DRAM Errors in the Field: A Statistical Approach

Darko Zivanovic
darko.zivanovic@bsc.es
Barcelona Supercomputing Center
Barcelona, Spain

Pouya Esmaili Dokht
pouya.esmaili@bsc.es
Barcelona Supercomputing Center

Sergi Moré
sergi.more@bsc.es
Barcelona Supercomputing Center

Javier Bartolome
javier.bartolome@bsc.es
Barcelona Supercomputing Center

Paul M. Carpenter
paul.carpenter@bsc.es
Barcelona Supercomputing Center

Petar Radojković
petar.radojkovic@bsc.es
Barcelona Supercomputing Center

Eduard Ayguadé
eduard.ayguade@bsc.es
Barcelona Supercomputing Center,
Universitat Politècnica de Catalunya

ABSTRACT

This paper summarizes our two-year study of corrected and uncorrected errors on the MareNostrum 3 supercomputer, covering 2000 billion MB-hours of DRAM in the field. The study analyzes 4.5 million corrected and 71 uncorrected DRAM errors and it compares the reliability of DIMMs from all three major memory manufacturers, built in three different technologies.

Our work has two sets of contributions. First, we illustrate the complexity of in-field DRAM error analysis and demonstrate the limitations of various widely-used methods and metrics. For example, we show that average error rates, *errors per MB-hour* and *mean time between failures* can provide volatile and unreliable results even after long periods of error logging, leading to incorrect conclusions about DRAM reliability. Second, we present formal statistical methods that overcome many of the limitations of the current approaches. The methods that we present are simple to understand and implement, reliable and widely accepted in the statistical community.

Overall, our study alerts the community about the need to, firstly, question the current practice in quantifying DRAM reliability and, secondly, to select a proper analysis approach for future studies. Our strong recommendations are to focus on metrics with a practical value that could be easily related to system reliability, and to select methods that provide stable results, ideally supported with statistical significance.

CCS CONCEPTS

• **Computer systems organization → Reliability**; • **Mathematics of computing → Probability and statistics**.

KEYWORDS

Memory, Reliability, Large-scale systems, Statistical analysis

ACM Reference Format:
Darko Zivanovic, Pouya Esmaili Dokht, Sergi Moré, Javier Bartolome, Paul M. Carpenter, Petar Radojković, and Eduard Ayguadé. 2019. DRAM Errors in the Field: A Statistical Approach. In *MEMSYS 2019: The International Symposium on Memory Systems, Sept. 30–Oct. 3, 2019, Washington, DC, USA.* ACM, New York, NY, USA, 16 pages. https://doi.org/https://doi.org/10.1145/3357526.3357558

1 INTRODUCTION

In large-scale compute clusters, main memory is one of the principal causes of hardware failures [8]. These failures are especially costly in high-performance computing (HPC) systems, where a single tightly-coupled job may execute for days on thousands of nodes. If one of these nodes fails, the whole job is terminated. It is therefore important to understand memory system reliability, as it is an important limit on the ability to scale to larger systems.

This paper summarizes our study of corrected and uncorrected errors on the MareNostrum 3 supercomputer [2], covering 2000 billion MB-hours of DRAM in the field. MareNostrum is one of six Tier-0 HPC systems in Europe; at the time of the study, it comprised 3056 servers, with more than 25,000 memory DIMMs from all three major memory manufacturers, built in three different DRAM technologies. The study covers a period of more than two years, from October 2014 to November 2016, during which we detected 4.5 million corrected and 71 uncorrected DRAM errors.

The main objective of our study is to help the community to define standards for any future quantitative analysis of DRAM errors in the field. Our work has two sets of contributions. First, we illustrate the complexity of in-field DRAM error analysis and demonstrate the limitations of various widely-used methods. Understanding these limitations is important because, as we show, widely-accepted approaches for DRAM analysis provide volatile, unreliable and statistically insignificant results that may lead to incorrect conclusions about DRAM reliability. Second, we present formal methods widely accepted in the statistical community that overcome many of the limitations of these currently-used approaches.

This is the first study that clearly distinguishes between two different approaches for DRAM error analysis. The first approach is to partition the DIMMs into various **categories**, e.g. based on whether they did or did not experience an error. The second approach is to

MEMSYS 2019, Sept. 30–Oct. 3, 2019, Washington, DC, USA

analyze the **error rates**, i.e. to present the total number of errors relative to other statistics, typically the number of MB-hours or the duration of the observation period. Although both methods are valid, our results clearly show that they are not interchangeable and can lead to completely different conclusions. This finding is important because various previous studies interleave categorical and error rate analysis and the conclusions based on them (some examples are in Section 6).

As a part of the **categorical analysis**, we explain and use independence tests to confirm or reject, in terms of statistical significance, any differences observed among various categories. We use these tests to analyse the percentages of DIMMs that experience uncorrected or corrected errors, for the different manufacturers and DRAM technologies. These tests allow us to ascertain whether the observed differences are likely to be due to real differences or are explainable merely by chance. To the best of our knowledge, this is the first study of DRAM errors that uses statistical tests to confirm or reject the significance of its results.

Regarding the **error rates**, we show that the average errors per MB-hour and average mean time between failures (MTBF) were highly volatile over the course of the study, with the final values depending critically on the moment at which the study is terminated. It is intuitive to conclude that we have little confidence in how the results would have looked if we were able to continue the study, e.g. for another year. We perform a careful study of the causes of this volatility, and conclude that the primary reason differs between uncorrected and corrected errors. For uncorrected errors, the volatility is explained by the small number of observations; 71 uncorrected errors over the course of the study. For corrected errors, the volatility is explained by error burstiness in time. Moreover, we show that using the corrected errors and fault rates as an indicator of DRAM reliability may be misleading because they have completely different trends from the uncorrected errors, which are the only errors that lead to system failure.

Our work opens various doubts about the stability and usefulness of the DRAM error rates analysis. Clarification of these doubts is important because error and fault rates, such as errors per MB-hour and MTBF are the current standard for quantifying memory system reliability. Before using them in future studies, the community should understand whether indeed they are appropriate reliability metrics. Overall, we believe that our study will help the community to define formal and reliable methods for analysis of the DRAM errors in the field. Our strong recommendations are to focus on measurements with a practical value, and select proper analysis methods that provide stable results, ideally supported with statistical significance.

The rest of the paper is organized as follows: Section 2 provides the necessary background on DRAM failures, errors and faults and statistical significance. Section 3 describes MareNostrum 3, the source of our data, and it outlines the experimental methodology. Sections 4 and 5 analyze in detail the corrected and uncorrected errors and faults, using the categorical and error rate approaches respectively. Section 6 describes the related work. Finally, Section 7 concludes the paper.

2 BACKGROUND

In the last decade several studies have analyzed field DRAM errors. These studies have quantified the variations in error rates among DRAM manufacturers and technologies and analyzed the nature of DRAM errors, including their temporal and spatial distributions. It is not easy, however, to compare the conclusions of different studies or to combine their findings into a clear overall understanding of memory system reliability, for three main reasons. First, the studies use non-unified terminology, especially when classifying the error types. Second, the studies interchangeably use different quantitative approaches, categorical and error rate analysis, which can often lead to different conclusions. Finally, the studies give quantitative results without reporting whether or to what degree the reported results are statistically significant.

2.1 Taxonomy: Are corrected DRAM errors failures?

In MareNostrum 3, and in the server domain in general, main memory is protected with error correcting codes (ECC). In modern HPC systems, sophisticated ECCs are able to correct multiple corrupted bits in a data word, and even handle cases where an entire DRAM chip is corrupted [3]. Data correction is performed in parallel with data read, so corrected errors effectively have no impact on system performance and reliability. But, if the ECC cannot correct a given DRAM error, the job typically has to be terminated and the server is shut down. The server is not operational until the DIMM is replaced and the node has been tested. The overall impact is lower reliability, lower system throughput and worse system availability.

Due to the requirement for high system reliability, original equipment manufacturers (OEMs) thoroughly test DIMMs from various manufacturers to certify that they can be used in production. It is usual practice, however, to quantify DIMM reliability using corrected errors, rather than the more important uncorrected errors. Likewise, most of the DRAM error field studies focus their analysis on corrected errors, although *only* uncorrected errors have an impact on system reliability.

Most of the previous studies use the definitions of errors, faults and failures from Avizienis et al. [1]:

- **Failure** is an event that occurs when the delivered service deviates from correct service. For example, it is expected that a data read from memory delivers correct data stored on a given address. Deviation from this service is a failure.
- **Error** is the deviation of the system state (seen externally) from its correct service state. For example, the fact that a DIMM delivers to the memory controller data that do not match the ECC is the DRAM error.
- **Fault** is the adjudged or hypothesized root cause of the error. The cause of a DRAM error could be a particle impact, or a defect in the memory cell or circuit.

The problem is that the definitions of failures and errors are tightly coupled with the scope of the target system and its boundaries. For example, the memory system comprises the memory controller, DRAM devices and all the circuitry between them [12]. DRAM errors that are corrected by ECC in the memory controller **are not** errors or failures of the memory system, because the memory system still delivers correct data. Such corrected DRAM errors therefore have no impact on the service provided by the server and the overall HPC system. DRAM errors that cannot be ECC-corrected, however, propagate over the boundaries of the memory system, so they **are** also memory system errors and failures. In current HPC systems, such errors propagate even further, causing failures of the whole server and the affected HPC job(s). The essence of the error classification problem is whether or not to categorize corrected DRAM errors as failures.

Although most previous studies categorize them as such, we believe that reporting corrected DRAM errors as failures, and using them to compute statistics such as the *failure rates* or the *mean time between failures (MTBF)* could be highly misleading as it exaggerates the problem of HPC system reliability. For example, a statement that the MTBF in MareNostrum 3 during the observation period was 14 seconds, based on the corrected error count in this study, could suggest that the system suffered frequent service interruptions. However, the provided number only states that at an average rate of once every 14 seconds, one out of eight memory controllers in one out of 3056 servers performs an ECC correction. The service is not interrupted and performance is not affected. So, in the HPC domain, it is very difficult to understand the practical value of this number. On the other hand, the mean time between uncorrected DRAM errors in MareNostrum 3 was 10 days (approximately 1 million seconds), meaning that on average every 10 days a single job is terminated, a single node is shut down and single DIMM is replaced.

Overall, it is important that DRAM error studies and the research motivated by them are clear as to whether the presented failure rates and MTBF values are based on corrected or uncorrected errors, or both. On MareNostrum 3, the difference between MTBF values calculated using corrected vs. uncorrected failures was five orders of magnitude; i.e. 14 seconds vs. 10 days. And we would strongly suggest to present numbers that have a practical value.

2.2 Quantitative DRAM errors analysis: Different approaches

Quantitative analysis of DRAM errors in the field can be performed with two approaches, which are often used interchangeably. The first approach is **categorical analysis**. This approach analyzes the errors at the DIMM level, and partitions the DIMMs into various categories, e.g. based on whether they *did* or *did not* experience an error of a given type. The categorical analysis does not consider the number of errors that occurred on a given DIMM; the DIMM is categorized as soon as the first error of that type is detected and any further errors do not change the DIMM's category. It is typically used to show the proportion of the DIMMs that experienced errors, or that were replaced from the system. The second approach is to analyze the **error rates**. In this approach, the total number of errors is presented relative to other statistics, typically the amount of the MB-hours or the duration of the observation period.

