Feng Shui of Supercomputer Memory

Positional Effects in DRAM and SRAM Faults

Vilas Sridharan
RAS Architecture
Advanced Micro Devices, Inc.
Boxborough, MA
vilas.sridharan@amd.com

Jon Stearley
Scalable Architectures
Sandia National Laboratories[1]
Albuquerque, New Mexico
jrstear@sandia.gov

Nathan DeBardeleben
Ultrascale Systems Research Center
Los Alamos National Laboratory[2]
Los Alamos, New Mexico
ndebard@lanl.gov

Sean Blanchard
Ultrascale Systems Research Center
Los Alamos National Laboratory[2]
Los Alamos, New Mexico
seanb@lanl.gov

Sudhanva Gurumurthi
AMD Research
Advanced Micro Devices, Inc.
Boxborough, MA
sudhanva.gurumurthi@amd.com

ABSTRACT

Several recent publications confirm that faults are common in high-performance computing systems. Therefore, further attention to the faults experienced by such computing systems is warranted. In this paper, we present a study of DRAM and SRAM faults in large high-performance computing systems. Our goal is to understand the factors that influence faults in production settings.

We examine the impact of aging on DRAM, finding a marked shift from permanent to transient faults in the first two years of DRAM lifetime. We examine the impact of DRAM vendor, finding that fault rates vary by more than 4x among vendors. We examine the physical location of faults in a DRAM device and in a data center; contrary to prior studies, we find no correlations with either. Finally, we study the impact of altitude and rack placement on SRAM faults, finding that, as expected, altitude has a substantial impact on SRAM faults, and that top of rack placement correlates with 20% higher fault rate.

[1]Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under Contract DE-AC04-94AL85000. This document's Sandia identifier is 2013-3402C.

[2]A portion of this work was performed at the Ultrascale Systems Research Center (USRC) at Los Alamos National Laboratory, supported by the U.S. Department of Energy contract DE-FC02-06ER25750. The publication has been assigned the LANL identifier LA-UR-13-22888.

SC13 November 17-21, 2013, Denver, CO, USA

1. INTRODUCTION

Recent studies have confirmed that faults are common in memory systems of high-performance computing systems [23]. Moreover, the U.S. Department of Energy (DOE) currently predicts an exascale supercomputer in the early 2020s to have between 32 and 100 petabytes of main memory, a 100x to 350x increase compared to 2012 levels [6]. Similar increases are likely in the amount of cache memory (SRAM) in an exascale system. These systems will require comparable increases in the reliability of both SRAM and DRAM memories to maintain or improve system reliability relative to current systems. Therefore, further attention to the faults experienced by memory sub-systems is warranted. A proper understanding of hardware faults allows hardware and system architects to provision appropriate reliability mechanisms, and can affect operational procedures such as DIMM replacement policies.

In this paper we present a study of DRAM and SRAM faults on two large high-performance computer systems. Our primary data set comes from Cielo, an 8,500-node supercomputer located at Los Alamos National Laboratory (LANL). A secondary data set comes from Jaguar, an 18,688-node supercomputer that was located at Oak Ridge National Laboratory. In Cielo, our measurement interval is a 15-month period from mid-2011 through early 2013, comprising 23 billion DRAM device-hours of data. In Jaguar, our measurement interval is an 11-month period from late 2009 through late 2010, comprising 17.1 billion DRAM device-hours of data. Both systems were in production and heavily utilized during their respective measurement intervals.

There are several contributions of this research:

- We study the impact of aging on the DRAM fault rate. In contrast to previous studies [21], we find that the composition of DRAM faults changes substantially during the first two years of DRAM lifetime, shifting from primarily permanent faults to primarily transient faults.
- We examine the impact of DRAM vendor and device

choice on DRAM reliability. We find that overall fault rates vary among DRAM devices in our study by up to 4x, and transient fault rates vary by up to 7x.

- We study the physical location of faults in a DRAM device. With the exception of one device-specific fault mode, we find an approximately uniform distribution of faults across DRAM row, column and bank addresses, in contrast to previous studies.

- We study the impact of location in a datacenter on DRAM fault rates. We find that correlations with datacenter location are fully explained by the mix of DRAM device across location. We conclude that analyses of external factors on DRAM reliability (e.g. the effects of temperature on DRAM reliability) must correct for the mix of devices in the data set or else they may lead to erroneous conclusions.

- We examine the impact of altitude and position in the data center on SRAM faults. We find that, as expected, altitude has a significant effect on the fault rate of SRAMs in the field. We also find that SRAM devices experience 20% higher transient fault rates when placed in "top of rack" nodes.

The rest of this paper is organized as follows. Section 2 defines the terminology we use in this paper. Section 3 discusses related studies and describes the differences in our study and methodology. Section 4 explains the system and DRAM configurations of Cielo and Jaguar. Section 5 describes the data we analyzed and the methodology for that analysis. Section 6 presents results on aggregate DRAM fault rates across the entire Cielo system. Section 7 looks at DRAM fault modes, the fault distribution in a DRAM device, and the impact of placement in a data center. Section 8 discusses location effects on SRAM fault rates, including placement in a data center and altitude. Finally, Section 9 discusses implications of our findings and presents our conclusions.

2. TERMINOLOGY

In this paper, we distinguish between a fault and an error as follows [3]:

- A fault is the underlying cause of an error, such as a stuck-at bit or high-energy particle strike. Faults can be *active* (causing errors), or *dormant* (not causing errors).

- An error is an incorrect portion of state resulting from an active fault, such as an incorrect value in memory. Errors may be *detected* and possibly *corrected* by higher-level mechanisms such as parity or error correcting codes (ECC). They may also go *uncorrected*, or in the worst case, completely *undetected* (e.g. silent).

Computers typically log error detections (indicating time and location), not fault activations. Therefore, one active fault can result in many error messages if the faulty location is accessed multiple times. For the remainder of this paper, a DRAM fault corresponds to the first observed error message per DRAM device. Additional details are given in Section 5.

Hardware faults can further be classified as [9]:

- *Transient faults*, which cause incorrect data to be read from a memory location until the location is overwritten with correct data. These faults occur randomly and are not indicative of device damage [5]. Particle-induced upsets ("soft errors"), which have been extensively studied in the literature [5][26], are one type of transient fault.

- *Hard faults*, which cause a memory location to consistently return an incorrect value (e.g., a stuck-at-0 fault). Generally, hard faults can be repaired only by disabling the component in question or by replacing the faulty device [10].

- *Intermittent faults*, which cause a memory location to sometimes return incorrect values. Unlike hard faults, intermittent faults occur only under specific conditions such as elevated temperature [9]. Unlike transient faults, however, an intermittent fault is indicative of device damage or malfunction.

