Hardware Failure Prediction for Cloud Service Reliability

Qiao Yu [1] , Min Zhou [2], Ronglong, Wu, Zhirong Shen[3]

[1]Technical University of Berlin [2]Huawei Cloud [3]Xiamen University

https://hwcloud-ras.github.io/HW_RAS-tutorial-HPCA-2025.github.io/

Agenda

Part 1. INTRODUCTION (Min Zhou, 13:10 - 13:30)

1. Reliability Challenges for Huawei Cloud in the AI era
2. Hardware Failure Prediction Progress in Huawei Cloud

Part 2. Memory Failure Prediction (Qiao Yu, 13:30 – 14:00)

1. Background of memory failure
2. Hierarchical memory failure prediction
3. Conclusion and future work

Part 3. HBM Failure Prediction and Reliable Storage System (Zhirong Shen, 14:00 – 14:30)

1. Introduce analysis of HBM errors in the field
2. Introduce HBM failure prediction framework
3. Introduce some techniques for reliable storage

Part 4. Smartmem Competition (Min Zhou, 14:30 – 15:00)

1. Overview of the SmartMem Competition
2. Attempts to unified memory prediction solution
3. Future work

Coffee Break (15:00 – 15:30)

Part 5. Hands-on Competition (15:30 – 17:00)

Overview of Huawei Cloud

800 +
e-Government cloud

500 +
Financial customers

90%
Internet (China)

90%
Top 30 Chinese automakers

90%
Top 50 Chinese e-commerce companies

90%
Top 50 Chinese Game Enterprises

300 +
SAP cloudification customers

120 +
Carriers

Infrastructure as a Service

Build a single network for global storage and computing, enabling global service reachability.

Technology as a Service

Bringing innovation within reach and accelerating application modernization

Experience as a Service

Replicate excellent performance and enable industry cloudification.

HUAWEI CLOUD

Ranked in 10 Leaders Quadrants

Win 21 + 1st

240+ cloud services

45,000+ partners

6+ million developers

Reliability Challenges for Huawei Cloud in the AI era

Hardware Fault | **Silent Data Corruption** | **Training/Inference**

software bugs | data loss or corruption | Insufficient redundancy | unstable network

Huawei Cloud

operate ~1,000,000 servers

offer 240+ cloud services

one of the leading global providers in cloud computing

Reliability Challenges

AI/LLM

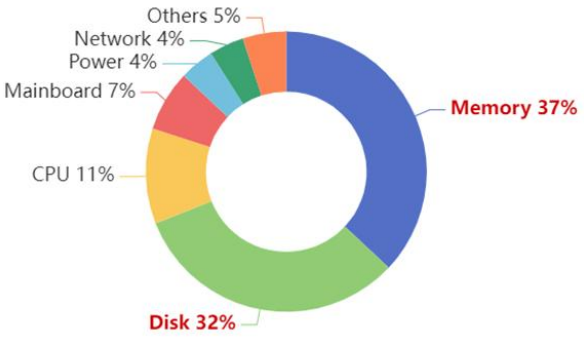
high computational resource requirement

long training time

Large-scale parameters

Hardware reliability is one of the most important operational concerns in Cloud AI Infra

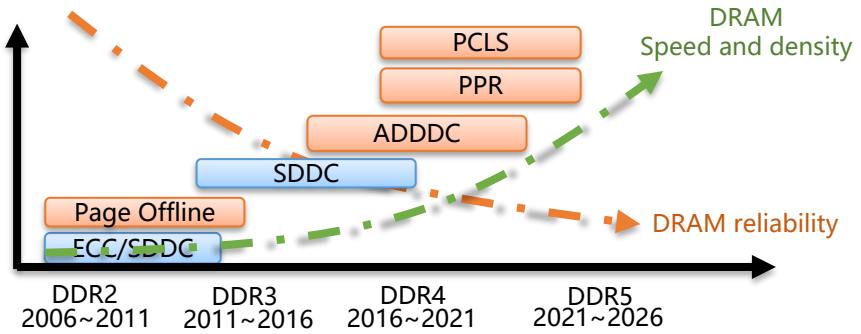
- Hardware faults in are leading causes of server crashes in datacenters, stem from imperfections in fabrication processes, incomplete manufacturing testing, device variability, circuit aging.
- Hardware reliability is one of the most important operational concerns in the AI tide, with the synchronous nature of current frontier training techniques as well as the component complexity .



DISTRIBUTION OF HARDWARE FAILURE IN DATA CENTERS (INTEL'S REPORT)

Component	Category	Interruption Count	% of Interruptions
Faulty GPU	GPU	148	30.1%
GPU HBM3 Memory	GPU	72	17.2%
Software Bug	Dependency	54	12.9%
Network Switch/Cable	Network	35	8.4%
Host Maintenance	Unplanned Maintenance	32	7.6%
GPU SRAM Memory	GPU	19	4.5%
GPU System Processor	GPU	17	4.1%
NIC	Host	7	1.7%
NCCL Watchdog Timeouts	Unknown	7	1.7%
Silent Data Corruption	GPU	6	1.4%
GPU Thermal Interface + Sensor	GPU	6	1.4%
SSD	Host	3	0.7%
Power Supply	Host	3	0.7%
Server Chassis	Host	2	0.5%
IO Expansion Board	Host	2	0.5%
Dependency	Dependency	2	0.5%
CPU	Host	2	0.5%
System Memory	Host	2	0.5%

DISTRIBUTION OF FAILURE IN LLM TRAINING (META'S REPORT)



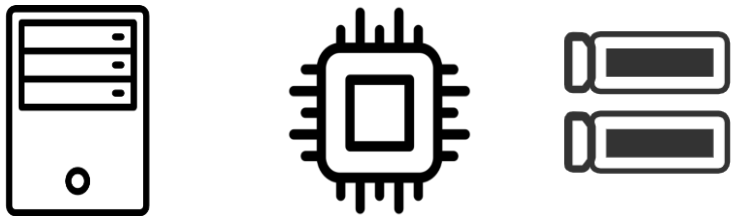
FAILURE RATE INCREASES WITH THE DENSITY AND SPEED

CNET Your guide to a better future

Culture

Google: Computer memory flakier than expected

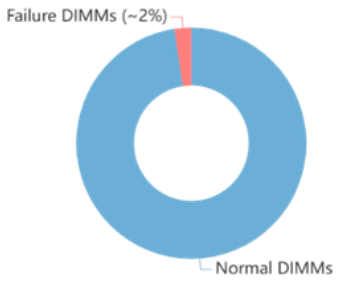
After studying most of its servers for more than two years, Google finds memory failures are much more common than expected, and debunks some other myths, too.

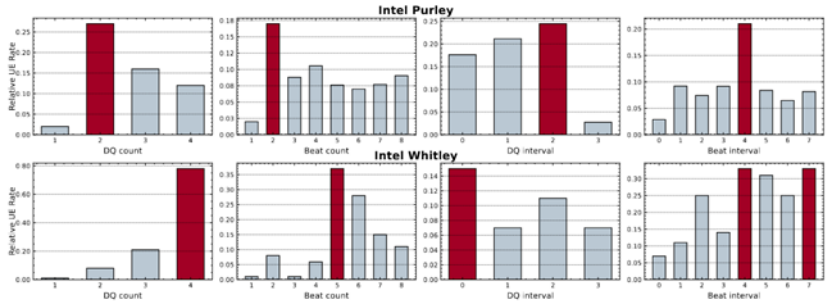


[1] Intel mca mfp stable efficient cloud services. https://www.intel.com/content/www/us/en/software/docs/intel-mca-mfp-jd-stable-efficient-cloud-services.html
[2] The Llama 3 Herd of Models | Research - AI at Meta. https://ai.meta.com/research/publications/the-llama-3-herd-of-models/

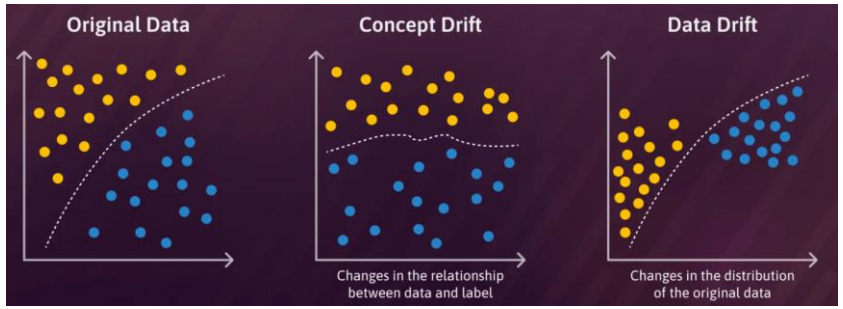
Challenges of hardware failure predictions

- **Complex Failure Mechanisms:** Hardware failures can result from a combination of factors, including environmental conditions, manufacturing defects, and usage patterns
- **Poor Data Quality and Availability：** Data is often sparse, imbalanced, or incomplete and label noisy is common
- **Lacking of fine-grain telemetry**: Lack indicators with standardized and fine-grain metrics, i.e. memory failure prediction largely depends on the coarse-grained CE logs.
- **Heterogeneous data sources:** differences in reported information and fault management capabilities of platforms; differences characteristics exist in hardware generations such as DDR4, DDR5, and HBM , or SATA, SAS, SDD hard disk.
- **Evolving data:** data shift and concept shifts due to the update of data collecting agent, hardware aging， and recovery activity





Failure modes across different CPU architectures

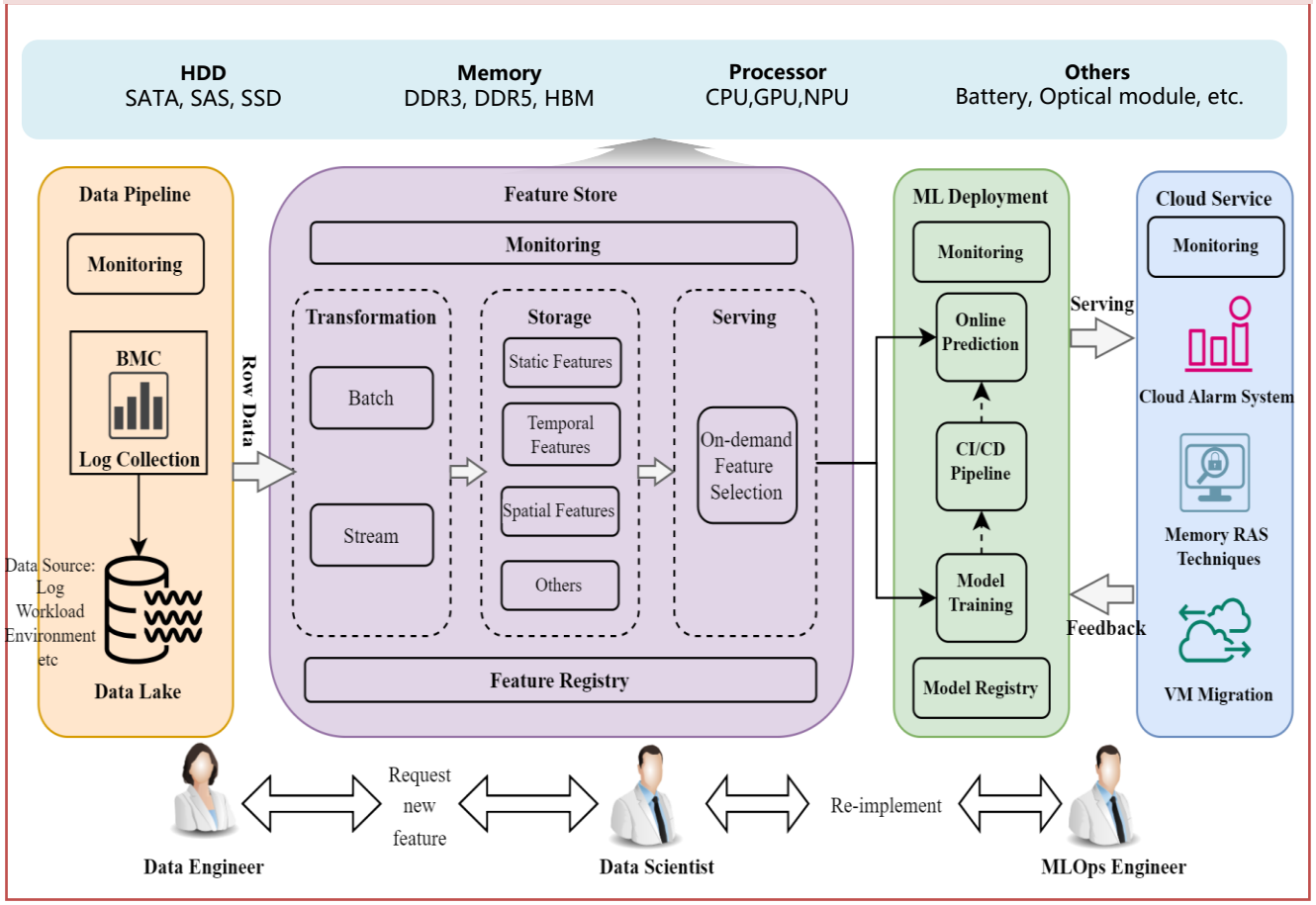


Source: Internet

Hardware Failure Prediction Progress in Huawei Cloud

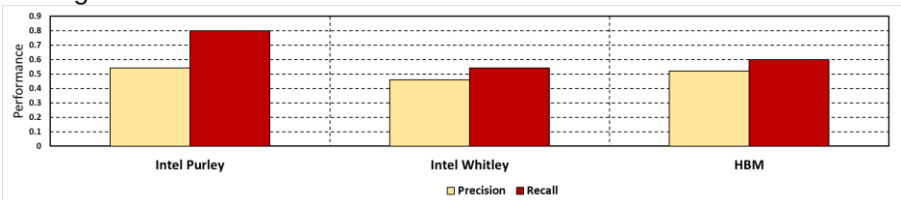
- **MLOps**-enabled hardware failure prediction system to assists the rapid, continuously iterative of knowledge and experience transfer cross different applications.
- **Collaborative team** includes both domain expert and data scientists. **Hardware fault analysis** system with both fault mechanism analysis and data-driven attempts.

Overall Architecture



Key Modules

- **Benchmark Dataset**: Rich data sources, unified evaluation indicators, compatibility with multiple algorithms and scenarios.
- **Feature Store**: Feature store ranging from bit-level fault pattern to macro-level region information
- **Model deployment:** Batch and stream to reduce the delay in E2E data collection and consumption and improving the algorithm efficiency
- **Retrospective Analysis**: Visualizes failure cases, monitors and analyzes data anomalies, and closes the loop between data and algorithms



Future Works

- **Fine-grained failure predictor** to support different level fault recovery needs(e.g. isolating, page-off, and replacing)
- **Robust and knowledge transferable system** to against the concept drift and adjust emerging hardware technology
- **A Deeper Look into failure' s sensitives** to reveal the cause of failures, providing insights to the feature extraction, new telemetry system, as well as future attacks and defenses.

