SAND2013-3402C

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
ndebard@lanl.gov

Sean Blanchard
Ultrascale Systems Research Center
Los Alamos National Laboratory
seanb@lanl.gov

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

David Robinson
Data Analysis and Informatics
Sandia National Laboratories[1]
Albuquerque, New Mexico
drobin@sandia.gov

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 choice, finding that fault rates vary by more than 4x between vendors. We examine the physical location of faults within a DRAM device and within a data center, and contrary to prior studies, find no correlations with either. We study the impact of altitude and rack placement on SRAM faults. Finally, we use our data to develop statistical models to simulate real-world fault behavior.

1. INTRODUCTION

Recent studies have confirmed that faults are common in memory systems of high-performance computing systems [19]. 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 [3]. 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 in order 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 impact 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, a 8500 node supercomputer located at Los Alamos National Laboratory (LANL). A secondary data set comes from Jaguar, a 18,688 node supercomputer that was located at Oak Ridge National Laboratory. On Cielo, our measurement interval is a 15 month period from mid 2011 through early 2013, comprising 14.6 billion DRAM hours of data. On Jaguar, our measurement interval is an 11 month period from late 2009 through late 2010, comprising 17.1 billion DRAM 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 [17], we find that the composition of DRAM faults changes substantially over 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 between 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 within a DRAM device. With the exception of one device-specific fault mode, we find an approximately uniform distribution

[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 SAND2012-2144C.

of faults across DRAM row, column and bank addresses, in contrast to previous studies.

- We study of the impact of location within 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 datacenter location and altitude on SRAM faults. We find that SRAM devices experience 20% higher transient fault rates when placed in "top of rack" nodes. We also find that, as expected, altitude has a significant effect on the fault rate of SRAMs.
- Finally, we use our dataset to fit models to both DRAM and SRAM faults. These models can be used by architects and system designers to estimate the expected fault behavior of the memory subsystem of a high-performance computer system.

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 configuration of Cielo as well as 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 within a DRAM device, and the impact of placement within a datacenter. Section 8 discusses location effects on SRAM fault rates, including placement within a data center and altitude. Section 9 uses our data to derive a set of fault models that can be used to mimic the expected behavior of memory subsystems. Finally, Section 10 discusses implications of our findings and presents our conclusions.

2. TERMINOLOGY

In this paper, we distinguish between a fault and an error as follows [2]:

- A fault refers to an underlying cause of a failure, such as a particle-induced bit flip or a stuck-at bit.
- An error is a symptom of a fault. A fault will return an error when it is read if the node provides higher-level detection such as parity or ECC. Note that one fault can cause many errors if the failed locations are accessed multiple times.

All data and analysis presented in this paper refers to faults, not errors.

Hardware faults can further be classified as [6]:

- *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 [11]. Particle-induced upsets ("soft errors"), which have been extensively studied in the literature [1] [11], 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). Hard faults are permanent and can be repaired only by replacing the faulty device [12].
- *Intermittent faults*, which cause a memory location to sometimes return incorrect values. Unlike hard faults, intermittent faults occur only under specific conditions (e.g., elevated temperature) [10]. 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 few years, several studies have been published studying DRAM failures in the field. In 2006, Schroeder and Gibson published a study on failure data from high-performance computer systems at Los Alamos National Labs [16]. In 2007, Li et al. published a study of memory errors on three different data sets, including a server farm of an Internet service [12]. In 2009, Schroeder et al. published a large-scale field study using Google's server fleet [17]. In 2010, Li et al. published an expanded study of memory errors on an Internet server farm and other sources [11]. In 2012, Hwang et al. published an expanded study on Google's server fleet as well as two IBM Blue Gene clusters [10], Sridharan and Liberty presented a study of DRAM failures in a high-performance computing system [19], and El-Sayed et al. published a study on temperature effects of DRAM in data center environments [8]. In 2013, Siddiqua et al. presented a study of DRAM failures from client and server systems [18].

