deploying it**

Conclusions

- **It is not often one gets to see field studies in HPC**
- **We have shown the value of:**
  - Collaborating with vendors to interpret the data (vendor “special sauce”)
  - Analyzing reliability based on vendor choice
  - Studying positional effects of faults in a data center
- **SRAM would benefit from more advanced ECC**
- **SRAM field data lines up relatively well with accelerated testing**
- **DDR3 address and command parity check is useful**
- **Some of the smallest SRAM structures are the most problematic on Cielo due to the low error protection provided in them**
- **With quality ECC, altitude effects are largely mitigated**
- **GPGPUs are an evolving story**

What Kind of ECC Is There? What Do We Use?

- **TERMS:**
  - Correct – found an error, fixed it, reports it to the system, the correct data is returned to you (small performance impact).
  - Detect – found an error, you can't have the correct data – we don't know what it is, today usually this crashes the system, next gen systems this will likely just kill the associated application
  - Silent Data Corruption (SDC) – we don't know these rates, highly application dependent, requires an application that can self-check
- **Takeaway:**
  - Some of these events are benign
  - Some of these events cause your applications to crash (or the node, or middleware)
  - Some of these events can cause SDC
  - **ALL** of these events are based on memory access patterns, usage

What Kind of ECC Is There? What Do We Use?

- **TERMS:**
  - Correct – found an error, fixed it, reports it to the system, the correct data is returned to you (small performance impact).
  - Detect – found an error, you can't have the correct data – we don't know what it is, today usually this crashes the system, next gen systems this will likely just kill the associated application
  - Silent Data Corruption (SDC) – we don't know these rates, highly application dependent, requires an application that can self-check
- **Parity – can tell if 1 bit changes, cannot correct**
- **SECDED (Single Error Correct, Double Error Detect)**
- **DECTED (Double Error Correct, Triple Error Detect)**

Cielo SRAM, Varies by Structure

- **Chipkill:**
  - Single-chipkill detect ("x4")
  - Single-chipkill correct, double-chipkill detect ("x8")
  - Double chip sparing
  - Double-chipkill correct

Cielo DRAM Sometimes

Cielo DRAM Usually