Agenda

Part 1. INTRODUCTION (Min Zhou, 13:10 - 13:30)

1. Reliability Challenges for Huawei Cloud in the AI era
2. Hardware Failure Prediction Progress in Huawei Cloud

Part 2. Memory Failure Prediction (Qiao Yu, 13:30 – 14:00)

1. **Background of memory failure**
2. **Hierarchical memory failure prediction**
3. **Conclusion and future work**

Part 3. HBM Failure Prediction and Reliable Storage System (Zhirong Shen, 14:00 – 14:30)

1. Introduce analysis of HBM errors in the field
2. Introduce HBM failure prediction framework
3. Introduce some techniques for reliable storage

Part 4. Smartmem Competition (Min Zhou, 14:30 – 15:00)

1. Overview of the SmartMem Competition
2. Attempts to unified memory prediction solution
3. Future work

Coffee Break (15:00 – 15:30)

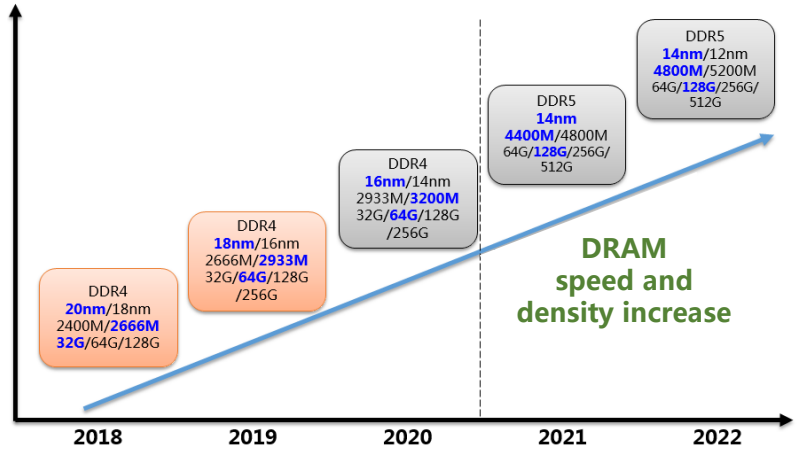
Part 5. Hands-on Competition (15:30 – 17:00)

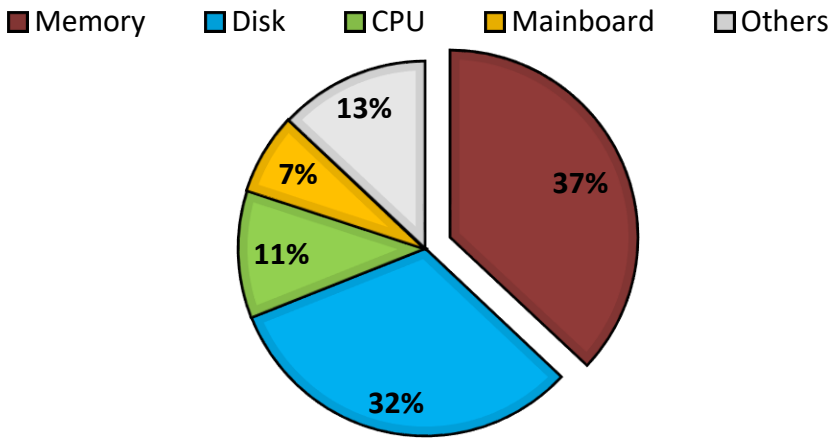
Memory Failure Prediction

1) **Introduction**
2) Related Work
3) Hierarchical Intelligent Memory Failure Prediction
4) Conclusion and Future Work

Memory Failure

- ***Background:*** With the increasing demand for cloud computing**, DRAM (Dynamic random access memory) failure is *one of leading causes*** of server crashes. *DRAM failure rate has increased* each generation **as the density and speed increase**.





DISTRIBUTION OF HARDWARE FAILURE IN DATA CENTERS (INTEL'S REPORT)

CNET Your guide to a better future

Culture

Google: Computer memory flakier than expected

After studying most of its servers for more than two years, Google finds memory failures are much more common than expected, and debunks some other myths, too.

How memory failure prediction keeps data centers and the digital economy up and running

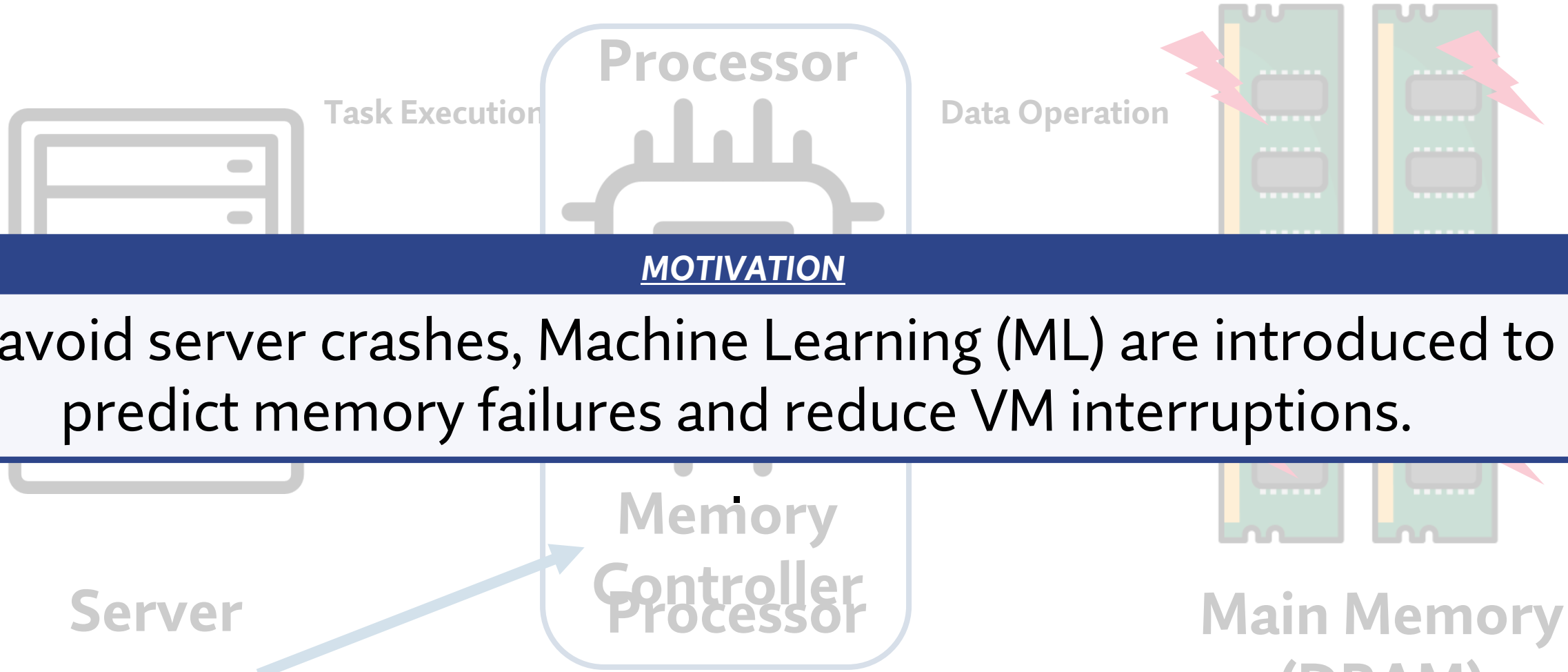
By Jeff Klaus
general manager of data center management solutions
Intel

It was approximately four years ago when I wrote in an industry publication, "Today's data centers are the modern equivalent of railroad infrastructure and the world's business rides upon its rails."

Looking back, one can't help but think how understated, if not quaint, the idea seems now.

Just consider the historic surge in digital services we've witnessed as global populations were forced to work, study, socialize, conduct retail transactions, entertain themselves and even meet with healthcare providers, all from home. As Microsoft CEO Satya Nadella famously said roughly

How Memory Failure Causes Server Crashes?



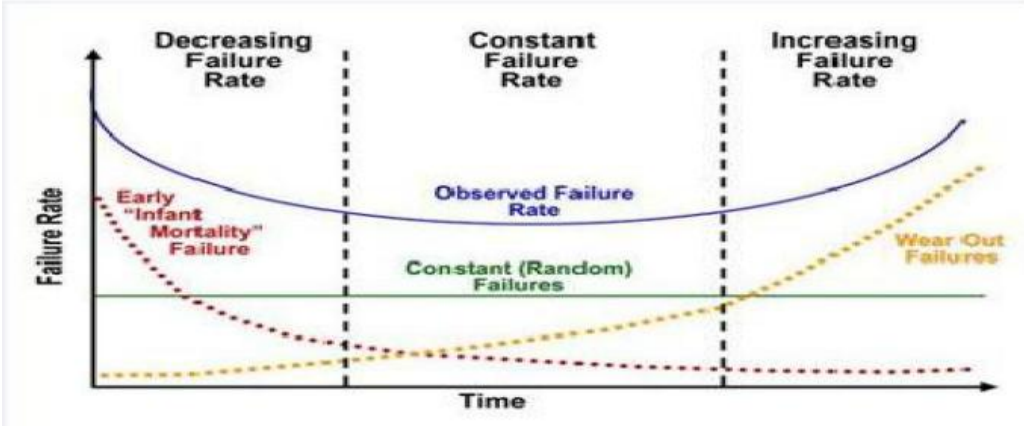
MOTIVATION

To avoid server crashes, Machine Learning (ML) are introduced to predict memory failures and reduce VM interruptions.

Memory Failure Category

- **Memory failures** can be categorized into early failures (due to design flaws or quality control issues), constant random failures (due to manufacturing particle control or design stress conflicts), and wear-out failures (due to fatigue, aging, or wear-related misoperation).

Failure Category	Characteristics	Factors
Early Failure	Typically occurs within about half a year	Design flaws, process-related issues, quality control problems, etc.
Constant Random Failure	Persists throughout the entire lifecycle and cannot be eliminated	inadequate particle control in manufacturing, conflict between design strength and actual operating stress, etc
Wear-Out Failure	leading to a rising failure rate	fatigue, aging, wear-related misoperation, etc.



These three failure modes form the bathtub curve

Memory Faults, Errors and Failures

Memory Faults:

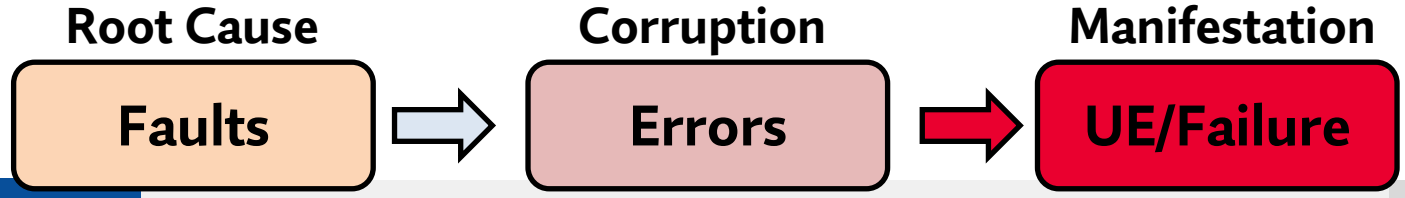
- The underlying causes of an error in DRAM.
  - **Soft Faults**: particle impacts, cosmic rays. --recoverable
  - **Hard Faults**: ware-out, manufacture defect. --repeatable

Memory Errors:

- **Correctable Error (CE):** Errors that can be corrected by ECC mechanism.
- **Uncorrectable error (UE):** uncorrected by ECC mechanism.
  - **Sudden UE**: UEs that instantly corrupt data.
  - **Predictable UE**: UEs that start as correctable errors and evolve to severe UEs.
    - **As-yet-unconsumed (AYU) UCE:** UCEs that have been **detected but not yet consumed** by the system or processors, e.g., Intel's UCNA, Patrol Scrubbing UCE.
    - **Critical UCE:** UCEs that typically **lead to service interruptions**.

Memory Failures:

- A critical event in which the delivered service **deviates from the correct service**. **UEs typically lead to catastrophic failures**, resulting in a crash.



Outline

1) Introduction

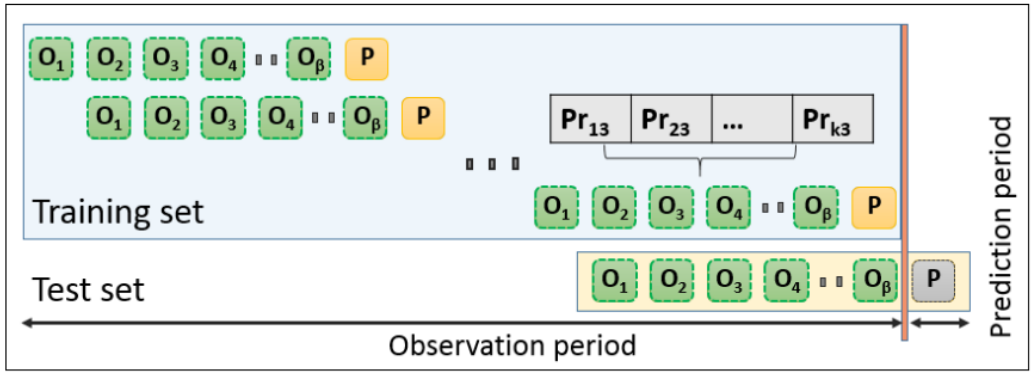
2) **Related Work**

3) Hierarchical Intelligent Memory Failure Prediction

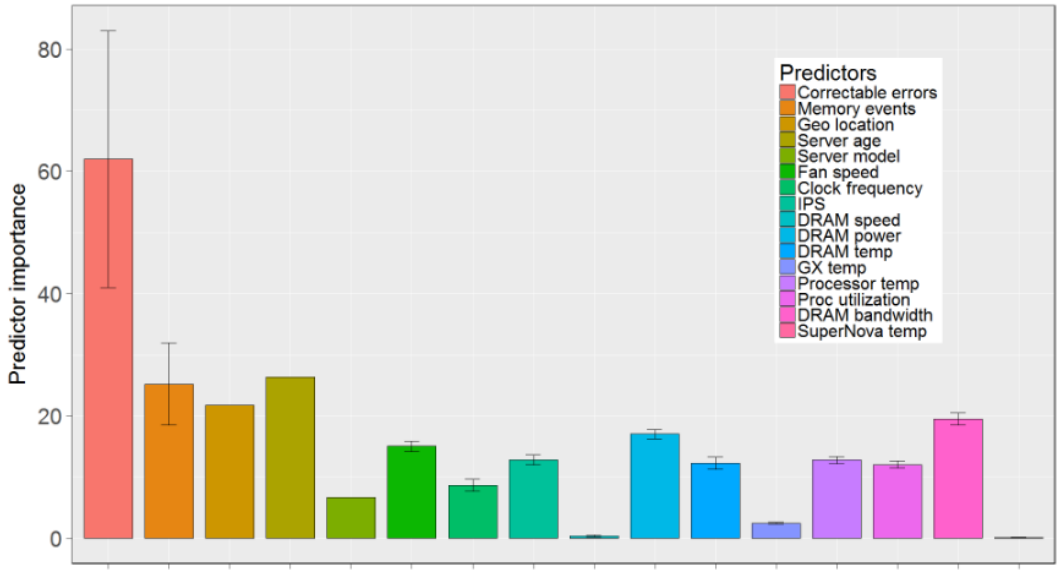
4) Conclusion and Future Work

Predicting DRAM reliability in the field

- Giurgiu et al. [***IBM, Middleware'17***] proposed the first predictive model for DRAM failure prediction utilizing **correctable errors, event logs, and sensor metrics**.



Training and prediction on methodology for sliding window model



Predictor importance across errors and sensors

Ioana Giurgiu et.al. Predicting DRAM reliability in the field with machine learning. In Proceedings of the 18th ACM/IFIP/USENIX Middleware Conference, 2017

Predicting DRAM reliability in the field

- Giurgiu et al. [***IBM, Middleware'17***] proposed the first predictive model for DRAM failure prediction utilizing **correctable errors, event logs, and sensor metrics**.

Table: Prediction results over a 2-week window

Cut-off threshold															
0.5			0.6			0.7			0.8			0.9			AUC
Prec.1	Rec.1	ACC	Prec.1	Rec.1	ACC	Prec.1	Rec.1	ACC	Prec.1	Rec.1	ACC	Prec.1	Rec.1	ACC	
0.38	0.18	0.59	0.56	0.13	0.56	0.81	0.1	0.55	0.91	0.08	0.54	0.96	0.07	0.53	0.79
0.52	0.2	0.6	0.75	0.15	0.58	0.86	0.12	0.57	0.93	0.1	0.55	0.96	0.05	0.53	0.8
0.6	0.23	0.62	0.78	0.18	0.6	0.85	0.14	0.59	0.94	0.1	0.55	0.96	0.06	0.53	0.8
0.66	0.26	0.65	0.84	0.23	0.63	0.87	0.17	0.61	0.93	0.11	0.58	0.94	0.07	0.56	0.81
0.6	0.31	0.67	0.83	0.25	0.65	0.89	0.19	0.63	0.95	0.13	0.6	0.92	0.08	0.58	0.83

Ioana Giurgiu et.al. Predicting DRAM reliability in the field with machine learning. In Proceedings of the 18th ACM/IFIP/USENIX Middleware Conference, 2017

Cost-Aware UE Prediction

- Boixaderas et al. [***BSC, SC'20***] presented Random Forest model with CE features at the DIMM, socket, and node levels, and **evaluated their UE prediction using a cost-aware method**.