Distinguishing a hard fault from an intermittent fault in a running system requires knowing the exact memory access pattern to determine whether a memory location returns the wrong data on every access. In practice, this is impossible in a large-scale field study such as ours. Therefore, we group intermittent and hard faults together in a category of *permanent* faults.

3. RELATED WORK

During the past several years, multiple studies have been published examining DRAM failures in the field. In 2006, Schroeder and Gibson studied failures in high-performance computer systems at LANL [20]. In 2007, Li et al. published a study of memory errors on three different data sets, including a server farm of an Internet service provider [16]. In 2009, Schroeder et al. published a large-scale field study using Google's server fleet [21]. In 2010, Li et al. published an expanded study of memory errors at an Internet server farm and other sources [15]. In 2012, Hwang et al. published an expanded study on Google's server fleet as well as two IBM Blue Gene clusters [14], Sridharan and Liberty presented a study of DRAM failures in a high-performance computing system [23], and El-Sayed et al. published a study on temperature effects of DRAM in data center environments [12]. In 2013, Siddiqua et al. presented a study of DRAM failures from client and server systems [22].

Our study contains analyses not performed in many of these previous studies, including: the effects of DRAM vendor choice on DRAM faults; the effect of aging on the rate of transient and permanent DRAM faults; and an examination of SRAM faults in the field. In addition, some previous studies use corrected error rates, rather than fault rates, as a metric [14][21]. This makes it difficult to compare our results to these studies. Moreover, chipkill ECC, which is prevalent in high-performance computing and cloud datacenters, allows any error from a single DRAM device (i.e. any error from a single fault) to be corrected. An uncorrected error will result only when two or more faults overlap in the same ECC word. Therefore, the relevant question for datacenter operators is not where the next error will come from, but where the next fault will come from.

There also has been significant accelerated testing work on DRAM devices dating back several decades [7][17][18][19].

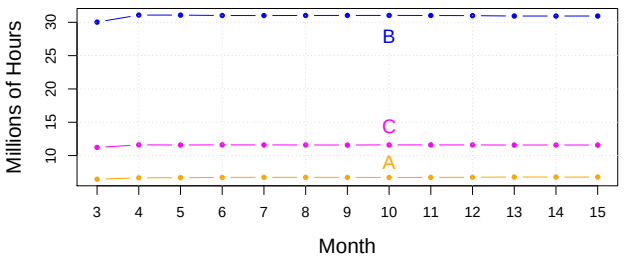


Figure 1: DRAM use per month was roughly constant for manufacturer A, B, and C. Aggregate totals are given in Figure 3(a). The first two months are omitted as explained in Section 6.2.

Of particular interest are the studies by Borucki and Quinn that identified significant variation in per-vendor and per-device fault modes and rates in a neutron beam. As far as we are aware, ours is the first study to examine this effect in the field.

4. SYSTEMS CONFIGURATION

We examine two systems in this paper: Cielo, a supercomputer located in Los Alamos, New Mexico at around 7,300 feet in elevation; and Jaguar, a supercomputer located in Oak Ridge, Tennessee, at approximately 875 feet in elevation.

Cielo contains approximately 8,500 compute nodes. Each Cielo *node* contains two 8-core AMD Opteron™ processors, each with eight 512KB L2 and one 12MB L3 cache. Each Cielo compute node has eight 4GB DDR-3 DIMMs for a total of 32GB of DRAM per node.

Cielo contains DRAMs from three different memory vendors. We anonymize DRAM vendor information in this publication and simply refer to DRAM vendors A, B, and C. As shown in Figure 1, the relative compositions of these DRAM manufacturers remain constant through the lifetime of Cielo.

During our measurement interval, Jaguar (which was upgraded in 2012 and is now named Titan) contained 18,688 nodes. Each node contained two 6-core AMD Opteron processors, each with six 512KB L2 and one 6MB L3 caches. Each Jaguar node has eight 2GB DDR-2 DIMMs for a total of 16GB of DRAM per node. We do not have DRAM vendor information for Jaguar.

The nodes in both machines are organized as follows. Four nodes are connected to a *slot* which is a management module. Eight slots are contained in a *chassis*, of which there are three mounted bottom-to-top (numerically) in a *rack*. Cielo has 96 racks, arranged into 6 *rows* each containing 16 racks.

At 7,320 feet in altitude, the Cielo system at LANL is subject to a higher flux of cosmic ray-induced neutrons than Jaguar at ORNL at 850 feet. The average flux ratio between the two locations due to altitude, longitude and latitude without accounting for solar modulation is 4.39 [1].

4.1 DRAM and DIMM Configuration

In Cielo, each DDR-3 DIMM contains two *ranks* of 18 DRAM devices, each with four data (DQ) signals (known as an x4 DRAM device). In each rank, 16 of the DRAM devices are used to store data bits and two are used to store check (ECC) bits. A *lane* is a group of DRAM devices on different ranks that shares data (DQ) signals. A memory *channel* has 18 lanes, each with two ranks (i.e. one DIMM per channel). DRAMs in the same lane also share a strobe (DQS) signal, which is used as a source-synchronous clock signal for the data signals. Each DRAM device contains eight internal *banks* that can be accessed in parallel. Logically, each bank is organized into *rows* and *columns*. Each row/column address pair uniquely identifies a 4-bit *word* in the DRAM device.

Physically, all DIMMs on Cielo (from all manufacturers) are identical. Each DIMM is double-sided. DRAM devices are laid out in two rows of nine devices per side. There are no heatsinks on any DIMMs in Cielo.

In Jaguar, each DDR-2 DIMM contains one rank of 18 x4 DRAM devices. Each memory channel contains 18 lanes with two ranks (i.e. two DIMMs per channel). The internal DRAM logical organization is similar to that of DRAMs on Cielo. Physically, each DIMM contains a single row of nine DRAM devices per side.

5. EXPERIMENTAL SETUP

For our analysis we use two different data sets - corrected error messages from console logs and hardware inventory logs.

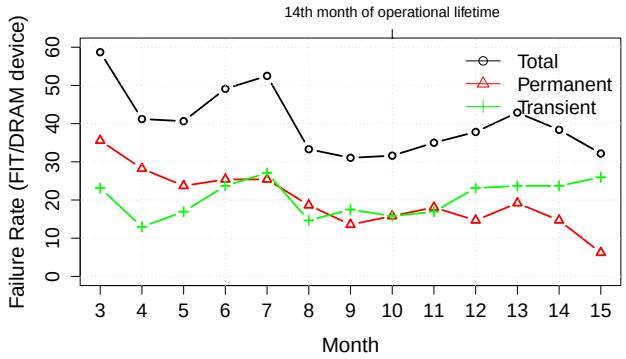
Corrected error logs contain events from nodes at specific time stamps. Each node in the system has a hardware memory controller that logs corrected error events in registers provided by the x86 machine-check architecture (MCA) [2]. Each node's operating system is configured to poll the MCA registers once every few seconds and record any events it finds to the node's console log.

