Removing Obstacles before Breaking Through the Memory Wall: A Close Look at HBM Errors in the Field

Ronglong Wu, Shuyue Zhou, Jiahao Lu, Zhirong Shen, and Zikang Xu, *Xiamen Key Laboratory of Intelligent Storage and Computing, Xiamen University;* Jiwu Shu, *Xiamen Key Laboratory of Intelligent Storage and Computing, Xiamen University, and Minjiang University;* Kunlin Yang and Feilong Lin, *Huawei Technologies Co., Ltd;* Yiming Zhang, *Xiamen Key Laboratory of Intelligent Storage and Computing, Xiamen University*

https://www.usenix.org/conference/atc24/presentation/wu-ronglong

This paper is included in the Proceedings of the 2024 USENIX Annual Technical Conference.

July 10–12, 2024 • Santa Clara, CA, USA

978-1-939133-41-0

Open access to the Proceedings of the 2024 USENIX Annual Technical Conference is sponsored by

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

Removing Obstacles before Breaking Through the Memory Wall: A Close Look at HBM Errors in the Field

Ronglong Wu[1], *Shuyue Zhou*[1], *Jiahao Lu*[1], *Zhirong Shen*[1]*, *Zikang Xu*[1], *Jiwu Shu*[1,3], *Kunlin Yang*[2], *Feilong Lin*[2], *and Yiming Zhang*[1]
[1]*Xiamen Key Laboratory of Intelligent Storage and Computing, Xiamen University*
[2]*Huawei Technologies Co., Ltd* [3]*Minjiang University*

Abstract

High-bandwidth memory (HBM) is regarded as a promising technology for fundamentally overcoming the memory wall. It stacks up multiple DRAM dies vertically to dramatically improve the memory access bandwidth. However, this architecture also comes with more severe reliability issues, since HBM not only inherits error patterns of the conventional DRAM, but also introduces new error causes.

In this paper, we conduct the first systematical study on HBM errors, which cover over 460 million error events collected from nineteen data centers and span over two years of deployment under a variety of services. Through error analyses and methodology validations, we confirm that the HBM exhibits different error patterns from conventional DRAM, in terms of spatial locality, temporal correlation, and sensor metrics which make empirical prediction models for DRAM error prediction ineffective for HBM. We design and implement Calchas, a hierarchical failure prediction framework for HBM based on our findings, which integrate spatial, temporal, and sensor information from various device levels to predict upcoming failures. The results demonstrate the feasibility of failure prediction across hierarchical levels.

1 Introduction

The performance gap between the computing power and the memory bandwidth is continuously widening in modern computing systems (a.k.a., memory wall), which becomes one of the major obstacles in training ever-larger machine learning models. Extensive efforts have been made to mitigate the impact of the memory wall, including leveraging access locality in data prefetching [16, 18, 46, 67, 73], exploring vectorization [38, 45], exploiting data compression [10, 11, 21, 70], and designing new system architectures (e.g., processing-in-memory architecture [42, 53, 77]).

Recently, *high bandwidth memory* (HBM) receives tremendous attentions and is considered as a promising technology to fundamentally overcome the memory bottleneck. For instance, the first 8-high 24 GB HBM3 Gen2 memory released by Micron can achieve the bandwidth greater than 1.2 TB/s [2], which reduces more than 30% of the training time for large language models and also the total cost of ownership. The rationale behind HBM is to stack up multiple planar DRAM chips vertically through *through-silicon vias* (TSVs) and microbumps [53, 54, 79]. HBM further provides several independent interfaces (called *pseudo channels*) to support parallel data access to different sets of banks, hence achieving high aggregated access bandwidth.

While being fast and power-efficient, HBM is more vulnerable to unexpected errors due to the following reasons. First, as HBM is constructed by stacking DRAM dies, it not only inherits the error characteristics of conventional DRAM, but also poses new error causes, such as more severe soft errors under the higher bit density [52] and TSV faults (including data TSV faults, command TSV faults, and power TSV faults [12, 55, 68]). Second, HBM usually equips weaker error correction codes than conventional DRAM due to cost and complexity considerations [60]. Therefore, understanding the error characteristics of HBM beforehand becomes extremely vital when designing techniques to guarantee the reliability of the stored data.

Extensive in-depth analyses of DRAM errors have already been conducted in recent years, which mainly study the error characteristics [13, 15, 22, 71, 82], identify the root causes of errors [23, 41, 48, 51, 61], and further attempt to predict the upcoming failures [14, 17, 25, 29, 31, 58, 87, 89]. However, after careful analyses and validations, we uncover that most of the findings and implications made from DRAM error analyses cannot be directly applied to HBM, since HBM has a more complex 3D-stacked architecture (e.g., system-in-package fabrication [74] and TSVs [43, 52]) that finally results in new fault modes and error patterns.

In this paper, we fill this blank by performing an in-depth data-driven analysis to study the error characteristics of HBM.

*Corresponding author: Zhirong Shen (shenzr@xmu.edu.cn)

We look into a large-scale dataset [1] collected from nineteen data centers over two years, which comprises error logs (e.g., *correctable errors* (CE) and *uncorrectable errors* (UE)), temperature logs (e.g., the temperature of DSA devices), and power logs (e.g., the average power). We perform a series of analysis, in terms of the spatial analysis (e.g., spatial locality and structure analysis), temporal analysis (e.g., CE storm), power analysis (e.g., power trends of errors), and temperature analysis (e.g., temperature distribution analysis), which deliver nine findings in total. Based on the above analysis, we perform two unsuccessful attempts, which try to use the CE rate (defined as the number of CEs occurring within a monitored time interval) and historical CE information to predict the upcoming UERs. We finally design Calchas, which is a hierarchical, comprehensive, and non-intrusive failure prediction framework for HBM. Based on our analysis results, Calchas integrate information at different levels while considering the impact of various errors to predict upcoming failures.

More specifically, we make the following contributions:

(1) In-depth analyses for HBM errors collected from production clusters. We collect over 460 million error events in total from nineteen data centers with the time duration of over two years. We also collect auxiliary information, including temperature logs and power logs, to facilitate the analysis. We first perform the spatial analysis to study whether the emerging errors of HBM still follow the spatial locality and investigate the impact of the 3D-stacked architecture on the errors (§3.2). We then carry out temporal analysis to study the intervals between CEs and UEs to seek the possibility of using historical CEs to predict upcoming UEs (§3.3). We finally study the impact of temperature and power on the error occurrence (§3.4).

To the best of our knowledge, this paper is the first effort that systematically analyzes the error characteristics of HBM. Moreover, the dataset we study contains over 460 million error events, which is orders of magnitude more than the largest datasets of DRAM errors that have been released till now (e.g., the dataset with 75.1 million CEs collected from production data centers at Alibaba [22]).

(2) Lessons learned from unsuccessful attempts. We have tried two unsuccessful attempts to build prediction models based on empirical approaches in DRAM error prediction [22, 31]. We learn the following lessons from these attempts. First, the effectiveness of previous CE rate indicator [29, 65] in predicting UEs is questionable for HBM, as there is no clear correlation between the CE rate and the occurrence of UERs (§4.1). Second, the CE-based predictor trained by historical CEs is also impractical to predict the future occurrence of UERs, as most UERs tend to coincide with other UERs rather than with CEs (§4.2).

(3) A hierarchical failure prediction framework for HBM. With the lessons and insights gained from previous unsuccessful attempts, we finally design and implement Calchas, a hierarchical failure prediction framework based on random forest [19] for HBM. It is the first prediction model for predicting failures in HBM. It combines features specific to HBM (e.g., stack features) with those used in traditional DRAM (e.g., component features) for prediction (§5.1). Moreover, Calchas employs different prediction timing (i.e., period-based approach for server-level prediction and event-driven approach for micro-level prediction) to realize the elastic and adaptive failure prediction (§5.3).

[1] The dataset is published at: https://github.com/wrl297/Calchas.

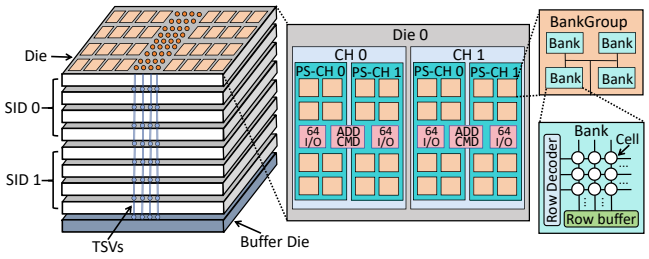


Figure 1: The HBM architecture with eight DRAM dies (§2.1).

2 Background

2.1 High Bandwidth Memory

High bandwidth memory (HBM) has a specific 3D-stacked structure that sets it apart from traditional planar memory architectures. It offers significantly higher data transfer rates compared to conventional DRAM. For instance, HBM2 (the second-generation of HBM) can achieve a bandwidth of around 256 GB/s per stack [24, 44], which is far larger than the bandwidth in DDR4 (around 25 GB/s per channel [59]).

HBM can be constructed via either a 4Hi stack (with four DRAM dies) or an 8Hi stack (with eight DRAM dies). Under the 8Hi configuration, every four dies can be packed to form an SID. Figure 1 illustrates the architecture of HBM with an 8Hi stack, which comprises two SIDs with eight dies. A DRAM die consists a number of *channels* (CH). In the pseudo-channel mode, a channel can be further divided into two *pseudo-channels* (PS-CH), which is composed of several *bank groups* (BG). A BG is consisted of four banks, each of which comprises multiple rows and columns.

Fundamentally, HBM utilizes *through-silicon vias* (TSVs) and microbumps to connect multiple layers of DRAM cells, which are then combined with a logic layer and a memory controller. At the bottom of the stack, there is a *buffer die* (or logic die) that serves as the control and communication hub for the entire HBM stack. The communication and data transfers among the stacked DRAM dies are then realized by the TSVs, which are vertical electrical connections that run through the silicon substrate, providing high-speed data pathways. While providing higher access bandwidth, HBM is limited in capacity (e.g., 5–10× smaller than DRAM [66])

due to cost and power considerations [53, 83].

In our platform, we pack multiple HBM chips within a single DSA (Domain Specific Architecture) device (e.g., GPU [4, 6, 7], TPU [49], and NPU [62]), and multiple DSA devices finally make up a server.

2.2 Terminology

To facilitate comprehension of the analyses conducted and approaches proposed in this paper, we provide a list of frequently used terminologies below.

