Supercomputer Field Data –

DRAM, SRAM, and Projections for Future Systems

Nathan DeBardeleben, Ph.D. (LANL)
Ultrascale Systems Research Center (USRC)

6th Soft Error Rate (SER) Workshop
Santa Clara, October 16, 2014
Acknowledgements

**Collaborators**
- Vilas Sridharan, AMD RAS, Advanced Micro Devices Inc.
- Sudhanva Gurumurthi, AMD Research, Advanced Micro Devices Inc.
- Jon Stearley, Scalable Architectures, Sandia National Laboratories
- Kurt Ferreira, Scalable Architectures, Sandia National Laboratories
- John Shalf, Computational Research Division, Lawrence Berkeley National Laboratory

**Thanks to**
- Many folks at Cray
- Many folks at LANL, LBNL, SNL, ORNL

**This is a team effort**
- By no means is this entirely my own work!
Cielo Data – A More Detailed Analysis Than Is Often Possible

- Cielo affords us more data sources than are often available
- It’s very easy to jump to wrong conclusions with the kind of data usually available
- Such as . . .
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.
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 closer to 30/70 permanent/transient

Sridharan et al., Feng Shui of Supercomputer Memory, SC 2013
What About Hopper?

- Cielo = 8.5k nodes, Hopper = 6k nodes
- Cielo at 7.3k feet elevation
- Hopper at 43 feet elevation
- Same DRAM vendor IDs, approximately same relative concentration in entire system
- Vendor A higher fault rate in Cielo, likely attributable to altitude

Currently under peer review . . .
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
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.
Where do faults Occur?

- **Data from Hopper**
  - 12,000 CPU sockets
  - 12MB L3 cache / socket
  - 3MB L2 cache / socket
  - ~1.5 years of observation

- **Faults occur often**
  - Even small structures (TLBs) see faults

- **Exascale systems will:**
  - Have 4-5x the number of compute sockets
  - Have much more SRAM per socket
  - Have more faults!

---

Caveat: vendors must pay attention to reliable design
Accelerated Testing Comparison

- Studies of DOE supercomputers compared to AMD accelerated testing
- Accelerated testing remains a good proxy for what is seen in the field
- We would expect lower field FIT rates than accelerated testing due to workload differences, faults being overwritten, etc.

![Fault rate per bit (a.u.) vs. L2\$ Data, L3\$ Data, L2\$ Tag, L3\$ Tag, Accelerated Testing](chart.png)
What will SRAM errors look like in exascale?

SRAM UNCORRECTED ERROR RATE RELATIVE TO CIELO

- **Two potential systems**
  - Small: 50k nodes
  - Large: 100k nodes

- **Same fault rate as 45nm**
  - Sky is falling

- **Scale faults per current trend**
  - Sky falls more slowly
  - Switch to FinFETs may make this even better

- **Add some engineering effort**
  - Sky stops falling

SRAM faults are unlikely to be a significantly larger problem than today

Attribution: Vilas Sridharan, AMD
What about DRAM?
AND OTHER EXTERNAL MEMORY SUBSYSTEMS

- **DRAM faults are...weird**
  - Affect multiple rows/columns/chip
  - Not just simple particle strikes...

- **Many permanent faults**
  - Entirely unlike SRAM

<table>
<thead>
<tr>
<th>Fault Mode</th>
<th>% Faulty DRAMs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-bit</td>
<td>67.7%</td>
</tr>
<tr>
<td>Single-word</td>
<td>0.2%</td>
</tr>
<tr>
<td>Single-column</td>
<td>8.7%</td>
</tr>
<tr>
<td>Single-row</td>
<td>11.8%</td>
</tr>
<tr>
<td>Single-bank</td>
<td>9.6%</td>
</tr>
<tr>
<td>Multi-bank</td>
<td>1.0%</td>
</tr>
<tr>
<td>Multi-rank</td>
<td>1.1%</td>
</tr>
</tbody>
</table>

Attribution: Vilas Sridharan, AMD
Projecting to Exascale

- **Uncorrected error rate**
  - 10-70x error rate of current systems
  - Is the sky falling?

- **This is not just a problem for exascale**
  - Cost problem for data centers / cloud
  - Reliability problem in client?

- **Solutions are out there**
  - Including for die-stacked DRAM
  - Lots of people working on this...

- **Historical example**
  - Chipkill vs. SEC-DED

Caveat: DRAM subsystems need higher reliability than today, but will likely get it

Attribution: Vilas Sridharan, AMD
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
  - Analyzing reliability based on vendor choice
  - Studying positional effects of faults in a data center
- SRAM would benefit from more advanced ECC
- Accelerated testing is useful
- DDR3 address and command parity check is useful
- Exascale trends are a mixed bag:
  - Sky is probably not falling
  - But there is no doubt that user experienced uncorrectable error rates will increase
- Always interested in collaboration!
- I haven’t said anything about silent data corruption here