The console logs contain a variety of information, including the physical address associated with the error, the time the error was recorded, and the ECC syndrome associated with the error. These events then are decoded further using memory controller configuration information to determine the DRAM location associated with the error. For this analysis we decoded the location to show the DIMM, as well as DRAM bank, column, row, rank, and lane.

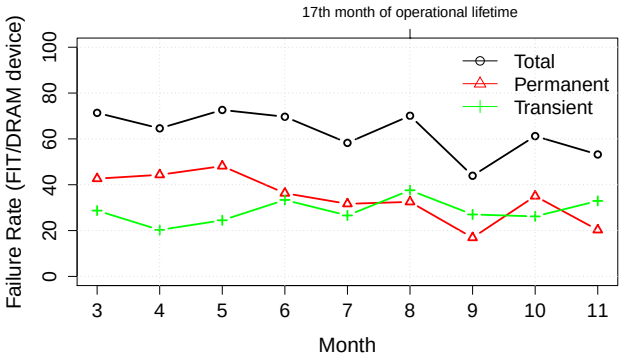
Hardware inventory logs are separate logs and provide snapshots of the hardware present in Cielo at different points in its lifetime. We analyzed 217 hardware inventory logs that covered a span of approximately two years from early 2011 to 2013. Each log file consists of more than 1.3 million lines of explicit description of each host's hardware. For our analysis, this provided detailed information about each DRAM DIMM attached including the manufacturer, part number, and much more.

These two types of logs provided the ability to map error messages to specific hardware present in the machine at that point in time. All the DIMM manufacturer data presented in this paper has been anonymized to protect interested parties.

All data and analyses presented in this paper refer to faults, not errors. Our observed fault rates indicate that fewer than two DRAM devices will suffer multiple faults within our observation window. Therefore, similar to previous field studies, we make the simplifying assumption that each DRAM device experiences a single fault during our observation interval [23]. The occurrence time of each DRAM fault corresponds to the time of the first observed error message per DRAM device. We then assign a specific type and mode to each fault based on the associated errors in the



(a) Cielo DDR3 DRAM device fault rates per month (30-day period); 23 billion DRAM hours total.



(b) Jaguar DDR2 DRAM device fault rates per month (30-day period); 17.1 billion DRAM hours total.

Figure 2: DRAM device fault rates over time.

% Faulty DRAMs	0.038%
% Faulty DIMMs	1.32%
Fault Rate (FIT/Mbit)	0.044
Fault Rate (FIT/device)	40.33

Table 1: DRAM Fault Rates.

System	0	1	2	3
Cielo	90.07%	9.10%	0.75%	0.08%
Jaguar	94.07%	5.48%	0.39%	0.06%

Table 2: Percentage of hosts with 0, 1, 2, or 3 faulty DRAMs.

console logs. We use a similar methodology (based on fault rates) for SRAM faults.

Because both Jaguar and Cielo include hardware scrubbers in DRAM, L2 and L3 caches, we can identify permanent faults as those faults that survive a scrub operation. Thus, we classify a fault as permanent when a device generates error messages in multiple scrub intervals, and transient when it generates errors in only a single scrub interval. In Cielo, the DRAM scrub interval is 24 hours, the L2 SRAM scrub interval is 10 seconds, and the L3 SRAM scrub interval is 129 seconds.

6. DRAM FAULT RATES

In this section, we present data on aggregate DRAM fault rates. We also examine the distribution of transient and permanent faults, and the impact of vendor and device on fault rates.

6.1 Aggregate Fault Rates

Table 1 shows aggregate fault rates for DRAM in Cielo, including the fault rate per megabit and fraction of DRAMs and DIMMs experiencing a fault. The table shows that 1.32% of DIMMs, and 0.04% of DRAM devices, experienced a fault during the experiment. The calculated fault rate of 0.044 FIT/Mbit translates to one fault approximately every 11 hours across the Cielo system. These results are similar to fault rates and "corrected error incidence per DIMM" reported by other field studies on DDR-2 DRAM [21][23]. This is important because it provides a data point showing that DRAM fault rates are similar across at least two technology generations.

Table 2 shows the fraction of nodes in Cielo and Jaguar with zero, one, two, and three DRAM faults. Slightly more than five percent of nodes on Jaguar experienced at least one faulty DRAM during our measurement interval, versus just under ten percent on Cielo, possibly due to altitude effects (see Section 8.3). The table shows that, in both systems, the number of hosts experiencing one, two, or three faulty DRAMs decreases by roughly an order of magnitude at each level, suggesting that faults are independent among DRAMs.

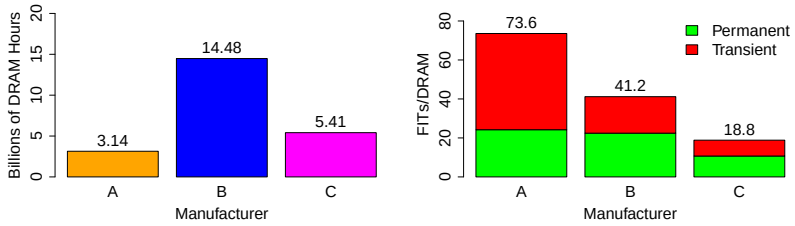
6.2 Fault Rates over Time

Figure 2(a) shows the total number of DRAM faults per month (30-day period) in Cielo. We omit the first two months of the data set because this would result in "over-counting" permanent faults that developed between the beginning of the system's lifetime and the start of our measurement interval. The figure shows that Cielo experienced a declining rate of DRAM faults during our measurement interval, matching results found by other studies that take place towards the beginning of a system's lifetime [23]. The figure further shows that this declining total rate of faults is comprised of an approximately constant rate of transient faults and a rapidly declining rate of permanent faults (similar to the trend shown by Siddiqua et al. [22]). The crossover point between permanent and transient faults occurs near the tenth month of the data set, which represents the fourteenth operational month of the Cielo system.

Figure 2(b) shows the same data for the Jaguar system. This figure shows a similar declining trend in the permanent fault rate of the DDR-2 DRAM in Jaguar. In Jaguar, we see the crossover point between permanent and transient faults near the eighth month of the data set, which represents the seventeenth operational month of the Jaguar system.

6.3 Fault Rates by DRAM Vendor

Figure 3(a) shows the aggregate number of DIMM-hours per DRAM vendor in Cielo during our observation period. Our observation period consists of 3.14, 14.48, and 5.41 billion device-hours for DRAM vendors A, B, and C, respectively. Therefore, we have enough operational hours on each vendor to make statistically meaningful measurements of each vendor's fault rate.