Error and fault: We call an error occurs if finding that the data read from HBM is inconsistent with the originally stored data (usually detected by ECC [47]). A fault refers to the fundamental cause of an error, which is often induced by hardware faults (e.g., memory wear-out [76]) and external factors (e.g., cosmic rays [41]). Note that a fault can be *active* (causing errors) or *dormant* (not causing errors) [81]. For example, writing a bit '1' to a cell with a stuck-at-1 fault (i.e., a hardware defect where a bit is permanently set to '1') will not introduce an error when reading the stored bit from this cell.

Failure: Failure refers to the inability for HBM chips to support normal read and write operations, resulting in service unavailability. For example, DRAM failure is one of the major causes of server crashes [58].

ECC: *Error correcting code* (ECC) encodes data to generate additional parity bits, so that errors can be identified and corrected. ECCs have been extensively used in DRAM to resist bit errors, including SEC-DED [61], Chipkill [26], and SDDC [20]. It is reported that the Intel Stratix 10 HBM2 Controller supports SEC-DED, with 64-bits of data and 8-bits of ECC code [3].

CE and UE: We classify the errors into *correctable errors* (CE) and *uncorrectable errors* (UE) based on the number of bit errors and the correction capability offered by ECC. CE refers to the errors within the correction capability of ECC and hence they can be successfully restored, while UE refers to those that exceed the correction capability of ECC. We then take a step further to categorize UE into another two branches based on the necessity to take actions for addressing the appearing errors: (i) UER (**U**ncorrectable **E**rror action **R**equired), which implies that the UE occurs during the system's runtime and immediate actions (e.g., replacement of DRAM DIMM, row remapping, and dynamic page offlining) must be taken to prevent server crashes and service interruptions; and (ii) UEO (**U**ncorrectable **E**rror action **O**ptional), which is usually discovered by the periodic memory scrubbing over HBM chips and does not affect the system runtime. In particular, the UEO can be addressed by taking pages offline.

2.3 Data Collection

Dataset description: We collect the error information of HBM2 chips from nineteen data centers, which comprise over 60,000 HBM2 chips across approximately 15,000 DSA devices. The monitoring duration spans over two years (from 2021-03-22 to 2023-06-21). The data centers run a variety of workloads, including autonomous driving, structural biology analysis, high-performance computing, and public cloud services. We elaborate on the data collection in details as follows.

Collection methodologies: The HBM status information is periodically collected by the *baseboard management controller* (BMC), which is a specialized processor embedded on the motherboard of a server and is responsible for recording the operational statistics of various hardware components [35]. The BMC logs are finally collected by the on-site engineers once a day.

Two approaches are employed to collect the error information with different preferences and hence we have two error logs, namely `ErrLog_Cycle` and `ErrLog_Occurrence`, respectively. Specifically, each CE entry in `ErrLog_Cycle` contains the physical address where the CE occurred, along with the accumulated times of CE occurrences at this address within the monitoring cycle. However, `ErrLog_Cycle` does not record the explicit timestamp when the CE occurred, but gives priority to the timestamp to collect the error and the error count. Hence, even if an address experiences multiple CEs in the monitoring cycle, it will be recorded as a single entry in the `ErrLog_Cycle`. On the other hand, each entry in `ErrLog_Occurrence` records a CE event captured by the BMC, which contains the explicit timestamp when the CE occurs and the corresponding physical address. For each UE, both `ErrLog_Cycle` and `ErrLog_Occurrence` record the occurrence timestamp and the physical address. However, due to the limited storage capacity of BMC, recording error entry in `ErrLog_Occurrence` may potentially be overwritten by subsequent errors, yielding error information loss, while `ErrLog_Cycle` does not miss any CE occurrences.

Sensor information: In addition to the error logs collected by BMC, we also collect the following sensor information to facilitate the identification of error root causes: (i) the temperature log, which collects the temperature every ten minutes using the temperature sensors embedded in the DSA devices; and (ii) the power log, which is collected by the power sensors embedded in the server and records the power information of the server every 10 minutes, including the peak power and the average power over the past ten minutes, as well as the transient power at the time of collection.

3 Analysis

3.1 Dataset Overview

We first provide an overview about the spatial distributions of different error types (i.e., CE, UEO, and UER) during the monitoring period. The results presented in Table 1 deliver the following implications.

Total number of different errors			
	CE	**UEO**	**UER**
Error Count	466,236,831	49,297	3,334
Number of components with errors			
	CE	**UEO**	**UER**
HBM	808	64	229
SID	817	66	231
PS-CH	847	69	233
BG	909	76	354
Bank	993	87	555
Column	2,001	258	923
Row	8,279	35,543	2,140
Cell	10,817	37,010	2,192

Table 1: Dataset overview (§3.1). The total number of errors across different device levels does not equate to the total count of device levels with errors, as multiple errors may occur within the same device level.

First, we observe over 460 million CEs in HBM chips during the two-year monitoring duration, whose number is several times higher than that reported in existing studies (e.g., 75.1 million in [22]). Additionally, it is worth noting that the total number of UEOs (i.e., 49,297) is significantly higher than that of UERs (i.e., 3,334). The reason is that UEOs are observed during periodic memory scrubbing. Hence, if a component has a fault, its error information may be repeatedly collected for multiple times during the memory scrubbing.

Second, for each device level (e.g., HBM, SID, and PS-CH), the number of components exhibiting different types of errors varies. Specifically, the number of components experiencing CEs is significantly higher than those experiencing UEOs or UERs. For example, 808 HBM chips have encountered CEs, while only 64 HBM chips have experienced UEOs and 229 HBM chips have exhibited UERs. However, we observe a different distribution for row and cell levels. This is due to column failures (e.g., TSV failures [68]) being prevalent in HBM (see Finding 5 in §3.2), and when a bank experiences a column failure, performing memory scrubbing will uncover numerous cells or rows with UEOs.

3.2 Spatial Analysis

Finding 1. *If an error occurs in one cell of a device level, there is a high probability of experiencing subsequent errors in another cell.*

Spatial locality. We start our analysis by calculating the percentage of errors in multiple cells across different device levels. For each device level, we examine the proportions of errors that occur across multiple cells and within a single cell, respectively. Figure 2 illustrates that the majority of device levels experiencing errors are found across multiple cells. Specifically, only 12.0% of device levels experience errors within a single cell on average. We also observe that 86.1% of errors in a row happen within a single cell, indicating that a single cell in a row — once emerges an error — will be likely to repeatedly experience subsequent errors. This observation

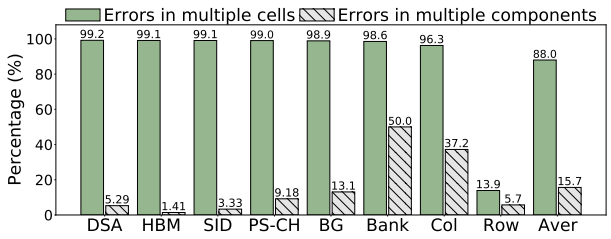


Figure 2: Spatial locality (§3.2). This figure shows percentage of cells and components across different device levels.

does not hold for other device levels, since we find that only 1.40% of errors in other device levels appear within a single cell. This analysis indicates that there exists a strong correlation between the error mode and the device levels for HBM, which can be leveraged to guide the error prediction in §5.

Finding 2. *While errors are common across multiple cells in most device levels, only the bank exhibits errors in multiple components.*

We further investigate whether the errors are correlated among the components within the same device level. Given a device level, we measure the percentages of errors occurring in multiple components. For example, consider a DSA device with four HBM chips, we measure the proportion of DSA device that experience errors in multiple HBM chips. Figure 2 also shows that for most levels (e.g., DSA, HBM, and SID), they usually have only a component containing errors. This analysis implies that even though the errors are found in multiple cells, they usually manifest within a single component rather than spreading across multiple components. However, we observe a distinctive phenomenon at the bank level, where 50.0% of banks experience errors in multiple components (i.e., rows or columns). As a memory cell is organized into rows and columns, errors occurring in multiple rows or columns will inherently involve multiple cells. This is why we observe errors occurring in multiple cells for most levels, but they are typically confined to a single component.

Hierarchical analysis. For HBM, we are interested in whether the 3D-stacked structure makes the errors vary across different components within the same device level. Here, we analyze the number of components experiencing errors at various device levels.

Finding 3. *The effects of crosstalk in the HBM may result in data loss, specifically in the 7-th, 15-th, 23-rd, and 31-st columns of a bank.*

Figure 3(a) shows the number of components with errors in various positions. Two key observations are made. First, for the columns with errors, those occurring in the 7-th, 15-th, 23-rd, and 31-st columns (marked in dashed boxes) are 338.4% higher on average than the errors in other column positions. After communicating with our device engineers, we suspect the reason is that crosstalk-induced faults [78] make errors more prevalent in specific columns. Moreover, we observe that, for the SIDs whose positions are closer to the buffer die

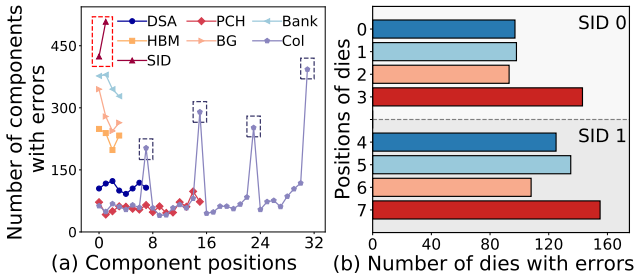


Figure 3: Impact of HBM structure. (§3.2). The figures depict (a) the distribution of components with errors across various positions and (b) the number of dies with errors in different positions.

(i.e., SID_1), exhibit a 20.0% higher susceptibility to errors compared to SID_0 (i.e., higher SID). We conjecture that the observed difference in error occurrence is due to poor heat dissipation in the lower SID (i.e., SID_1).

Finding 4. *Lower SIDs (i.e., SID_1) exhibit a higher susceptibility to errors.*

In Figure 3(b), we count the number of dies with errors in different positions to further demonstrate the impact of the 3D-stacked structure. When examining two dies located at the same positions but with distinct SIDs, such as Die_0 and Die_4, there is a higher probability of errors occurring in the die associated with the lower SID compared to the die linked to the higher SID. For example, errors in Die_5 and Die_6 are 28.9% and 37.8% more prevalent than those in Die_1 and Die_2, respectively. These findings affirm that the lower SIDs in HBM are more susceptible to error occurrence.