Our study contains analyses not performed in many of these previous studies, including: the effects of DRAM device and vendor 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 [17][10]. This makes it difficult to compare 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 only result 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 has also been significant accelerated testing work on DRAM devices dating back several decades [13][14][4][15]. Of particular interest are the studies by Borucki and Quinn which 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.

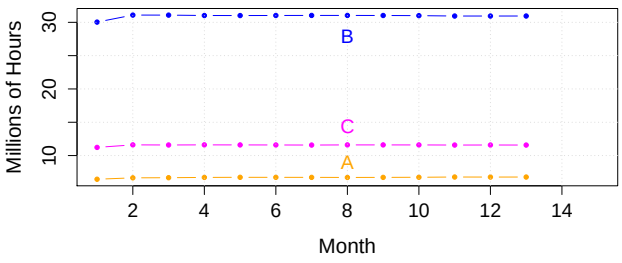


Figure 1: DRAM use per month was roughly constant for each manufacturer. Aggregate totals are given in Figure 4(a).

4. SYSTEM 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 8500 compute nodes. Each *node* of Cielo contains two AMD Opteron^{TM}8-core processors. Each Cielo compute node has 8 4GB DDR-3 DIMMs for a total of 32GB of DRAM per node.

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

During our measurement interval, Jaguar (which was taken offline in 2012) contained 18,688 nodes. Each node contained two AMD Opteron^{TM}6-core processors. Each Jaguar node has 8 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 of these slots are contained within a *chassis* of which there are three mounted bottom-to-top (numerically) in a *rack*. Sixteen racks are aligned into a *row* and there are 6 rows.

At 7320 feet in altitude, the Cielo system at Los Alamos National Laboratory is subject to a higher flux of cosmic ray induced neutrons than Jaguar at Oak Ridge National Laboratory 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 [**?**].

4.1 DRAM Configuration

On 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, sixteen 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 that share 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.

On 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.

5. DATA AND METHODOLOGY

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

Correctable DRAM 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) [1]. 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 are then further decoded 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.

Based on the methodology used in previous field studies, we associated each error in the logs with a specific type and mode of fault [19]. Since both systems include a hardware DRAM scrubber, we are able to identify permanent faults as those faults that survive a scrub operation.

Hardware inventory logs are separate logs and provide snapshots of the hardware that was present on Cielo at different points in its lifetime. We analyzed 217 hardware inventory logs over a span of approximately two years from early 2011 to 2013. Each log file consists of over 1.3million 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 together provided us the ability to map DRAM error messages to specific hardware that was present in the machine at that point in time. In total we analyzed over 14 billion DRAM hours. All of the DIMM manufacturer data presented in this paper has been anonymized to protect interested parties.

6. DRAM FAULT RATES

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

% Faulty DRAMs	0.05%
% Faulty DIMMs	1.86%
Fault Rate (FIT/Mbit)	0.051
Fault Rate (FIT/device)	46.9

Table 1: DRAM Fault Rates.

System	0	1	2
cielo	87.65%	10.96%	1.20%
jaguar	87.16%	9.76%	1.31%

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

permanent faults, and examine 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 1.78% of DIMMs, or 0.1% of DRAM devices, experienced a fault during the experiment. The calculated fault rate of 0.052 FIT/Mbit translates to one fault approximately every 21 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 [19] [17]. This is important because this 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 with 0, 1, 2, and 3 DRAM faults. 10% of all nodes in the system had a DRAM fault at some point during our observation period. No node ever had more than 3 faulty DRAMs. The figure shows that a node with an existing DRAM fault or faults has the same likelihood of developing another DRAM fault as a node with no DRAM faults. This strongly implies that DRAM faults are independent across devices.

6.2 Fault Rates Over Time

Figure 2 shows the total number of DRAM faults per month (defined as a 30-day period) in Cielo. We omit the first month of the data set since this would result in "overcounting" permanent faults that developed between the beginning of the system's lifetime and the start of our observation period. The figure shows that Cielo experiences a declining rate of DRAM faults over our observation period, matching results found by other studies that take place towards the beginning of a system's lifetime [19]. 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. The crossover point between permanent and transient faults occurs in the ninth month of the data set, which represents the fourteenth month of production for the Cielo system.