To the best of our knowledge, our study is the first to distinguish between approaches for DRAM error analysis, employ both of them and compare their results on the same data. Our results clearly show that categorical and error rates analyses are not interchangeable and that they may differ greatly in stability and often lead to different conclusions.

2.3 Statistical significance

Statistical significance means that a result from testing or experimenting has a low probability of occurring randomly or by chance, allowing us to conclude with confidence that it is likely to have a specific cause. Previous field studies of DRAM errors often claim that their findings are statistically significant because the analysis covers years of data on machines with thousands of servers, totaling thousands of billions of MB-hours.

Unfortunately, these claims are misleading. Statistical significance is defined by the probability of the outcome happening by chance, not the amount of data. Therefore the significance has to be confirmed or rejected using a carefully designed statistical test that considers the type and distribution of the data under study. As we show in this paper, a large-scale experiment with a large number of observations, e.g. millions of corrected DRAM errors, does not *per se* guarantee statistical significance.

In addition to providing exploratory analysis, e.g. plotting the error rates for different memory manufacturers, we perform statistical significance tests for each finding that we present. Our analysis shows that various widely-accepted approaches for comparing DIMMs from different categories, e.g. different manufacturers, provide data that appear to support an interesting conclusion, but are not statistically significant, meaning that there is insufficient evidence to conclude that it is not merely the result of chance. We hope that these conclusions will encourage future work to analyze their data using formal statistical methods.

2.4 Replaced DIMMs

It is commonly believed that corrected DRAM errors can be used as early signals of failing devices. This reasoning is used by system protection mechanisms that, in order to prevent future uncorrected errors, retire potentially failing memory pages [9, 17, 28, 30] or replace the affected DIMMs [7, 11, 16, 22].

DIMM replacement causes bias when analyzing the dependency between corrected and uncorrected errors. DIMM replacement is based on an assumption that uncorrected DRAM errors can be predicted based on corrected errors. This assumes that probability of an uncorrected error is higher if the DIMM experienced preceding corrected errors, which, by definition, means that the two variables are statistically dependent. In systems that employ DIMM replacement, the potentially failing DIMMs could be replaced before an uncorrected error is detected. The error log of this DIMM would show corrected errors that were not followed by an uncorrected error. This input would suggest **no dependency** or even a **negative dependency** between corrected and uncorrected errors, both intuitively and in the statistical analysis. A successful DIMM replacement is actually a self-defeating prophecy— it predicts that a DIMM will fail, and therefore decides to replace it; However, since the DIMM is replaced it does not fail, so the prophecy is defeated. The bias introduced by DIMM replacement can be significant because the number of replaced DIMMs is large relative to the number of uncorrected errors. In our study, for instance, we detected 71 uncorrected errors, but an extra 51 DIMMs were replaced due to pre-failure alerts. [1]

Removing the bias due to DIMM replacement is not trivial. The main problem is that we typically do not know the effectiveness of the DIMM replacement policy, i.e., we cannot estimate the probability that a replaced DIMM would have eventually experienced an uncorrected error. In order to consider a potential bias due to DIMM replacement, we use two data-sets to analyze the dependency between corrected and uncorrected errors. First, we analyze the data logs with all the monitored events, e.g. all Corrected Errors (CEs) and Uncorrected Errors (UEs). This approach, taken by the previous studies, considers the whole system as is, without any further analysis whether additional CEs or UEs would have occurred in the case of a different system management policy, e.g. no DIMM replacement policy. Second, we analyze the data log without the replaced DIMMs, i.e., during the lot pre-processing we remove *all* the information about the DIMMs that we eventually replaced from the system. Third option would be to monitor CEs and UEs of the replaced DIMMs once that they are removed from the production. This analysis, however, is not covered in this paper because we could not obtain any monitoring data after the DIMMs were replaced.

3 ENVIRONMENT DESCRIPTION

3.1 MareNostrum 3

Our analysis is based on measurements of the memory errors on the MareNostrum 3 supercomputer [2] over a period of more than two years, from October 2014 to November 2016. In that period the MareNostrum 3 supercomputer [2] was one of the six Tier-0 (largest) HPC systems in the Partnership for Advanced Computing in Europe (PRACE) [20]. It comprised 3056 compute nodes, each with two eight-core Intel Sandy Bridge-EP E5-2670 sockets with a 2.6 GHz nominal clock frequency. In addition to the compute nodes, MareNostrum 3 also included login and test nodes. However, to mitigate the impact of different workloads, we report and analyze only the DRAM errors on the compute nodes. MareNostrum 3 included more than 25,000 DDR3-1600 DIMMs, and during the observation period we collected measurements on more than 2000 billion MB-hours. The main workloads executed on MareNostrum 3 were large-scale scientific HPC applications and the typical system utilization exceeded 95%.

We analyze DIMMs from all three major memory manufacturers, built in three different DRAM technologies. All the DIMM manufacturers presented in this study have been anonymized to protect the interested parties. In this paper, we will refer to the different memory manufacturers as *Manufacturer A*, *B* and *C*.[2] Similarly, technologies in the DIMMs under study are also anonymized, and we show only the first of two digits of the nanometer technology. $\overline{3x}$ nm, $\overline{2y}$ nm and $\overline{2z}$ nm represent three different DRAM technologies in descending order, i.e., $\overline{3x}$ nm > $\overline{2y}$ nm > $\overline{2z}$ nm.

MareNostrum 3 uses a Single Device Data Correction (SDDC) ECC scheme which can correct all errors coming from a single x4 device, usually referred to as Chipkill. For x8 devices it can correct up to 4-bit errors coming from the same DRAM chip. Each time the data is requested from the memory, the demand memory scrub checks whether the accessed data correspond to the ECC. If this is not the case, the ECC will performs error correction or report an uncorrected DRAM error. The system also includes patrol scrubbing that periodically traverses the whole physical memory, performing an ECC check on each location. If the scrubber detects any errors that are corrected by the ECC, it fixes the errors and writes the correct data back to the same memory location.

System monitoring software of MareNostrum 3 also includes pre-failure alerts that inform system administrators of the DIMMs with early signals of failing. The potentially failing DIMMs are then replaced in order to prevent future uncorrected errors.

3.2 Data collection

In Intel server architectures, the memory errors that are corrected by the ECC are recorded in the machine-check architecture (MCA) registers [13]. To log the **corrected DRAM errors**, we designed a daemon, based on the mcelog Linux kernel module [13], that periodically, each 100 ms, accesses the MCA registers, extracts the information of interest and logs them into a file. The log file contains information about the error time stamp, server and DIMM id, and the exact physical location of the error in the DIMM including rank, bank, row, column and DQ pin. Also, the daemon can distinguish whether the correction was done on an application memory read or by patrol scrubbing. If more than one error occurred in the 100 ms time interval, the MCA registers record the number of errors, but they provide detailed information only for one error in the interval. Therefore, our logs contain the exact number of corrected errors that occurred in our system, while the detailed error information is available for a statistical sample (sampled in time) of all the errors.

If multiple DRAM errors occurred in the exactly the same physical location, they are counted as a single **fault**. The faults can be extracted only from the errors with known exact physical location. Increasing the frequency of daemon access to the MCA registers would increase the sample of errors with detailed information and the sample of observed faults. However, this would also increase the performance penalty of the error logging daemon. The 100 ms time interval was selected as the shortest time interval that causes less then 1% overhead to the production applications. Previous studies perform similar readings of the memory error registers with a period of a few seconds [25–27] or once per hour [15].

On a node restart, the daemon logs the DIMM locations, manufacturer information, and a serial number unique for each DIMM. This information enables us to keep the DIMM error and fault history, even if the servers or the DIMMs are moved.

Uncorrected errors are logged by the IBM firmware [10], which is part of the MareNostrum 3 monitoring software. For each uncorrected error, the log specifies the DIMM that failed and the cause of the error, i.e. whether the error happened during an application memory read or patrol scrubbing. After an uncorrected error is reported, the corresponding DIMM is removed from production and exposed to a stress test. If additional errors are detected during

[1] Coincidentally, it matches the number of DIMMs that experienced uncorrected errors. In total, 51 DIMMs experienced uncorrected errors and additional 51 DIMMs were replaced due to the pre-failure alerts.

[2] There are 6717, 13,419 and 5247 DIMMs from anonymized *Manufacturers A*, *B* and *C* respectively.

testing, the DIMM is retired. If no errors are detected, the DIMM is returned to production. In our study, we detected 71 uncorrected errors from 51 DIMMs.

The MareNostrum 3 system monitoring log also contains information about 51 DIMMs that were replaced due to the **pre-failure alerts**. The log specifies the DIMM, the date and time of the replacement.

4 CATEGORICAL ANALYSIS

This section analyzes the percentage of DIMMs that experience errors, and evaluates whether there is a significant difference among the manufacturers and DRAM technologies. The presented analysis is formally referred to as a *categorical analysis* because the population of all DIMMs is divided into different categories based on, e.g., whether they *did* or *did not* experience an error of a given type.

4.1 Uncorrected errors

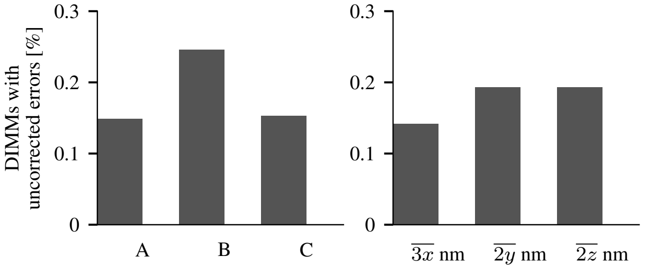
The results for the uncorrected errors are summarized in Figure 1. Figure 1(a) compares different DRAM manufacturers. Only 0.15% of *Manufacturers A* and *C* DIMMs experience uncorrected errors. For *Manufacturers B*, this percentage is somewhat higher, 0.25%. Figure 1(b) shows the technology comparison. The $\overline{3x}$ nm technology shows the best reliability with 0.14% of DIMMs experiencing uncorrected errors. For $\overline{2y}$ nm and $\overline{2z}$ nm technology the percentage of DIMMs with errors increases to 0.19%.

Overall we could conclude that the percentage of DIMMs that experience uncorrected errors is low, with some differences among the manufacturers and DRAM technologies. However, based solely on the results presented in Figure 1, we have no evidence as to whether these differences are *statistically significant.*

To verify significance of the results, we introduce formal statistical tests. Since, to the best of our knowledge, this is the first time that statistical tests are applied to the analysis of the DRAM errors, we explain the rationale behind them as well as their step-by-step application.

Essentially we want to understand whether probability of UE occurrence in a given DIMM *depends* on the DIMM manufacturer. This dependency can be checked with a statistical tests of independence. These tests can help us to determine whether there really is a difference; e.g. whether the DIMMs from *Manufacturer B* indeed have a higher probability of an uncorrected error, or whether our results show the typical variations due to chance that would be expected even without differences among the manufacturers and DRAM technologies.