Error modes. Our analysis reveals that multiple errors usually occur simultaneously within the same bank (see Finding 2). We further study the error modes at the bank level. We find that the errors emerged in the 1,430 banks exhibit the following nine modes. Figure 4 also plots nine typical error modes for easy comprehension.

- Single-cell mode: Errors only occur in one cell of an HBM bank (Figure 4(a)).
- Two-cell mode: Errors only occur in two cells of an HBM bank, where the two cells are neither in the same column nor the same row (Figure 4(b)).
- Single-row mode: Errors occur in multiple cells within a single row (Figure 4(c)).
- Two-row mode: Errors occur in two rows, each of which contains multiple cells with errors (Figure 4(d)).
- Single-column mode: Errors are found in multiple cells within a single column (Figure 4(e)).
- Two-column mode: Errors are concentrated in two columns, each of which has multiple cells with errors (Figure 4(f)).
- Row-dominant mode: Errors manifest across various rows, with over 80% concentrated in specific ones (Figure 4(g)).
- Column-dominant mode: It is the dual mode of the row-dominant mode. In this mode, errors are identified across several columns, with over 80% of them concentrated in specific columns (Figure 4(h)).
- Irregular mode: It contains the modes that deviate from the aforementioned eight categories (Figure 4(i)).

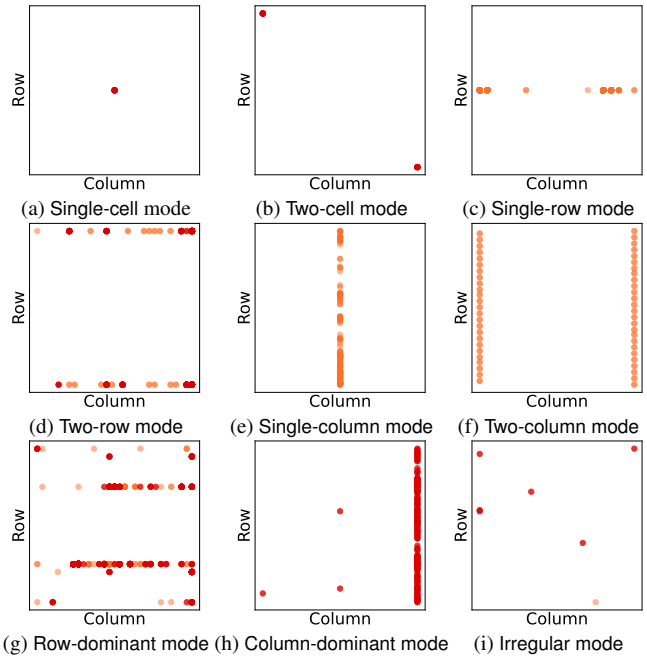


Figure 4: Examples of error modes (§3.2). The figures show (a) single-cell, (b) two-cell, (c) single-row, (d) two-row, (e) single-column, (f) two-column, (g) row-dominant, (h) column-dominant, and (i) irregular error modes. Markers in orange (light) represent the occurrence of CEs, while the markers in red (dark) denote the occurrence of UEs.

Table 2 further gives the percentages of the nine error modes under different combinations of HBM errors, which delivers the following finding.

Finding 5. *Column-related error modes are more commonly observed in HBM, as opposed to the prevalent row-dominated error modes in conventional DRAM [15, 81].*

We find that 50.0% of faulty banks have errors in just one cell. Furthermore, 29.2% of faulty banks show the column-related modes (including the single-column, two-column, and column-dominant modes), while only 4.97% exhibit the row-related modes (including single-row, two-row, and row-dominant modes), implying that column-related error modes occur more frequently than row-related ones, deviating from the observations reported in prior analyses of DRAM errors [15, 81]. We suspect the root cause is TSV failures in HBM, which may manifest as column failures [68]. In addition, the sense amplifier failure may also exhibit the symptom of three error modes that closely relate to the column failures.

Besides, we also observe that the distribution of error modes is different when considering a single error type and combinations of error types. In particular, when considering a single error type within a bank, we find that column-related

Error Mode	Error Types						
	All of the Error Types	Single Error Type			Combinations of Error Types		
		CE	UER	UEO	CE&UER	UEO&UER	CE&UEO&UER
Single-cell mode	50.0%	55.9%	47.6%	8.05%	0%	2.27%	0%
Two-cell mode	4.69%	3.62%	11.3%	2.30%	9.38%	2.27%	0%
Single-row mode	0.629%	0.504%	0.180%	0%	2.50%	4.55%	0%
Two-row mode	2.94%	2.72%	0.721%	4.60%	20.6%	25.0%	50.0%
Single-column mode	23.7%	27.2%	16.2%	29.9%	10.0%	18.2%	12.5%
Two-column mode	2.65%	0.201%	1.80%	25.3%	4.38%	15.9%	3.13%
Row-dominant mode	1.40%	1.71%	0.721%	8.05%	10.0%	22.7%	21.9%
Column-dominant mode	2.87%	1.61%	3.96%	14.9%	6.88%	6.82%	9.38%
Irregular mode	11.1%	6.55%	17.5%	6.90%	36.3%	2.27%	3.13%

Table 2: Percentage of different error modes. (§3.2). The table displays the percentage of the nine error modes across various error types. **All of the Error Types** gives the percentage of nine error modes for all banks experiencing errors (falling into one of the three error types).

error modes (40.4% on average) appear more frequently than the row-related error modes (averaging 6.34%). However, if we analyze the banks with a combination of error types, we notice that the row-related fault modes exhibit a 23.3% higher probability. Motivated by this observation, we perform predictions for upcoming UERs at various micro-level components (e.g., rows and columns) in §5.

3.3 Temporal Analysis

Although the CE dominates the errors that are captured in most component levels (see Table 1), the UER is the root cause to trigger HBM failures. In this analysis, we pay close attention to the properties of UERs. We first study the temporal correlation among UERs at the bank level. Our objective is to leverage the temporal correlation to guide the prediction designs (see §5), such that memory systems can take timely actions to cope with the upcoming UERs once receiving a UER alert. After that, we also investigate the temporal correlation among UERs and CEs, which is expected to figure out a way for the prediction of UERs based on historical CEs/UEs.

Bank with UERs: We first provide a statistical overview of the banks used in the analysis. Out of the 1,430 banks experiencing errors in total, about 91.0% (1,302 banks) have valid error events for analysis [2], where 525 banks have ever encountered UER(s). Among the 525 banks, we further identify that 219 banks only experience a single UER throughout the monitored time interval and hence are unqualified for the analysis. As a result, we conduct the correlation analysis over the remaining 306 banks, with 202 banks that contain UER(s) only and another 104 banks that emerge combinations of UER(s) and other error types.

Time between errors: After picking out the banks qualified for analysis, we start with the analysis on the time between failures. Recall that there are three error types (i.e., CE, UEO, and UER), where the occurrence time of the UEO is not accurately recorded, since it is captured by the periodic memory scrubbing over HBM chips (i.e., its occurrence time

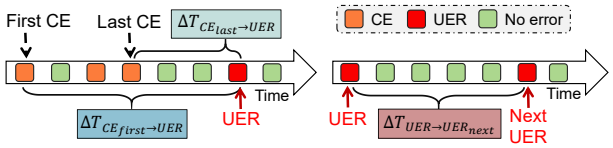


Figure 5: Example of time between errors (§3.3). The figure illustrates the three intervals, including the time between the first/last CE and its subsequent UERs (i.e., $\triangle T_{CE_{\text{first}} \to UER}$ and $\triangle T_{CE_{\text{last}} \to UER}$), as well as the period between two successive UERs(i.e., $\triangle T_{UER} \to T_{UER_{\text{next}}}$).

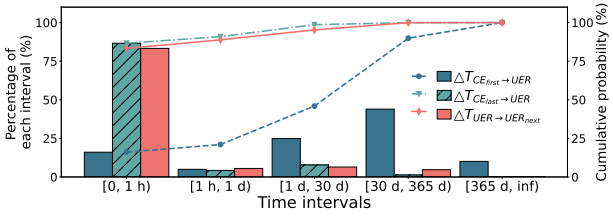


Figure 6: Time between errors (§3.3). The figure illustrates the distribution and cumulative probability for each interval. The bars represent the distribution of $\triangle T_{CE_{\text{first}} \to UER}$, $\triangle T_{CE_{\text{last}} \to UER}$, and $\triangle T_{UER} \to T_{UER_{\text{next}}}$, while the lines depict their cumulative probability.

should be earlier than the recorded time). Hence, our analyses abandon the UEO information and aims to find the answers for the following questions: (i) how soon will the first CE [3] evolve to the first UER within a bank (i.e., $\triangle T_{CE_{\text{first}} \to UER}$ in Figure 5); (ii) how much time remains for the system administrator to prevent system crashes when the last CE occurs (i.e., $\triangle T_{CE_{\text{last}} \to UER}$ in Figure 5)? and (iii) how long will the next UER occur once encountering a UER (i.e., $\triangle T_{UER \to UER_{\text{next}}}$ in Figure 5)?

Finding 6. *There is a significant probability that two successive UERs occur within one hour.*

Figure 6 answers the above three questions by plotting the cumulative probability of different time intervals (i.e., $\triangle T_{CE_{\text{first}} \to UER}$, $\triangle T_{CE_{\text{last}} \to UER}$, and $\triangle T_{UER \to UER_{\text{next}}}$) between errors. We make the following observations.

[2] We exclude some error events because their timestamps fall outside the monitoring period.

[3] Our analysis treats the first CE appeared in our dataset as the real first CE, but the real first CE may occur prior to the collection of the dataset.

Threshold	Servers with CE storms	Servers with no CE storm
10	26.47%	15.85%
20	21.73%	16.49%
30	37.50%	18.18%

Table 3: Impact of CE storm (§3.3). The table displays the percentage of servers experiencing UERs for those with CE storms compared to the servers without CE storms. The threshold values for identifying the occurrence of a CE storm are set between 10 and 30.

First, two successive UERs within a bank may appear in one hour with a high probability (i.e., 83.3%). This observation indicates that once a UER is detected, the prediction algorithm should be frequently launched (e.g., performing the prediction within an hour) to infer the newly emerging UERs at the micro-level components (e.g., banks). Additionally, a remapping method (e.g., row-mapping [87]) is necessary to offline the page when a UER is observed and replicate data to a new page.