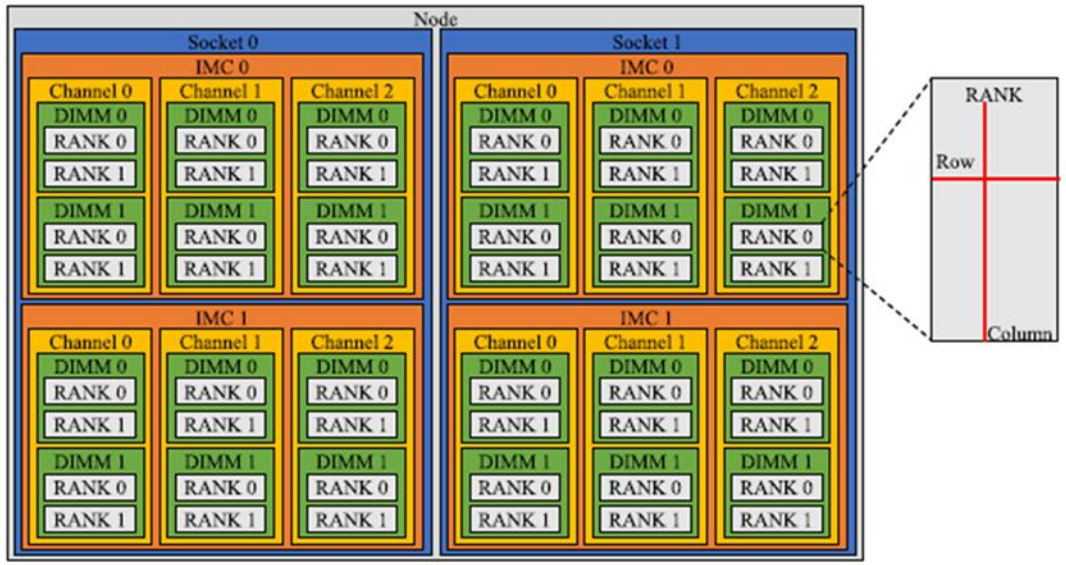


Fig. The DRAM hierarchical topology

Feature for prediction method
Per DIMM:
Number of corrected errors (CEs)
Number of ranks, banks, rows and columns with CEs
Average and standard deviation of errors per rank, bank, row and col.[a]
Number of uncorrected errors (UEs)
Number of warnings
Time since the DIMM was installed in its current position
Number of times the DIMM has changed its position in production
Manufacturer[b]
DIMM Capacity: 4, 8 or 16 GB[b]
Chip Capacity: 2 Gbit or 4 Gbit[b]
Data Width: x4 or x8[b]
Per socket:
Number of DIMMs with corrected errors
Sum of corrected errors in all the DIMMs
Number of DIMMs with uncorrected errors
Sum of uncorrected errors in all the DIMMs
Number of DIMMs with warnings
Sum of warnings in all the DIMMs
Per node:
Sum of each per-socket feature across sockets in the node
Number of node boots (starts) in the last minute, hour and day

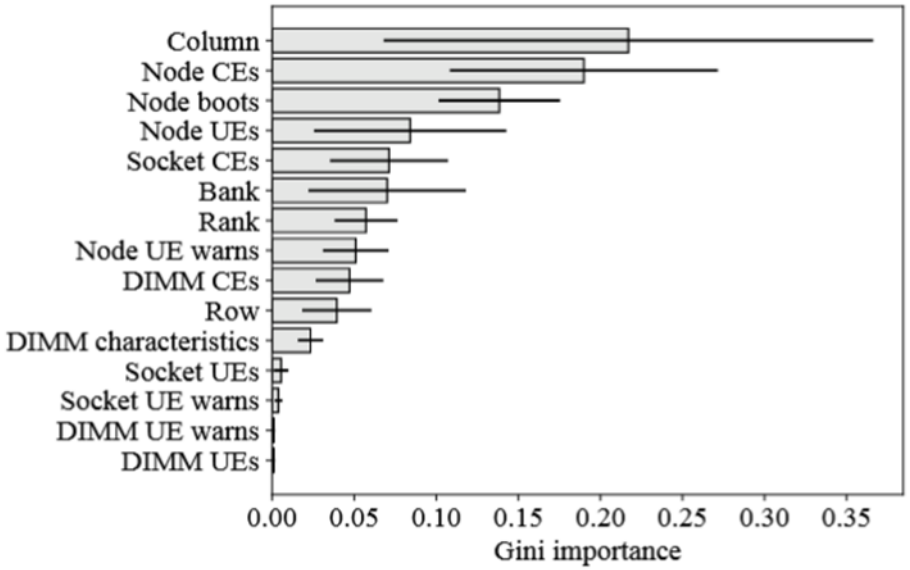
[a] Considering only the ranks, banks, rows and columns with errors.
[b] Transformed into features using one-hot encoding.

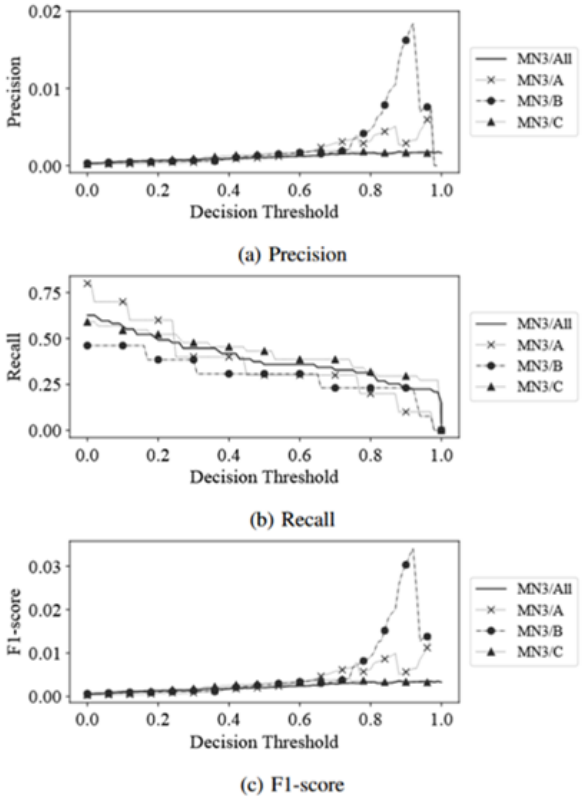
Table. Observation features used for the prediction

I. Boixaderas et al., "Cost-Aware Prediction of Uncorrected DRAM Errors in the Field," SC20: International Conference for High Performance Computing, Networking, Storage and Analysis, Atlanta, GA, USA, 2020

Cost-Aware UE Prediction

- Boixaderas et al. [***BSC, SC'20***] presented Random Forest model with CE features at the DIMM, socket, and node levels, and **evaluated their UE prediction using a cost-aware method**.





I. Boixaderas et al., "Cost-Aware Prediction of Uncorrected DRAM Errors in the Field," SC20: International Conference for High Performance Computing, Networking, Storage and Analysis, Atlanta, GA, USA, 2020

Workload-Aware DRAM Failure Prediction

- X. Wang et al. [Alibaba VTS'20] integrated system-level workload indicators with CE log for DRAM failure prediction.

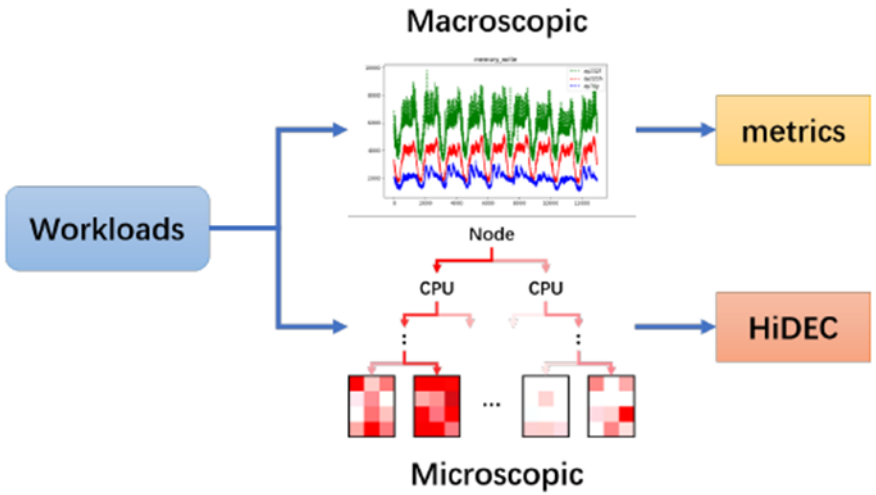


Fig. The macroscopic and microscopic workload features

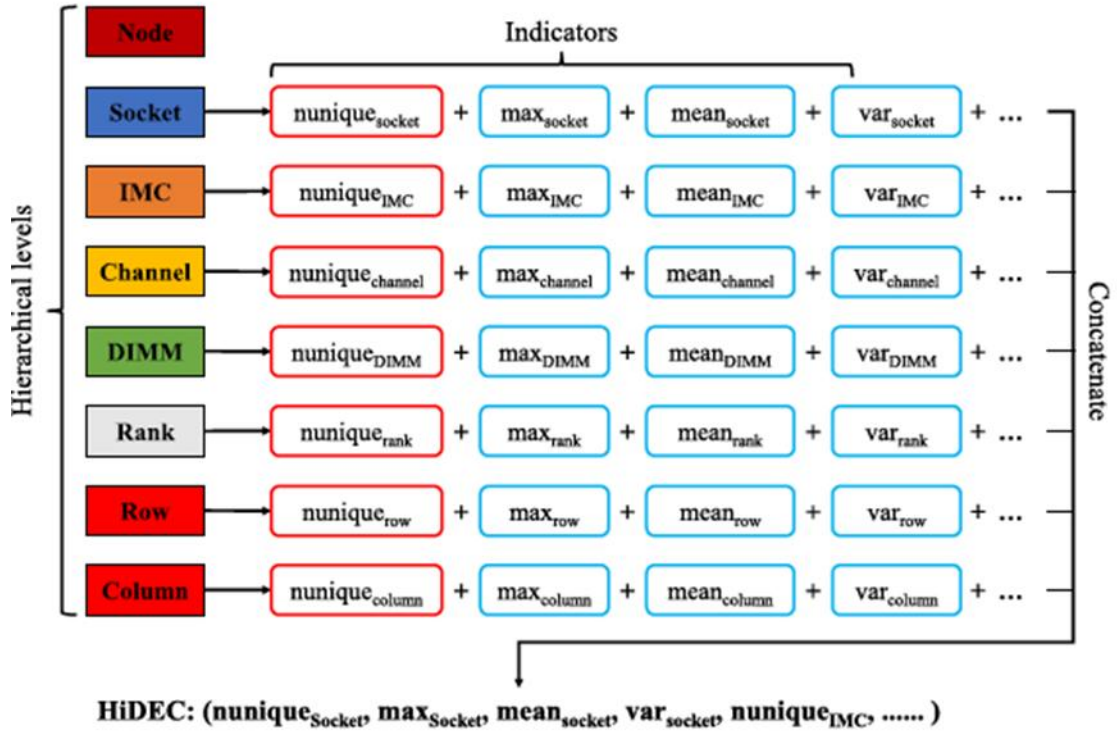


Fig. HiDEC construction procedure

X. Wang et al., "On Workload-Aware DRAM Failure Prediction in Large-Scale Data Centers," 2021 IEEE 39th VLSI Test Symposium (VTS), San Diego, CA, USA, 2021

Workload-Aware DRAM Failure Prediction

- X. Wang et al. [Alibaba VTS'20] integrated system-level workload indicators with CE log for DRAM failure prediction.

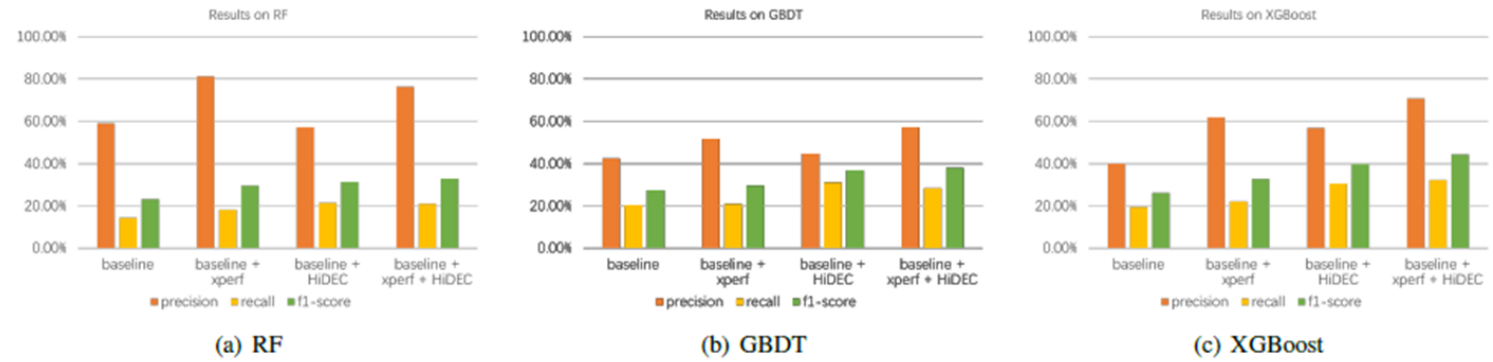


Fig. Precision, recall and f1-score of DRAM failure prediction across different workload feature combinations

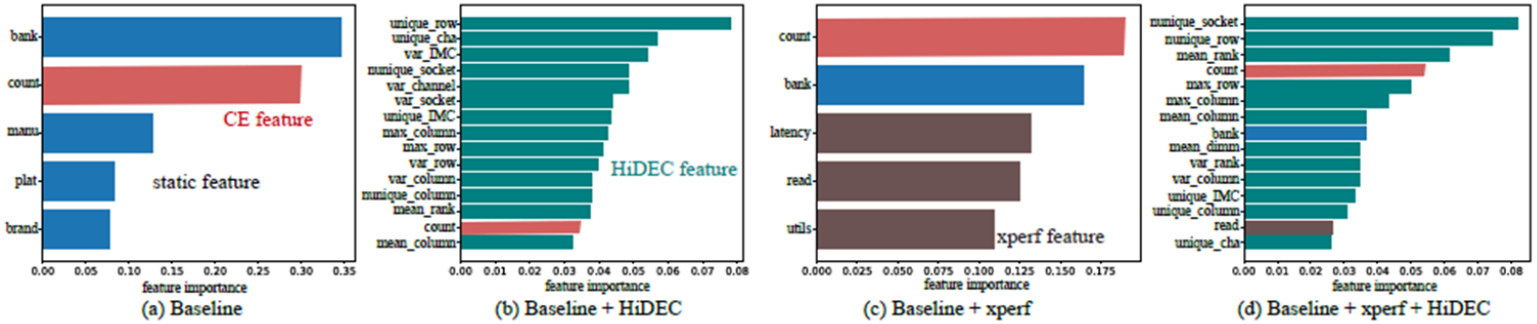


Fig. Feature importance of XGBOOST

X. Wang et al., "On Workload-Aware DRAM Failure Prediction in Large-Scale Data Centers," 2021 IEEE 39th VLSI Test Symposium (VTS), San Diego, CA, USA, 2021

Predicting DRAM-Caused Node Unavailability

- Zhang et al. [***SJTU, DSN'22***] proposed to predict DRAM-caused Node unavailability , considering **both UE and CE storm** from the ***Alibaba datacenter***.

- DRAM-caused Node unavailability (DCNU)

  - Uncorrectable error
  - CE Storm: a burst of CEs occurs in a short period
  - DIMM communication loss: the communication between the DIMM and hardware system is lost.