In our case in particular, we use independence tests that analyze the relation between two *categorical variables*: whether a DIMM experienced an uncorrected error (first categorical variable), and the DIMM manufacturer (second variable). Once that the categorical variables are defined, each DIMM is classified into one of the categories. In our example, the DIMMs can be classified into six categories based on the DRAM manufacturer (*Manufacturer A, B* or *C*) and the error occurrence (the DIMM *did* or *did not* experience an uncorrected error), as illustrated in Table 1. Table 1 is referred to as a *contingency table*, and it shows the number of DIMMs that belong to each category. The contingency table is the input to a statistical test of independence. The test assumes the *null hypothesis* that the categorical variables are independent. The test output is the *p-value*, which is the probability of obtaining a result equal to or more extreme than what was actually observed, assuming the null hypothesis is true. If p-value is small, then we can conclude that the null hypothesis can be rejected, i.e., there is enough evidence to claim dependency between the variables. We use an $\alpha = 0.05$ cutoff for accepting or rejecting the null hypothesis, which is a standard value used in academia. In this paper, we use **Pearson's chi-square test**, the most widely used test for the independence between two categorical variables.[3]



(a) Manufacturer comparison

(b) Technology comparison

Figure 1: Percentage of DIMMs with uncorrected errors: Manufacturer and technology comparison.

	DIMMs w/ UEs	DIMMs wo/ UEs
Manufacturer A	10	6,707
Manufacturer B	33	13,386
Manufacturer C	8	5,239

Table 1: Contingency table: Comparison between different manufacturers by the number of DIMMs that experienced uncorrected error (UE). The statistical test indicates no dependency, p-value = 0.24, so we cannot claim any statistically significant difference in the probability that DIMMs from *Manufacturers A, B* and *C* will experience uncorrected errors.

The test applied to our data shows no statistically significant dependence between the number of DIMMs that experienced uncorrected error and the memory manufacturer. The outcome of the test is p-value $= 0.24$, meaning that although the results may seem to provide convincing evidence of a difference, we would expect similar or more extreme results to appear 24% of the time by chance, even if there were no differences at all. Also, in contrary to the common belief in the community is that as the technology scales down, the probability of DRAM errors increases, measurements show no statistically significant decrease in the reliability for the three generations of DIMM technology used in our system, p-value $= 0.93$.

[3] If cell values in the contingency table are small, it is recommended to use **Fisher's exact test**. Fisher's exact test is similar to Pearson's chi-square test, and a rule of thumb is to use it instead of a chi-squared test if more than 20% of the values in contingency table are lower than five. In all our results, we employ both Pearson's chi-square test and Fisher's exact test, using the chisq.test() and fisher.test() functions from the R programming language, respectively. Even for small cell values both tests have the same conclusions about accepting or rejecting the null hypothesis. Therefore, in the rest of the paper, we report p-values from Pearson's chi-square test.

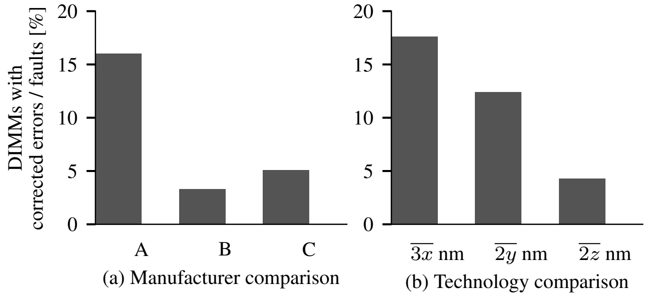


Figure 2: Percentage of DIMMs with corrected errors: Manufacturer and technology comparison.

	DIMMs w/ CEs	DIMMs wo/ CEs
DIMMs w/ UEs	23	28
DIMMs wo/ UEs	1,764	23,722

Table 2: Contingency table: Dependency between the DIMMs (all manufacturers) experiencing corrected (CEs) and uncorrected errors (UEs). The statistical test indicates strong dependency, p-value $< 2.2 \times 10^{-16}$; i.e. we can claim that DIMMs that experienced CEs have higher probability of also experiencing UEs.

4.2 Corrected errors

The results for the corrected errors are summarized in Figure 2. Figure 2(a) compares different DRAM manufacturers. *Manufacturers B* and *C* have 3.3% and 5.1% of DIMMs with corrected errors, respectively, while for *Manufacturers A* it reaches 16%. Figure 2(b) shows the technology comparison. Contrary to the common belief in the community is that as the technology scales down the DRAM reliability decreases, our measurements show the opposite trend: $\overline{3x}$ nm technology has the highest percent of DIMMs with errors, followed by $\overline{2y}$ nm and $\overline{2z}$ nm technology, respectively.

Again we use the statistical test of independence to validate whether the detected differences between the manufacturers and technology are statistically significant. The test applied to our data confirms a statistically significant difference among DRAM manufacturers (p-value $< 2.2 \times 10^{-16}$) and technologies (p-value $= 1.51 \times 10^{-15}$).

4.3 Corrected vs. Uncorrected errors

System reliability is affected only by uncorrected memory errors. However, current practice in academia and industry is to analyze corrected errors and faults as DIMM reliability indicator [15, 17, 22, 23, 25–27], In this section, we perform statistical tests to analyze the dependency between the DIMMs that experienced corrected and uncorrected errors. The contingency table for this statistical test of independence is shown in Table 2. In different rows of the table, we show the number of DIMMs *with* and *without* an uncorrected error (UE). Similarly, the table columns show the number of DIMMs *with* and *without* at least one corrected error (CE). The independence test indicates a strong dependency between the DIMMs experiencing corrected and uncorrected errors, p-value $< 2.2 \times 10^{-16}$. We repeated the tests for each manufacturer separately and observed the same conclusion. For *Manufacturers A, B* and *C*, the p-values are 0.013, $< 2.2 \times 10^{-16}$ and 0.006, respectively.

The results presented in Table 2 include all CEs and UEs monitored in our system, including the ones proceeding from the DIMMs that were replaced due to pre-failure alerts. In order to consider a potential bias due to the DIMM replacement, as described in Section 2.4, we remove all the data logs related to the replaced DIMMs and repeat the analysis. Again, the independence test indicates a strong dependency between the DIMMs experiencing corrected and uncorrected errors, p-value $< 2.2 \times 10^{-16}$. Also, the tests for each manufacturer separately lead to the same conclusion: for *Manufacturers A, B* and *C*, the p-values are 0.0007, $< 2.2 \times 10^{-16}$ and 0.0007, respectively.

4.4 Errors vs. Faults

A couple of studies [25–27] argue that the DIMMs should be compared in terms of DRAM faults rather than errors. The categorical analysis presented in this section would directly apply to the faults as well. This is due to the inherent dependency between the errors and faults—a DIMM that experienced an error at the same time experienced a fault; while a DIMM with no errors, has no faults neither. Actually, all the contingency tables, p-values and conclusions for the DRAM faults would remain precisely the same as the ones that we presented for the errors.

4.5 Summary

In this section we perform a categorical analysis of the DIMMs that experienced corrected or uncorrected memory errors. Our study is the first to support its quantitative analysis of the error logs with statistical tests that confirm or reject the significance of the findings.

The percentage of DIMMs that experience uncorrected errors is very small, which is consistent with previous studies (see Section 6). We notice some differences among manufacturers and DRAM technologies. We are the first, to the best of our knowledge, to use statistical tests to validate our findings, and the first to show a lack of their statistical significance.

We repeat the analysis for the corrected errors and, unlike for the uncorrected errors, the independence tests confirm with high confidence that these differences are statistically significant. Contrary to the common belief that scaling down the technology reduces DRAM reliability, our measurements show that the fraction of DIMMs that experience errors has reduced significantly in each DRAM generation.

Finally, we show a statistically significant dependence between the DIMMs experiencing corrected and uncorrected errors, and we find that the probability of an uncorrected error is higher if the DIMM previously experienced corrected errors. This result formally confirms the possibility to predict upcoming uncorrected DRAM errors based on preceding corrected errors which motivates further work on pre-failure predictions.

5 ERROR RATE ANALYSIS

In addition to the categorical analysis, DIMM reliability can be quantified and compared using the error rates: *errors per MB-hour* or *mean time between failures (MTBF)*. The errors per MB-hour metric considers the total number of errors, and the capacity and production time of each DIMM in a given category:

$$\text{Errors per MB-hour} = \frac{\text{Total number of errors}}{\sum(\text{DIMM capacity [MB]} \times \text{Production time [hours]})} \quad (1)$$

The mean time between failures (MTBF) for a given DIMM category, e.g., a specific manufacturer, is computed as the ratio of the sum of DIMM production times divided by the total number of detected failures for all the DIMMs of the target manufacturer:

$$\text{MTBF [hours]} \quad = \quad \frac{\sum \text{Production time [hours]}}{\text{Total number of failures}} \quad (2)$$

MTBF and per MB-hour metrics are the de facto standard for quantifying DIMM reliability, and are used by previous studies to analyze faults and corrected/uncorrected errors. However, to the best of our knowledge, no prior study has supported its findings based on these metrics with a statistical significance, nor confirmed that they provide stable and reliable results.

5.1 Uncorrected errors

5.1.1 Distribution. Figure 3 shows the incidence of uncorrected errors over time. The x-axis shows time, in months from the beginning of the study (October 2014) and the y-axis shows the number of uncorrected errors per day, across all DIMMs of a given manufacturer. On most days we detect no errors, on a few days we detect one error, and very occasionally we detect two or three errors on the same day. In total during the observation period of 25 months we detect only 71 errors.

Based on the observed errors in Figure 3, the number of errors per MB-hour is calculated using Equation 1. Figure 4 the evolution of this empirical uncorrected error rate over time. The x-axis is again the time, in months from the beginning of the study. The y-axis is now the average number of uncorrected errors per billion MB-hour, based only on the measurements done until that point. We can see that the mean error rate is highly volatile, even after more than a year of observation.

The errors per MB-hour evolve in time as an *impulse and down-ramp* function. The impulses in the average errors per MB-hour are caused by the observed errors. For example, if we consider *Manufacturer A*, the impulses in months 1, 8, 9, 12, etc., are perfectly aligned with the errors detected in Figure 3. This happens because the total number of errors is small, so the mean error rate changes significantly each time a new error is detected. For example, *Manufacturer A* experiences the first error in the 1^{st} month of the observation, and then experiences no errors in the following eight months. At the 8^{th} month, the second error is detected, which causes both the total error count and the errors per MB-hour to double. In the 9^{th} month, we detect three more errors in the same day, so the number of observed errors changes from two to five, with the result that the total number of errors and the errors per MB-hour are both multiplied by 2.5. A similar phenomenon is seen in the 12^{th} month of the observation period. We detect a similar behavior for the DIMMs produced by *Manufacturers B* and *C*.