Figure 3 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. On Jaguar, we see the crossover point between permanent faults and transient faults in the sixth month of the data set, which represents

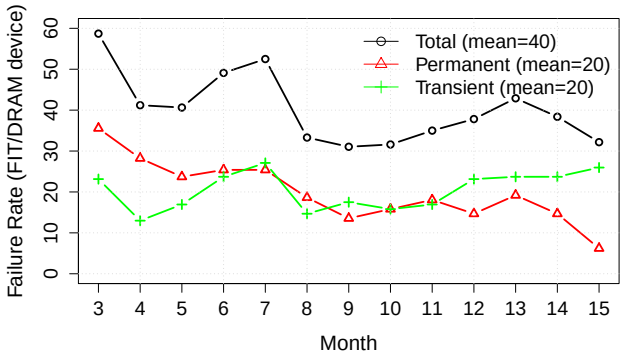


Figure 2: Cielo DDR3 DRAM device fault rates per month (30-day period); 14.6 billion DRAM hours total.

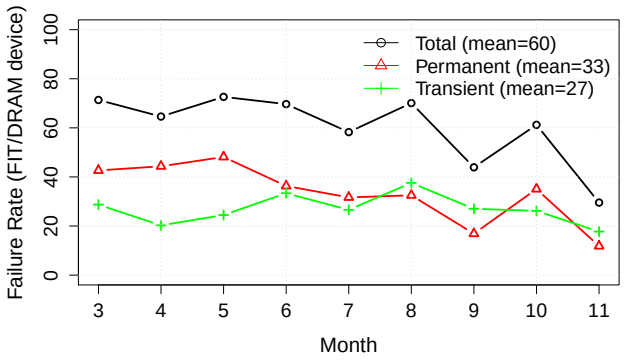


Figure 3: Jaguar DDR2 DRAM device fault rates per month (30-day period); 17.1 billion DRAM hours total.

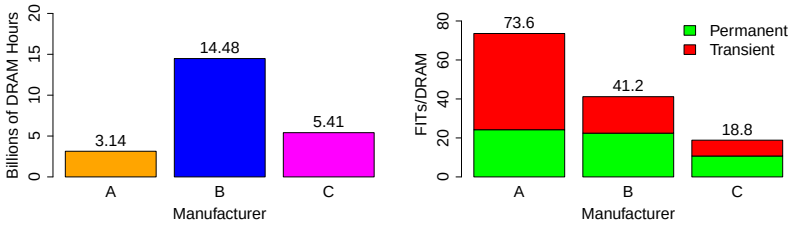
the sixteenth month of lifetime for the majority of the DIMMs in the system.

6.3 Fault Rates by DRAM Vendor

Figure 4(a) shows the aggregate number of DIMM-hours per DRAM vendor on Cielo in 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.

Figure 4(b) shows the fault rate experienced by each vendor over this time period, divided into transient and permanent faults. The figure shows a substantial difference between 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 between vendors, from 22.8 FIT to 10.1 FIT, while the transient fault rate varies by over 6x between vendors, from 46.5 FIT to 7.6 FIT. 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 fault rates than transient fault rates.

Previous publications have pointed to permanent faults as the primary source of faults in modern DRAM [17][19]. However, our data indicates that this conclusion depends heavily on the mix of DRAM vendor in the system(s) under test. Another interesting result in the figure is that transient and permanent fault rates vary together, so the vendor with the



(a) Operational hours

(b) Permanent and transient fault rates

Figure 4: Operational hours and FIT rate by vendor.

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, overall DRAM fault rates appear to be similar across DDR generations, and DRAM faults appear to be independent across devices. Second, we see a decrease in the permanent fault rate over time (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* over the first year and a half of system operation. This is in marked contrast to previous studies looking at *error* rates, which found no significant changes until past the 18 month timeframe [17]. 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 significantly by DRAM vendor and device, and that the choice of DRAM vendor and device will play a major role in the overall fault rate and types of fault experienced by the system.