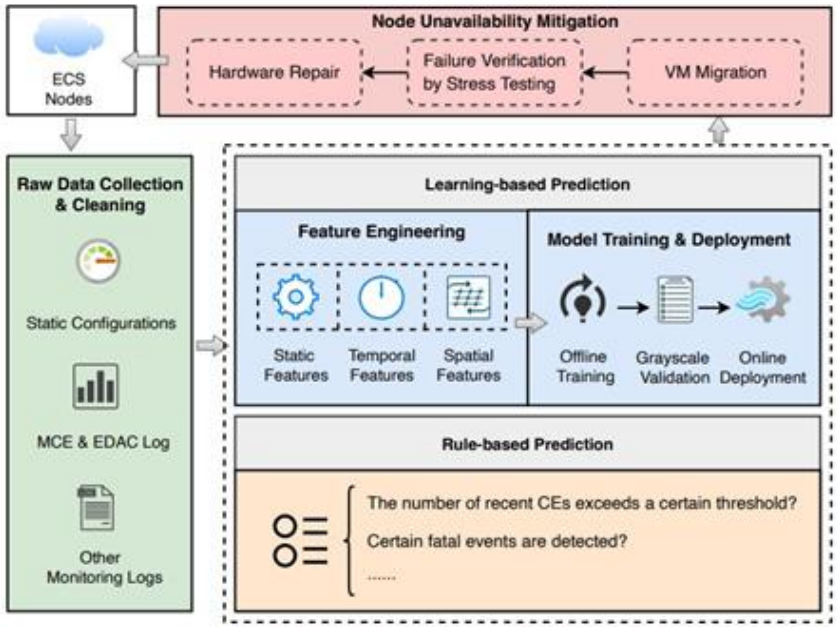
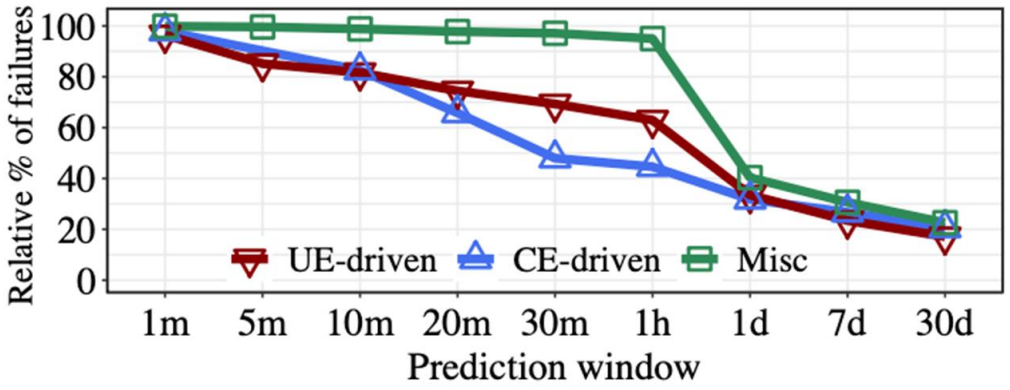


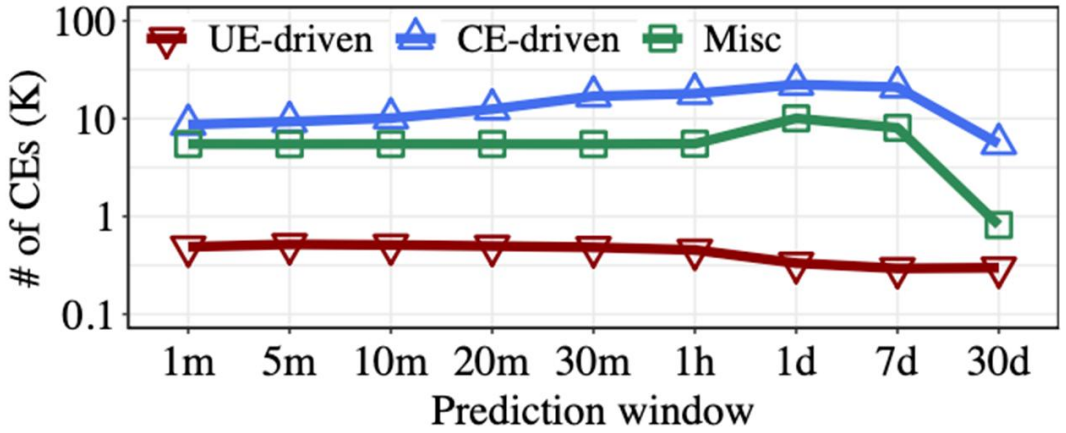
Fig. The workflow of failure prediction

P. Zhang et al., "Predicting DRAM-Caused Node Unavailability in Hyper-Scale Clouds," 2022 52nd Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)

Correlative Study Between DRAM Errors and Server Failures

- Cheng et al. [***CUHK, SRDS'22***] proposed to predict server failures, considering **both UE-driven and CE-driven failures** from the ***Alibaba datacenter***.

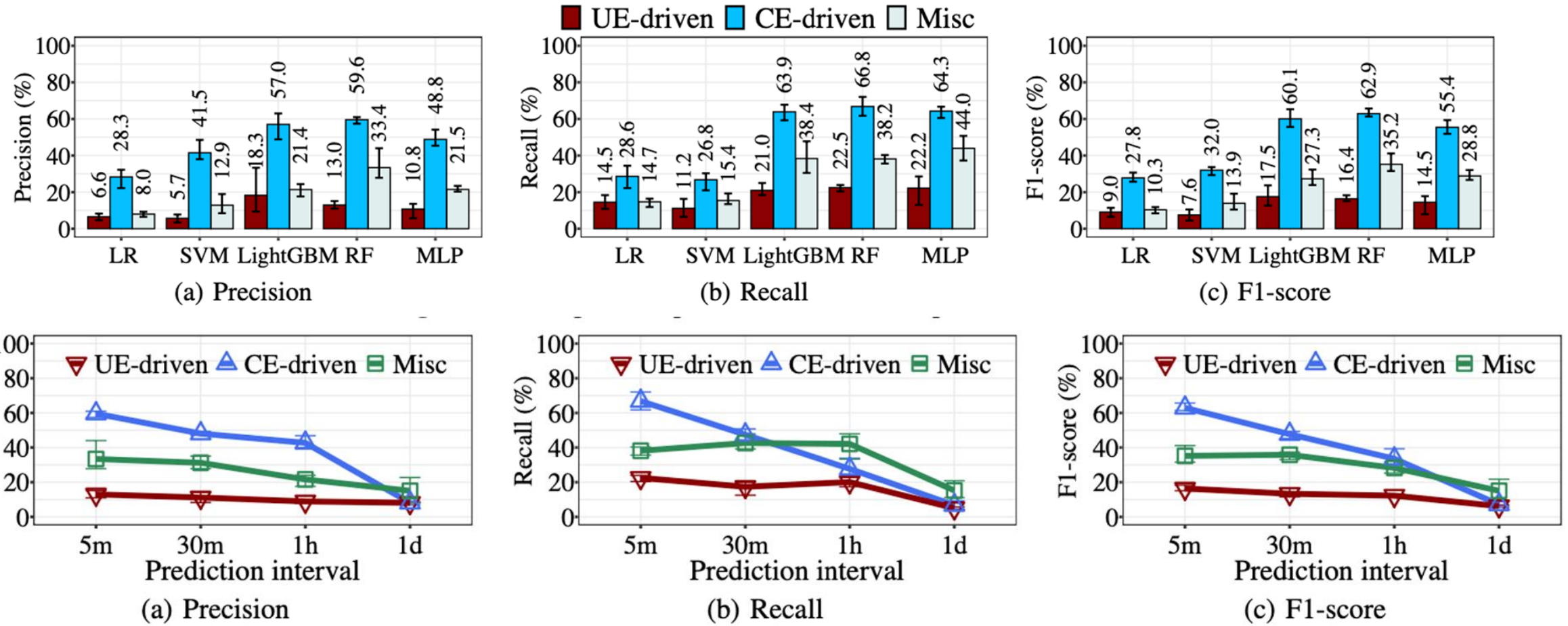




Z. Cheng et.al., "An In-Depth Correlative Study Between DRAM Errors and Server Failures in Production Data Centers," 2022 41st International Symposium on Reliable Distributed Systems (SRDS), Vienna, Austria, 2022

Correlative Study Between DRAM Errors and Server Failures

- Cheng et al. [***CUHK, SRDS'22***] proposed to predict server failures, considering **both UE-driven and CE-driven failures** from the ***Alibaba datacenter***.



What Error Bits Tell

- Li et al. [***Intel&ByteDance, SC'22***] presented an empirical study correlating CEs and UEs using **error bit patterns** and DIMM part numbers.

TABLE I: Characteristics of the DRAM error dataset.

DIMM mfr.	Part number	DIMMs w/ CEs	DIMMs also w/ UEs	DIMMs w/ UEs in %
A	A1	3115	392	12.6%
	A2	705	67	9.5%
	A3	496	24	4.8%
	A4	211	10	4.7%
	A5	102	8	7.8%
	A6	118	5	4.2%
	A7	28	3	10.7%
	A8	306	3	1.0%
	A9	72	2	2.8%
	A10	164	1	0.6%
	A11∼A14	41	10	0%
	Total	5358	515	9.6%
B	B1	1931	24	1.2%
	B2	28	5	19.7%
	B3	177	4	2.3%
	B4	703	3	0.4%
	B5	386	2	0.5%
	B6	10	2	20.0%
	B7	10	2	20.0%
	B8	27	1	3.7%
	B9∼B15	633	0	0%
	Total	3905	43	1.1%
C	C1	1017	10	1.0%
	C2	337	3	0.9%
	C3	132	3	2.3%
	C4∼C11	98	0	0%
	Total	1584	16	1.0%

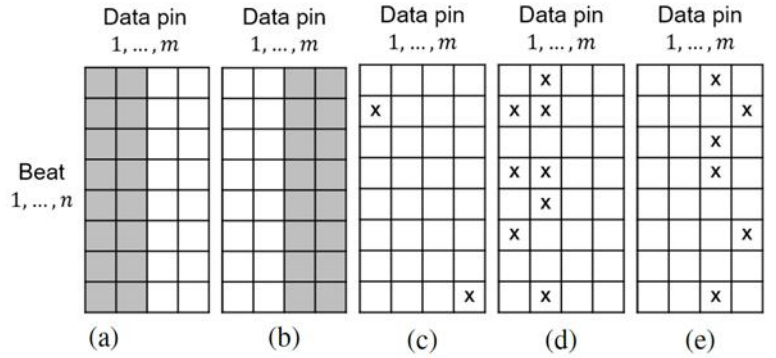


Fig. 2: Examples of (a) a fully correctable error-bit pattern; (b) another fully correctable error-bit pattern; (c) an actual error-bit pattern from a risky CE; (d) an actual error-bit pattern from a CE which is not risky; and (e) another actual error-bit pattern from a CE which is not risky.

C. Li et al., "From Correctable Memory Errors to Uncorrectable Memory Errors: What Error Bits Tell," SC22: International Conference for High Performance Computing, Networking, Storage and Analysis, Dallas, TX, USA, 2022

What Error Bits Tell

- Li et al. [***Intel&ByteDance, SC'22***] presented an empirical study correlating CEs and UEs using **error bit patterns** and DIMM part numbers.

A1: Risky_CE $\wedge$ Column($\theta_c = 8$) $\rightarrow$ UE
A2: Risky_CE $\rightarrow$ UE
A1: Risky_CE $\rightarrow$ UE

(a)

Risky_CE $\wedge$ Bank($\theta_x = 8, \theta_y = 8$) $\rightarrow$ UE

(b)

B1: Risky_CE $\wedge$ Bank($\theta_x = 8, \theta_y = 8$) $\rightarrow$ UE
Risky_CE $\wedge$ Bank($\theta_x = 32, \theta_y = 8$) $\rightarrow$ UE

(c)

C1: Risky_CE $\wedge$ Bank($\theta_x = 4, \theta_y = 8$) $\rightarrow$ UE

(d)

Fig. 11: Examples of typical decision lists learned including (a) one list learned for DIMMs from manufacturer A; (b) one list learned for DIMMs from manufacturer B; (c) another list learned for DIMMs from Manufacturer B; and (d) one list learned for DIMMs from manufacturer C.

TABLE IX: Without the risky CE indicator, a complex fault indicator with more conditions [14], [16] can sometimes recover part of the information for better UE prediction.

Rule	Precision
A1: Column($\theta_c = 8$) $\rightarrow$ UE	42.3%
A1: $\text{Column}_{\text{complex}}(\theta_c = 8, l_c = 4096)$ $\rightarrow$ UE	52.6%
A1: Risky_CE $\wedge$ Column($\theta_c = 8$) $\rightarrow$ UE	**63.1%**

C. Li et al., "From Correctable Memory Errors to Uncorrectable Memory Errors: What Error Bits Tell," SC22: International Conference for High Performance Computing, Networking, Storage and Analysis, Dallas, TX, USA, 2022

What Error Bits Tell

- Li et al. [***Intel&ByteDance, SC'22***] presented an empirical study correlating CEs and UEs using **error bit patterns** and DIMM part numbers.

TABLE VIII: Breakdown of prediction result for DIMMs from manufacturer B to different part numbers. (Those 'N/A's come from 0/0.)

Part number	DIMMs w/ UEs	Precision	Recall
B1	24	77.8%	58.3%
B2	5	0%	0%
B3	4	75.0%	75.0%
B4	3	75.0%	100.0%
B5	2	50.0%	50.0%
B6	2	N/A	0%
B7	2	100.0%	50.0%
B8	1	100.0%	100.0%
B9∼B15	0	N/A	N/A

TABLE VII: Breakdown of prediction result for DIMMs from manufacturer C to different part numbers. (Those 'N/A's come from 0/0.)

Part number	DIMMs w/ UEs	Precision	Recall
C1	10	100.0%	40.0%
C2	3	N/A	0%
C3	3	N/A	0%
C4∼C11	0	N/A	N/A

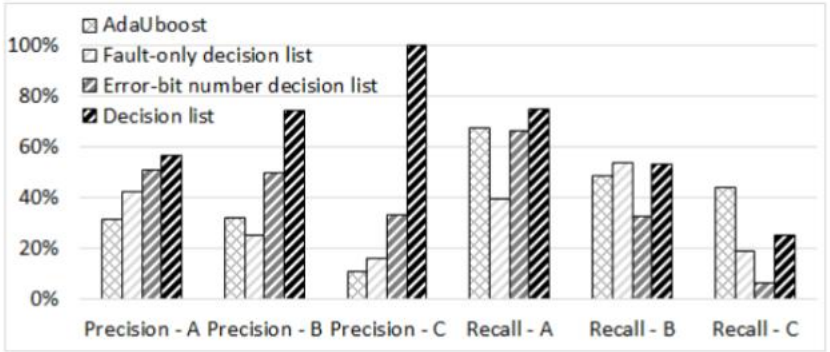


Fig. 9: Comparing with the AdaUBoost approach [16], the decision list without the new risky CE indicator, and the decision list using the indicators based on number of error bits.

C. Li et al., "From Correctable Memory Errors to Uncorrectable Memory Errors: What Error Bits Tell," SC22: International Conference for High Performance Computing, Networking, Storage and Analysis, Dallas, TX, USA, 2022

Outline

1) Introduction

2) Related Work

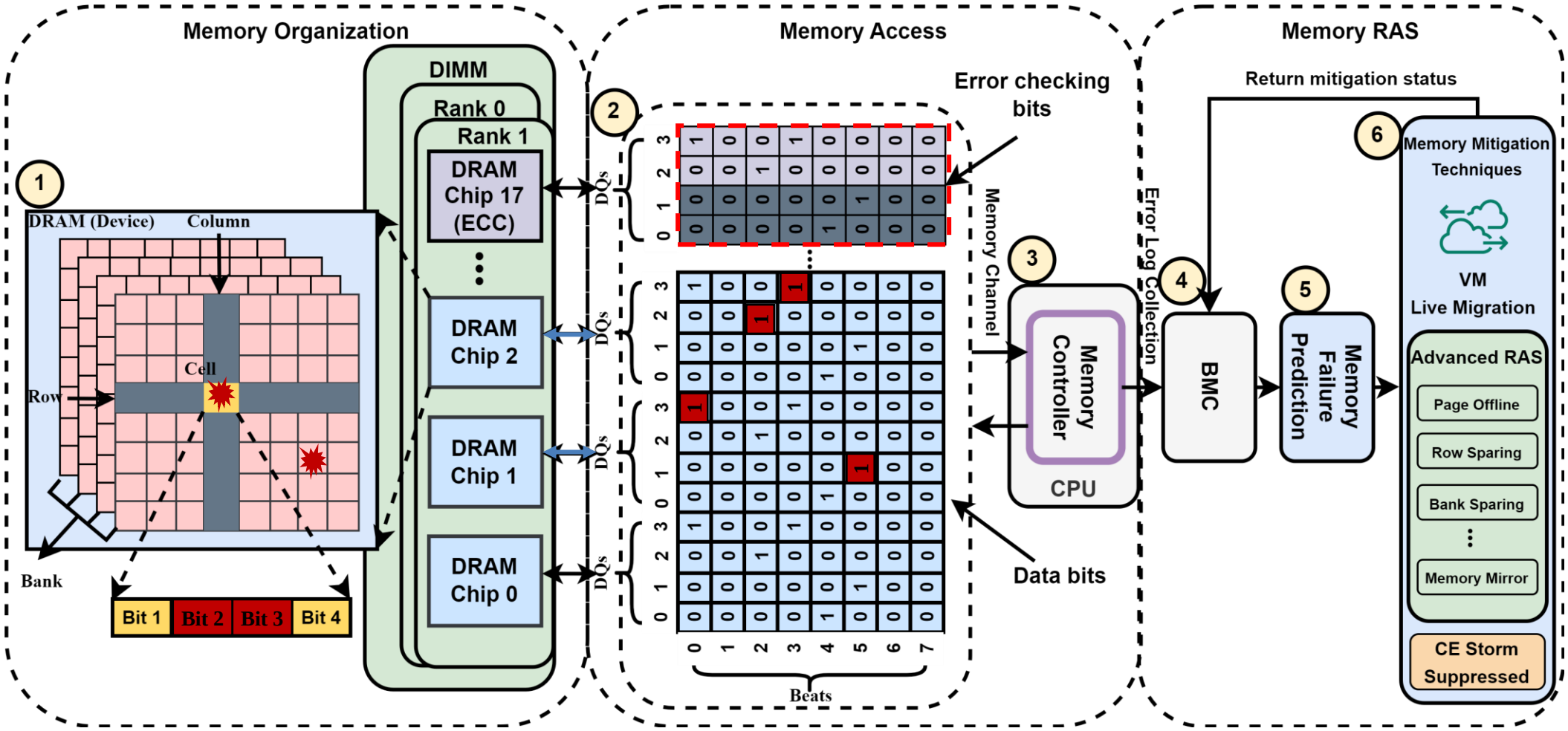
3) **Hierarchical Intelligent Memory Failure Prediction**