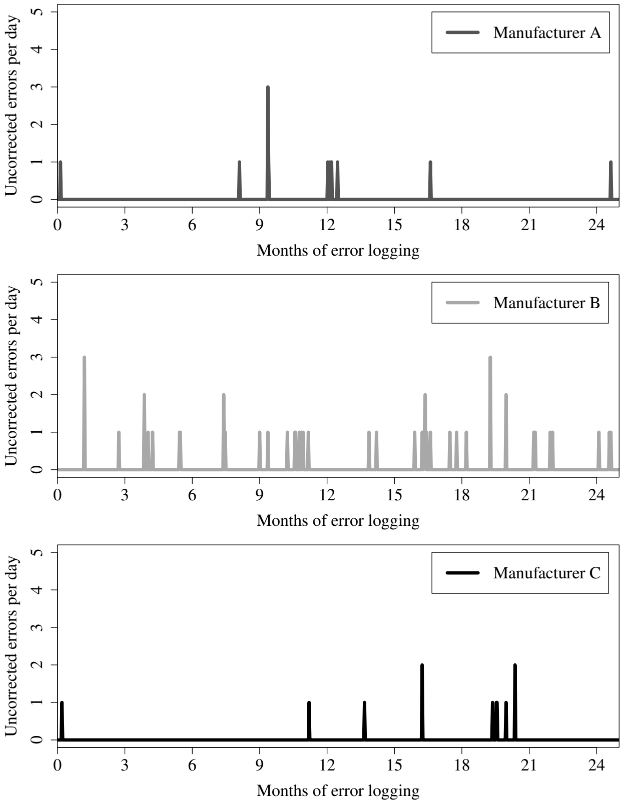


Figure 3: Uncorrected errors per day. On most days we detect no errors, on a few days we detect one error, and very occasionally we detect two or three errors on the same day.

The down-ramp segments observed in the plot of errors per MB-hour correspond to periods in which we detect no errors. In these periods the cumulative number of errors remains constant while the observation time increases. Therefore, the shape of the errors per MB-hour function is proportional to $1/t$, where t is the observation time. This is well illustrated for *Manufacturer A* between the first and eighth months of the observation period and for *Manufacturer C* between the second and eleventh months.

As a consequence of the high variability in the error rates, the ranking of manufacturers switches several times. For example, 12 months into the study *Manufacturer A* had 90% higher error rate than *Manufacturer C*. At the end of the study, at month 25, *Manufacturer C* now had 60% higher error rate than *Manufacturer A*. It is intuitive to conclude that we have little confidence in how the results would have looked if we were able to continue the study for another year.

5.1.2 Other DRAM categories and MTBF. Figure 4 illustrates the volatility of the conclusions when the error rates are used for comparison of different DRAM manufacturers. We repeat the whole analysis by looking into other DRAM categories, MTBF metric and the DRAM faults.

DRAM categories: We repeated the analysis for the three different DRAM technologies (rather than manufacturers) and reach the same conclusions. Since the total number of DRAM errors is the

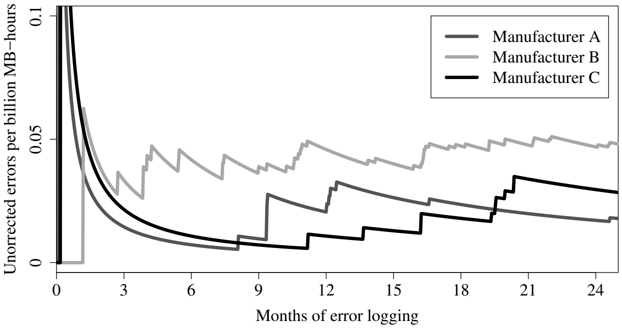


Figure 4: Average uncorrected errors per MB-hour: each point is the running average number of uncorrected errors per MB-hour observed up to that point. The error rates can vary significantly, by tens of percents each time a new error is detected. Depending on the moment of observation, we reach different conclusions about the ranking of the different DIMM manufacturers.

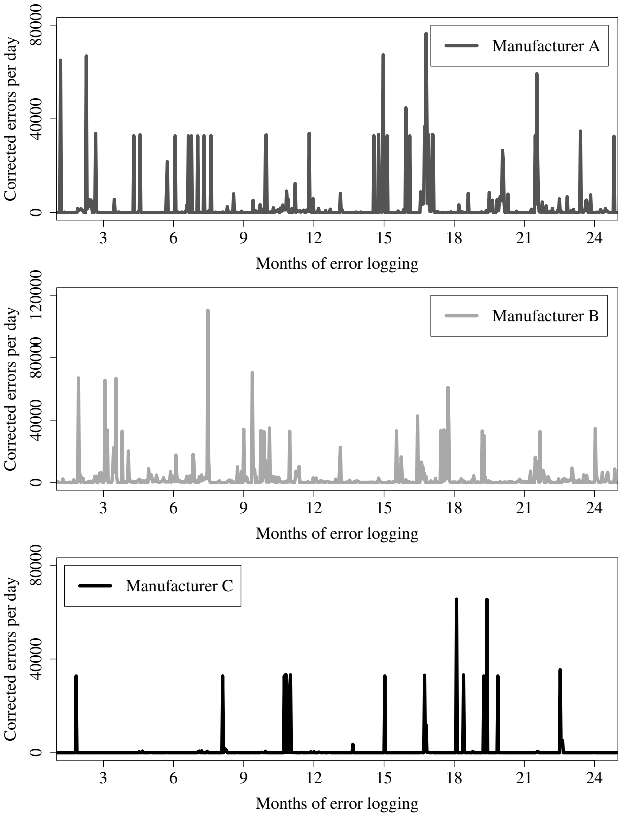


Figure 5: Corrected errors per day. On most days we detect zero or close to zero corrected errors; but on a few days there are very large numbers of corrected errors, up to about 110,000 errors.

same as before, the number of non-zero observations contributing to the calculation of the mean is still small. We strongly argue that the same problem is present when the error rates are compared for various DIMMs categories, such as DIMMs located in different datacenters, different racks or servers, or DIMMs running different workloads.

MTBF: We used the same approach to test the MTBF metric, and the conclusions are the same—we detect huge variability in the MTBF, even after long periods of error logging.

5.2 Corrected errors

We repeat the above analysis for the corrected errors and, we obtain precisely the same conclusions: the average corrected errors per MB-hour has a large variance even after more than two years of error logging.

5.2.1 *Distribution.* Figure 5 shows the number of corrected errors over time. As before, the x-axis shows time, in months from the beginning of the study, and the y-axis shows the total number of corrected errors per day, for a single manufacturer. This figure clearly illustrates the error distribution: on most days there are zero, or close to zero, corrected errors, but on a few days there are very large numbers of corrected errors, up to about 110,000 (*Manufacturer B*, Month 7). When repeated at a finer granularity, by measuring not per day, but per hour, minute or second, we detected the same trend: >99% of the observations had no errors, and again a very small proportion of observations had very high values.

Similarly to the uncorrected error case, to illustrate the difficulty in measuring the mean, Figure 6 plots the evolution of the empirical mean error rate, per MB-hour, for the three manufacturers, over time. This figure shows that corrected errors per MB-hour also evolve in time as an *impulse and down-ramp* function. An intensive error burst, as seen in Figure 5, *significantly increases* the number of detected errors in a short time interval, causing an impulse in the average errors per MB-hour function. For example, if we observe *Manufacturer C*, the impulses in Figure 6, e.g. in the observation months 2, 8, 11, 15, etc., are perfectly aligned with the intensive error burst in Figure 5. The down-ramp segments of the errors per MB-hour function correspond to the periods in which we detect few errors.

Figure 6 clearly shows the volatility when calculating the mean errors per MB-hour. The error rates can vary significantly, by tens of percents in just a few days. We detect this behavior for all three manufacturers. Also, we detect this behavior not only at the beginning of the study, where it might be expected due to the small observation period, but also after long periods e.g. well after one year of the study. As a consequence of the high variability in the error rates, the ranking of manufacturers switches several times. Actually, during the observation period, *Manufacturer A* and *B* switched order eight times. At month 4, *Manufacturer B* had 40% higher error rate than *Manufacturer A*, but at month 17, *Manufacturer A* had 25% higher error rate than *Manufacturer B*.

Overall, our results show high variability of the results despite the millions of corrected errors observed over the course of the study. Even after long periods of error logging, any comparison of different DRAM manufacturers based on the errors per MB-hour may support different conclusions depending on the moment in which the measurements are finalized.

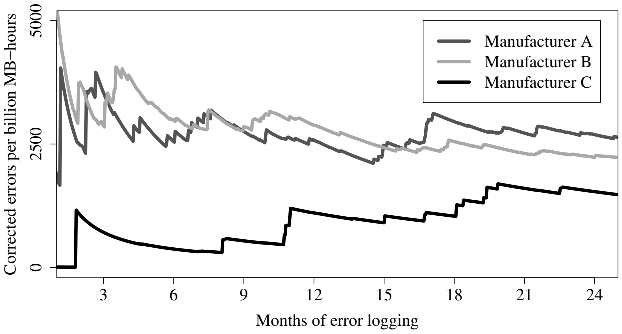


Figure 6: Average corrected errors per MB-hour: each point is the running average observed up to that point. Depending on the moment of observation, we reach different conclusions about the ranking of the DIMM manufacturers. During the observation period, *Manufacturer A* and *B* switched order eight times.

5.2.2 Other DRAM categories and MTBF. As for the uncorrected errors, we repeat the analysis for different **DRAM technologies**, the **MTBF metric** and **DRAM faults**, and the conclusions are the same—the error rates can vary significantly, by tens of percents in just a few days, not only at the beginning of the study, but also after long observation periods.

5.3 On the volatility of the error rates

In the previous section we have seen that intensive bursts of corrected errors can cause significant changes in the average error rates even after long observation periods. Next, we quantify and explore in more detail the bursty behavior of corrected errors.

5.3.1 Quantifying burstiness of observed errors. A recent study of Goh and Barabási [5] explores burstiness in wide range of systems, and identifies two causes for it: (1) The inter-event time distribution, which, in our study corresponds to the distribution of the interval lengths between consecutive errors, and (2) Memory, i.e. correlation between pairs of consecutive inter-event interval lengths. The study also defines two parameters that quantify these causes of burstiness: the *burstiness parameter*, B, and the *memory coefficient*, M.

The **burstiness parameter B** is based on the inter-event time distribution and it is defined as the ratio $B = \frac{\sigma - m}{\sigma + m}$, where σ is the empirical standard deviation and m is the empirical mean, calculated from the measured times between errors. The value of B is in the bounded range $[-1, 1)$, and its magnitude correlates with the signal's burstiness: $B \longrightarrow 1$ for the most bursty signals, $B = 0$ is neutral (e.g. for Poisson sequence), and $B = -1$ corresponds to a completely regular (periodic) signal.

The **memory coefficient M** quantifies the correlation between pairs of consecutive inter-event interval lengths. It is based on the Pearson correlation between two samples of the same size, $X = x_1, \cdots, x_N$ and $Y = y_1, \cdots, y_N$ [24]:

$$\rho_{X,Y} = \frac{1}{N} \sum_{i=1}^{N} \frac{(x_i - m_X)}{\sigma_X} \frac{(y_i - m_Y)}{\sigma_Y}$$

where N is the size of the samples, m_X and m_Y are the sample means of X and Y, respectively, and σ_X and σ_Y are the sample standard deviations. The Pearson correlation is in the range $[-1, 1]$. A value greater than zero implies that samples X and Y are positively correlated, i.e. x_i and y_i tend to deviate from their means in the same direction. A value less than zero implies that X and Y are negatively correlated, while a value of 0 implies no correlation.