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 within a device and multiple ranks within a lane [19][10]. In this section, we examine the prevalence of these different fault modes in our dataset, and examine their likelihood of occurrence by location within the DRAM and by vendor. We also examine correlations between a node's location within 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 rate of single-bit faults in Cielo's DDR-3 memory than found in Jaguar's DDR-2 memory by Sridharan and Liberty [19] (65.3% versus 49.7%), but this may be due to external factors such as altitude or vendor rather than to inherent differences in DDR-3 memory.

Fault Mode	Total Faults	Transient	Permanent
Single-bit	67.8%	34.9%	39.8%
Single-word	0.2%	0.2%	0%
Single-column	8.8%	3.8%	4.9%
Single-row	11.6%	5.7%	6.1%
Single-bank	9.5%	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.

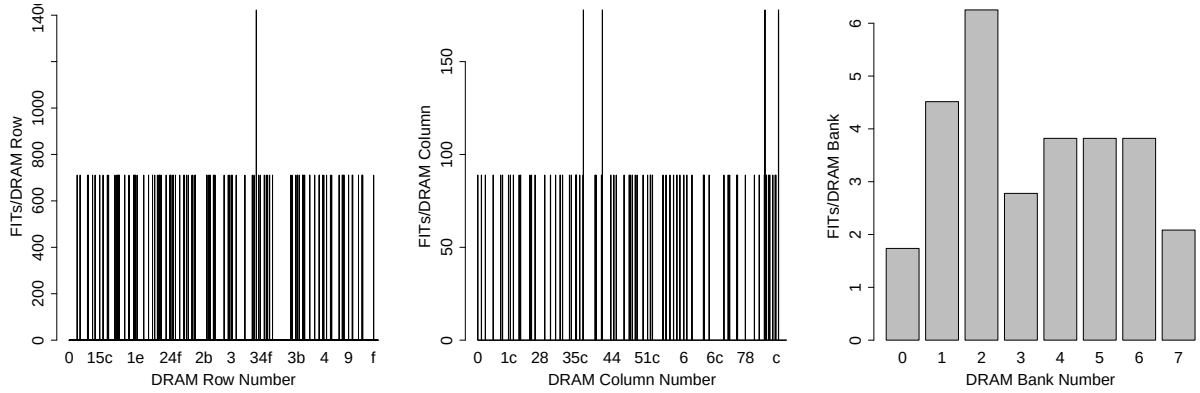
Fault Mode	Vendor A	Vendor B	Vendor C
Single-bit	64.6%	69.5%	58.4%
Single-word	0%	0.3%	0%
Single-column	8.7%	8.8%	11.9%
Single-row	12.2%	10.6%	14.9%
Single-bank	13.5%	7.8%	9.9%
Multiple-bank	1.3%	0.7%	2.0%
Multiple-rank	1.3%	3.0%	3.0%

Table 4: DRAM fault modes by vendor.

Table 4 shows the same data but broken down by vendor. The table shows taht 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 has shown that these fault modes are most likely to lead to an uncorrected error [19]. Therefore, vendor C's low overall fault rate may not translate into a low rate of uncorrected errors. The table also shows that only vendor B experienced single-word faults. Because 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.