4) Conclusion and Future Work

Exploring Error Bits for Memory Failure Prediction

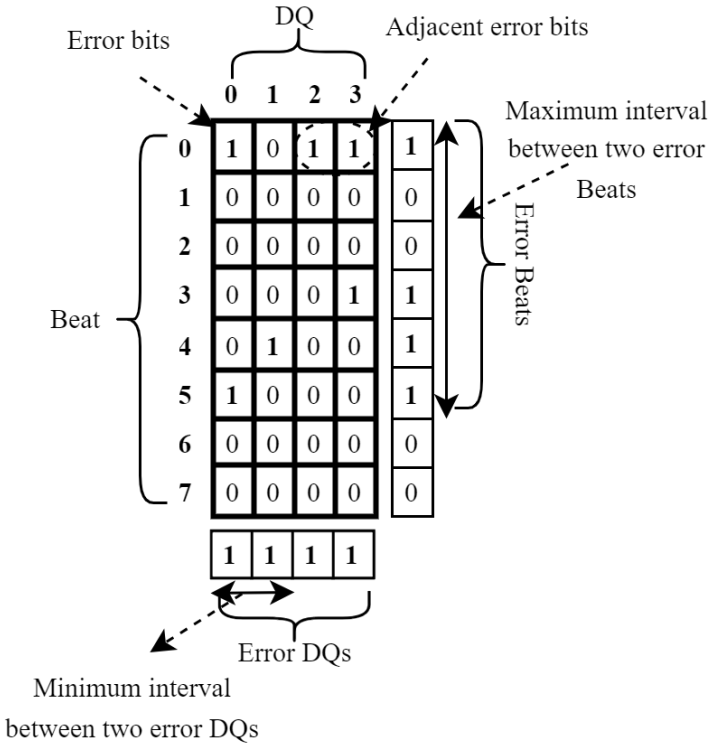
Q. Yu, W. Zhang, J. Cardoso and O. Kao, "Exploring Error Bits for Memory Failure Prediction: An In-Depth Correlative Study," 2023 IEEE/ACM International Conference on Computer Aided Design (ICCAD)

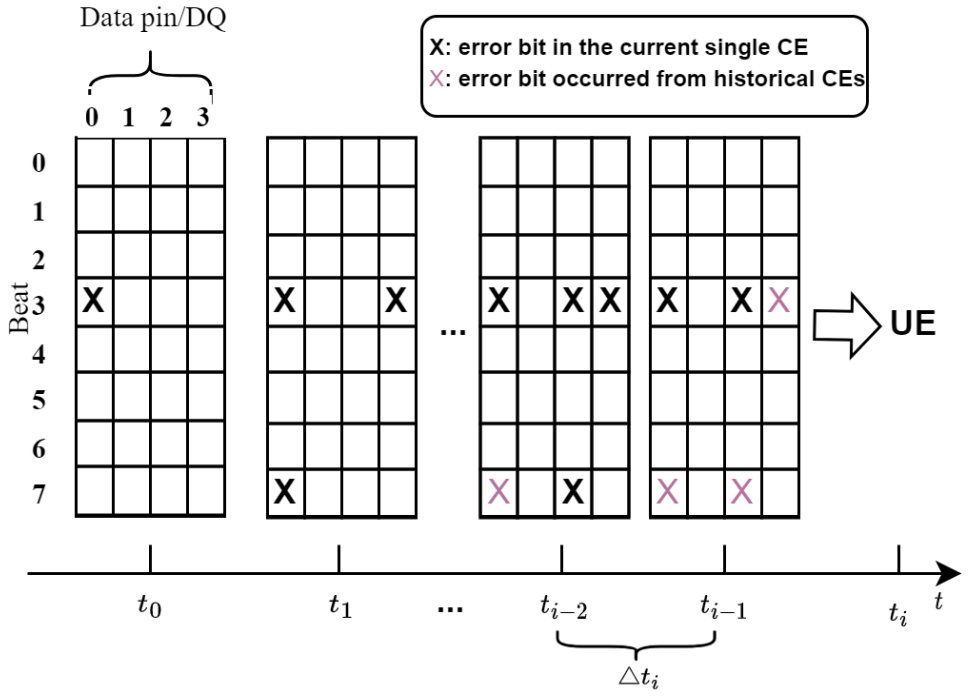
Exploring Error Bits for Memory Failure Prediction



Q. Yu, W. Zhang, J. Cardoso and O. Kao, "Exploring Error Bits for Memory Failure Prediction: An In-Depth Correlative Study," 2023 IEEE/ACM International Conference on Computer Aided Design (ICCAD)

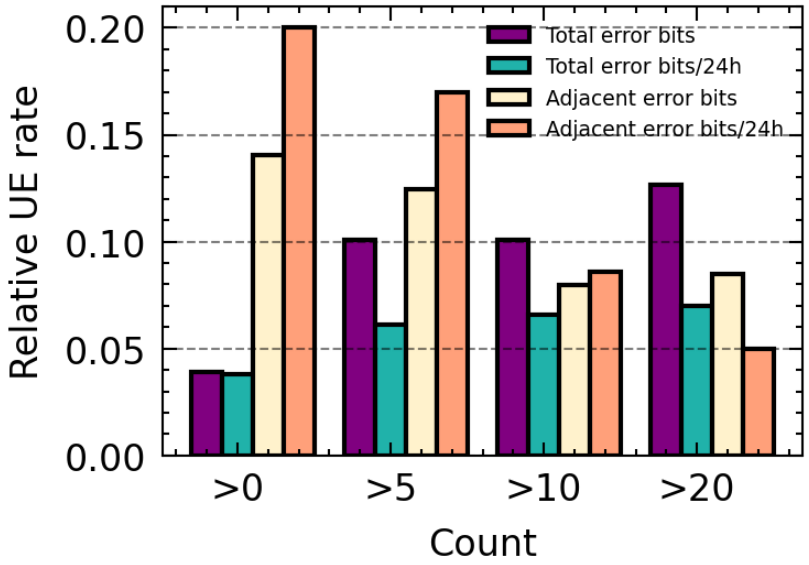
Error Bit Distribution





Q. Yu, W. Zhang, J. Cardoso and O. Kao, "Exploring Error Bits for Memory Failure Prediction: An In-Depth Correlative Study," 2023 IEEE/ACM International Conference on Computer Aided Design (ICCAD)

Error Bit Analysis

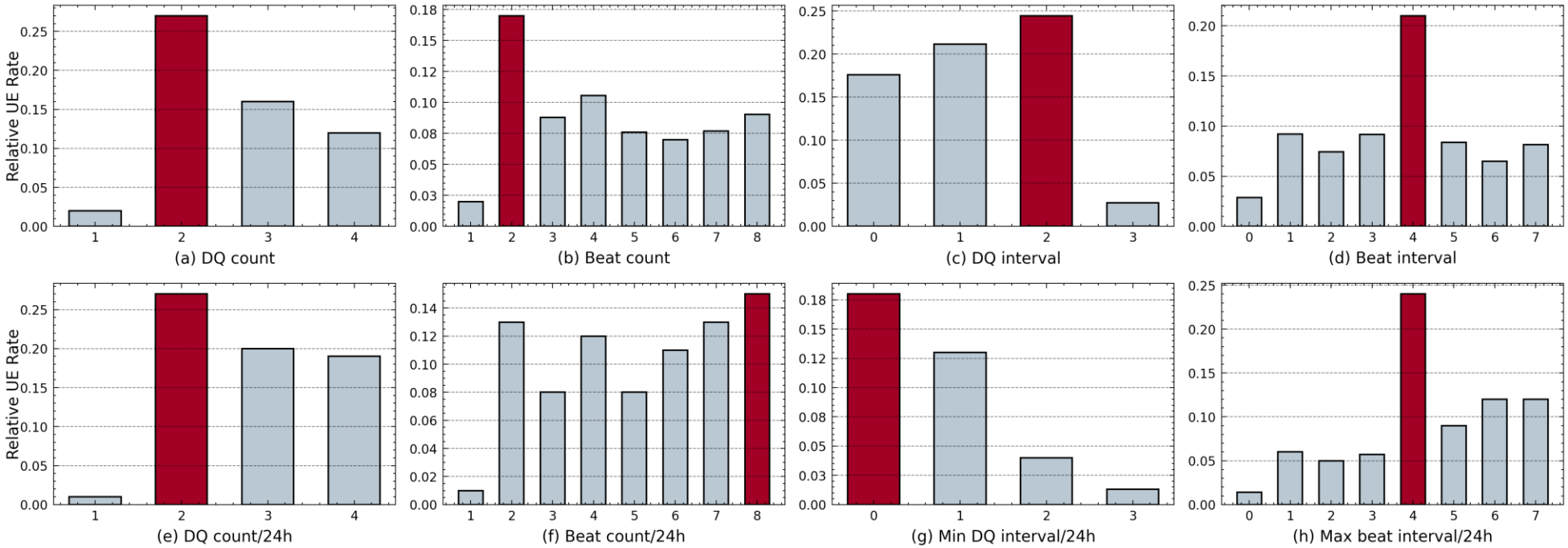


FINDING

The total number of error bits exhibits weaker correlation of UE compared to adjacent error bits. Even a small number of adjacent bits can lead to UE occurrence.

Q. Yu, W. Zhang, J. Cardoso and O. Kao, "Exploring Error Bits for Memory Failure Prediction: An In-Depth Correlative Study," 2023 IEEE/ACM International Conference on Computer Aided Design (ICCAD)

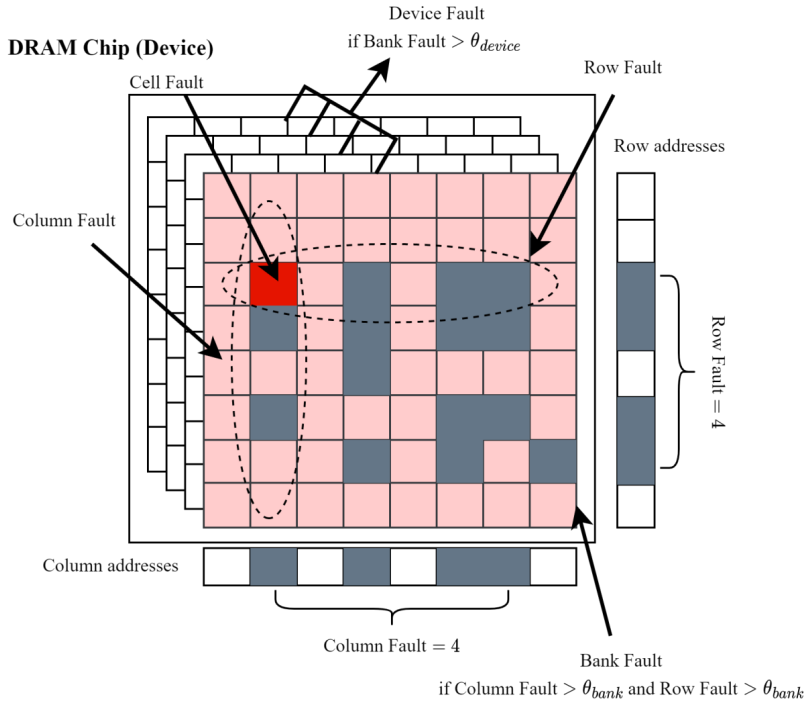
Error Bit Analysis

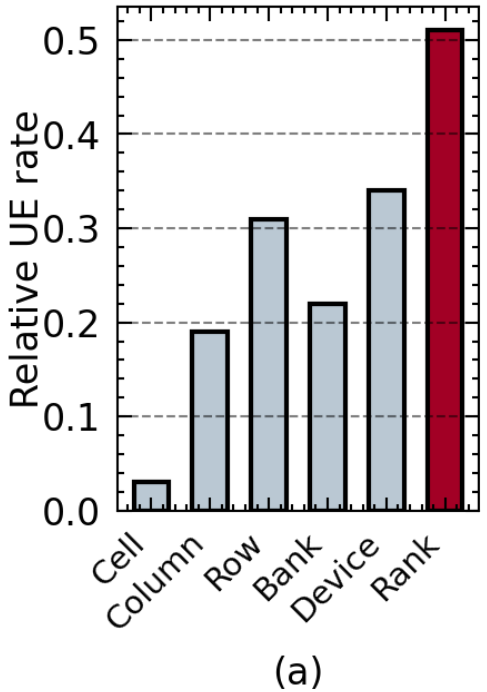


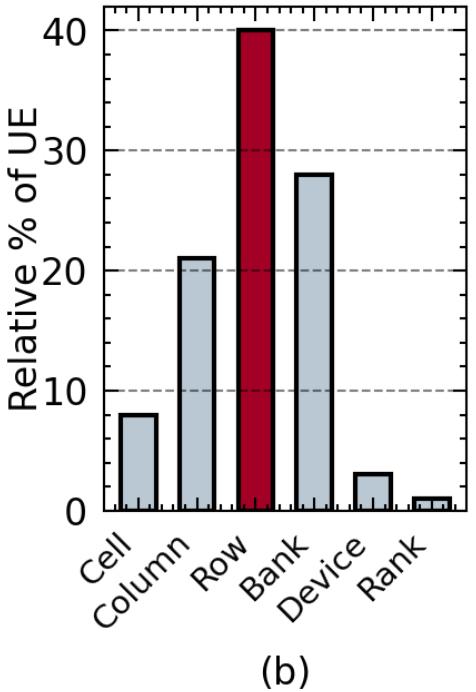
FINDING

Our analyses reveal that both spatial and temporal error bits in DQs and beats play a significant role in distinguishing UE occurrences.

Component Fault Analysis



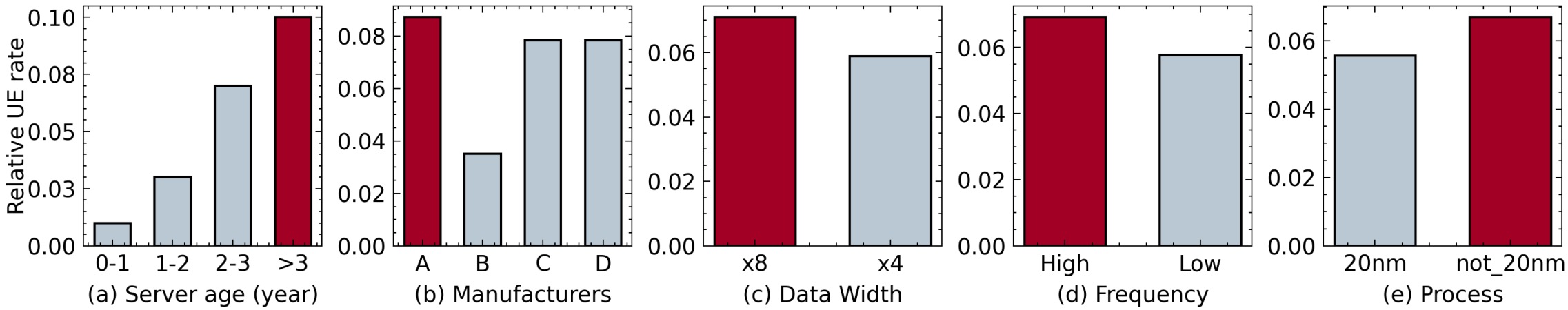




FINGDING

While higher-level faults may have a higher likelihood of causing UEs, Row and Bank faults account for the majority of UEs in the system.

System Configuration Analysis



FINDING

The UE rate varies across server age, manufacturers, data width, frequency and process, while we did not observe significant differences in the capacity of the DIMM.

Hierarchical Intelligent Memory Failure Prediction (HiMFP)

Q. Yu, et al., "HiMFP: Hierarchical Intelligent Memory Failure Prediction for Cloud Service Reliability," 2023 53rd Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN), Porto, Portugal, 2023,

Dataset and Performance Measures

Large-Scale dataset from **Huawei Cloud datacenters**

- We collected datasets across CPU architectures including **Intel Skylake/Cascade Lake, Icelake and Huawei ARM K920 architectures servers.**
  - Error logs w/ **static configuration, MCE log and memory events.**
  - Over 90,000 DDR4 DIMMs (Samsung, Hynix, Micron, etc) are examined.