Applying this idea to find the correlation between consecutive inter-event interval lengths gives [5]:

$$M = \frac{1}{n-1} \sum_{i=1}^{n-1} \frac{(\tau_i - m_X)}{\sigma_X} \frac{(\tau_{i+1} - m_Y)}{\sigma_Y}$$

where τ_i is the i-th inter-arrival interval and n is the total number of errors. Then, m_X and σ_X are the sample mean and standard deviation of $\tau_1, \cdots, \tau_{n-1}$ (sequence X) and m_Y and σ_Y are the sample mean and standard deviation of $\tau_2, \cdots, \tau_n$ (sequence Y). The range of values and interpretation of the memory coefficient corresponds to the Pearson correlation. When a short inter-event interval tends to be followed by another short interval and long by long, we get a positive memory coefficient, $0 < M \leq 1$, indicating positive correlation. When short intervals are more likely to be followed by *long* intervals and vice versa, the consecutive interval lengths are negatively correlated, so we get $-1 \leq M < 0$. When the interval lengths are uncorrelated, such as, e.g. for a Poisson process, we have $M = 0$.

5.3.2 Corrected error volatility. For Manufacturer A, B, and C, respectively, the inter-event time distribution of our corrected errors has the burstiness parameter of 0.94, 0.86, and 0.95, while the memory coefficient equals 0.18, 0.29 and 0.29. These B and M values are very high, and seldom seen in experimental contexts related to natural and human phenomena [5].[4]

In order to get an intuitive understanding of the importance of the inter-event distribution (quantified by B) and correlation (quantified by M), we plot the evolution of the real corrected error rate in time alongside two synthetic timelines exploring the two causes of the error rates volatility (see Figure 7). As before, in all three charts, the x-axis is the time, in months from the beginning of the study, and the y-axis is the average number of corrected errors per billion MB-hour, based only the measurements done until that point.

Figure 7(a) is the real timeline. Figure 7(b) plots a synthetic timeline with the real inter-event distribution (B is unchanged), but with these intervals randomly permuted in time ($M = 0$). We see immediately that Figure 7(b) has much lower volatility than Figure 7(a), illustrating the impact of correlation on the volatility of the results. The original distribution has a high positive memory parameter, and the error rates vary significantly, by tens of percents in just a few days, even after long periods. By only making the inter-error time intervals independent, i.e. by setting memory parameter $M = 0$, the presented error rates become stable only after a few months of the measurements.

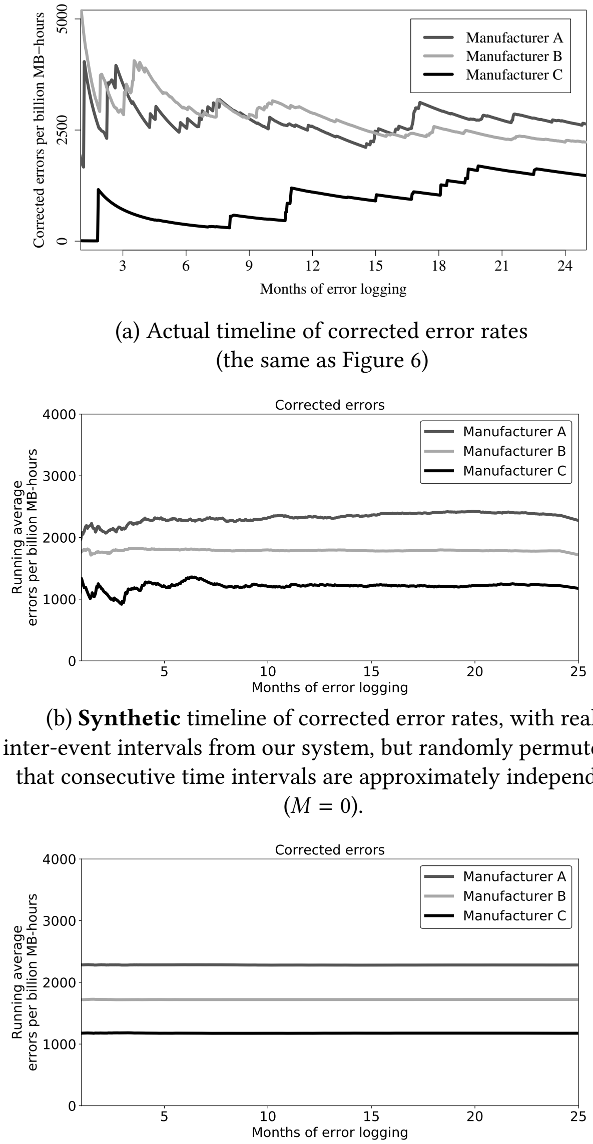
As the final step of the analysis, we explore the impact of the inter-event distribution. Figure 7(c) plots synthetic data assuming

[4] Goh and Barabási [5] also provide a detailed interpretation of the B and M parameters, and use the $(B,\ M)$ phase diagram to plot and compare various human activities and natural phenomena.

Poisson arrivals ($M = 0$, $B = 0$), at a rate equal to the empirical mean from the actual measurements. In this case the sample average rapidly converges, even in the first day, to very close to the correct value.

Note that Figures 7(b) and (c) each plot a single trial of a synthetic timeline assuming independent inter-event intervals and Poisson inter-event intervals, respectively. We performed the experiment several times and obtained the same general behavior.

5.3.3 Uncorrected error volatility. In Section 5.1 we illustrated the impact of the small number of observations on the volatility of the uncorrected error rate. We extend this explanation with a formal analysis as to whether part of this volatility is due to burstiness of the uncorrected errors. As for the corrected errors, we compute the B and M parameters for the different manufacturers, [5] and analyze the impact of these causes on the volatility of the *errors per MB-hour* metric. Figure 8 includes three charts: the actual measured results from our system (Figure 8(a)). the same data with an independent, randomly permuted inter-error periods ($M = 0$) (Figure 8(b)), and data assuming Poisson arrivals ($M = 0$, $B = 0$) at a rate equal to the empirical mean from the actual measurements (Figure 8(c)). Even when the data is plotted assuming Poisson arrivals with $M = 0$ and $B = 0$, the error rates have large variation after even long observation periods, and there is no qualitative difference between the empirical error rates plotted in the three subfigures. Also, a small number of observations causes significant variation in the shape of the data plotted in Figure 8(b) and (c), from one trial to another (we plot only one of several trials that we performed). Therefore we conclude that a large part of the justification for the volatility in the average number of uncorrected errors is simply due to the small number of uncorrected errors in the sample.



(a) Actual timeline of corrected error rates (the same as Figure 6)

(b) **Synthetic** timeline of corrected error rates, with real inter-event intervals from our system, but randomly permuted so that consecutive time intervals are approximately independent ($M = 0$).

(c) **Synthetic** timeline of corrected error rates, with synthetic inter-event intervals assuming Poisson arrivals ($M = 0$, $B = 0$), at a rate equal to the empirical mean from the actual measurements.

Figure 7: Average corrected errors per MB-hour: each point is the running average observed up to that point. The memory coefficient M and burstiness parameter B of the inter-error time distribution are the main causes of the average error rate variability.

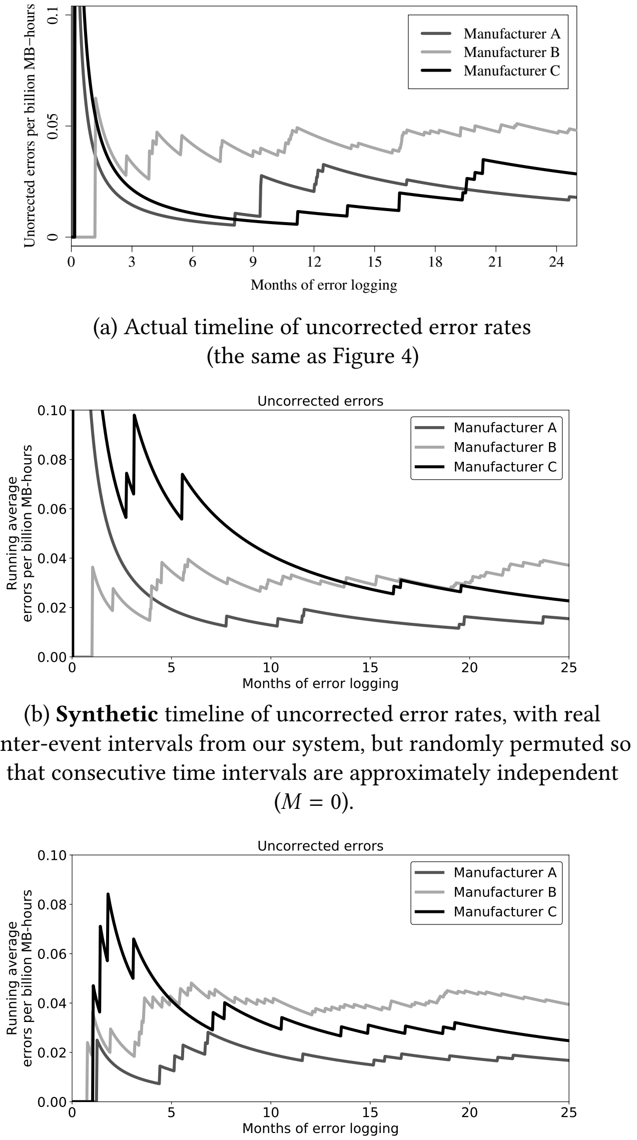
5.4 Statistical significance

Errors per MB-hour and MTBF are standard metrics for measuring DRAM reliability in large-scale systems. In the previous sections, however, we show that their values can vary significantly even after long observation periods, which can lead to unreliable conclusions that depend on the moment in which the measurements are finalized. It is therefore essential that the community looks for alternative approaches to quantify DRAM errors rates and reliably compare them between different categories.

In this section we consider the choice of statistical methods for comparison of different error rate distributions. Whereas the categorical analysis in Section 4, which used Pearson's chi-square test, only needed to assume that the observations of different DIMMs are random, independent and identically distributed, the analysis of numerical data usually requires assumptions about the underlying distribution.

The most common tests for statistical significance in numerical data are the t-test (for two sets of data) and ANOVA (its generalization to three or more sets of data). Both tests assume that the sample means have a normal distribution. In many circumstances, this assumption is justified, either because the population itself is close to normal or because the average of a large sample is close to normal by the Central Limit Theorem. Unfortunately, however, if the data being averaged, in our case the error occurrences, has heavy tails or are not independent, then convergence of the sample mean to normal is very slow, requiring an extreme sample size. Our

[5] Manufacturer A, B, and C, respectively, have the burstiness parameters of 0.17, 0.14, and 0.25, and the memory coefficient of 0.25, −0.17 and 0.34.



(a) Actual timeline of uncorrected error rates (the same as Figure 4)

(b) **Synthetic** timeline of uncorrected error rates, with real inter-event intervals from our system, but randomly permuted so that consecutive time intervals are approximately independent ($M = 0$).