(a) Operational hours per vendor (b) Fault rates per vendor

Figure 3: Operational hours and fault rate by vendor.

Figure 3(b) shows the fault rate experienced by each vendor during this period, divided into transient and permanent faults. The figure shows a substantial difference among vendors. Vendor A has a 3.9x higher fault rate than Vendor C. This figure also shows that the permanent fault rate varies by 2.3x among vendors, from 24.2 FIT to 10.7 FIT per DRAM device, while the transient fault rate varies by more than 6x among vendors, from 49.4 FIT to 8.1 FIT per DRAM device. The figure also shows that Vendor A's transient fault rate is larger than its permanent fault rate, while the other two vendors have higher permanent than transient fault rates.

In Cielo, over 50% of the faults are transient, in contrast to previous studies that have pointed to permanent faults as the primary source of faults in modern DRAM [21][23]. Our data indicates that this conclusion depends heavily on the mix of DRAM vendors in the system under test. Another interesting result in the figure is that transient and permanent fault rates vary together, so the vendor with the highest transient fault rate also has the highest permanent fault rate. It is unclear why this should be the case, but may indicate shared causes between transient and permanent faults.

6.4 Conclusions

Our data leads to three main conclusions. First, although we see a slightly lower fault rate in Cielo than in Jaguar, overall DRAM fault rates appear to be similar across the DDR-2 and DDR-3 generations. This confirms previous accelerated testing findings that DRAM vendors maintain an approximately constant fault rate per device [7].

Second, we conclude that fault rates, and not error rates, are the appropriate metric to examine aging effects in DRAM. Prior studies that looked at error rates compared to DRAM age either did not find significant effects of aging or found effects only beyond the first 10-18 months of lifetime [21][22]. In contrast, we find a decrease in the permanent fault rate over time (i.e., the leading edge of the bathtub curve) but no corresponding decrease in transient fault rate. This implies that the primary fault type experienced by DRAMs depends on the age of the DRAM, with a shift from permanent to transient faults during the first year and a half of system operation. Continued observation of our systems may indicate a shift back to permanent faults as the DRAMs reach the rising edge of the bathtub curve.

Third, we find that both transient and permanent fault rates vary by DRAM vendor, with the overall fault rate varying by 4x among vendors. Therefore, we conclude that the choice of DRAM vendor and device plays a major role in the overall fault rate and types of fault experienced by a system.

Fault Mode	Total Faults	Transient	Permanent
Single-bit	67.7%	34.9%	32.8%
Single-word	0.2%	0.2%	0%
Single-column	8.7%	3.8%	4.9%
Single-row	11.8%	5.7%	6.1%
Single-bank	9.6%	4.0%	5.5%
Multiple-bank	1.0%	0.2%	0.8%
Multiple-rank	1.1%	0.5%	0.5%

Table 3: DRAM fault modes.

7. DRAM LOCATION EFFECTS

Previous studies have confirmed the existence of faults that affect a single bit, word, column, row, and bank, as well as faults that affect multiple banks in a device and multiple ranks within a lane [14][23]. In this section, we examine the prevalence of these different fault modes in our dataset and their likelihood of occurrence by location in the DRAM and by vendor. We also examine correlations between a node's location in the datacenter and the rate of DRAM faults it experiences.

7.1 Fault Modes

Table 3 shows a breakdown of fault modes in Cielo, for both permanent and transient faults. The table shows that 67.7% of faults in Cielo are single-bit faults, while 32.3% are larger multi-bit faults. Cielo's DDR-3 DRAM experiences all the same fault modes observed by prior DDR-2 field studies, indicating that DDR-3 devices remain susceptible to all of these fault modes. Our data show a higher percentage of single-bit faults in Cielo's DDR-3 memory than found in Jaguar's DDR-2 memory by Sridharan and Liberty (67.7% versus 49.7%) [23]. Our data also show that almost 50% of all single-column, single-row, and single-bank faults in Cielo are transient. By contrast, over 90% of these faults are permanent in Jaguar [23]. Because Cielo and Jaguar differ in many ways, it is impossible to determine whether these differences are due to inherent differences between DDR-2 and DDR-3 or to external factors such as altitude or DRAM vendor.

Table 4 shows the same data broken down by vendor. The table shows that not all vendors experience fault modes at the same rates. For instance, devices from Vendor B are more likely to experience single-bit faults than devices from Vendor C. Similarly, Vendor C is the most likely to experience multiple-bank and multiple-rank faults. Prior work showed that these fault modes are most likely to lead to an uncorrected error [23]. Therefore, Vendor C's low overall fault rate may not translate into a low rate of uncorrected er-

Fault Mode	Vendor A	Vendor B	Vendor C
Single-bit	64.6%	69.5%	58.4%
Single-word	0%	0.3%	0%
Single-column	8.7%	8.8%	11.9%
Single-row	12.2%	10.6%	14.9%
Single-bank	13.5%	7.8%	9.9%
Multiple-bank	1.3%	0.7%	2.0%
Multiple-rank	1.3%	3.0%	3.0%

Table 4: DRAM fault modes by vendor.

rors. The table also shows that only Vendor B experienced single-word faults. Because these faults are rare, and we have substantially more operational hours on devices from Vendor B, it is impossible to determine whether this is a real effect or whether this is simply due to statistical variation.

7.2 Fault Distribution in a Device

In this section, we examine the distribution of DRAM faults in a DRAM device.

A previous field study by Hwang et al. showed a correlation between error rate and location in a DRAM device [14]. However, the error rate from a given DRAM location is affected both by the fault rate and the memory access pattern of that node. As a result, the data presented by Hwang is consistent with either of two assertions: (1) faults are non-uniformly distributed; or (2) memory accesses (error detections) are non-uniformly distributed. The data presented by Hwang cannot be used to prove or disprove assertion (1) due to confounding from (2). As far as we are aware, ours is the first study to examine the distribution of faults in a DRAM device (i.e. to prove or disprove assertion (1)).

To perform this study, we plot the locations associated with each fault in our Cielo dataset. For single-bit faults, we plot the row, column, and bank address in which each fault occurred. For single-row faults, we plot the row address in which each fault occurred. For single-column faults, we plot the column address of each fault, and for single-bank faults, we plot the bank address of each fault.

Figure 4 plots the location of single-row, single-column, and single-bank faults, respectively. For single-row and single-column faults (Figure 4(a) and 4(b)), most locations had no faults or one fault, while a few locations had two faults. This is consistent with a uniform random distribution throughout the DRAM device. Therefore, we conclude that the rate of row and column faults has no relationship to DRAM location.