Second, we find that 16.1% of $\triangle T_{CE_{\text{first}} \rightarrow UER}$ and 86.6% of $\triangle T_{CE_{\text{last}} \rightarrow UER}$ are shorter than one hour. Besides, over 44.0% of $\triangle T_{CE_{\text{first}} \rightarrow UER}$ range from 30 days to 365 days. This phenomenon indicates that banks with both CEs and UERs may have encountered CEs as early as a long time ago (e.g., 365 days), and these banks may continue to experience CEs until the time approaches (e.g., one hour) the UER occurrence. Therefore, predicting UERs based on historical CEs seems like a promising approach. We make this attempt in §4.

Finding 7. *Servers that have experienced CE storms may increase the probability of encountering UERs.*

Impact of CE storm. The CE storm refers to the phenomenon that numerous CEs on a server occur within a short period (typically one minute [29, 89]). In this analysis, we try to infer if the CE storm has an impact on the occurrence of UERs. We first classify the servers into two categories based on the occurrence of CE storms: (i) the servers with CE storms; and (ii) the servers with no CE storm. We then calculate the ratio of the servers that encounter at least an UER for the two categories. We vary the *threshold value* (defined as the number of CEs suddenly occurred in one minute) to identify the occurrence of a CE storm from 10 to 30 [9, 28, 29], and show the results in Table 3. We can discover that the servers with CE storms have a higher probability (11.7% on average) to produce UERs than the servers with no CE storm, indicating that the occurrence of CE storms can be leveraged for UER prediction.

In addition, we also find that configuring the threshold value to 30 can help identify the servers with UERs with the largest probability (37.50%), yet there is still a challenge to balance the trade-off between the accuracy (in predicting the upcoming UERs) and the sample size (i.e., the number of servers with CE storms) when selecting a threshold value to identify the CE storm. The rationale is that a larger threshold value will reduce the sample size, making the prediction accuracy more sensible to the variation of positive samples (i.e., the servers whose CE storms finally translate to UERs). For example, in our dataset, setting the threshold value to 30 makes the sample size smaller than 10; in this case, any change in the number of positive samples will introduce more than 10% of change to the ratio of the positive samples. Finally, in the dataset, we suggest configuring the threshold value to 10, which can achieve a considerable ratio (i.e., 26.47% in Table 3) with the most sample size.

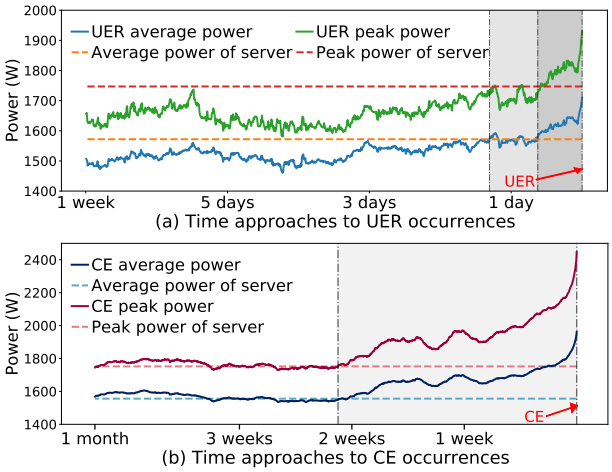


Figure 7: Impact of power (§3.4). The figures show both peak power and average power as the time approaches (a) the occurrence of UERs and (b) the occurrence of CEs for the servers.

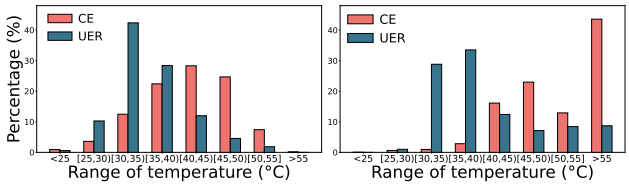
3.4 Analysis of Sensor Information

In addition to the error logs, we also collect auxiliary information from sensors that report the power and temperature information every ten minutes. Prior studies have revealed that sensor information (e.g., temperature [50] and voltage [51]) has an impact on HBM errors and can be a benefit for failure prediction [37], we also analyze the impacts of both temperature and power on the error occurrence.

Impact of power: We first assess the impact of the power on the emergence of UERs. Each entry of the power log contains two pieces of information: (i) the average and the peak power in the last ten minutes; and (ii) the transient power at the time of the collection. Since the latter only reflects the power at a certain time point, we use the average and the peak power in the following analysis.

Finding 8. *The average and peak power of a server both exhibit a rapid increase when approaching the error occurrence.*

To learn the impact of the power, we first pick out the servers that have ever experienced either CEs or UERs. For each error event, we examine a specific time interval—seven days for the UER and thirty days for the CE— preceding the occurrence of the error. As we collect power information every ten minutes, we roughly depict the trends of peak power and average power for both CE events and UE events in Figure 7.



(a) Average temperature (b) Maximum temperature

Figure 8: Impact of temperature (§3.4). The figures illustrate the distribution of (a) average temperature and (b) maximum temperature within one day preceding the occurrence of CEs and UERs, respectively.

Figure 7(a) shows the power trends of servers encountering UERs. Both peak power and average power exhibit a noticeable increase as time approaches the occurrence of UERs. More specifically, we observe that the power starts to rise around the average power of the server 31 hours prior to the UER events and surpasses the average power of the server 15 hours ahead of the occurrence of UERs. As the power may increase as the time approaches the occurrence of UERs, we can utilize this as a feature to predict potential UERs.

We also analyze the power trends of servers experiencing CEs in Figure 7(b). The power exhibits a more significant increase as time approaches the occurrence of CEs compared to UERs. We attribute this to the fact that CEs are typically triggered by soft errors resulting from memory-intensive access (e.g., RowHammer [48, 69]). In contrast, UERs are usually caused by hardware errors (e.g., TSV failures [68]). Given that memory-intensive access tends to increase server power [88], the average power of servers increases as time approaches the occurrence of CEs.

Finding 9. *The temperature distribution of CEs and UERs exhibits significant differences before their occurrence.*

Impact of temperature. We then analyze the influence of temperature on the occurrence of CEs and UERs. We examine the temperature distribution within one day preceding the error events for the DSA devices that have experienced errors. For each DSA device, we calculate both the average temperature and the maximum temperature one day before the occurrence of CEs or UERs. Figure 8 illustrates the temperature distributions for the DSA devices with errors. We observe that the average temperature for DSA devices with CEs is higher than that in DSA devices with UERs. Specifically, 60.57% of the DSA devices with CEs have an average temperature (refer to Figure 8(a)) exceeding 40°C one day before the CE events, while the average temperature of DSA devices with UERs is more prevalent (70.76%) within the range of 30°C to 40°C. Furthermore, DSA devices with CE events are more likely (58.92% higher) to experience a maximum temperature exceeding 40°C compared to the DSA devices with UERs. We speculate that the reason is that higher temperatures increase charge leakage [76], which may lead to data loss but can be restored by the ECC (i.e., recorded as a CE). Hence, we observe a higher temperature as the time approaches the occurrence of CEs.

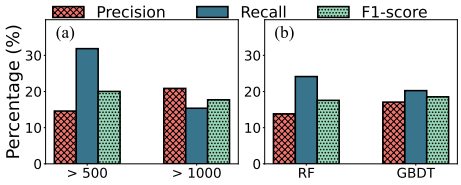


Figure 9: Unsuccessful attempts. (§4). The figures show the precision, recall, and F1-score for (a) CE rate indicator and (b) CE-based predictor, respectively. We evaluate the CE rate indicator using two thresholds (i.e., 500 and 1000) and implement the CE-based predictor with RF and GBDT models.

Summary: We observe that both temperature and power correlate with the occurrence of errors, though their contributions to HBM errors vary. Specifically, frequent access to HBM can lead to increased power consumption, potentially raising the probability of errors. However, power itself does not directly contribute to error occurrence. In contrast, temperature increases directly correlate with error occurrence, as higher temperatures can lead to increased charge leakage [76], potentially resulting in errors.

4 Unsuccessful Attempts

In prior studies [22, 31], historical CEs have been utilized to predict future UERs. Our analysis in §3.3 also indicates the promise of using CEs for predicting potential UERs. As our objective is to precisely predict potential UERs, we employ two methods based on historical CEs in this section.

4.1 Attempt 1: CE Rate Indicator

We first employ a CE rate indicator (also utilized in previous studies [29, 30, 56, 65, 89]), which utilizes the CE rate to indicate whether a UER occurs in the near future. Specifically, the CE rate indicator monitors the number of CEs (denoted as N_{CE}) within the past one day. If N_{CE} exceeds a predefined threshold (e.g., 500 [31]), we consider that a UER may occur in the future. Given that errors are common across multiple components within a bank (Finding 2 in §3.2), we focus the prediction at the bank level.

Limitation. Although we try our best effort to adjust the threshold of the CE rate indicator, its performance is still unsatisfactory (e.g., averaging 20.5% in Figure 9(a)). We suspect that the limited correlation between the CE rate and UERs is the underlying reason, given the significant gap between the total number of CEs (more than 460 million) and the total number of UERs (only 3,334). Therefore, utilizing the CE rate to predict subsequent UERs proves ineffective.

4.2 Attempt 2: CE-based Predictor

The limitation of our first attempt lies in solely considering the number of CEs. To address this, we employ a CE-based

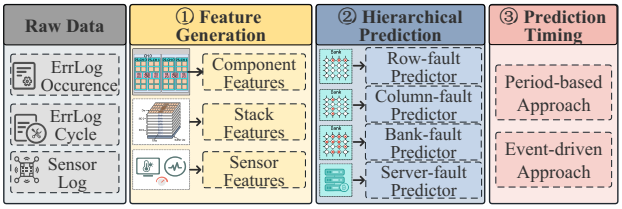


Figure 10: The overview of Calchas (§5). Calchas extracts features from raw data (①) and feeds them into hierarchical predictors (②). Finally, Calchas employs different prediction timings at different levels (③).

predictor that leverages component features related to CEs (including numbers of components with CEs at different device levels) for predicting upcoming UERs. We utilize machine learning algorithms for the CE-based predictor, which is also applied in previous studies [17, 22, 27, 84]. The prediction is made at the bank level. For the predictor, we utilize the all features related to CEs to predict the upcoming UERs. Figure 9(b) presents the results of the CE-based predictor when employing Random Forest (RF) [19] and Gradient Boosting Decision Tree (GBDT) [36] models. On average, the precision, recall, and F1-score of the CE-based predictor are 22.2%, 15.4%, and 18.0%, respectively.