(c) **Synthetic** timeline of uncorrected error rates, with synthetic inter-event intervals assuming Poisson arrivals ($M = 0$, $B = 0$), at a rate equal to the empirical mean from the actual measurements.

Figure 8: Average uncorrected errors per MB-hour. Volatility of the uncorrected error rates is caused by the small number of observations. Even when the data is plotted assuming Poisson arrivals ($M = 0$, $B = 0$), we detect large variation after long observation periods.

experiments indicate that even with 2000 billion MB-hours and two years, the sample mean of the corrected errors is unlikely to be normal. In fact, applying the Kolmogorov–Smirnoff test to the actual numbers of corrected errors, with one sample per DIMM, gives a p-value $< 10^{-60}$ for all three manufacturers, indicating that if the distributions were normal, then the observed results would be highly unlikely. Due to the small number of non-zero samples for the uncorrected errors, it is difficult to determine whether or not this is also the case for uncorrected errors.[6]

There are, however, additional statistical tests that do not assume a normal distribution, or any parametrized distribution (such as normal, whose parameters are the mean and variance). Such *non-parametric* tests include the Mann–Whitney U test (for two sets of data) and Kruskal–Wallis (its generalization to three or more sets of data). These tests are commonly thought to compare population medians, rather than means, but this is not strictly true. In our case comparing population medians would be useless since, for all DRAM manufacturers and technologies, >99% of MB-hours had zero (un)corrected errors, so the median number of (un)corrected errors per MB-hour is zero.

We applied the Kruskal–Wallis test and could not conclude that there is any statistically significant difference among the distributions (p-value $< 2.2 \times 10^{-16}$). It is important to realize that even if we had found a statistically significant difference, the strongest conclusion that we could have made would have been that the DRAM manufacturers have different distributions, not that one has a higher mean or median than another. To conclude the latter would have required an assumption of statistical dominance; i.e. that the cumulative distribution functions do not cross, but our experiments (not presented) show that it is quite likely that they do.

5.5 Corrected vs. Uncorrected errors

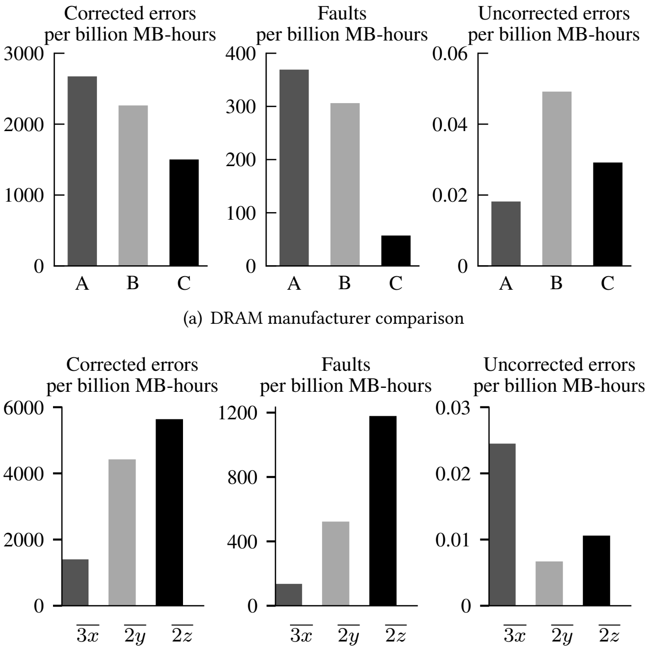
Next, we compare the errors per MB-hour and MTBF metrics based on DRAM faults, corrected errors and uncorrected errors. In case that these trends were similar, we could conclude that measurements based on the corrected errors and faults might be used as an indirect indicator of the DRAM reliability. Our results, however, clearly show that the corrected errors and faults rates show trends that are **completely different** from the uncorrected errors.

In Figure 9(a), we compare the errors per MB-hour for various manufacturers.[7] Corrected errors, faults and uncorrected errors results are presented in separate charts, while different bars refer to the different DRAM manufacturers. When counting the corrected errors, the highest error rate, 2665 errors per billion MB-hours, is measured for *Manufacturer A*, followed by *Manufacturers B* and *C* with 15% and 44% lower error rates, respectively. Fault rates follow a similar trend, *Manufacturer A* has the highest fault rate followed by *Manufacturers B* and *C*. The uncorrected error rates, however, follow a **different trend**: *Manufacturer A* shows the lowest rate, followed by *Manufacturer C* (1.6× increment) and *Manufacturer B* (2.7× increment).

We get the same conclusion when comparing DIMMs with different technologies, see Figure 9(b). The corrected error and fault rates increase as the technology scales down from $\overline{3x}$ nm to $\overline{2y}$ nm and $\overline{2z}$ nm. The uncorrected error rates, again, follow a **different trend**: $\overline{2y}$ nm technology shows the lowest error rates followed by the $\overline{2z}$ nm (1.6× increment) and $\overline{3x}$ nm technology (3.7× increment).

[6] The p-values are 0.01, 5×10^{-6} and 0.06.

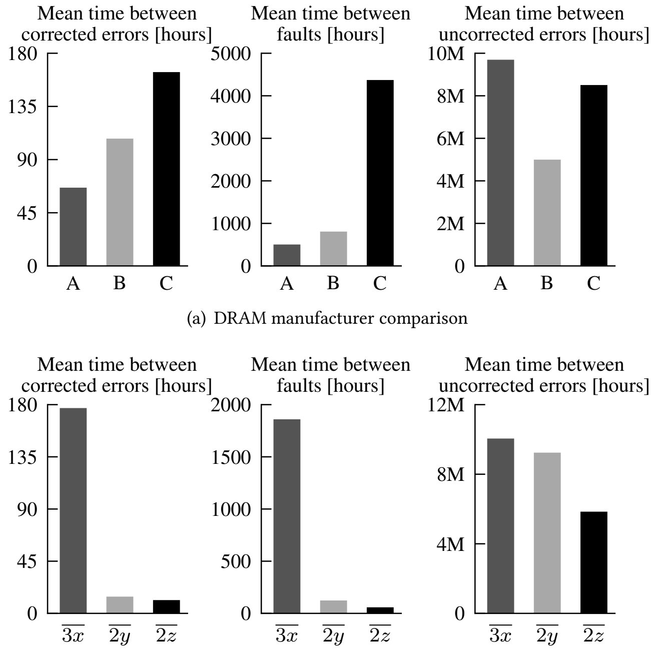
[7] Note that because of the big difference in error rates, charts in Figures 9 and 10 have different scales on the y-axis.



(a) DRAM manufacturer comparison

(b) DRAM technology comparison

Figure 9: Corrected errors, faults and uncorrected errors per billion MB hours. The corrected error and fault rates have the same trend, but the uncorrected error rates exhibit a different trend.



(a) DRAM manufacturer comparison

(b) DRAM technology comparison

Figure 10: Mean time between corrected errors, faults and uncorrected errors. The corrected error and fault rates have the same trend, but the uncorrected error rates exhibit a different trend.

Similarly to the analysis of errors per MB-hour, we compare the MTBF metric based on DRAM faults, corrected and uncorrected errors. We perform the analysis for different DRAM manufacturers and DIMM technologies, presented in Figures 10(a) and 10(b), respectively. As in the errors per MB-hour analysis, MTBF based on the corrected error and fault show similar trends, that are **completely different** from the uncorrected errors.

The results presented in Figures 9 and 10 include all corrected and uncorrected errors monitored in our system, including the ones proceeding from the DIMMs that were replaced due to pre-failure alerts. In order to account for a potential pre-failure alert bias, as explained in Section 2.4, we remove all the data logs of the replaced DIMMs and repeat the analysis. Again, our results show that the corrected errors and faults errors rates show trends that are completely different from the uncorrected errors. Therefore, we conclude that the errors per MB-hour and MTBF metrics based on DRAM faults and corrected errors **cannot be used** even as an indirect indicator of the DRAM reliability.

5.6 Error rates vs. Categorical analysis

As the final step of our study, we compare the findings of the categorical and the error rates analysis. Our results clearly show that although quantitative DRAM error analysis may be performed with both approaches, they are **not interchangeable** and could lead to different conclusions. We illustrate this with three examples based on the data used in this study.

Figure 11(a) compares the uncorrected DRAM errors for three technologies under study, $\overline{3x}$ nm, $\overline{2y}$ nm and $\overline{2z}$ nm. The left figure shows the uncorrected error rates per MB-hour, while the right figure shows the categorical analysis, the percent of DIMMs that experienced an uncorrected error. It is clear that the trends on the figures are **completely different**, e.g. $\overline{3x}$ nm technology has the highest rate of the errors per MB-hour, while it has the lowest percent of the DIMMs with uncorrected errors.

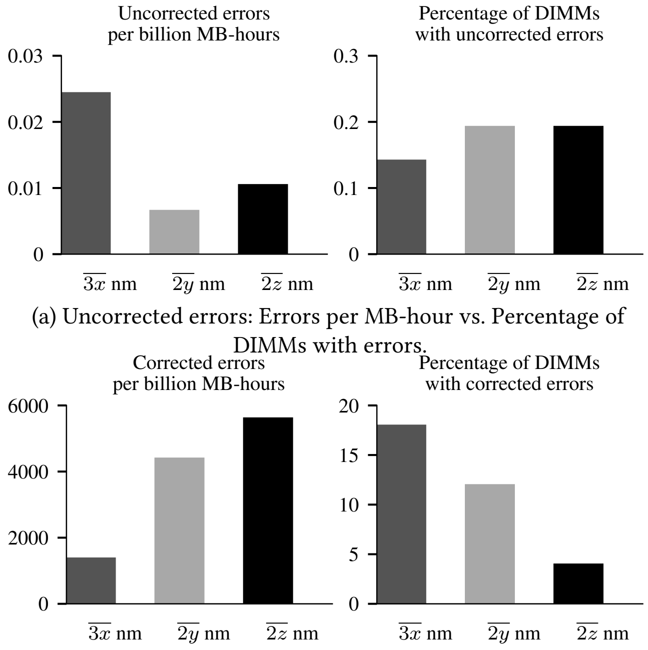
Figure 11(b) illustrates the same for the corrected errors. Again, the trends on the are completely different depending on whether we compare the DRAM technologies based on the error rates per MB-hour or the percent of DIMMs that experienced an error.

Finally, our categorical analysis confirms the strong dependency (p-value $< 2.2 \times 10^{-16}$) between the DIMMs experiencing corrected and uncorrected DRAM error, see Section 4.3. However, analysis of the same error logs showed that the per-MB-hour and MTBF metrics based on corrected DRAM faults and errors have trends that are **completely different** from the uncorrected errors, see Section 5.5.

5.7 Summary

In this section, we analyzed the DRAM error distributions and the variability of errors per MB-hour and MTBF over the course of the 25-month observation period.

First, we show that average errors rates, errors per MB-hour and MTBF, have a **large variance** even after **more than two years**



(a) Uncorrected errors: Errors per MB-hour vs. Percentage of DIMMs with errors.

(b) Corrected errors: Errors per MB-hour vs. Percentage of DIMMs with errors.