Performance Measures

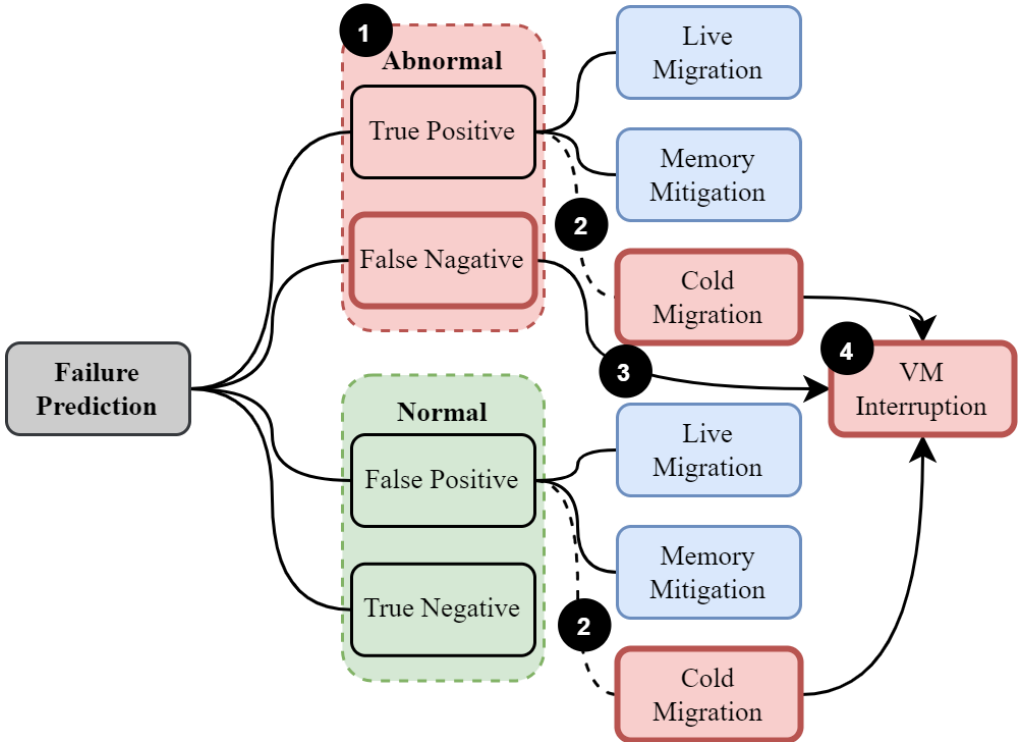
- **Traditional performance metrics:** $\boldsymbol{Precision = \frac{TP}{TP+FP}, Recall = \frac{TP}{TP+FN}\ and\ F_1 = \frac{2 \times Precision \times Recall}{Precision+Recall}}$, where
  - **True Positive (TP):** failure predicted in the prediction window.
  - **False Positive (FP):** failure not in the prediction window.
  - **False Negative (FN):** failure happens without prediction.
  - **True Negative (TN):** no failure happens with no prediction.

- **Production Impact:** *VM Interruption Reduction Rate (VIRR)* calculates VM interruptions saved from memory failure prediction in the production environment.

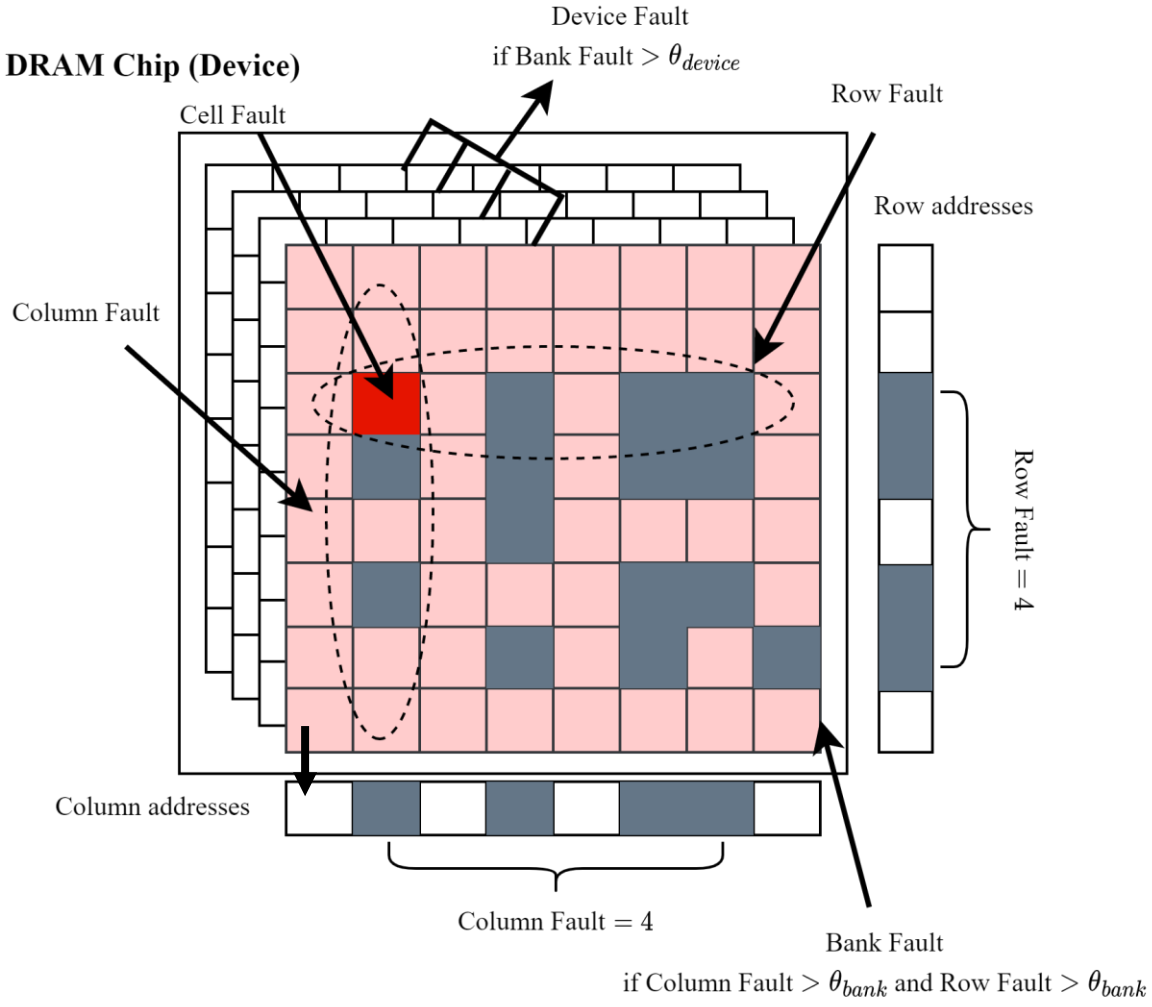
Dataset and Performance Measures

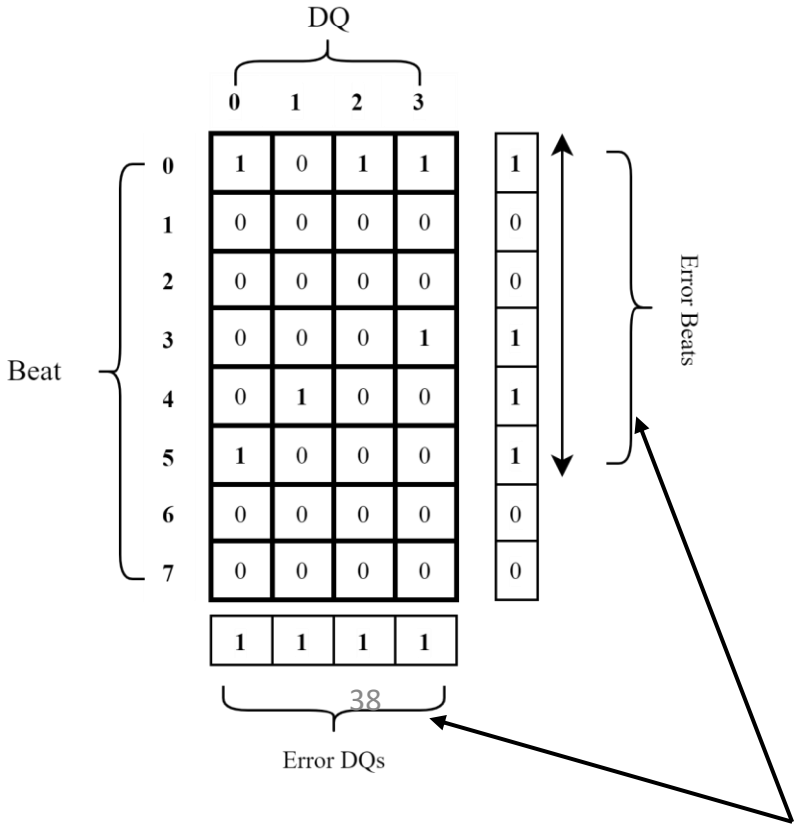
Virtual Machine Interruption Rate

- **Production Impact:** *VM Interruption Reduction Rate (VIRR)* calculates VM interruptions saved from memory failure prediction in the production environment.



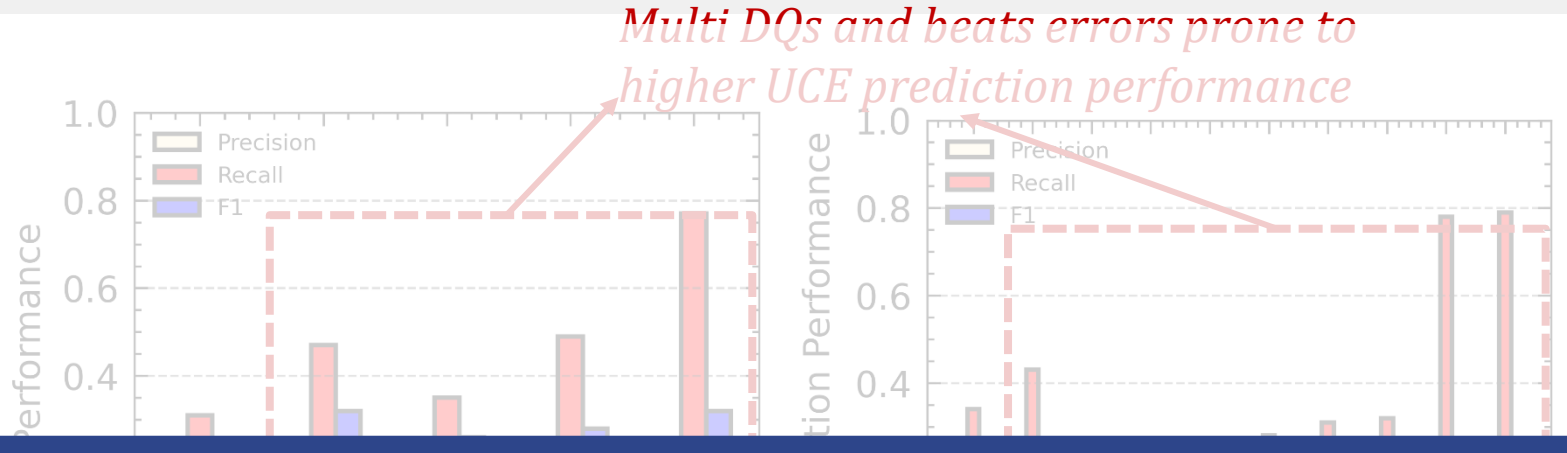
Micro-level Evaluation





Multi-DQ-Beat Fault

Micro-level Evaluation



KEY OBSERVATION

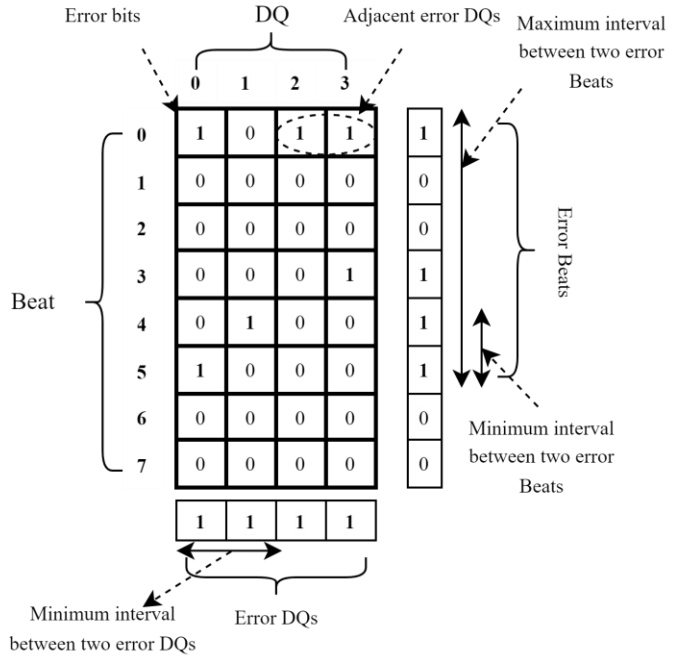
Error DQ and beats play significant role in distinguishing UE occurrences.

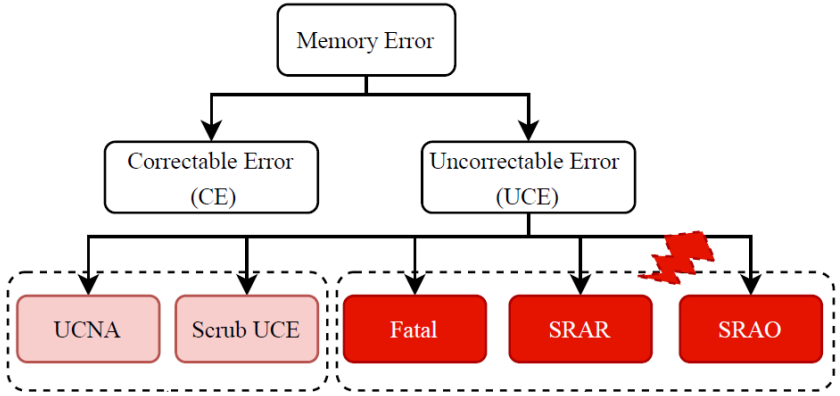
DIMM-level Features

Seven groups of spatio-temporal features are characterized for DIMM-level prediction.

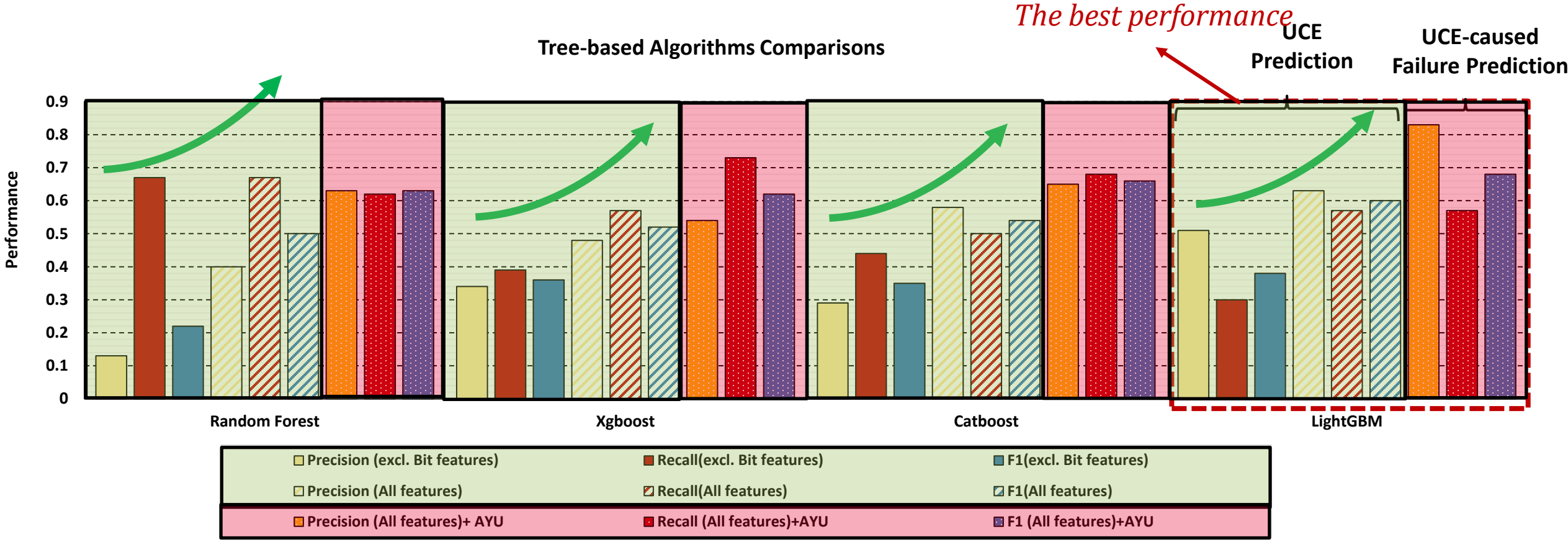
1. Static Features: DIMM characteristics including **Manufacturer, Data width, Frequency and Chip process**.
2. CE Error Rate: the number of CEs and their **occurrence frequency**.
3. Fault Counts: cumulative number of **micro-level faults (cell, row, column, bank and device)**.
4. Memory Events: **CE storm, CE storm suppressed notification, CE overflow**, etc.