Limitation. The performance of CE-based predictors is disappointing in our dataset, even when considering different models (e.g., RF and GBDT) and various configurations (e.g., prediction windows). The reasons for this are twofold. First, we observe that only a small fraction of banks (11.2%) exhibit a combination of UERs and CEs. Second, the banks experiencing both CEs and UERs are commonly distributed across single or multiple rows. As a result, the CE-based predictor frequently mislabels normal banks. This suggests that we need to leverage additional information (e.g., UERs and UEOs) that has been recorded and consider row-level predictors for banks with multiple errors for failure predictions.

5 Calchas Design

With the insights obtained from previous analyses and attempts, we propose Calchas, a hierarchical, comprehensive, and non-intrusive failure prediction framework for HBM. Calchas can use the row-level, column-level, and bank-level predictors to identify the upcoming UERs for micro-level components in time. It also employs the server-level predictor to periodically predict the potential server-level failures.

Overview of Calchas. Figure 10 illustrates the workflow of Calchas, which includes three successive steps. We first collect the raw error logs from the data centers and generate a sample of features (e.g., the number of cells with UERs) to perform prediction (§5.1). Next, we use the generated features to train four hierarchical predictors (i.e., row-level, column-level, bank-level, and server-level predictors §5.2). We finally select a suitable approach to trigger prediction (§5.3).

5.1 Feature Generation

Before applying prediction models, we first generate representative features for model training. Recall that the collected logs contain the following information: (i) the addresses of errors, (ii) the time of error occurrence, (iii) the error types, and (iv) the sensor logs (see §3.1). Here, we generate 43 features in total from the raw logs and classify them into three categories.

- *Component features*: We first count the number of components experiencing errors (falling into one of three error types) across nine device levels (i.e., cell, row, column, bank, BG, PS-CH, SID, HBM, and DSA). Since different error types within a component may exhibit distinct error modes (Finding 5 in §3.2), we then present statistics for the number of components with CEs, UEOs, and UERs, respectively. We finally assess whether the servers have encountered a CE storm in the past. In total, we consider 37 component features.
- *Stack features*: We have confirmed that the 3D-stacked structure of HBM has an impact on the error emergence, where the SID near the buffer die is more likely to appear errors (Finding 4 in §3.2). Hence, we convert the positions of SIDs into features using one-hot encoding [8] and generate two stack features.
- *Sensor features*: We obtain four sensor features based on the values of maximum temperature, average temperature, peak power, and average power. Given the significant differences in temperature and power before error occurrences (Findings 8 and 9 in Section 3.4), we generate two features from the temperature sensors and another two features from the power sensors.

We generate a collection of features when performing predictions for various predictors. Specifically, we choose component features and sensor features to predict potential UERs at the server level. Furthermore, we turn to use component features, sensor features, and stack features to predict errors at the micro-level components (e.g., rows, columns, and banks).

5.2 Hierarchical Prediction

After generating the features, we start the model training for prediction. Recall that the correlations between the features and the errors may vary at different levels. For example, we observe that errors are common across multiple components within a bank (Finding 2 in §3.2), whereas the CE storms may increase the likelihood of the emergence of UERs in a server (Finding 7 in §3.3). Hence, we establish hierarchical predictors to forecast the upcoming UERs as follows:

- *Row-level predictor*: Given that the row-related error modes are prevalent (52.4%) when multiple error types are observed within the same bank (§3.2), we build a row-level predictor to forecast UERs within a row.
- *Column-level predictor*: We have pointed out that the column-related error mode is more prevalent than the

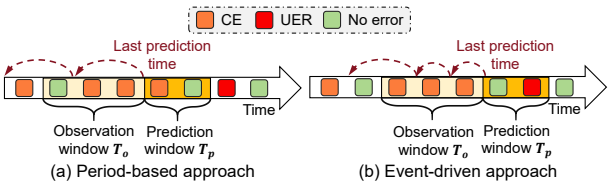


Figure 11: The methods used to trigger predictions (§5.3).

row-related error mode (Finding 5 in §3.2). Hence, we establish the column-level predictor to capture the potential UERs within a column.

- *Bank-level predictor*: We also build a bank-level predictor based on the observation that a bank, which has ever experienced an error before, is more likely to appear errors in the future (Finding 6 in §3.3). This bank-level predictor can work cooperatively with the row-level and column-level predictors to help uncover two-cell and irregular error modes of banks.
- *Server-level predictor*: We finally establish a server-level predictor that leverages the impact of the CE storms (Finding 7 in §3.3) and auxiliary sensor information (Finding 8 and Finding 9 in §3.4), with the aim of capturing the symptom of server failure.

Specifically, if we detect a UER within a *prediction window* (i.e., the time interval for which a prediction is made) at a specific device level, we label the sample as positive; otherwise, it is considered as negative. For each predictor, we initiate the process by generating a set of features and subsequently training a model for prediction.

5.3 Prediction Timing

While hierarchical predictors can be employed to predict potential UERs, a challenge arises: when is the appropriate time to make these predictions? To answer this question, we explore two approaches for triggering predictions as follows.

- *Period-based approach* [22, 87]: It refers to performing prediction periodically. Figure 11(a) shows the period-based approach, which comprises two windows: an *observation window* (represented by $\triangle T_o$), which denotes the duration of the past period under observation for prediction, while a prediction window (represented by $\triangle T_p$), which represents the time interval during which a failure can be predicted. To monitor each time interval until a UER alarm, we set the period of performing prediction the same as the prediction window.
- *Event-driven approach* [86, 87]: It refers to the prediction triggered by error events. As shown in Figure 11(b), the event-driven approach also utilizes the same two windows as the period-based method. Given the potential for frequent errors in HBM (such as CE storms in §3.3), we constrain the minimum interval between two predictions to alleviate the overhead of prediction.

The period-based approach relies on fixed intervals for feature generation and failure prediction, which benefits the capture of abnormal sensor metrics but lacks flexibility in handling irregular events [86]. Conversely, the event-based approach provides a rapid response to errors once they occur. However, it is ineffective when employed for a long-term prediction (e.g., a 16-hour observation window in [86]). We utilize the advantages of both approaches in Calchas. Specifically, we adopt the period-based approach for server-level predictor to monitor the fluctuation of sensor metrics. Meanwhile, we employ the event-driven approach to predict micro-level component failures for leveraging error events in time.

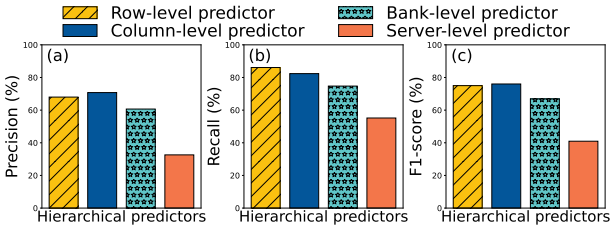


Figure 12: Exp#1 (Performance of Calchas). The figures show the (a) precision, (b) recall, and (c) F1-score of Calchas.

6 Evaluation

6.1 Experimental Setup

To assess effectiveness of Calchas, we split the dataset into a proportion of 7:3, with 70% used for training the model and the remaining 30% for testing [39, 85]. We assess the following metrics.

- *Precision*: The ratio of correctly predicted UER events to the total number of events predicted as UERs.
- *Recall*: The ratio of correctly predicted UER events to the total number of actual UER events.
- *F1-score*: The harmonic mean of precision and recall [34], providing a balanced measure of both metrics.

We first study the performance for hierarchical predictors of Calchas (Exp#1, Figure 12). We then evaluate three machine learning algorithms (Exp#2, Figure 13), including Random Forest (RF) [19], Support Vector Machine (SVM) [64] and Gradient Boosting Decision Tree (GBDT) [36], and eventually choose RF as our default model. These models are implemented in Python (v3.6.8) using the scikit-learn library (v0.24.2) [72].

We employ a one-day observation window and a one-hour prediction window for predicting the potential UERs in micro-level components to leverage the finding that two successive UERs may occur within a short-term period (Finding 6 in §3.3). As error events trigger the micro-level prediction (§5.3), we set a minimum interval to perform prediction as five minutes [22, 87]. For server-level prediction, the observation window and prediction window are set to thirty days and one day, respectively, aiming to better capture the long-term impact of power and temperature (§3.4). Additionally, we explore the

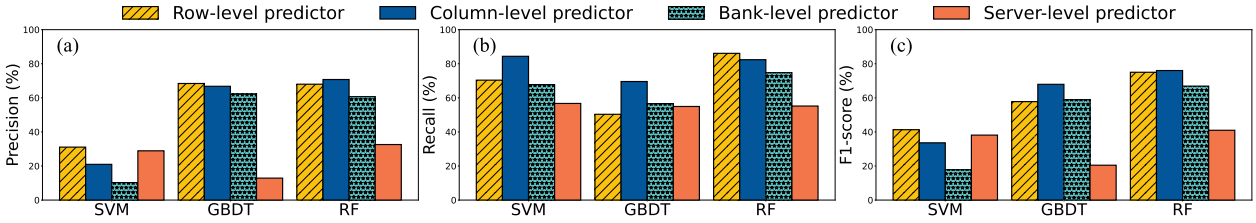


Figure 13: Exp#2 (Impact of prediction models). The figures show the (a) precision, (b) recall, and (c) F1-score of four predictors when using the **SVM**, **GBDT**, and **RF**, respectively.

impact of various sizes (ranging from one hour to thirty days) for observation windows in Exp#3 (Figure 14) and prediction windows in Exp#4 (Figure 15), respectively.

6.2 Evaluation Results

Exp#1 (Performance of Calchas). We first evaluate the performance (including precision, recall, and F1-score) of the four hierarchical predictors of Calchas. Figure 12 illustrates the precision, recall, and F1-score of predictors across different device levels. Specifically, we achieve an average precision of 58.0%, recall of 74.6%, and F1-score of 64.7% across various predictors. In Figure 12(a) and (c), we observe that the column-level predictor (i.e., the second bar) outperforms the other three predictors in the precision (70.8%) and F1-score (76.0%). These high precision and F1-score indicate that column-level predictors rarely mislabel normal columns as failures. The reason is that column-related faults are more prevalent in HBM banks (Finding 5 in §3.2). We can identify the column-related modes by counting the number of cells within the column (a piece of component features) to improve the prediction. Moreover, we notice that the row-level predictor (i.e., the first bar) provides a high recall (82.4%) in Figure 12(b). Given the prevalent use of row isolation methods in deployment systems [5,87], the high recall can facilitate the advanced isolation of the faulty row.