Figure 11: Analysis of the same DRAM error data with different approaches, categorical and the error rates, can lead to completely different conclusions.

of the error logging. The findings are the same for corrected and uncorrected errors and for both comparisons: DRAM manufacturer and technology. We also show that errors per MB-hour and MTBF show different conclusions depending on the moment in which the measurements are finalized. It is intuitive to conclude that we have **little confidence** in how the results would have looked if we were able to continue the study, e.g. for another year.

Second, we explore in more detail the causes of the corrected and uncorrected error rate volatility. For uncorrected errors, a large part of the justification for this volatility is simply due to the small number of uncorrected errors observed during the study. For corrected errors, the significant changes in the average error rates are caused by intensive error bursts, up to about 110,000 errors in only few days. We further explore the causes of this burstiness, and conclude that it is a consequence of inter-event time distribution and a strong correlation between consecutive inter-error interval lengths.

Third, we carefully consider the options for statistical significance tests when comparing the average DRAM error rates from DIMMs in different categories. We conclude that this would require statistical dominance, i.e. that the cumulative distribution functions from different categories do not cross, supported with the non-parametric tests that could confirm statistical difference among the distributions. We are aware that this is a very strict requirement, and we would encourage future work that would explore less conservative approaches to this problem.

Also, we show that using the corrected error rate and fault rate, errors per MB-hour or MTBF, as an indicator of DRAM reliability is **misleading** because the uncorrected error trends can be **completely different**. This is one more example that shows how important it is to understand the relation between DRAM faults, corrected errors and uncorrected errors. Any metrics based on corrected errors or faults should be used as a DRAM reliability indicator **only** if there is a clearly understood relationship with uncorrected errors.

As the final step of our study, we compare the findings of the categorical and the error rates analysis. Our results clearly show that although quantitative DRAM error analysis may be performed with both approaches, they are **not interchangeable** and could lead to different conclusions.

Error and fault rates are the **de facto standard** for measuring DRAM reliability in both academia and industry. To the best of our knowledge this is the first study that analyzes the limitations of these approaches and demonstrates that they can provide volatile and unreliable results, leading to incorrect conclusions about DRAM reliability. It is therefore essential to question the current practice in quantifying DRAM reliability and to select a proper analysis approach. Our strong suggestion would be to select the method that provides the most stable and, ideally, statistically significant results. Another important requirement is that the selected method provides numbers with a practical value that could be easily related to HPC system reliability.

6 RELATED WORK

In recent years, various studies have analyzed corrected and uncorrected DRAM errors and faults in the field.

Uncorrected DRAM errors and whole system resiliency: Schroeder et al. [21], Martino et al. [16] and Gupta et al. [6] analyze the impact of DRAM errors on the resiliency of large-scale compute clusters. The authors consider numerous causes of the system failures including hardware components, software and environment. The analysis of the DRAM errors is only a small part of their studies.

Schroeder et al. [21] analyze failures of the Los Alamos National Laboratory HPC systems between 1996 and 2005. The authors report that uncorrected memory errors account for 20% of all hardware failures, and were the root cause of 30% of the server failures.

Martino et al. [16] analyze failures of the Blue Waters supercomputer during 261 days. The supercomputer includes general purpose computing nodes with chipkill protected DDR3 and GPU accelerators with SEC-DED protected DDR5. The authors detect 1.5 million corrected and 28 uncorrected DRAM errors, and report that DRAM is the cause of 44% of all server failures.

Gupta et al. [6] compare and contrast the reliability characteristics of five production HPC systems deployed at the Oak Ridge National Laboratory. The study analyzes error logs gathered during more than eight years and covering more than one billion compute node hours. The study shows that hardware-related errors, such as uncorrected errors in the CPU caches and main memory, are equally or more dominant than the software errors, such as those in the file system and kernel. The authors emphasizes the importance of the CPU and GPU memory errors, and advocate for increasing the reliability of these components by better memory provisioning and replication.

The previous studies are very important for two reasons. First, these studies show that DRAM is one of the main causes of hardware failures, and they quantify the impact of these failures on system reliability. This positions DRAM failures in the overall picture of large-scale compute cluster reliability. Second, when quantifying system reliability, the studies focus on uncorrected DRAM errors. Although the message is not as explicit as it could be, it is very clear: system reliability is driven by uncorrected errors not corrected errors.

Corrected vs. Uncorrected DRAM errors: Few studies mention the dependency between corrected and uncorrected DRAM errors.

The results of Schroeder et al. [22] indicate that two months after a corrected error the DIMM has higher probability to experience an uncorrected one. The authors also present the idea of an early replacement policy, where a DIMM is replaced after experiencing a significant number of corrected errors, rather than waiting for the first uncorrected one.

Sridharan and Liberty [26] confirm that the probability of an uncorrected fault may increase if the server had preceding corrected faults, especially if the corrected faults affected various ranks and banks of a given DIMM.

Levy et al. [14] analyze SRAM and DRAM failure data collected during the entire life-time of the Cielo supercomputer. Unlike all previous studies, this work concludes that corrected DRAM faults **are not predictive** of future uncorrected faults. The presented data show only a weak temporal relation between corrected DRAM faults and a subsequent uncorrected DRAM fault at the same server. The study also analyzes the average fraction of servers with corrected and uncorrected DRAM faults per day, and concludes that there is no strong correlation between these two statistics. Finally, the authors compute the average number of corrected faults per server, and show a small difference between the nodes that did and did not experience uncorrected faults.

Comparing and combining the conclusions of these studies is not trivial. First, although the studies analyze the same phenomena, they do not analyze the same problem. Schroeder et al. [22] use a **categorical approach** to analyze the probability of corrected and uncorrected **error** occurrence in each **DIMM**. Sridharan and Liberty [26] interchangeably use terms DRAM **error** and **fault**. They use a **categorical approach** to analyze the probability of DRAM errors/faults occurrence at different **servers**. Levy et al. [14] also analyze the DRAM **faults** at the **server level**. This study, however, uses the **error rate analysis** to explore correlation between the rates of corrected and uncorrected DRAM faults. Finally, although all the studies agree that the number of observed uncorrected errors (or faults) is very low, none of the presented quantitative results are supported by statistical tests.

Predicting uncorrected DRAM errors: Giurgiu et al. [4] propose a machine learning random forest model for prediction of uncorrected DRAM errors. The prediction is based on the previously-detected corrected errors and measurements from over 100 sensors that monitor system functioning. The model achieves very high precisions of up to 96%; meaning that up to 96% of the predicted uncorrected DRAM errors indeed occur in future. The model still has problems to predict uncorrected errors which have no preceding corrected ones. Predicting these errors is very important because typically they correspond to a majority of all uncorrected DRAM errors [4, 14].

Corrected DRAM errors: Most DRAM error studies focus their analysis on corrected errors. These studies cover various large-scale compute systems, with DDR1, DDR2, DDR3 and FBDIMM DIMMs.

Schroeder et al. [22] present the first large-scale study of DRAM memory errors in the field. The study covers 2.5 years (Jan 2006–June 2008) of DRAM error logging of the Google fleet with six different platforms using DDR1, DDR2 and FBDIMM memory with SEC-DED and chipkill ECC. The study analyzes corrected and uncorrected error probabilities and rates, and correlates them with different factors, such as chip capacity, temperature, utilization, aging and DIMM generation.

Li et al. [15] report on nine months of DRAM error collection from various platforms with a total of 800 GB of DDR2 memory. The authors pay special attention to a comparison of transient and non-transient errors, and the study discovers a significant number of non-transient errors, with multiple errors often occurring in the same row or column.

Hwang et al. [9] study data on DRAM errors collected from four different environments: SciNet HPC cluster (Canada), the IBM Blue Gene/L at Lawrence Livermore National Laboratory, the Blue Gene/P from the Argonne National Laboratory, and 20,000 machines from Google's data centers. In total, this work covers nearly 300 terabyte-years of main memory utilization in the field. The study distinguishes between soft transient errors and hard DRAM errors caused by the device defects and provides a detailed analytical study of both error types. The authors also propose memory page retirement policy as a protection mechanism that would prevent a large number of corrected DRAM errors in production systems, while sacrificing only a negligible fraction of the total DRAM.

Sridharan and Liberty [26] analyze 11 months (Nov 2009–Oct 2010) of DRAM error logs from the Jaguar supercomputer located at the Oak Ridge National Laboratory. The study covers DDR2 DIMMs with chipkill ECC and presents detailed analysis of the corrected errors and fault types: permanent, transient, single-bit, multi-bit, row, column, bank, multi-bank and multi-rank. Sridharan et al. [27] extend this study with the analysis of the error logs from the Cielo supercomputer located at the Los Alamos National Laboratory. These logs observe 15 months (mid-2011 to early-2013) of chipkill-protected DDR3 DIMMs. This study focuses on DRAM faults (corrected errors faults) and presents a detailed analysis of fault types, similarly to Sridharan and Liberty [26]. The study also analyzes fault rates as a function of the DRAM vendor, physical location of the fault in the DRAM device, location of the DRAM device in the data-center, and the data-center altitude. Sridharan et al. [25] extend these studies with the 18 months (April 2011 to January 2013) of the error logs from the Hopper supercomputer located at the Lawrence Berkley Labs. The study covers DDR3 DIMMs with a chipkill ECC scheme. These studies [25–27] strongly argue that "system health" should be measured in terms of DRAM faults rather than errors. The term "system health" is not explicitly defined, but if it is a synonym for system reliability, then we argue instead that it should be quantified by uncorrected DRAM errors, rather than corrected errors or faults.

Siddiqua et al. [23] analyze DRAM errors logs collected from 30,000 servers over a period of three years. This is the first study that uses the pattern of the error addresses to distinguish between errors caused by the memory module, memory controller, memory channel and bus. The authors conclude that memory module faults are by far the most dominant fault type. Meza et al. [17] extend this work, and distinguish between errors caused by the DIMM bank, row, column and cell. They analyze 14 months of DRAM error logs from the Facebook fleet comprising DDR3 DIMMs, and conclude that 85% of memory errors are not caused by the DIMM, but by the socket and memory channel, which is opposite to the conclusions of Siddiqua et al. [23]. Meza et al. [17] also analyze corrected error rates as a function of DIMM manufacturer, DIMM architecture, technology, workload characteristics, CPU and memory utilization.

These studies present many quantitative results on corrected DRAM errors and faults, and their analysis is extremely valuable for the understanding of DRAM error and fault rates, distributions, and correlated factors. However, it is important to not forget that *only* uncorrected DRAM errors have an impact on system reliability. Our results, as well as the previous studies [4, 14, 22, 26] show there is no direct relation between corrected and uncorrected errors and faults. So, it is not straightforward to quantify whether, and to what extent, the analyses focused on corrected DRAM errors improve our understanding of the memory system reliability.

GPU errors in the field: Extensive use of GPUs in HPC motivated studying and improving of their reliability, and few recent studies analyze GPU errors in the field. Tiwari et al. [29] analyze GPU error logs from February 2013 to August 2014 (over 18 months) of the Titan supercomputer comprising 18,688 K20X GPUs. Nie et al. [18] continue this work with the analysis of the GPU-error related data on the same system from February 2015 to June 2015. The follow-up study from the same team [19] proposes and evaluates several machine learning-based models for the GPU error prediction. The studies reveal interesting insights about the temporal and spatial distribution of GPU errors, their correlation with temperature, GPU power consumption and workload characteristics.

