Field Studies of DRAM Errors

- AMD Boxborough
- University of Toronto & Google
AMD Study on Jaguar

• Carried out over 11 months – approx. 50 M DIMM-days.

• Jaguar – 18,866 nodes each with
  – Two six-core AMD Opterons
  – Four 72-bit DDR2 channels, 2 DIMMs/channel, 1 Rank/DIMM, 18 x4 chips/rank (16 for data + 2 for SSC-DSD)
  – memory scrubber
  – Memory controllers log error events every 5 seconds.
Taxonomy

- **Fault**
  - Underlying failure mode (e.g. stuck at fault or particle induced bit flip)
- **Error**
  - Visible symptom of a fault (e.g. ECC mismatch)
- **Failure**
  - Transition from a period of correct service to incorrect service.

**Types of Faults**

- **Hard Faults**
  - Causes a memory location to consistently return incorrect data
- **Transient Fault**
  - Wrong data until it is overwritten
- **Intermittent fault**
  - Sometimes sends out wrong data (e.g. under elevated temperature)

Intermittent + Hard Faults -> *Recurring Faults*
Transient Faults -> *Non-recurring Faults*
Methodology : Error Logging

- Memory controller logs data from the Machine Check Architecture (MCA) registers every 5 seconds.

- Log has physical address, time stamp, corrected/uncorrected error, ECC codeword.

- MCA regs retain their values across warm resets – thus uncorrected errors that cause such resets can also be logged.

- Overflow bit to indicate that at least one error was not logged.
Methodology: Classifying faults

- Determine the fault type from the error logs.

- Node experiences errors in one scrub interval only
  - Indicates non-recurring fault

- If a node logs errors from a single DRAM device in multiple scrub intervals
  - Indicates either a single recurring fault or multiple non-recurring faults.
  - The study sees very little occurrence of multiple non-recurring faults in a device.
Observations

Average no. of faults/month = 927.5
Average no. of errors per month = 250,000

Figure 3. Corrected errors per month across the Jaguar system.

Figure 4. Faults experienced per month across the Jaguar system.
Observations

**DRAM Failure Incidents:** 2.95 % of DIMMs or 5.9% of nodes

In line with Schroeder study.

**DRAM Fault Rate:** 1 DRAM fault every six to seven hours.

Not rare.
Observations

26.8% of errors manifested only in one scrub interval. A total of 29.6% of all errors are due to non-recurring faults. Recurring faults main culprit – 70%.
Patterns of Failing Addresses

<table>
<thead>
<tr>
<th>Failure Pattern</th>
<th>% Faulty Nodes</th>
<th>Failure Pattern</th>
<th>% Faulty Nodes</th>
<th>Failure Pattern</th>
<th>% Faulty Nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Bit</td>
<td>47.6%</td>
<td>1 Column</td>
<td>10.5%</td>
<td>1 Lane</td>
<td>4.8%</td>
</tr>
<tr>
<td>2 Bits</td>
<td>0.7%</td>
<td>1 Row</td>
<td>12.0%</td>
<td>1 Rank</td>
<td>0.2%</td>
</tr>
<tr>
<td>3 Bits</td>
<td>0.05%</td>
<td>1 Bank</td>
<td>16.2%</td>
<td>2 DRAMs</td>
<td>1.1%</td>
</tr>
<tr>
<td>1 Word</td>
<td>2.4%</td>
<td>1 DRAM</td>
<td>2.4%</td>
<td>1 Channel</td>
<td>0.1%</td>
</tr>
<tr>
<td>2 Words</td>
<td>0.3%</td>
<td>2 Columns</td>
<td>0.5%</td>
<td>1 Node</td>
<td>0.4%</td>
</tr>
<tr>
<td>3 Words</td>
<td>0.1%</td>
<td>2 Rows</td>
<td>0.9%</td>
<td>Total</td>
<td>100%</td>
</tr>
</tbody>
</table>

• % represents nodes which show all errors on the same location.
• 47.6% same bit
• 38.7% to same DRAM row, col, bank.
• 4.8% of errors on the same lane – i.e. either a stuck DQ or DQS pin.
DRAM Errors in the Wild

• Study on Google’s fleet of servers spanning 2.5 years.