Exp#2 (Impact of prediction models). We then examine the impact of prediction models by implementing three classical machine learning algorithms, including RF [19], SVM [64], and GBDT [36]. Figure 13 shows the results of utilizing different prediction models for each of the four hierarchical predictors. Our observations are as follows. First, while predictors based on SVM achieve high recalls (69.8% on average), their precision falls below 31.1%. Hence, the predictors based on SVM may bring a considerable probability of mislabeling the normal components and servers as failures. Second, the tree-based prediction models (i.e., RF and GBDT) achieve a 25.3% higher F1-score for predicting the upcoming UREs in the micro-level components (i.e., rows, columns, and banks) compared with the SVM. Moreover, we also observe that RF achieves a high precision, recall, and F1-score. These results are aligned with the existing studies in predicting failure in traditional DRAM [17, 22]. Therefore, we select it as the default model for each predictor.

Exp#3 (Impact of the observation window). We further

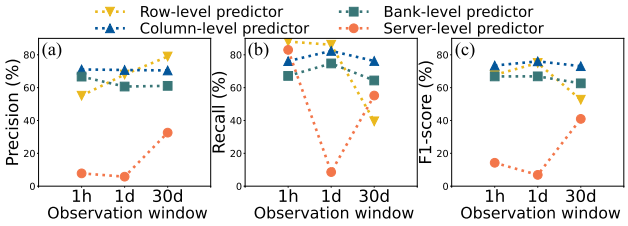


Figure 14: Exp#3 (Impact of the observation window). The figures show the performance of different predictors when the observation window is set to one hour (**1h**), one day(**1d**), and thirty days (**30d**).

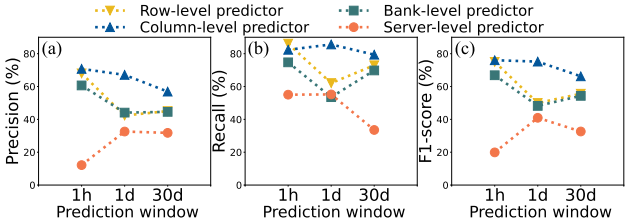


Figure 15: Exp#4 (Impact of the prediction window). The figures show the (a) precision, (b) recall, and (c) F1-score when setting the prediction window to one hour (**1h**), one day(**1d**), and thirty days (**30d**) across different predictors.

study the impact of the observation window by varying its interval from one hour to thirty days. Figure 14 depicts precision, recall, and F1-score for different observation windows across various predictors. We make the following observations. First, our results reveal that the three micro-level predictors (i.e., row, column, and bank) exhibit the highest F1-score when the observation window is set to one day. We employ an observation window interval of one day to achieve a better trade-off between precision and recall. Second, we observe that the performance of the server-level predictor significantly varies with the change of prediction windows. We suspect the reason is that the prediction at a server level requires a long-term monitor to capture the variation of sensor metrics.

Exp#4 (Impact of the prediction window). We finally assess the impact of the prediction window. We vary it from one hour to thirty days for the four predictors. Figure 15 illustrates the performance of hierarchical predictors with different prediction windows. A notable observation is that the performance of both the row-level and column-level predictors deteriorate when the prediction window is set to one day. We suspect the reasons are twofold. First, there is a slight probability (5.53%) of the occurrence of two successive UERs within

the time interval ranging from one hour to one day. Second, the time intervals between CEs and their subsequent UERs are observed to be more prevalent within one hour (51.3%) and more than one day (44.0%). Therefore, compared to a one-day prediction window, the performance improves when adopting a prediction window of one hour or thirty days.

7 Discussion

7.1 Fault Tolerate

We offer an in-depth analysis of errors in HBM and introduce Calchas for predicting the upcoming UERs at different device levels. In this section, we will explore effective strategies to ensure reliability and availability following prediction.

EC-based operator. To enhance the reliability of micro-level components predicted to experience future UERs, we suggest utilizing EC-based operators, which integrate operators with erasure coding [40, 63]. Specifically, if Calchas predicts an upcoming UER in a specific memory region, we continuously monitor this region and employ EC-based operators when writing data to it. This approach allows us to avoid suspending the training task and enables data recovery using the parity of EC when UERs actually occur. We have implemented EC-based operators on servers with CPUs, observing only a 3% performance overhead in the CPU. Given that DSA devices possess stronger parallel computing capabilities [4] and higher bandwidth for data access [44] than CPUs, the overhead of EC-based operators may be even lower than that on CPUs. Therefore, implementing EC-based operators in DSA devices is feasible and has a negligible performance degradation.

Transparent migration. The DSA devices in multiple servers are interconnected using specific techniques (e.g., NVLink [57]). In the event of a server failure, the loss of parameters may occur across all servers. Therefore, it is essential to implement a reliability scheme once potential failures in the servers are predicted. As Calchas achieves a long-term monitor (i.e., thirty days) for prediction at the server level, we have enough time to deal with the upcoming failures. Therefore, we suggest adopting an elastic scaling approach, initiating an upward scale (i.e., increasing the number of servers for training tasks) upon predicting a potential failure. Following the scaling operation, we identify and remove servers that may occur failures. This strategy ensures that the training performance remains unaffected, even in cases where Calchas mislabels a normal server.

7.2 Limitations

Collection challenges. As a service provider, we face challenges in obtaining detailed information about the workload, a factor known to influence failures according to existing studies [32, 37]. To address this gap, we explore the impact on power and identify it as a viable metric for assessing the workload (§3.4). Real-time collection of sensor information for each error event is impractical due to hardware overhead; thus, we collect sensor data collection every ten minutes. Fortunately, our observation (§3.4) reveals that both temperature and power exhibit long-term (e.g., one day in Finding 9) effects on UERs. As a result, collecting sensor information every ten minutes can also be beneficial for predicting upcoming UERs.

Generalizability. The findings are derived from the analysis of HBM2 in our data centers; therefore, their generalizability is limited to HBM standards. Although HBM2 is commonly utilized in commercial DSA devices (e.g., NVIDIA V100 [7], NVIDIA P100 [6], and AMD MI50 [1]), it may not be applicable in cases where HBM2E is considered [4]. We specifically focus on analyzing the impacts of common features (3D-stacked structures and sensor metrics) of different HBM standards on failures. Similar findings may be applicable across various HBM standards.

8 Related Work

HBM failure study. With HBM being extensively employed in commercial DSA devices, both academia and industry have focused their attention on exploring the impacting factors of HBM failures. Several studies [50, 51] pay close attention to exploring the impact of operating conditions; they find the rising temperature causes increment of retention errors [50] and a higher rate of bit flips when reducing the voltage [51]. Researchers [43, 52, 54] from SK Hynix unveil that package joints may affect the reliability of HBM. Moreover, TSV structure also brings new faults [12, 55, 68], with potential impacts on large portions of memory [68]. Similar to DRAM dies, HBM dies also face severe disturbance issues (e.g., RowHammer [48, 69]) that degrade reliability. Different from existing work that studies the characteristics of HBM to guide hardware design, our work provides an in-depth analysis to assist in failure prediction.

DRAM failure analysis. Given the escalating severity of faults and errors in DRAM, substantial efforts have been made to analyze the root causes of faults [23, 41, 60, 61]. Memory faults can be caused by various factors, such as particle strikes [61], cosmic ray [41], and defects in the memory circuit [60]. Prior work [15, 65, 75, 76] also analyses the DRAM errors within production systems to study the characteristics. These studies share important insights regarding the fault modes [15, 81], the impacting factors [80–82], and the correlation between error and memory region [33, 71]. Our study is different from theirs as we focus on the analysis of the impact factors of HBM failures based on the dataset collected from deployed DSA devices.

DRAM failure prediction. For identifying the potential failures in advance, some studies [32, 37] reveal that employing a combination of CE information and server metrics

(e.g., sensor metrics [37]) can improve prediction performance. Zhang et al. [89] and Cheng et al. [22] concentrate on the correlation between DRAM and server failures.The researchers [27, 30, 31] from Intel further perform a failure prediction at the micro-level and propose an ensemble predictor for adapting to diverse environments [29]. Moreover, some studies focus on enhancing the effectiveness of prediction by employing a deep learning algorithm [84], leveraging yet-to-be-consumed uncorrectable errors [87], and adopting new metrics [17]. Calchas distinguishes from them by combining spatial, temporal, and sensor features for predicting potential UERs across various device levels of HBM.

9 Conclusions

This paper presents the first in-depth analysis of HBM errors based on the large-scale dataset, which is collected fromnineteen data centers for over two years. We perform the spatial and temporal analyses to understand the properties of HBM errors. We then make two attempts that aim to predict imminent UEs based on the CE rate and the historical CEs, which are proved to be unsuccessful because of the new error causes introduced by the 3D-stacked architecture of HBM. We finally present Calchas, an approach that utilizes spatial locality, temporal correlation, and sensor information to predict upcoming UERs. Calchas achieves accurate and comprehensive predictions at various device levels.

Acknowledgement

We would like to thank the reviewers' thorough and insightful comments on improving our manuscript. This work was supported by the National Key Research and Development Program of China under Grant No.2022YFB4500302, the Major Research Plan of the National Natural Science Foundation of China (No. 92373114), the Natural Science Foundation of China (No. 62072381), and the Natural Science Foundation of Fujian Province of China (No. 2023J06001). We thank Huawei for providing the HBM error dataset used in this work.

References

[1] Amd radeon instinct™ mi50. https://www.amd.com/system/files/documents/radeon-instinct-mi50-datasheet.pdf.

[2] Hbm3 gen2. https://investors.micron.com/node/45591/pdf.

[3] Intel® stratix® 10. https://www.intel.com/content/www/us/en/docs/programmable/683189/21-3-19-6-1/hbm2-in-intel-stratix-10-devices.html.

[4] Nvidia a100 tensor core gpu. https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/a100/pdf/a100-80gb-datasheet-\update-nvidia-us-1521051-r2-web.pdf.

[5] Nvidia gpu memory error management. https://docs.nvidia.com/deploy/a100-gpu-mem-error-mgmt/index.html.

[6] Nvidia tesla p100. https://images.nvidia.com/content/pdf/tesla/whitepaper/pascal-architecture-whitepaper.pdf.