For single-bank faults, shown in Figure 4(c), we see the number of faults in a bank vary between five (bank 0) and 18 (bank 2). In Jaguar (not shown in the figure), the number of faults per bank varies between 15 (bank 3) and 34 (bank 5). These variations may be indicative of a pattern, but are likely to be statistical noise due to the relatively small number of single-bank faults in the system (fewer than 100 in Cielo).

Figure 5 shows the row, column, and bank distribution of all single-bit faults. Figures 5(a) and 5(c) show fault rates per row and bank consistent with a uniform random distribution. Figure 5(b), by contrast, shows a significant spike in the fault rate in column 0, while the remainder of the data appears to be distributed randomly across columns. Figure 6 "zooms in" on eight column addresses, including column 0, and further breaks down the data by vendor and fault type. The figure demonstrates that the spike in column 0 is dominated by transient faults due to a single DRAM vendor (this fault mode accounts for 10 of vendor A's 73 total FITs shown in Figure 3(b)). Because a DRAM column spans all DRAM addresses in a bank, the spike in column 0 would not manifest as being towards the "top" or "bottom" of a node's physical address range, but instead would be distributed across all addresses in the bank.

7.3 Location in the Data Center

The physical conditions in a large machine room can vary widely. For example: poor cooling may lead to hot spots, or an improperly installed circuit may lead to voltage spikes. The LANL data center is carefully designed and heavily monitored to minimize such effects. We examined Cielo fault data with respect to physical location to verify there were no facilities-based effects.

Most observed variances across physical location in the LANL machine room were uninteresting or statistically inconclusive. However, there is one notable exception to the lack of variance, shown in Figure 7(a). Lower-numbered racks show significantly higher DRAM fault rates than higher-numbered racks (with faults aggregated across rows). Without any further information, this trend could be attributed to temperature or other environmental differences across racks.

However, when examining operational hour data by vendor in Figure 7(b), it is clear that lower-numbered racks had significantly more operational hours from Vendor A than higher-numbered racks, which had more operational hours from Vendor C than lower-numbered racks. (Racks 3, 5, 7, and 9 show fewer operational hours because they contain visualization nodes with different hardware, which we omitted from this study.) As shown in Figure 3, Vendor A has a higher overall fault rate than Vendor C. Therefore, racks with DRAM from Vendor A will naturally experience a higher fault rate than racks with DRAM from Vendor C. In the case of Cielo, this translates to lower-numbered racks having higher fault rates than higher-numbered racks.

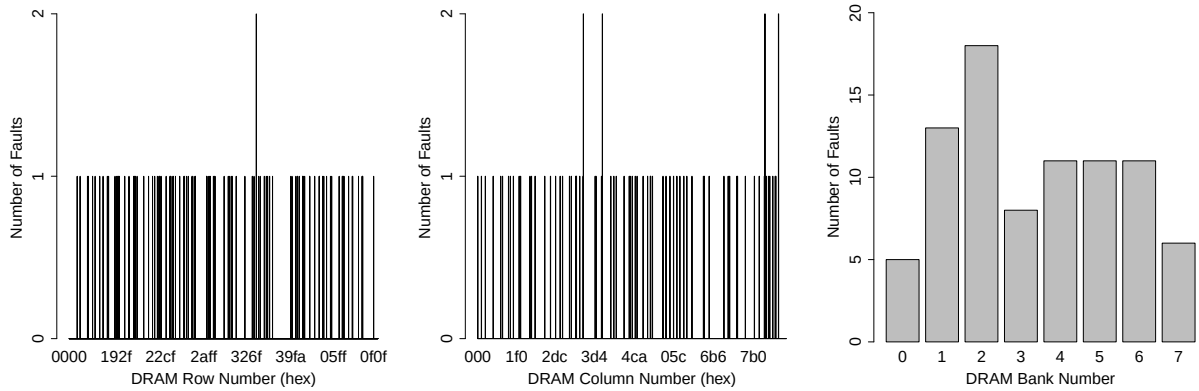
When we examine the by-rack fault rates by vendor (Figure 7(c)), the by-rack fault rate trend essentially disappears. There is a slight trend in the fault rates across racks for Vendor B that is currently unexplained, but this is a very weak effect and may be due to statistical variation rather than a true effect.

7.4 Conclusions

We draw several conclusions from this data. First, we find that all DRAM fault modes identified in DDR-2 DRAMs remain a concern in DDR-3 devices in the field and are present across multiple DRAM vendors. Therefore, we conclude that these fault modes are an inherent consequence of DRAM organizaton, and will likely be present in any DRAM device.

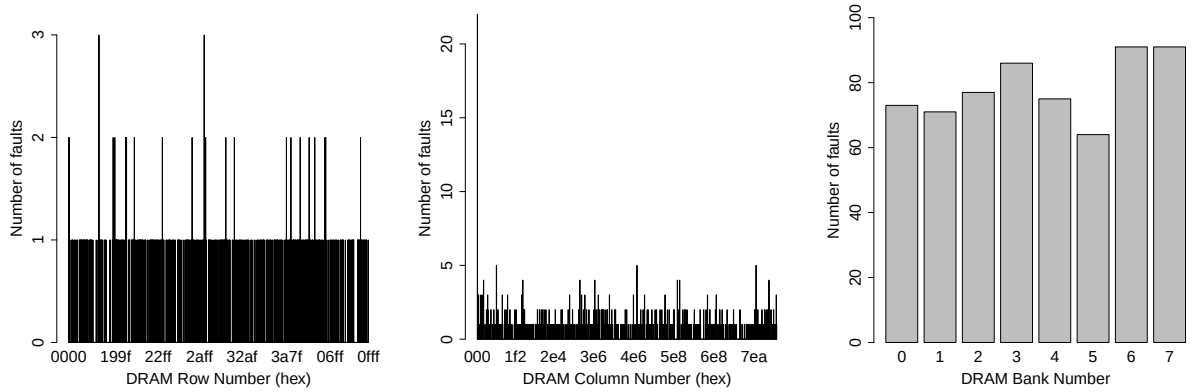
Second, with one vendor-specific exception, we see no evidence of correlation between fault rate and location in a DRAM device. DRAM faults in our system are consistent with a uniform random distribution of faults in a device, implying that DRAM faults are equally likely to occur in any region of a DRAM device.