5. DQ-Beat Error Bits Features: the **spatial and temporal distribution of error bits** in DQs and Beats.
6. Error bits pattern features: **error-bit patterns** which are more likely to **encounter UCE**.
7. AYU UCE rule: **As-yet-unconsumed UCEs (e.g., UCNA, Patrol Scrub UCE)** for UCE-caused failure prediction.





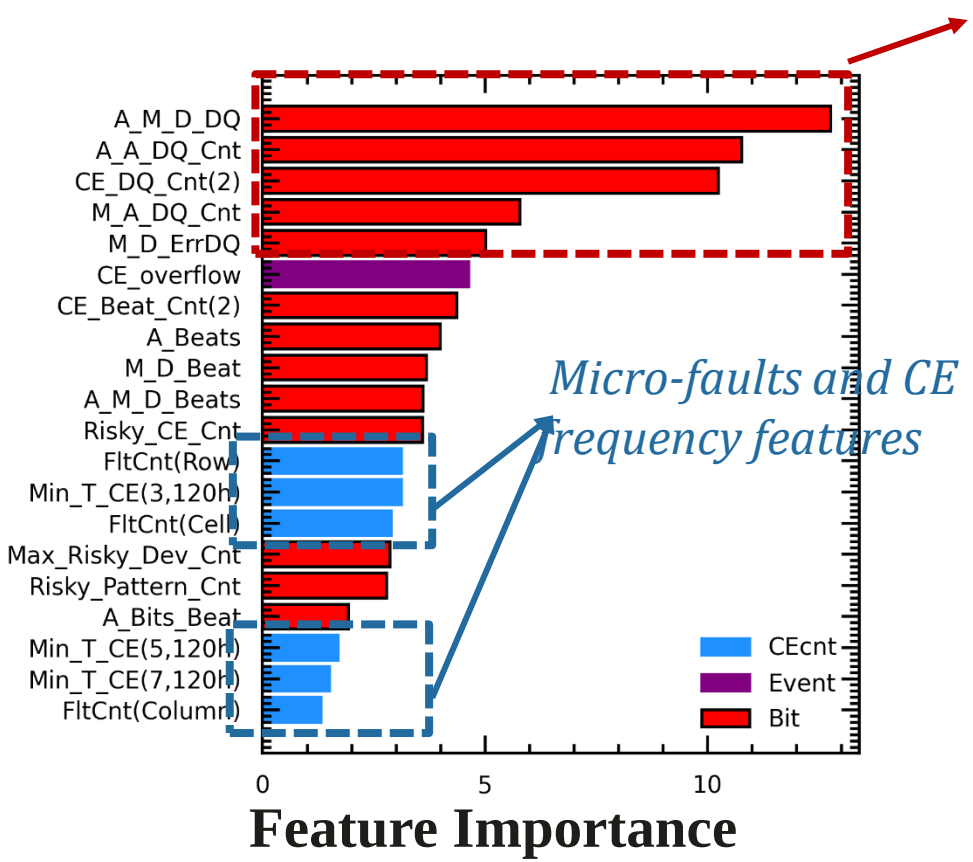
DIMM-level Evaluation

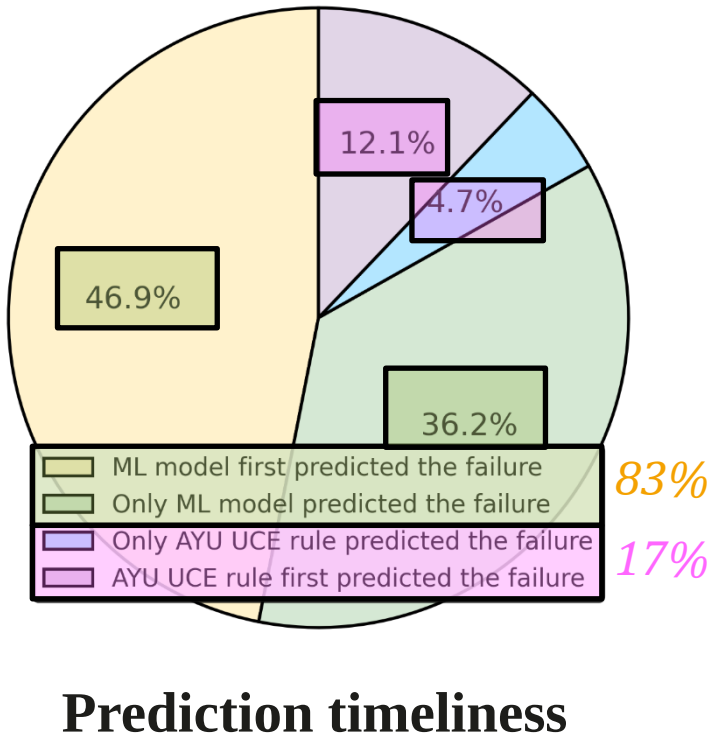


SUMMARY

Proposed Bit-level features and AYU UCE demonstrate significant generality across all algorithms, while LightGBM stands out as the best performance.

Model Analysis



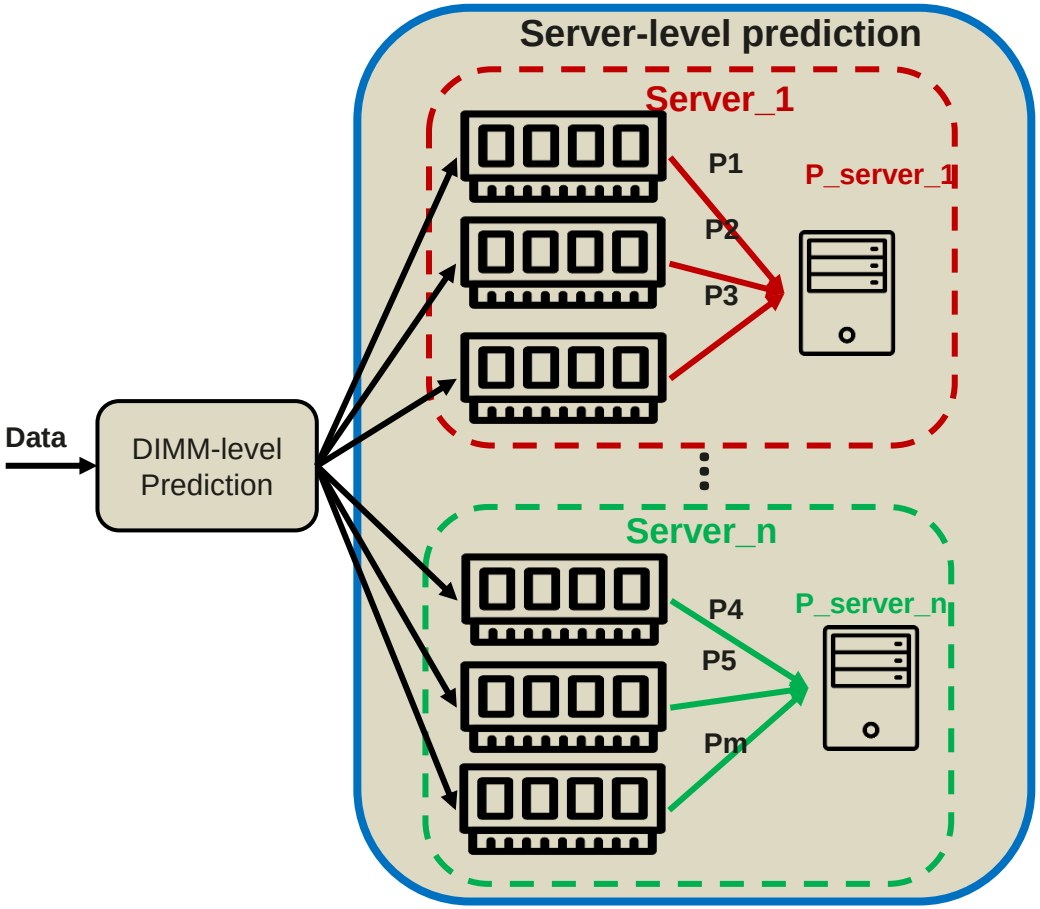


KEY OBSERVATION

Bit-level ones hold predominant importance among all features, while the combination of the ML model and AYU UCE showcases the best timeliness.

Server-level Prediction

Server-level memory failure prediction provides a comprehensive health assessment for all DIMMs in the server.



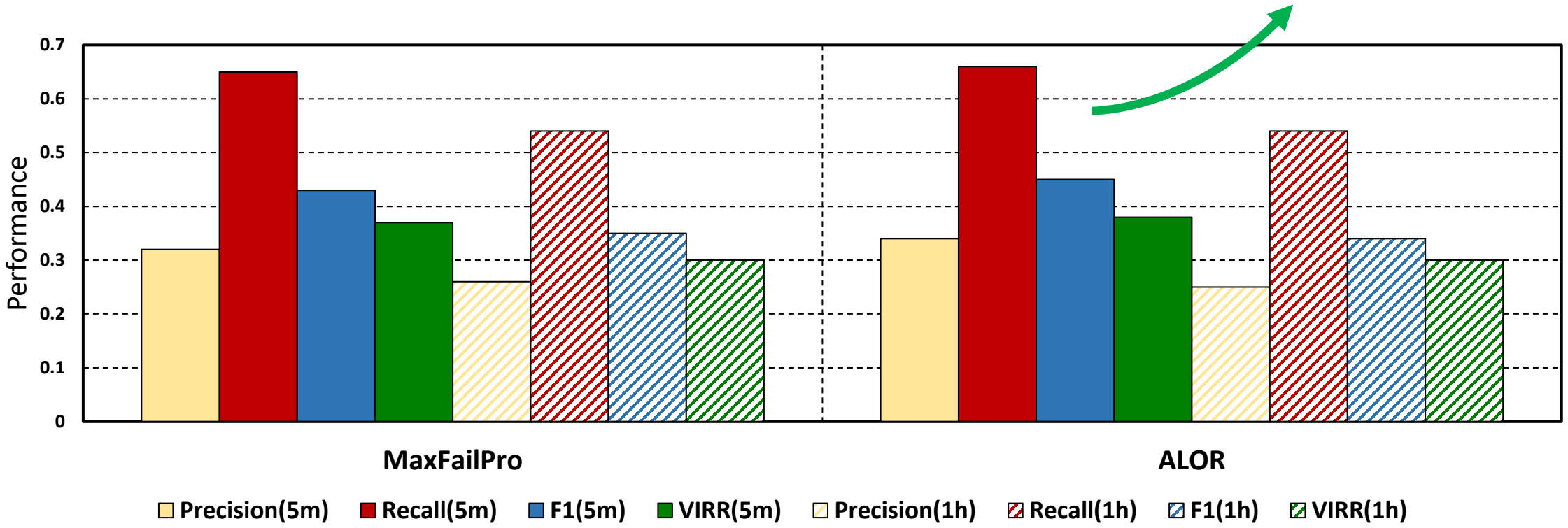
'Server_id': 210611xxxC000281, **'Advanced RAS configuration': {'Bank sparing': enabled, 'DIMM mirror': enabled}**, 'timestamp': 1653085843, 'alarm time': 2022-05-20 22:30:47, **'server-level probability': 0.72**,
{'DIMM_id': 163xxx5E, 'timestamp': 1653085801, 'probability': 0.40, 'faulty type': {bank fault, column fault, row fault}, Recovery status: {None}},
{'DIMM_id': 160xxxB3, 'timestamp': 1653085781, **'probability': 0.53**, **'faulty type': {bank fault, Multi-DQ fault, DIMM fault}**, **Recovery status: {Bank sparing: succeeded, DIMM mirror: succeeded}}**

(a)

'Server_id': 21061111xxx000053, 'Advanced RAS configuration': {'Banking Sparing': enabled}, 'timestamp':1652721781, 'alarm time': 2022-05-16 17:23:01, **'server-level probability': 0.56**,
{'DIMM_id': 1D4xxxA1, 'timestamp': 1652721763, **'probability': 0.25,** 'faulty type': {bank fault}, Recovery status: {None},
{'DIMM_id': 1F8xxx68, 'timestamp': 1652721721, **'probability': 0.21,** 'faulty type': {column fault, multi-DQ fault}, Recovery status: {None}},
{'DIMM_id': 206xxxAF, 'timestamp': 1652721724, **'probability': 0.26**, 'faulty type': {column fault, multi-DQ fault}, Recovery status: {None}}

(b)

Performance in Server-level Prediction



The ALOR approach performs better in the 5-minute lead time, while achieving similar results to the MaxFailPro prediction in the 1 hour lead time.

Conclusion and Insights

- **Conclusion:**
  - We introduce the **background of memory failures**, highlighting the various types and causes of failures that can occur in memory systems.
  - **Recent work on memory failure prediction** approaches has been reviewed, showcasing the advancements in predictive models and techniques.
  - We also present a **hierarchical memory failure prediction solution**, which integrates multiple levels of prediction to enhance accuracy and reliability.
- **Insights:**
  - **Observability:** The company also should focus on understanding the internal state or condition of a complex memory system based solely on knowledge of its external outputs.
  - **Open source:** Utilizing open datasets and code can significantly contribute to the growth of this field.
  - **ML and System intersection**: Collaboration between machine learning and system researchers is essential.

Agenda

Part 1. INTRODUCTION (Min Zhou, 13:10 - 13:30)

1. Reliability Challenges for Huawei Cloud in the AI era
2. Hardware Failure Prediction Progress in Huawei Cloud

Part 2. Memory Failure Prediction (Qiao Yu, 13:30 – 14:00)

1. Background of memory failure
2. Hierarchical memory failure prediction
3. Conclusion and future work

Part 3. HBM Failure Prediction and Reliable Storage System (Zhirong Shen, 14:00 – 14:30)

1. **Introduce analysis of HBM errors in the field**
2. **Introduce HBM failure prediction framework**
3. **Introduce some techniques for reliable storage**

Part 4. Smartmem Competition (Min Zhou, 14:30 – 15:00)

1. Overview of the SmartMem Competition
2. Attempts to unified memory prediction solution
3. Future work

Coffee Break (15:00 – 15:30)

Part 5. Hands-on Competition (15:30 – 17:00)

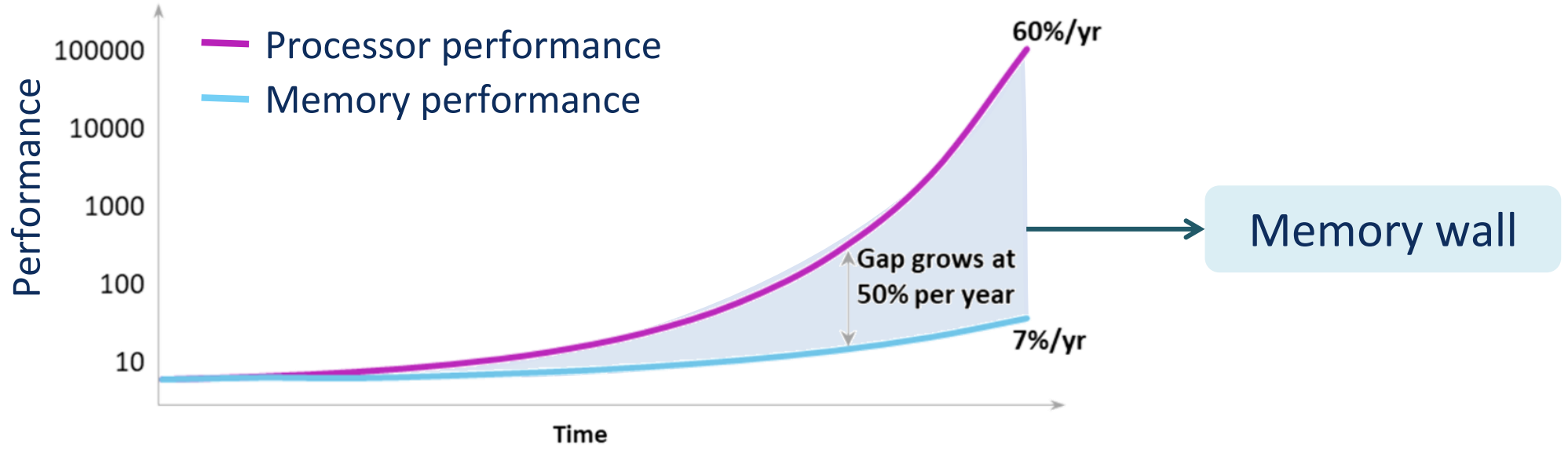
Outline

- Background
- Analyses and Findings
- Calchas: Hierarchical Prediction Framework
- Conclusion

Background: Memory Wall

The gap between computing power and memory bandwidth is continuously widening in modern systems[1]