[7] Nvidia tesla v100 gpu architectureu. https://images.nvidia.com/content/volta-architecture/pdf/volta-architecture-whitepaper.pdf.

[8] Python one-hot encoder. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html.

[9] Xfusion. https://www.xfusion.com/cn/service/fusioncare-basic-service.

[10] Alaa R Alameldeen and David A Wood. Adaptive cache compression for high-performance processors. *ACM SIGARCH Computer Architecture News*, 32(2):212, 2004.

[11] Alexandra Angerd, Angelos Arelakis, Vasilis Spiliopoulos, Erik Sintorn, and Per Stenström. Gbdi: Going beyond base-delta-immediate compression with global bases.

[12] Kwanho Bae and Jongsun Park. Efficient tsv fault detection scheme for high bandwidth memory using pattern analysis. In *Proceeding of the 2020 International SoC Design Conference (ISOCC)*, 2020.

[13] Elisabeth Baseman, Nathan DeBardeleben, Kurt Ferreira, Scott Levy, Steven Raasch, Vilas Sridharan, Taniya Siddiqua, and Qiang Guan. Improving dram fault characterization through machine learning. In *Proceedings of 2016 46th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshop (DSN-W)*, 2016.

[14] Elisabeth Baseman, Nathan Debardeleben, Kurt Ferreira, Vilas Sridharan, Taniya Siddiqua, and Olena Tkachenko. Automating dram fault mitigation by learning from experience. In *Proceedings of 2017 47th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W)*, 2017.

[15] Majed Valad Beigi, Yi Cao, Sudhanva Gurumurthi, Charles Recchia, Andrew Walton, and Vilas Sridharan. A systematic study of ddr4 dram faults in the field. In *Proceedings of 2023 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, 2023.

[16] Rahul Bera, Anant V Nori, Onur Mutlu, and Sreenivas Subramoney. Dspatch: Dual spatial pattern prefetcher.

In *Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 2019.

[17] Isaac Boixaderas, Darko Zivanovic, Sergi Moré, Javier Bartolome, David Vicente, Marc Casas, Paul M Carpenter, Petar Radojković, and Eduard Ayguadé. Cost-aware prediction of uncorrected dram errors in the field. In *Proceedings of the 2020 International Conference for High Performance Computing, Networking, Storage and Analysis (SC)*.

[18] Peter Braun and Heiner Litz. Understanding memory access patterns for prefetching. In *Proceedings of the 2019 International Workshop on AI-assisted Design for Architecture (AIDArc)*, 2019.

[19] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[20] Chin-Long Chen. Error-correcting codes for byte-organized memory systems. *IEEE transactions on information theory*, 32(2):181–185, 1986.

[21] Xi Chen, Lei Yang, Robert P Dick, Li Shang, and Haris Lekatsas. C-pack: A high-performance microprocessor cache compression algorithm. *IEEE transactions on very large scale integration systems*, 18(8):1196–1208, 2009.

[22] Zhinan Cheng, Shujie Han, Patrick PC Lee, Xin Li, Jiongzhou Liu, and Zhan Li. An in-depth correlative study between dram errors and server failures in production data centers. In *Proceedings of the 2022 41st International Symposium on Reliable Distributed Systems (SRDS)*, 2022.

[23] Hyungmin Cho, Chen-Yong Cher, Thomas Shepherd, and Subhasish Mitra. Understanding soft errors in uncore components. In *Proceedings of the 52Nd Annual Design Automation Conference (DAC)*, 2015.

[24] Young-kyu Choi, Yuze Chi, Jie Wang, Licheng Guo, and Jason Cong. When hls meets fpga hbm: Benchmarking and bandwidth optimization. *arXiv preprint arXiv:2010.06075*, 2020.

[25] Carlos HA Costa, Yoonho Park, Bryan S Rosenburg, Chen-Yong Cher, and Kyung Dong Ryu. A system software approach to proactive memory-error avoidance. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC)*, 2014.

[26] T Dell. A white paper on the benefits of chipkill-correct ecc for pc server main memory. ibm microelectronics division. *Technical Report*, 1997.

[27] Xiaoming Du and Cong Li. Memory failure prediction using online learning. In *Proceedings of the 2018 International Symposium on Memory Systems (MEMSYS)*, 2018.

[28] Xiaoming Du and Cong Li. Combining error statistics with failure prediction in memory page offlining. In *Proceedings of the International Symposium on Memory Systems (MEMSYS)*, 2019.

[29] Xiaoming Du and Cong Li. Predicting uncorrectable memory errors from the correctable error history: No free predictors in the field. In *Proceedings of the 2021 International Symposium on Memory Systems (MEMSYS)*, 2021.

[30] Xiaoming Du, Cong Li, Shen Zhou, Xian Liu, Xiaohan Xu, Tianjiao Wang, and Shijian Ge. Fault-aware prediction-guided page offlining for uncorrectable memory error prevention. In *Proceedings of IEEE 39th International Conference on Computer Design (ICCD)*, 2021.

[31] Xiaoming Du, Cong Li, Shen Zhou, Mao Ye, and Jing Li. Predicting uncorrectable memory errors for proactive replacement: An empirical study on large-scale field data. In *Proceedings of 2020 16th European Dependable Computing Conference (EDCC)*, 2020.

[32] Yuyang Du, Hongliang Yu, Yunhong Jiang, Yaozu Dong, and Weimin Zheng. A rising tide lifts all boats: how memory error prediction and prevention can help with virtualized system longevity. In *Proceedings of the Sixth international conference on Hot topics in system dependability (HotDep)*, 2010.

[33] Kurt B Ferreira, Scott Levy, Joshua Hemmert, and Kevin Pedretti. Understanding memory failures on a petascale arm system. In *Proceedings of the 31st International Symposium on High-Performance Parallel and Distributed Computing (HPDC)*, pages 84–96, 2022.

[34] Peter Flach and Meelis Kull. Precision-recall-gain curves: Pr analysis done right. *Advances in neural information processing systems*, 28, 2015.

[35] Jessie Frazelle. Opening up the baseboard management controller. *Communications of the ACM*, 63(2):38–40, 2020.

[36] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[37] Ioana Giurgiu, Jacint Szabo, Dorothea Wiesmann, and John Bird. Predicting dram reliability in the field with machine learning. In *Proceedings of the 18th ACM/IFIP/USENIX Middleware Conference: Industrial Track (Middleware)*, 2017.

[38] Ameer Haj-Ali, Nesreen K Ahmed, Ted Willke, Yakun Sophia Shao, Krste Asanovic, and Ion Stoica. Neurovectorizer: End-to-end vectorization with deep reinforcement learning. In *Proceedings of the 18th ACM/IEEE International Symposium on Code Generation and Optimization (CGO)*, 2020.

[39] Zhulin Hao, Jianqiang Du, Bin Nie, Fang Yu, Riyue Yu, and Wangping Xiong. Random forest regression based on partial least squares connect partial least squares and random forest. In *Proceeding of the 2016 International Conference on Artificial Intelligence: Technologies and Applications (ICAITA)*, 2016.

[40] Yuchong Hu, Liangfeng Cheng, Qiaori Yao, Patrick PC Lee, Weichun Wang, and Wei Chen. Exploiting combined locality for {Wide-Stripe} erasure coding in distributed storage. In *Proceedings of 19th USENIX Conference on File and Storage Technologies (FAST)*, 2021.

[41] Andy A Hwang, Ioan A Stefanovici, and Bianca Schroeder. Cosmic rays don't strike twice: Understanding the nature of dram errors and the implications for system design. *ACM SIGPLAN Notices*, 47(4):111–122, 2012.

[42] Joe Jeddeloh and Brent Keeth. Hybrid memory cube new dram architecture increases density and performance. In *Proceedings of the 2012 symposium on VLSI technology (VLSIT)*, 2012.

[43] Hongshin Jun, Jinhee Cho, Kangseol Lee, Ho-Young Son, Kwiwook Kim, Hanho Jin, and Keith Kim. Hbm (high bandwidth memory) dram technology and architecture. In *Proceedings of the 2017 IEEE International Memory Workshop (IMW)*, 2017.

[44] Amit Kumar Kabat, Shubhang Pandey, and Venkatesh Tiruchirai Gopalakrishnan. Performance evaluation of high bandwidth memory for hpc workloads. In *Proceedings of the 35th International System-on-Chip Conference (SOCC)*, 2022.

[45] Vasileios Karakasis, Georgios Goumas, and Nectarios Koziris. Exploring the effect of block shapes on the performance of sparse kernels. In *Proceedings of the 2009 IEEE International Symposium on Parallel & Distributed Processing (IPDPS)*, 2009.

[46] Jinchun Kim, Seth H Pugsley, Paul V Gratz, AL Narasimha Reddy, Chris Wilkerson, and Zeshan Chishti. Path confidence based lookahead prefetching. In *Proceedings of the 49th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 2016.

[47] Jungrae Kim, Michael Sullivan, and Mattan Erez. Bamboo ecc: Strong, safe, and flexible codes for reliable computer memory. In *2015 IEEE 21st International Symposium on High Performance Computer Architecture (HPCA)*, 2015.

[48] Yoongu Kim, Ross Daly, Jeremie Kim, Chris Fallin, Ji Hye Lee, Donghyuk Lee, Chris Wilkerson, Konrad Lai, and Onur Mutlu. Flipping bits in memory without accessing them: an experimental study of dram disturbance errors. In *Proceeding of the 41st annual international symposium on Computer architecuture (SIGARCH)*, 2014.

[49] Haklin Kimm, Incheon Paik, and Hanke Kimm. Performance comparision of tpu, gpu, cpu on google colaboratory over distributed deep learning. In *Proceedings of 2021 IEEE 14th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC)*, 2021.

[50] Junhyeong Kwon, Shi-Jie Wen, Rita Fung, and Sanghyeon Baeg. Temperature estimation of hbm2 channels with tail distribution of retention errors in fpga-hbm2 platform. *Electronics*, 12(1):32, 2022.

[51] Seyed Saber Nabavi Larimi, Behzad Salami, Osman S Unsal, Adrián Cristal Kestelman, Hamid Sarbazi-Azad, and Onur Mutlu. Understanding power consumption and reliability of high-bandwidth memory with voltage underscaling. In *Proceedings of the 2021 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2021.