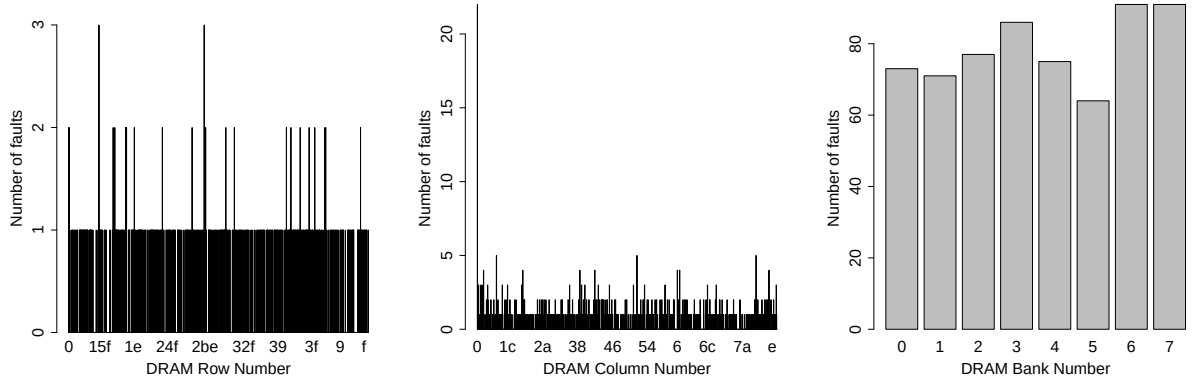
7.2 Fault Distribution Within a Device

Previous field studies have shown a correlation between error rate and location within a DRAM device [10]. 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 far as we are aware, no previous studies have examined the distribution of faults within a DRAM device.



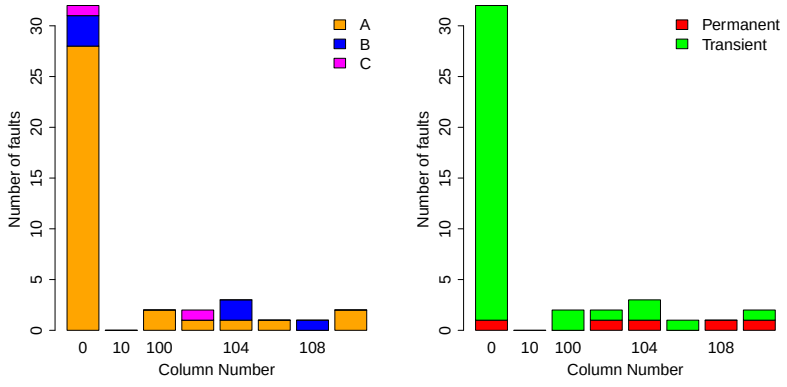
(a) Single-Row Faults by Row (b) Single-Column Faults by Column (c) Single-Bank Faults by Bank

Figure 5: Distribution of single-row, single-column, and single-bank faults within a DRAM device.



(a) Distribution over Rows (b) Distribution over Columns (c) Distribution over Banks

Figure 6: Distribution of single bit faults.



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

Figure 7: "Zooming in" on the first several column addresses. The spike of single-bit faults in column 0 is due to a vendor-specific spike in transient faults.

To perform this study, we plot the locations associated with each fault in our dataset. For single-bit faults, we plot the row and column 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 5 plots the location of single-row, single-column, and single-bank faults, respectively. The figure shows no significant relationship between fault rate and either row, column, or bank address; the faults appear to be randomly distributed throughout the DRAM device. Figure 6 shows the row, column, and bank distribution of all single-bit faults. This figure, by contrast, shows a significant spike in the fault rate in column 0. Figure 7 "zooms in" on the first 10 column addresses, and further break 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, with the remainder of the data appears to be randomly distributed across columns. As a DRAM column spans all DRAM addresses within 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.

7.3 Location Within the Data Center

The physical conditions within 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 within the LANL machine room were uninteresting or statistically inconclusive. However, there is one notable exception to the lack of variance, shown in Figure 8(a). Lower numbered racks show higher DRAM FIT rates than higher numbered racks. Without examining the data by vendor, this trend might be attributed to temperature or other environmental differences across racks. When examining the by-vendor data in Figure 8(a), however, it is clear that the variation by rack is primarily a result of a difference in the FIT contribution of each vendor within each rack. This trend is explained by the difference in operational hours for each vendor in each rack (Figure 8(b)). As seen in figure 6.3, manufacturer A's DIMMs appear to have a higher susceptibility to transient faults on Cielo. Therefore, racks with more operational hours from vendor A (the lower numbered racks) have higher overall FIT rates.