In K20X GPUs the error correction codes protect all the major memory structures including register files, caches and device memory. The studies analyze all the single-bit errors together, and do not distinguish between the errors in different memory structures. Therefore, it is not easy to understand whether and how the findings of these studies could be applied to DRAM resiliency. Tiwari et al. [29] actually report that 98% of the detected errors come from the L2 cache, while the register files, instruction cache, L1 cache, shared memory and device memory together contribute to only 2%.

Our study: The main focus of our study is not to understand the cause of DRAM errors in the field and the correlated factors, but rather to emphasize the complexity of this quantitative analysis and importance of the statistically sound methodology. Although our analysis detects various weaknesses in the quantitative DRAM error analysis performed by previous studies, our objective is not to discredit there studies nor their findings. Instead, our objective is to increase the awareness of the limitations of various widely-used methods, and to present methodology, statistical tests and examples that improve any future analysis of the DRAM errors in the field.

7 CONCLUSIONS

This paper summarizes our two-year study of corrected and uncorrected errors on the MareNostrum 3 supercomputer, covering 2000 billion MB-hours of DRAM in the field. The main objective of the paper is to help the community to define standards for any future quantitative analysis of DRAM errors. The paper has two sets of contributions. First, we illustrate the complexity of in-field DRAM error analysis and demonstrate the limitations of various widely-used methods. Understanding these limitations is important because, as we show, widely-accepted approaches for DRAM analysis provide volatile, unreliable and statistically insignificant results that may lead to incorrect conclusions about DRAM reliability. Second, we present formal statistical methods that overcome many of the limitations of the current approaches. The methods that we present are simple to understand and implement, reliable and widely accepted in the statistical community.

This is the first study that clearly distinguishes between the *categorical* and *error rate* analysis. Although both methods are valid, our results clearly show that they are not interchangeable and can lead to completely different conclusions. This is very important finding, because various previous studies interleave categorical and error rate analysis and the conclusions based on them.

As a part of the categorical analysis, we explain and use independence tests to confirm or reject, in terms of statistical significance, any differences observed among various categories. We use these tests to analyse the percentages of DIMMs that experience uncorrected or corrected errors, for the different manufacturers and DRAM technologies. These tests allow us to ascertain whether the observed differences are likely to be due to real differences or are explainable merely by chance. To the best of our knowledge, this is the first study of DRAM errors that uses statistical tests to confirm or reject the significance of its results.

Regarding the error rates, we show that the average errors per MB-hour and average MTBF were highly volatile over the course of the study, with the final values depending critically on the moment at which the study is terminated. It is intuitive to conclude that we have little confidence in how the results would have looked if we were able to continue the study, e.g. for another year. We perform a careful study of the causes of this volatility, and conclude that the primary reason differs between uncorrected and corrected errors. For uncorrected errors, the volatility is explained by the small number of observations; 71 uncorrected errors over the course of the study. For corrected errors, the volatility is explained by error burstiness in time. Moreover, we show that using the corrected errors and fault rates as an indicator of DRAM reliability may be misleading because they have completely different trends from the uncorrected errors, which are the only errors that lead to system failure.

Our study alerts the community about the need to question the current practice in quantifying DRAM reliability and to select a proper analysis approach for future studies. Our strong recommendations are: firstly, to focus on measurements with a practical value that can be easily related to system reliability; secondly, to select a proper analysis approach that provides reliable results, ideally supported with statistical significance. Overall, we believe that the analysis and guidelines summarized in this paper will help the community to define formal and reliable methods for analysis of the DRAM errors in the field.

ACKNOWLEDGMENTS

This work was supported by the Collaboration Agreement between Samsung Electronics Co., Ltd. and BSC, Spanish Government through Severo Ochoa programme (SEV-2015-0493), by the Spanish Ministry of Science and Technology through TIN2015-65316-P project and by the Generalitat de Catalunya (contracts 2014-SGR-1051 and 2014-SGR-1272). This work has also received funding from the European Union's Horizon 2020 research and innovation programme under EuroEXA project (grant agreement № 754337). Darko Zivanovic holds the Severo Ochoa grant (SVP-2014-068501) of the Ministry of Economy and Competitiveness of Spain.

REFERENCES

[1] Algirdas Avizienis, Jean-Claude Laprie, Brian Randell, and Carl Landwehr. 2004. Basic Concepts and Taxonomy of Dependable and Secure Computing. *IEEE Transactions on Dependable and Secure Computing* 1, 1 (jan 2004), 11–33.

[2] Barcelona Supercomputing Center. 2016. *MareNostrum 3 User's Guide.*

[3] Timothy J. Dell. 1997. *A White Paper on the Benefits of Chipkill-Correct ECC for PC Server Main Memory.* Technical white paper 4AA4-3490ENW. IBM.

[4] Ioana Giurgiu, Jacint Szabo, Dorothea Wiesmann, and John Bird. 2017. Predicting DRAM Reliability in the Field with Machine Learning. In *Proceedings of the 18th ACM/IFIP/USENIX Middleware Conference: Industrial Track.* 15–21.

[5] K.-I. Goh and A.-L. Barabási. 2008. Burstiness and memory in complex systems. *EPL (Europhysics Letters)* 81, 4 (Jan 2008). https://doi.org/10.1209/0295-5075/81/48002

[6] Saurabh Gupta, Tirthak Patel, Christian Engelmann, and Devesh Tiwari. 2017. Failures in Large Scale Systems: Long-term Measurement, Analysis, and Implications. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis.* 44:1–44:12.

[7] Hewlett Packard Enterprise 2016. *HPE ProLiant DL580 Gen9 Server User Guide.* Hewlett Packard Enterprise.

[8] HP. 2016. *How memory RAS technologies can enhance the uptime of HPE ProLiant servers.* Technical white paper 4AA4-3490ENW. Hewlett Packard Enterprise.

[9] Andy A. Hwang, Ioan A. Stefanovici, and Bianca Schroeder. 2012. 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).*

[10] IBM 2014. *System x iDataPlex dx360 M4 Types 7912 and 7913: Problem Determination and Service Guide.* IBM.

[11] Intel Server Products and Solutions 2017. *System Event Log (SEL) Troubleshooting Guide.* Intel Server Products and Solutions.

[12] Bruce Jacob, Spencer W. NG, and David T. Wang. 2008. *Memory Systems: Cache, DRAM, Disk.* Morgan Kaufmann.

[13] Andy Kleen. 2010. MCELOG: Memory Error Handling in User Space. In *International Linux System Technology Conference (Linux Kongress).*

[14] Scott Levy, Kurt B. Ferreira, Nathan DeBardeleben, Taniya Siddiqua, Vilas Sridharan, and Elisabeth Baseman. 2018. Lessons Learned from Memory Errors Observed over the Lifetime of Cielo. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis (SC '18).*

[15] Xin Li, Michael C. Huang, Kai Shen, and Lingkun Chu. 2010. A Realistic Evaluation of Memory Hardware Errors and Software System Susceptibility. In *Proc. of the USENIX Conference on USENIX Annual Technical Conference (USENIXATC).* 6–6.

[16] Catello Di Martino, Zbigniew Kalbarczyk, Ravishankar K. Iyer, Fabio Baccanico, Joseph Fullop, and William Kramer. 2014. Lessons Learned from the Analysis of System Failures at Petascale: The Case of Blue Waters. In *Proc. of the IEEE/IFIP International Conference on Dependable Systems and Networks (DSN).* 610–621.

[17] Justin Meza, Qiang Wu, Sanjeev Kumar, and Onur Mutlu. 2015. Revisiting Memory Errors in Large-Scale Production Data Centers: Analysis and Modeling of New Trends from the Field. In *Proc. of the IEEE/IFIP International Conference on Dependable Systems and Networks (DSN).* 415–426.

[18] B. Nie, D. Tiwari, S. Gupta, E. Smirni, and J. H. Rogers. 2016. A large-scale study of soft-errors on GPUs in the field. In *2016 IEEE International Symposium on High Performance Computer Architecture (HPCA).*

[19] B. Nie, J. Xue, S. Gupta, T. Patel, C. Engelmann, E. Smirni, and D. Tiwari. 2018. Machine Learning Models for GPU Error Prediction in a Large Scale HPC System. In *2018 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN).*

[20] PRACE. 2019. PRACE Research Infrastructure. http://www.prace-ri.eu.

[21] B. Schroeder and G. Gibson. 2010. A Large-Scale Study of Failures in High-Performance Computing Systems. *IEEE Transactions on Dependable and Secure Computing* 7, 4 (Oct 2010), 337–350.

[22] Bianca Schroeder, Eduardo Pinheiro, and Wolf-Dietrich Weber. 2009. DRAM Errors in the Wild: A Large-scale Field Study. In *Proc. of the International Joint Conference on Measurement and Modeling of Computer Systems (SIGMETRICS).* 193–204.

[23] Taniya Siddiqua, Athanasios Papathanasiou, Arijit Biswas, , and Sudhanva Gurumurthi. 2013. Analysis of Memory Errors from Large-Scale Field Data Collection. In *IEEE Workshop on Silicon Errors in Logic - System Effects (SELSE).*

[24] The Royal Society. 1895. *Proceedings of the Royal Society of London.* Number v. 58. Taylor & Francis.

[25] Vilas Sridharan, Nathan DeBardeleben, Sean Blanchard, Kurt B. Ferreira, Jon Stearley, John Shalf, and Sudhanva Gurumurthi. 2015. Memory Errors in Modern Systems: The Good, The Bad, and The Ugly. In *Proc. of the International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS).* 297–310.

[26] Vilas Sridharan and Dean Liberty. 2012. A Study of DRAM Failures in the Field. In *Proc. of the International Conference on High Performance Computing, Networking, Storage and Analysis (SC).* Article 76, 76:1–76:11 pages.

[27] Vilas Sridharan, Jon Stearley, Nathan DeBardeleben, Sean Blanchard, and Sudhanva Gurumurthi. 2013. Feng Shui of Supercomputer Memory: Positional Effects in DRAM and SRAM Faults. In *Proc. of the International Conference on High Performance Computing, Networking, Storage and Analysis (SC).* Article 22, 22:1–22:11 pages.

[28] D. Tang, P. Carruthers, Z. Totari, and M. W. Shapiro. 2006. Assessment of the Effect of Memory Page Retirement on System RAS Against Hardware Faults. In *International Conference on Dependable Systems and Networks (DSN'06).*

[29] D. Tiwari, S. Gupta, J. Rogers, D. Maxwell, P. Rech, S. Vazhkudai, D. Oliveira, D. Londo, N. DeBardeleben, P. Navaux, L. Carro, and A. Bland. 2015. Understanding GPU errors on large-scale HPC systems and the implications for system design and operation. In *2015 IEEE 21st International Symposium on High Performance Computer Architecture (HPCA).*

[30] D. Watts, R. Doughty, and I. Solovyev. 2018. *Lenovo System x3850 X6 and x3950 X6 Planning and Implementation Guide.* Lenovo Press.