[52] Chang Yeol Lee, Sungchul Kim, Hongshin Jun, Kyung Whan Kim, and Sung Joo Hong. Tsv technology and challenges for 3d stacked dram. In *Proceedings of the 2014 Symposium on VLSI Technology: Digest of Technical Papers (VLSIT)*, 2014.

[53] Dong Uk Lee, Kyung Whan Kim, Kwan Weon Kim, Hongjung Kim, Ju Young Kim, Young Jun Park, Jae Hwan Kim, Dae Suk Kim, Heat Bit Park, Jin Wook Shin, et al. 25.2 a 1.2 v 8gb 8-channel 128gb/s high-bandwidth memory (hbm) stacked dram with effective microbump i/o test methods using 29nm process and tsv. In *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*. IEEE, 2014.

[54] Dong Uk Lee, Kang Seol Lee, Yongwoo Lee, Kyung Whan Kim, Jong Ho Kang, Jaejin Lee, and Jun Hyun Chun. Design considerations of hbm stacked dram and the memory architecture extension. In *Proceedings of the 2015 IEEE Custom Integrated Circuits Conference (CICC)*, 2015.

[55] Jong Chern Lee, Jihwan Kim, Kyung Whan Kim, Young Jun Ku, Dae Suk Kim, Chunseok Jeong, Tae Sik Yun, Hongjung Kim, Ho Sung Cho, Sangmuk Oh, et al. High bandwidth memory (hbm) with tsv technique. In *Proceeding of the 2016 International SoC Design Conference (ISOCC)*, 2016.

[56] Scott Levy, Kurt B Ferreira, Nathan DeBardeleben, Taniya Siddiqua, Vilas Sridharan, and Elisabeth Baseman. Lessons learned from memory errors observed over the lifetime of cielo. In *Proceedings of the 2018 International Conference for High Performance Computing, Networking, Storage and Analysis (SC)*, 2018.

[57] Ang Li, Shuaiwen Leon Song, Jieyang Chen, Jiajia Li, Xu Liu, Nathan R Tallent, and Kevin J Barker. Evaluating modern gpu interconnect: Pcie, nvlink, nv-sli, nvswitch and gpudirect. *IEEE Transactions on Parallel and Distributed Systems*, 31(1):94–110, 2019.

[58] Cong Li, Yu Zhang, Jialei Wang, Hang Chen, Xian Liu, Tai Huang, Liang Peng, Shen Zhou, Lixin Wang, and Shijian Ge. From correctable memory errors to uncorrectable memory errors: what error bits tell. In *Proceedings of the 2022 International Conference for High Performance Computing, Networking, Storage and Analysis (SC)*, 2022.

[59] Shang Li, Dhiraj Reddy, and Bruce Jacob. A performance & power comparison of modern high-speed dram architectures. In *Proceedings of the International Symposium on Memory Systems (MEMSYS)*, 2018.

[60] Xin Li, Michael C Huang, Kai Shen, and Lingkun Chu. A realistic evaluation of memory hardware errors and software system susceptibility. In *Proceeding of 2010 USENIX Annual Technical Conference (USENIX ATC)*, 2010.

[61] Xin Li, Kai Shen, Michael C Huang, and Lingkun Chu. A memory soft error measurement on production systems. In *Proceedings of the 2007 USENIX Annual Technical Conference (USENIX ATC)*, 2007.

[62] Heng Liao, Jiajin Tu, Jing Xia, and Xiping Zhou. Davinci: A scalable architecture for neural network computing. In *Proceedings of 31st Hot Chips Symposium (HCS)*, 2019.

[63] Shiyao Lin, Guowen Gong, Zhirong Shen, Patrick PC Lee, and Jiwu Shu. Boosting {Full-Node} repair in {Erasure-Coded} storage. In *Proceedings of 2021 USENIX Annual Technical Conference (USENIX ATC)*, 2021.

[64] David Meyer, Friedrich Leisch, and Kurt Hornik. The support vector machine under test. *Neurocomputing*, 55(1-2):169–186, 2003.

[65] Justin Meza, Qiang Wu, Sanjeev Kumar, and Onur Mutlu. Revisiting memory errors in large-scale production data centers: Analysis and modeling of new trends from the field. In *Proceeding of 2015 45th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, 2015.

[66] Hongyu Miao, Myeongjae Jeon, Gennady Pekhimenko, Kathryn S McKinley, and Felix Xiaozhu Lin. Streambox-hbm: Stream analytics on high bandwidth hybrid memory. In *Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, 2019.

[67] Pierre Michaud. Best-offset hardware prefetching. In *Proceedings of the 2016 22nd International Symposium on High Performance Computer Architecture (HPCA)*, 2016.

[68] Prashant J Nair, David A Roberts, and Moinuddin K Qureshi. Citadel: Efficiently protecting stacked memory from tsv and large granularity failures. *ACM Transactions on Architecture and Code Optimization*, 12(4):1–24, 2016.

[69] Ataberk Olgun, Majd Osseiran, A Giray Yağlıkçı, Yahya Can Tuğrul, Haocong Luo, Steve Rhyner, Behzad Salami, Juan Gomez Luna, and Onur Mutlu. An experimental analysis of rowhammer in hbm2 dram chips. In *Proceedings of the 2023 53rd Annual IEEE/IFIP International Conference on Dependable Systems and Networks-Supplemental Volume (DSN-S)*, 2023.

[70] Sungbo Park, Ingab Kang, Yaebin Moon, Jung Ho Ahn, and G Edward Suh. Bcd deduplication: Effective memory compression using partial cache-line deduplication. In *Proceedings of the 26th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, 2021.

[71] Ayush Patwari, Ignacio Laguna, Martin Schulz, and Saurabh Bagchi. Understanding the spatial characteristics of dram errors in hpc clusters. In *Proceedings of the 2017 Workshop on Fault-Tolerance for HPC at Extreme Scale (FTXS)*, 2017.

[72] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine Learning research*, 12:2825–2830, 2011.

[73] Seth H Pugsley, Zeshan Chishti, Chris Wilkerson, Pengfei Chuang, Robert L Scott, Aamer Jaleel, Shih-Lien Lu, Kingsum Chow, and Rajeev Balasubramonian. Sandbox prefetching: Safe run-time evaluation of aggressive prefetchers. In *Proceedings of the 20th International Symposium on High Performance Computer Architecture (HPCA)*, 2014.

[74] Suresh Ramalingam. Hbm package integration: Technology trends, challenges and applications. In *2016 IEEE Hot Chips 28 Symposium (HCS)*, 2016.

[75] Bianca Schroeder and Garth A. Gibson. A large-scale study of failures in high-performance computing systems. *IEEE Transactions on Dependable and Secure Computing*, 7(4):337–350.

[76] Bianca Schroeder, Eduardo Pinheiro, and Wolf-Dietrich Weber. Dram errors in the wild: a large-scale field study. *ACM SIGMETRICS Performance Evaluation Review*, 37(1):193–204, 2009.

[77] Vivek Seshadri, Donghyuk Lee, Thomas Mullins, Hasan Hassan, Amirali Boroumand, Jeremie Kim, Michael A Kozuch, Onur Mutlu, Phillip B Gibbons, and Todd C Mowry. Ambit: In-memory accelerator for bulk bitwise operations using commodity dram technology. In *Proceedings of the 50th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 2017.

[78] Seyed Mohammad Seyedzadeh, Alex K Jones, and Rami Melhem. Mitigating wordline crosstalk using adaptive trees of counters. In *Proceedings of 2018 ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA)*. IEEE, 2018.

[79] Keeyoung Son, Subin Kim, Hyunwook Park, Seongguk Kim, Keunwoo Kim, Shinyoung Park, Boogyo Sim, Seungtaek Jeong, Gapyeol Park, and Joungho Kim. A novel through mold plate (tmp) for signal and thermal integrity improvement of high bandwidth memory (hbm). In *Proceeding of the 2020 IEEE MTT-S International Conference on Numerical Electromagnetic and Multiphysics Modeling and Optimization (NEMO)*, 2020.

[80] Vilas Sridharan, Nathan DeBardeleben, Sean Blanchard, Kurt B Ferreira, Jon Stearley, John Shalf, and Sudhanva Gurumurthi. Memory errors in modern systems: The good, the bad, and the ugly. *ACM SIGARCH Computer Architecture News*, 43(1):297–310, 2015.

[81] Vilas Sridharan and Dean Liberty. A study of dram failures in the field. In *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis (SC)*, 2012.

[82] Vilas Sridharan, Jon Stearley, Nathan DeBardeleben, Sean Blanchard, and Sudhanva Gurumurthi. Feng shui of supercomputer memory: Positional effects in dram and sram faults. In *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis (SC)*, 2013.

[83] JEDEC Standard. High bandwidth memory (hbm) dram. *Jesd235*, 16, 2013.

[84] Xiaoyi Sun, Krishnendu Chakrabarty, Ruirui Huang, Yiquan Chen, Bing Zhao, Hai Cao, Yinhe Han, Xiaoyao Liang, and Li Jiang. System-level hardware failure prediction using deep learning. In *Proceedings of the 56th Annual Design Automation Conference (DAC)*, 2019.

[85] Shiyang Xuan, Guanjun Liu, Zhenchuan Li, Lutao Zheng, Shuo Wang, and Changjun Jiang. Random forest for credit card fraud detection. In *Proceeding of the 2018 IEEE 15th international conference on networking, sensing and control (ICNSC)*, 2018.

[86] Li Yu, Ziming Zheng, Zhiling Lan, and Susan Coghlan. Practical online failure prediction for blue gene/p: Period-based vs event-driven. In *Proceedings of 2011 IEEE/IFIP 41st International Conference on Dependable Systems and Networks Workshops (DSN-W)*, 2011.

[87] Qiao Yu, Wengui Zhang, Paolo Notaro, Soroush Haeri, Jorge Cardoso, and Odej Kao. Himfp: Hierarchical intelligent memory failure prediction for cloud service reliability. In *Proceedings of the 2023 53rd Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, 2023.

[88] Kaiqiang Zhang, Dongyang Ou, Congfeng Jiang, Yeliang Qiu, and Longchuan Yan. Power and performance evaluation of memory-intensive applications. *Energies*, 14(14):4089, 2021.

[89] Pengcheng Zhang, Yunong Wang, Xuhua Ma, Yaoheng Xu, Bin Yao, Xudong Zheng, and Linquan Jiang. Predicting dram-caused node unavailability in hyper-scale clouds. In *Proceedings of the 2022 52nd Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, 2022.