• 6 different platforms defined by (motherboard + DIMM type combo)
  • DDR1, DDR2, DDR3 , FB-DIMM (1,2,4Gb)

• Distributed logging and analysis of errors

• Uncorrectable errors always lead to shutdown and DIMM replacement

• No distinction between hard and soft errors.
Errors per machine

- Avg no of correctable errors/year > 22000
- Highly variable no of errors for every platform
  - Coefficient of Variation between 3.4 and 20
  - 20% of machines contribute 90% of errors
- 93% of machines that see 1 correctable error see at least one more in the same year.
Errors per DIMM

<table>
<thead>
<tr>
<th>Platf.</th>
<th>Tech.</th>
<th>CE Incid. (%</th>
<th>CE Rate Mean</th>
<th>CE Rate C.V.</th>
<th>CE Median Aflct.</th>
<th>UE Incid. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>DDR1</td>
<td>21.2</td>
<td>4530</td>
<td>6.7</td>
<td>167</td>
<td>0.05</td>
</tr>
<tr>
<td>B</td>
<td>DDR1</td>
<td>19.6</td>
<td>4086</td>
<td>7.4</td>
<td>76</td>
<td>-</td>
</tr>
<tr>
<td>C</td>
<td>DDR1</td>
<td>3.7</td>
<td>3351</td>
<td>46.5</td>
<td>59</td>
<td>0.28</td>
</tr>
<tr>
<td>D</td>
<td>DDR2</td>
<td>2.8</td>
<td>3918</td>
<td>42.4</td>
<td>45</td>
<td>0.25</td>
</tr>
<tr>
<td>E</td>
<td>FBD</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.08</td>
</tr>
<tr>
<td>F</td>
<td>DDR2</td>
<td>2.9</td>
<td>3408</td>
<td>51.9</td>
<td>15</td>
<td>0.39</td>
</tr>
<tr>
<td>Overall</td>
<td></td>
<td>8.2</td>
<td>3751</td>
<td>36.3</td>
<td>64</td>
<td>0.22</td>
</tr>
</tbody>
</table>

• Avg DIMM sees > 4000 CEs a year
• Error incidences vary by platform type
  • but not DRAM technology type or by manufacturer
• Difference in mobo and DIMM design responsible ??
• Uncorrected errors high for C & D which do not have chipkill – but why is it high for F too ?
• For all platforms 20% of DIMMs contribute > 95% of errors
Correlation between Correctable Errors

- CEs in the same month lead to between 13X & 91X increase in CE probability
- CEs in the previous month lead to between 35X to 228X increase in CE probability
- The number of CEs in a month increases continuously based on the number of CEs in the previous month & is an order of magnitude higher than the CEs in the previous month
Correlation between Uncorrectable Errors

- Strong probability of UEs if there were CEs in the same month.
- Probability of UEs increases with observed CE rates in the same month.
- About 65-80% of UEs are preceded by CEs in the same month.
- Absolute UE probability (1.7-2.3%) is too low to use pre-emptive DIMM replacement.
Correlation between DIMMs on the same m/c

- If there are errors in one DIMM, there is some increase in the probability of errors in another DIMM – but correlation is not as high as in the previous figures.
- Environmental factors not so significant??
Effect of DIMM Capacity

• Doubling the capacity has –ve or no effect.
• But there is not a clear correlation between chip size and error rates/probabilities.
• Other confounding factors at work.
Effect of Temperature

• CE rates increase by 3X when the temperature increases by 20C for B,C&D and by 10C for A.

• Temperature could be a proxy for utilization, i.e. CPU activity and allocated memory capacity – not clear if the temp and error rate relationship is cause-effect type.

• Isolated the temperature effects (by dividing the utilization into deciles and reporting temp effects in each decile)
  • significantly smaller effect of temperature
Effect of Utilization

• With CPU utilization and allocated memory, CE rates grew logarithmically.

• Isolated the effect of utilization by measuring error rates in different temperature ranges
  • shows strong correlation between utilization and error rates

• High error rate for high utilization
  • maybe due to higher detection of errors
  • but platforms with memory scrubbers also show increase – indicating that these are maybe hard errors or errors induced on the motherboard or DIMM datapath
Effect of Age

• CE rates increase quickly as DIMM population ages beyond 10 months.
• This continues till 20 months and then the slope flattens out.
• Older DIMMs that did not have CEs in the past will not develop errors later on.
• Error rates vary similarly with age for all different types of DRAMs.
• Very little infant mortality – DIMMs are burnt-in prior to putting them into servers?
Conclusions

• A third of machines and 8% of DIMMs saw at least one CE per year, much higher than what lab studies of DIMMs have indicated.
• Chipkill enabled nodes have 4-10 times lower UE rates compared to SECDED ones.
• Memory errors are strongly correlated.
• Incidence of CEs increases with age and the incidence of UEs decrease with age (because the bad ones are replaced).
• No evidence that newer generation DIMMs are any worse than older ones.
• Temperature has a surprisingly low effect on memory errors.
• Error rates are strongly correlated with utilization.
• Error rates are unlikely to be dominated by soft errors.