7.4 Conclusions

We draw three main conclusions from this data. First, we conclude that all DRAM fault modes identified in DDR-2 generation DRAMs remain a concern in DDR-3 generation devices. This suggests that many of 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 clear trend in the distribution of faults within a DRAM device, implying that DRAM faults are equally likely to occur in any region of a DRAM device. Third, we find no conclusive effects due to position within a datacenter. While we find a significant correlation between rack position and fault rate, this is mostly explained by the mix of DRAM vendor within each rack. We conclude that analyses of external factors on DRAM reliability (e.g. the effects of location, temperature, or altitude) must correct for the mix of devices in the data set or else they may lead to erroneous conclusions.

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 within the LANL data center. We also compare the fault rates of Cielo and Jaguar in order to extract any effect caused by the 6500 foot difference in altitude between the locations of the two supercomputers.

8.1 CPUs Under Test

Both Cielo and Jaguar use AMD Opteron^{TM}processors. Both processors are based on the 45nm process technology node and share a common core microarchitecture. While core counts and cache sizes differ between processors, the SRAM cells within each cache (e.g. L3 cache) are similar across systems. Therefore, SRAM fault rates on Jaguar and Cielo can be compared as long as results are adjusted for cache size. In this section, we perform comparisons on a per-bit basis. All results are presented in arbitrary units.

8.2 Transient and Permanent Faults

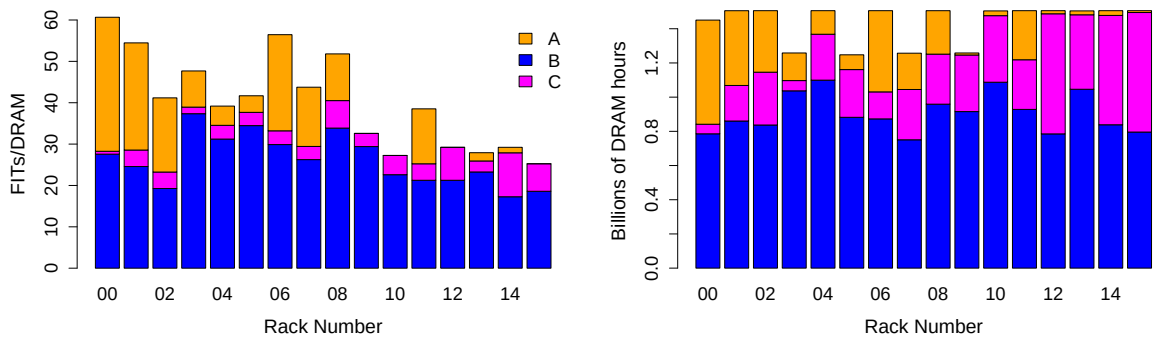
Much of the existing literature on SRAM faults assumes that transient faults are the dominant fault mode in SRAM devices. Figures 9(a) and 9(b) confirm this in the L2 and L3 caches on Cielo and Jaguar. Both figures are presented in arbitrary units. The figures show that over 98% of SRAM faults are transient in both L2 and L3 caches in both Jaguar and Cielo. Interestingly, both caches show a slight increase in SRAM transient faults over time. It is unclear what causes this increase. We address this finding in Section 9.

Both figures also show a low and declining rate of permanent faults over time in both systems. This is similar to the declining rate of permanent faults observed in DRAM and is indicative of a small number of early-life failures in the CPUs.

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. 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 for with lower machine room temperatures and higher air flow, data centers typically do not attempt to compensate for cosmic ray neutrons directly.