Finally, we conclude that any analysis of DRAM reliability (e.g. the effects of location, temperature, or altitude on fault



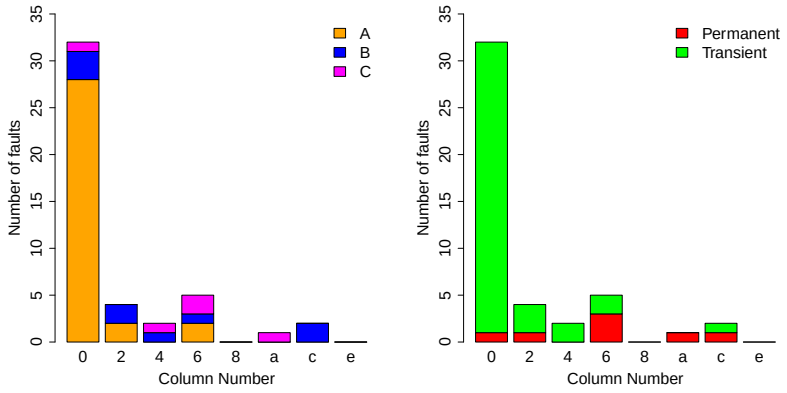
(a) Single-row faults by row (b) Single-column faults by column (c) Single-bank faults by bank

Figure 4: Distribution of single-row, single-column, and single-bank faults in a DRAM device in Cielo.



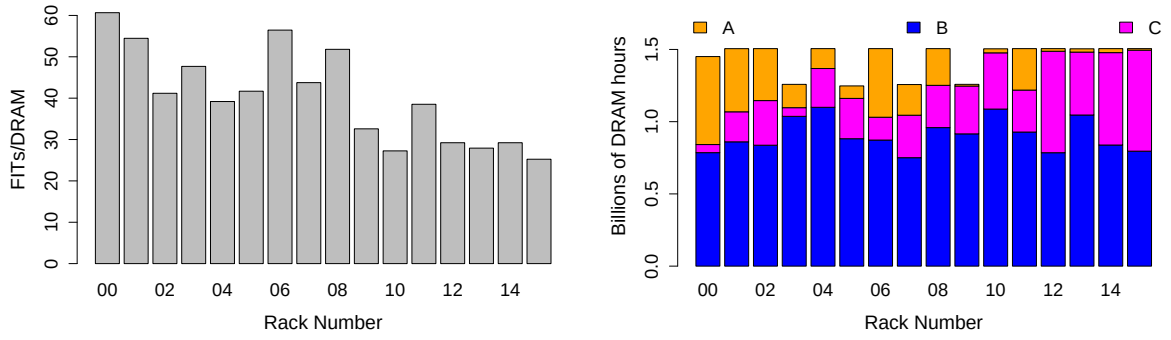
(a) Distribution over rows (b) Distribution over columns (c) Distribution over banks

Figure 5: Distribution of single-bit faults in a DRAM device in Cielo.



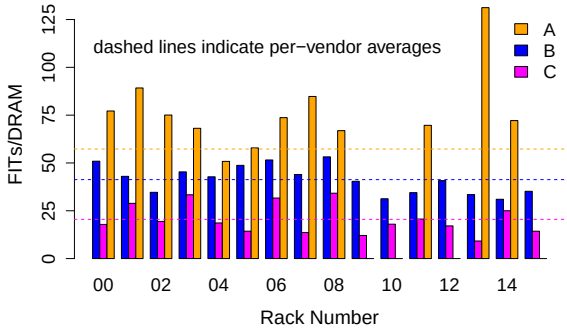
(a) Fault rate by vendor (b) Fault rate by fault type

Figure 6: "Zooming in" on the first several column addresses. The spike of faults in column 0 is due to a vendor-specific spike in transient faults.



(a) Fault rate per rack

(b) Operational hours per rack



(c) Fault rate per rack by vendor

Figure 7: DRAM fault rate positional effects by rack

rates) must correct for the mix of devices and vendors in the data set; otherwise, the study may lead to ambiguous or erroneous conclusions. Our study, for instance, found a correlation between fault rate and rack position in the data center, but this effect was explained by the mix of DRAM vendor in each rack. Studies that look at the impact of other external effects (e.g. the effect of temperature [12]) must also correct for vendor and device effects to properly identify trends.

8. SRAM FAULTS

In this section, we examine fault rates of SRAM. First, we examine the breakdown of transient and permanent faults in SRAM. We investigate a variance by physical location of faults in the LANL data center. We also compare the fault rates of Cielo and Jaguar to extract any effect caused by the 6,500-foot difference in altitude between the locations of the two supercomputers.

8.1 CPUs under Test

Both Cielo and Jaguar use AMD Opteron processors. Both processors are based on the 45-nm process technology node and share a common core microarchitecture. While core counts and cache sizes differ among processors, the SRAM cells within each cache (e.g. L3 cache) are similar across systems. Therefore, Jaguar and Cielo SRAM fault rates can be compared as long as results are adjusted for cache count and size.

All SRAM faults in our data set were corrected by on-board ECC and thus did not cause any failures in the systems. All rates in this section are presented in arbitrary units.

8.2 Transient and Permanent Faults

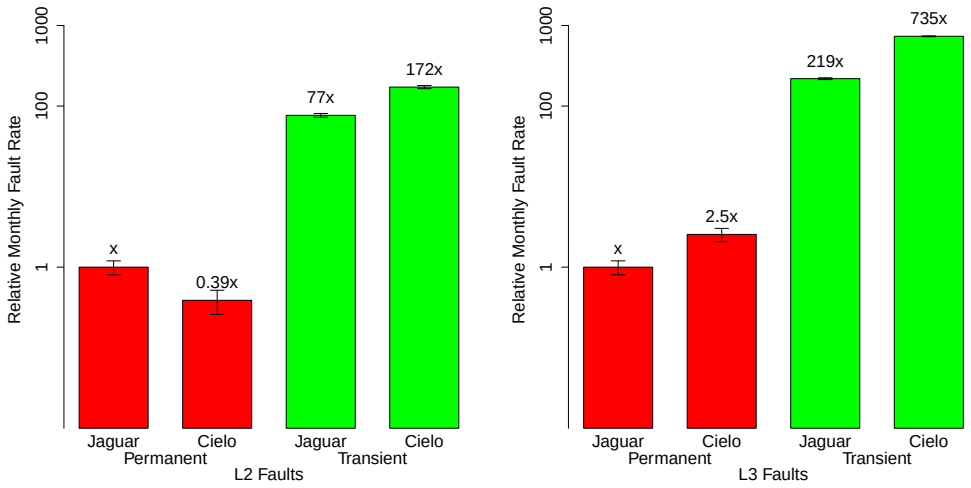
Much of the existing literature on SRAM faults assumes that transient faults are the dominant fault mode in SRAM devices. There has been significant work on beam testing to characterize particle-induced transient fault effects in SRAM devices (e.g., [11]). In addition, transient faults have caused very well-publicized events at customer sites for major vendors (e.g., [4]).

Therefore, it is well-established that particle-induced transient faults are an important component of SRAM faults. Further, we expect altitude to have an impact on SRAM faults due to the increased neutron flux at higher elevations. As far as we know, however, there is relatively little published data from production systems on altitude effects on SRAM. Further, there is little data comparing the rate of SRAM transient faults to the rate of SRAM permanent faults.