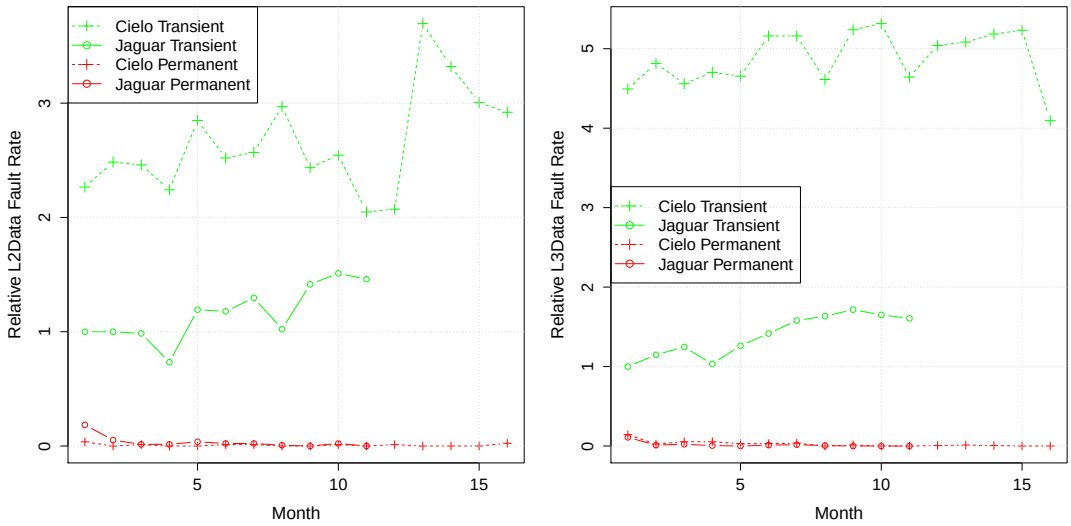
Figure 9 shows that Cielo experiences a 2.3-3.5x increase in the SRAM fault rate relative to Jaguar, depending on the structure in question. At 7320 feet in altitude, Los Alamos National Laboratory is subject to a higher flux of cosmic ray induced neutrons than Oak Ridge National Laboratory at 850 feet. The average flux ratio between the two locations



(a) DRAM fault rate per rack. Colorings for manufacturer A, B, and C indicate the *percentage* of the rack FIT rate.

(b) Operational hours per vendor per rack

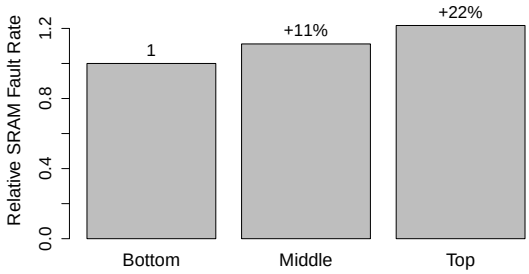
Figure 8: Fault rate positional effects by rack.



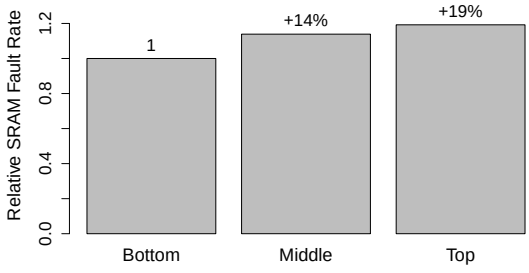
(a) L2

(b) L3

Figure 9: SRAM faults in Cielo and Jaguar (arbitrary units).



(a) Cielo



(b) Jaguar

Figure 10: SRAM shows an increased fault rate in the upper chassis.

due to altitude, longitude and latitude without accounting for solar modulation is 4.39 [**?**]. We attribute the increase in SRAM fault rates to the increase in particle flux experienced in Los Alamos.

8.4 Location Within 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 valid trend. SRAM fault rates for SRAM 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 in 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, 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 play an effect in the rate of faults [9] and it is a potential cause of the increased rate we observe.

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 are aligned vertically

System	Jaguar	Cielo
DRAM	**FIXME**	**FIXME**
SRAM	**FIXME**	**FIXME**

Table 5: Weibull shape parameters for DRAM and SRAM transient faults.

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 a small off center component that can cause scattered neutrons to be deflected and potentially not impinge upon hardware in a lower chassis [**?**]. Neutrons from cosmic rays have a mean path length of a few centimeters in the materials present inside each Cielo chassis. It is possible 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 of multiple devices in a neutron beam [7]. This effect is also similar (although at a different scale) to neutron shielding observed in 3D stacked devices, where the bottom layers experience a lower neutron flux due to shielding by the top layers in the stack [20].

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. [5]

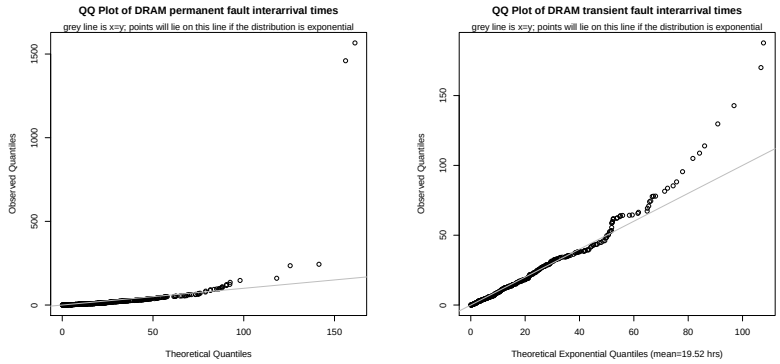
8.5 Conclusions

We draw three main 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 investigation is needed to determine the root cause of this difference. We do not see any comparable trend in DRAM, which may indicate that DRAM faults and SRAM faults have different root causes. Third, 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. MODELING MEMORY FAULTS

In this section, we provide information on how to model the DRAM and SRAM transient faults experienced by the systems during our observation interval. These models can be used by system operators to determine operational procedures; by system architects to estimate future system reliability; and by processor architects when determining the effectiveness of new memory reliability techniques.

Initially, our expectation was that the inter-arrival times of transient faults would be exponentially distributed, which would correspond to a uniform random process such as particle strikes. However, as show by Figure 11, the actual inter-arrival times follow a Weibull distribution. With **FIXME%** confidence, we can say that neither SRAM nor DRAM faults are exponentially distributed.



(a) DRAM Permanent Faults (b) DRAM Transient Faults

Figure 11: Distribution of fault inter-arrival times.

Table 5 provides the Weibull shape parameters for both SRAM and DRAM transients. The shape parameters are consistent with an increasing hazard rate; that is, the transient fault rate is increasing over time. It is uncertain what causes this increase. For example, the data are consistent with a rising fault rate due to aging (e.g. the "back half" of the bathtub curve), where the aging-related faults initially manifest as transients but would eventually become permanent (intermittent or hard) faults. Confirming or denying this hypothesis will require continued observation of Cielo. (Jaguar was decommissioned in 2011 and thus no further data is available.)

10. SUMMARY

This paper has 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 over the first two years of lifetime, changing from primarily permanent faults to primarily transient faults.
- We found a significant inter-vendor / inter-device effect on DRAM fault rates, with fault rates between vendors varying by up to 4x. A main conclusion that we draw from this result is that DRAM studies that do not adjust for device type 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.
- We demonstrated that transient fault inter-arrival times on our systems are approximated via a Weibull, not exponential, distribution, consistent with an increasing hazard rate over time.

Overall, we believe that reliability will continue to be a significant challenge in the years ahead. Understanding the nature of faults that are experienced in practice will be a benefit to 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.

11. ACKNOWLEDGEMENTS

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-TBD.

12. REFERENCES

[1] Amd64 architecture programmer's manual revision 3.17, 2011.

[2] 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.

[3] 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, K. Yelick, K. Bergman, S. Borkar, D. Campbell, W. Carlson, W. Dally, M. Denneau, P. Franzon, W. Harrod, J. Hiller, S. Keckler, D. Klein, P. Kogge, R. S. Williams, and K. Yelick. Exascale computing study: Technology challenges in achieving exascale systems peter kogge, editor & study lead, 2008.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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 seventeenth international conference on Architectural Support for Programming Languages and Operating Systems*, ASPLOS XVII, pages 111–122, New York, NY, USA, 2012. ACM.

[11] 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 6–6, Berkeley, CA, USA, 2010. USENIX Association.

[12] 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, CA, USA, 2007. USENIX Association.

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

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] T. Siddiqua, A. Papathanasiou, Athanasios amd 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.

[19] 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, CA, USA, 2012. IEEE Computer Society Press.

[20] 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, DC, USA, 2008. IEEE Computer Society.