Figures 8(a) and 8(b) plot the rate of SRAM transient faults to the rate of SRAM permanent faults in the L2 and L3 caches in Cielo and Jaguar. Both figures are presented in arbitrary units. The figures plot the mean fault rate; the error bars represent the standard deviation of the monthly fault count. The figures confirm that more than 98% of SRAM faults are transient in both L2 and L3 caches in both Jaguar and Cielo. Although it is difficult to see due to the log scale, there is more variation (larger standard deviations) in the month-to-month rate of transient faults than in the month-to-month rate of permanent faults.

8.3 Altitude Effects

It is well known that the altitude at which a data center resides has consequences with regards to machine fault rates.



(a) L2 (b) L3

Figure 8: SRAM faults in Cielo and Jaguar (arbitrary units).

The two primary causes of increased fault rates at higher altitude are reduced cooling due to lower air pressure and increased cosmic ray-induced neutron strikes. While the first can be corrected by lower machine room temperatures and higher air flow, data centers typically do not attempt to compensate for cosmic ray neutrons directly.

Figure 8 shows the SRAM transient and permanent fault rates relative to permanent SRAM faults on Jaguar. Depicted as whiskers on the graph, 95% of the data falls within one standard deviation of the mean fault rate. Figure 8 shows that Cielo experiences a 2.3x increase in the SRAM transient fault rate relative to Jaguar in L2, and a 3.4x increase relative to Jaguar in L3. The average flux ratio between LANL and ORNL without accounting for solar modulation is 4.39 [1]. Therefore, we attribute the increase in SRAM fault rates to the increase in particle flux experienced in Los Alamos. The fact that Cielo's increase in fault rate relative to Jaguar is less than that predicted by altitude alone indicates that there may also be other sources of SRAM faults, such as Alpha particles [5].

8.4 Location in the Data Center

We examined the distribution of SRAM faults across datacenter location for any statistically interesting trends. Most datacenter locations (rack, row, cabinet, slot) showed uniform fault rates. However, the fault rates of SRAM across chassis show a statistically significant trend, shown in Figure 9. SRAM fault rates on both Cielo and Jaguar show an approximately 20% increase in FIT from the bottom to top chassis of a rack.

Both Cielo and Jaguar are deployed such that chassis 0 is at the bottom of a rack and closest to the machine room floor. Chassis 2 is therefore the highest in any rack. We believe there are two possible causes for the differences seen by chassis that are related to their physical location.

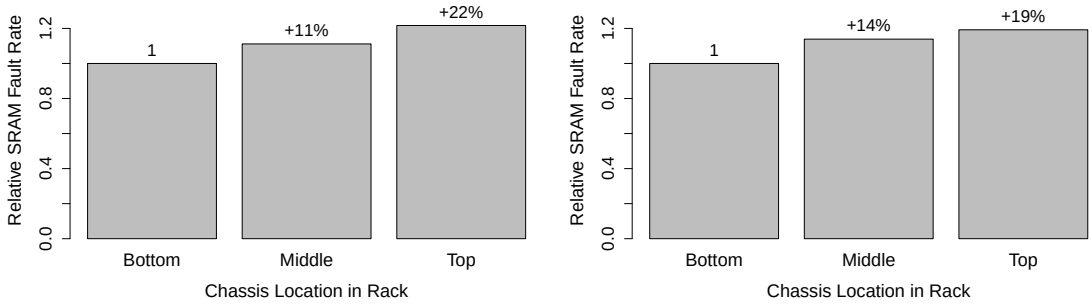
The Cray XE6 architecture of Cielo is such that cold air is extracted from the floor of the LANL machine room and passes up through all three chassis starting with 0 and ending with 2. As might be expected, hardware in chassis 2 is typically exposed to higher temperatures than hardware in chassis 0. Temperature has been shown to affect the rate of faults [13] and it is a potential cause of the increased rate we observe. We did not consult long-term temperature logs, but did measure several racks, observing an increase of 16° Celsius from bottom to top.

Another possible cause of the elevated fault rates in the higher chassis is neutrons from cosmic rays. Chassis 2 may be providing a small degree of shielding from cosmic ray neutrons to lower chassis. The hardware is aligned vertically such that incident neutrons must pass through the devices in an upper chassis before interacting with equivalent devices in a lower chassis. Neutron elastic scattering in the energy range of interest has an off-center component that can cause scattered neutrons to be deflected and potentially not impinge on hardware in a lower chassis [24]. Neutrons from cosmic rays have a mean path length of a few centimeters in the materials present inside each Cielo chassis. It is possible that enough neutrons are being deflected as they pass through the rack to account for some or all of the observed fault rate differences, similar to neutron scattering observed when testing multiple devices in a neutron beam [11]. This effect is also similar (although at a different scale) to neutron shielding observed in 3D stacked devices, in which the bottom layers experience a lower neutron flux due to shielding by the top layers in the stack [25].

Further experimentation, including heat and beam studies, are required to determine the cause of the measured differences in FIT rate throughout a rack of Cielo. It has been observed that temperature has an effect on the cross section of neutron-induced faults in silicon, so we may also be seeing a combined effect [8].

8.5 Conclusions

We draw several conclusions from this data. First, we find that SRAM faults in the field are dominated by transient faults, matching expectations based on prior literature. Second, we find that SRAM experiences 20% higher fault rates when placed at the top of rack relative to the bottom of rack. We postulate that this difference may be due to temperature or neutron shielding, but further investiga-



(a) Cielo (b) Jaguar

Figure 9: SRAM shows an increased transient fault rate in the upper chassis.

tion is needed to determine the root cause. We do not see any comparable trend in DRAM, which may indicate that DRAM faults and SRAM faults have different root causes. Finally, as expected, we see a significant altitude effect on SRAM fault rate, indicating that the dominant fault mode in SRAM is due to cosmic-ray induced neutrons.

9. SUMMARY

This paper presented a field study of DRAM and SRAM faults across two large high-performance computer systems. Our study resulted in several primary findings:

- In contrast to prior work, we found that the composition of DRAM faults shifts markedly during the first two years of lifetime, changing from primarily permanent faults to primarily transient faults.
- We found a significant inter-vendor effect on DRAM fault rates, with fault rates varying by up to 4x among vendors. A main conclusion that we draw from this result is that DRAM studies that do not adjust for vendor may lead to erroneous results.
- Again in contrast to prior work, we found no correlation between DRAM location and fault rates, except for one vendor-specific effect.
- We found that SRAM faults in the field are primarily transient, including expected altitude effects, and that SRAM seems to experience 20% higher fault rates when placed in top-of-rack nodes.

Overall, we believe that reliability will continue to be a significant challenge in the years ahead. Understanding the nature of faults experienced in practice can benefit all stakeholders, including processor and system architects, data center operators, and even application writers, in the quest to design more resilient high-performance computing systems.

10. ACKNOWLEDGMENTS

We thank Steve Johnson and Dave Londo at Cray for providing data collection and configuration information on the Jaguar system, and for Cielo information and data, Bob Ballance and John Noe from Sandia, and Kyle Lamb and Jeff Johnson from Los Alamos.

11. REFERENCES

[1] Flux calculator. http://seutest.com/cgi-bin/FluxCalculator.cgi.

[2] AMD64 architecture programmer's manual revision 3.17, 2011.

[3] A. Avizienis, J.-C. Laprie, B. Randell, and C. Landwehr. Basic concepts and taxonomy of dependable and secure computing. *Dependable and Secure Computing, IEEE Transactions on*, 1(1):11–33, 2004.

[4] R. Baumann. Soft errors in commercial semiconductor technology: Overview and scaling trends. In *IEEE Reliability Physics Tutorial Notes*, 2002.

[5] R. Baumann. Soft errors in advanced computer systems. *Design Test of Computers, IEEE*, 22(3):258–266, 2005.

[6] K. Bergman, S. Borkar, D. Campbell, W. Carlson, W. Dally, M. Denneau, P. Franzon, W. Harrod, J. Hiller, S. Karp, S. Keckler, D. Klein, R. Lucas, M. Richards, A. Scarpelli, S. Scott, A. Snavely, T. Sterling, R. S. Williams, and K. Yelick. Exascale computing study: Technology challenges in achieving exascale systems, Peter Kogge, editor & study lead, 2008.

[7] L. Borucki, G. Schindlbeck, and C. Slayman. Comparison of accelerated DRAM soft error rates measured at component and system level. In *Reliability Physics Symposium, 2008. IRPS 2008. IEEE International*, pages 482–487, 2008.

[8] A. Chugg, A. Burnell, P. Duncan, S. Parker, and J. Ward. The random telegraph signal behavior of intermittently stuck bits in sdrams. *Nuclear Science, IEEE Transactions on*, 56(6):3057–3064, 2009.

[9] C. Constantinescu. Impact of deep submicron technology on dependability of vlsi circuits. In *Dependable Systems and Networks, 2002. DSN 2002. Proceedings. International Conference on*, pages 205–209, 2002.

[10] C. Constantinescu. Trends and challenges in vlsi circuit reliability. *Micro, IEEE*, 23(4):14–19, 2003.

[11] A. Dixit, R. Heald, and A. Wood. Trends from ten years of soft error experimentation. In *Silicon Errors in Logic - System Effects (SELSE), 2009 IEEE Workshop on*, 2009.

[12] N. El-Sayed, I. A. Stefanovici, G. Amvrosiadis, A. A. Hwang, and B. Schroeder. Temperature management

in data centers: why some (might) like it hot. In *Proceedings of the 12th ACM SIGMETRICS/PERFORMANCE joint international conference on Measurement and Modeling of Computer Systems*, SIGMETRICS '12, pages 163–174, New York, NY, USA, 2012. ACM.

[13] M. Gadlage, J. Ahlbin, B. Narasimham, V. Ramachandran, C. Dinkins, B. Bhuva, R. Schrimpf, and R. Shuler. The effect of elevated temperature on digital single event transient pulse widths in a bulk cmos technology. In *Reliability Physics Symposium, 2009 IEEE International*, pages 170–173, 2009.

[14] A. A. Hwang, I. A. Stefanovici, and B. Schroeder. Cosmic rays don't strike twice: understanding the nature of dram errors and the implications for system design. In *Proceedings of the 17th international conference on Architectural Support for Programming Languages and Operating Systems*, ASPLOS XVII, pages 111–122, New York, NY, USA, 2012. ACM.

[15] X. Li, M. C. Huang, K. Shen, and L. Chu. A realistic evaluation of memory hardware errors and software system susceptibility. In *Proceedings of the 2010 USENIX conference on USENIX annual technical conference*, USENIXATC'10, pages 6–20, Berkeley, Calif., USA, 2010. USENIX Association.

[16] X. Li, K. Shen, M. C. Huang, and L. Chu. A memory soft error measurement on production systems. In *2007 USENIX Annual Technical Conference on Proceedings of the USENIX Annual Technical Conference*, ATC'07, pages 21:1–21:6, Berkeley, Calif., USA, 2007. USENIX Association.

[17] T. May and M. H. Woods. Alpha-particle-induced soft errors in dynamic memories. *Electron Devices, IEEE Transactions on*, 26(1):2–9, 1979.

[18] A. Messer, P. Bernadat, G. Fu, D. Chen, Z. Dimitrijevic, D. Lie, D. Mannaru, A. Riska, and D. Milojicic. Susceptibility of commodity systems and software to memory soft errors. *Computers, IEEE Transactions on*, 53(12):1557–1568, 2004.

[19] H. Quinn, P. Graham, and T. Fairbanks. Sees induced by high-energy protons and neutrons in sdram. In *Radiation Effects Data Workshop (REDW), 2011 IEEE*, pages 1–5, 2011.

[20] B. Schroeder and G. Gibson. A large-scale study of failures in high-performance computing systems. In *Dependable Systems and Networks, 2006. DSN 2006. International Conference on*, pages 249–258, 2006.

[21] B. Schroeder, E. Pinheiro, and W.-D. Weber. DRAM errors in the wild: a large-scale field study. *Commun. ACM*, 54(2):100–107, Feb. 2011.

[22] T. Siddiqua, A. Papathanasiou, A. Biswas, and S. Gurumurthi. Analysis of memory errors from large-scale field data collection. In *Silicon Errors in Logic - System Effects (SELSE), 2013 IEEE Workshop on*, 2013.

[23] V. Sridharan and D. Liberty. A study of DRAM failures in the field. In *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, SC '12, pages 76:1–76:11, Los Alamitos, Calif., USA, 2012. IEEE Computer Society Press.

[24] M. Walt and H. H. Barschall. Angular distributions of elastically scattered 1-mev neutrons. *Phys. Rev.*, 90:714–715, May 1953.

[25] W. Zhang and T. Li. Microarchitecture soft error vulnerability characterization and mitigation under 3d integration technology. In *Proceedings of the 41st annual IEEE/ACM International Symposium on Microarchitecture*, MICRO 41, pages 435–446, Washington, D.C., USA, 2008. IEEE Computer Society.

[26] J. Ziegler and W. Lanford. The effect of sea level cosmic rays on electronic devices. *Journal of Applied Physics*, 52(6):4305–4312, 1981.