- Processors are improving exponentially, but memory bandwidth is increasing slowly



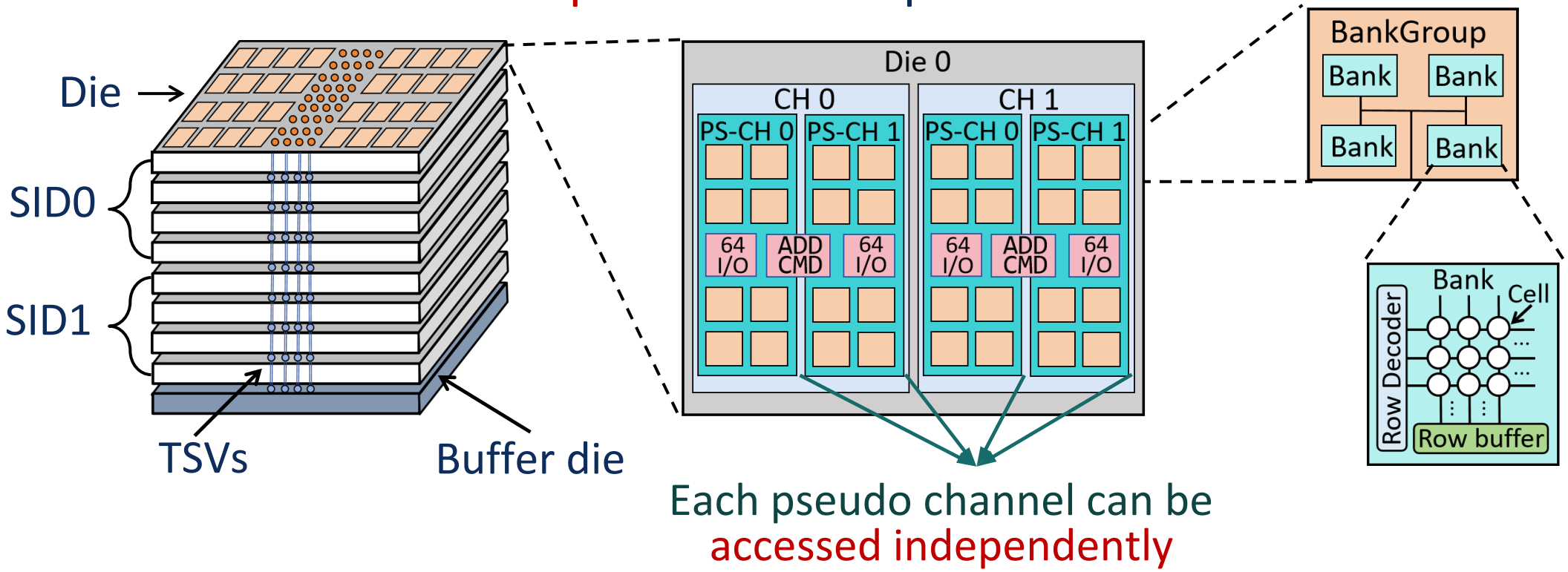
Memory wall becomes one of the major obstacles in training LLM models.

[1] Micron Inc., Micron's Perspective on Impact of CXL on DRAM Bit Growth Rate

Background: High-Bandwidth Memory

HBM is a hopeful technology to overcome the memory wall

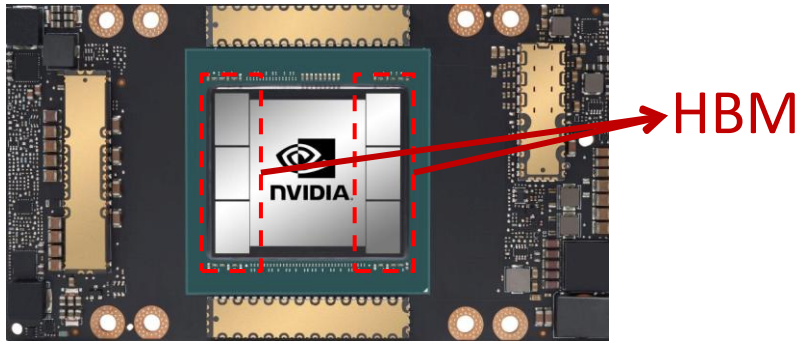
- Save massive physical space by stack vertically
- Offer significantly higher data transfer rates
- Introduce reduced power consumption



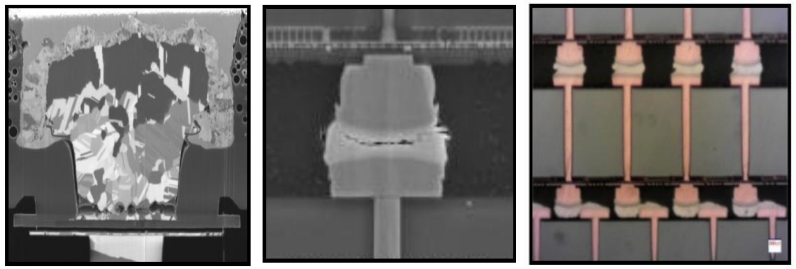
Background: HBM Errors

Disadvantages of HBM Memory

- Packed within a DSA devices ⟶ Less flexibility
- Complex manufacturing process ⟶ More special
- Equip weaker ECC ⟶ More errors



NVIDIA A100 GPU[2]



Example of various micro-bump joint failures in HBM[3]

Errors in HBM are more common and distinct compared to DRAM, but there is a lack of analysis on HBM errors in real production environments.

[2] NVIDIA A100 Tensor Core GPU. https://www.nvidia.com/en-us/data-center/a100/.
[3] Jun et al. HBM (High Bandwidth Memory) DRAM Technology and Architecture. Proc. of IMW, 2017.

Outline

- Background
- Analyses and Findings
- Calchas: Hierarchical Prediction Framework
- Conclusion

Analyses: Our Dataset

Over 15K DSA devices

- 60K+ HBM2 memory chips
- 460M+ errors
- 19 data centers: diverse services

Two years BMC log

- ErrLog_Cycle log: concise but fully recordable
- ErrLog_Occurrence log: detailed but difficult to fully retain
- Sensor log: temperature, power, etc.

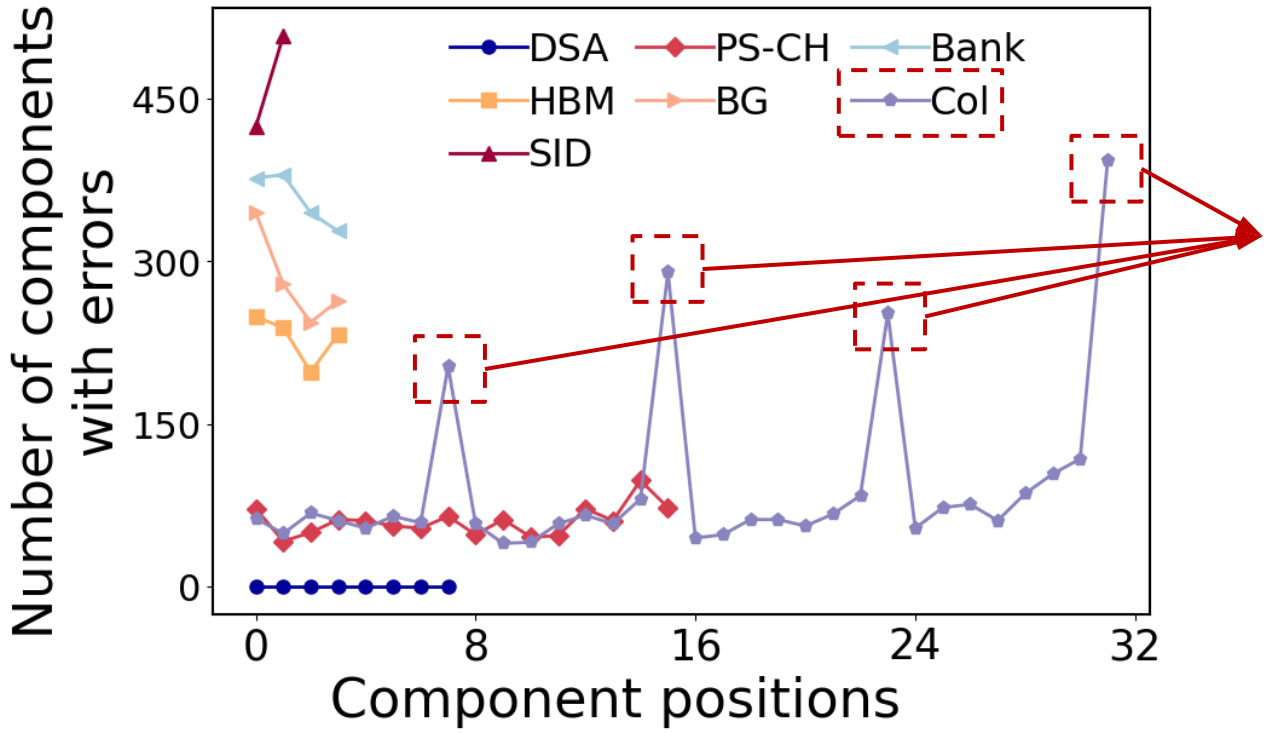
Test dataset released

- https://github.com/wrl297/Calchas

Analyses: Hierarchical Levels

Errors in different positions

- E.g., A BankGroup consists of four Banks, each located in different positions. Count the number of errors at each different bank position.

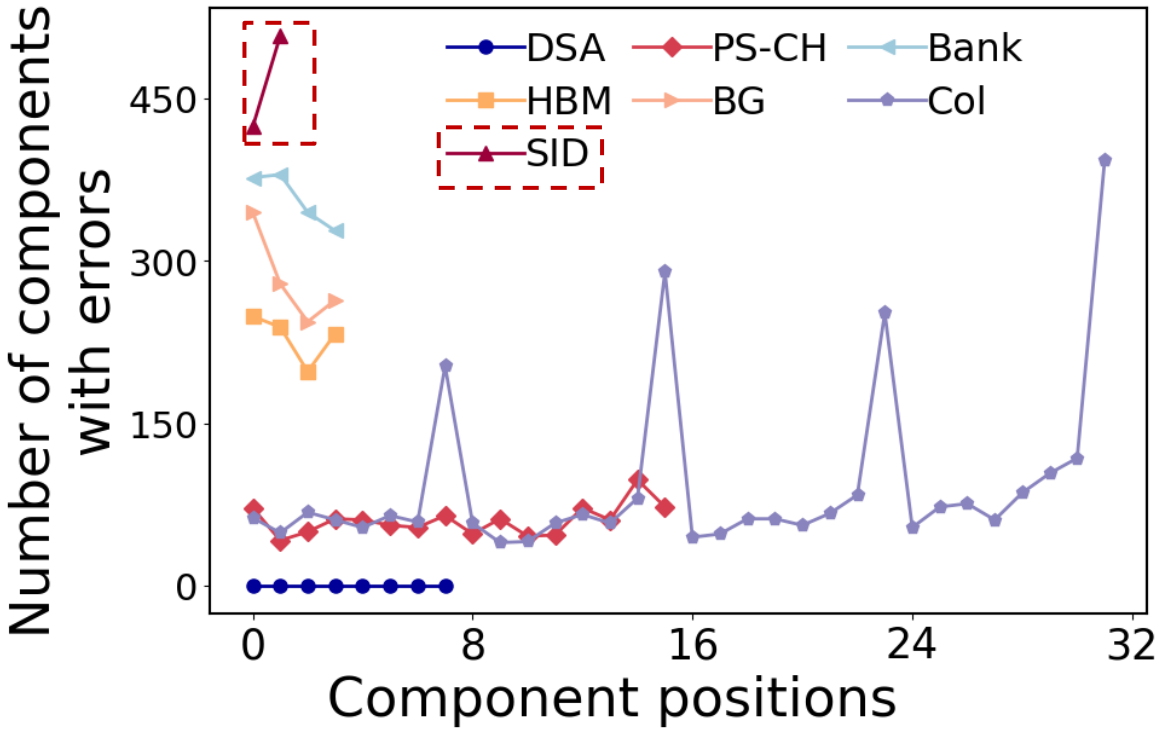


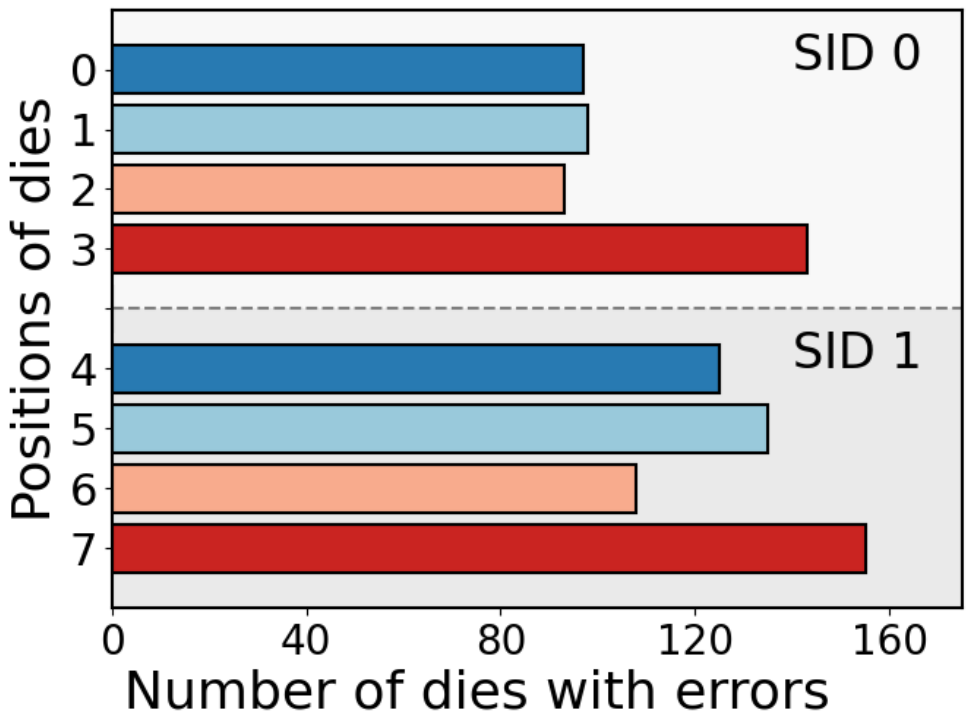
Finding: Errors occurring in the 7th, 15th, 23rd, and 31st columns are 338.4% higher on average than the errors in other column positions.

Root cause: The effects of **crosstalk** in the HBM may result in data loss

Analyses: Hierarchical Levels

Errors in different positions





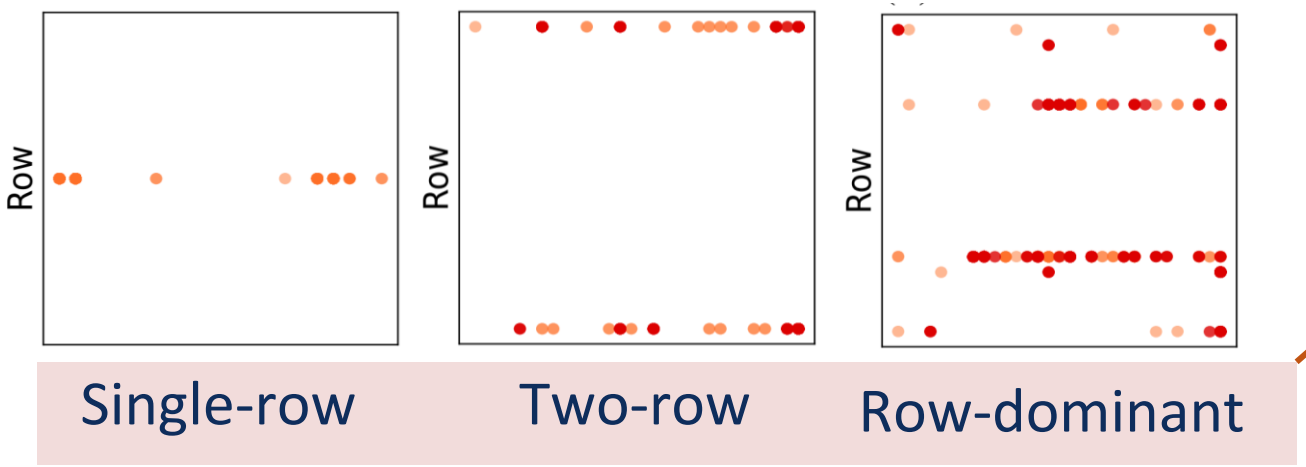
Top

Bottom

Finding: Lower SIDs (i.e., SID1) exhibit a higher susceptibility to errors due to the due to poor heat dissipation in the lower SID

Analyses: Error Modes

Typical error modes in a bank



Row-related modes (5.0%)

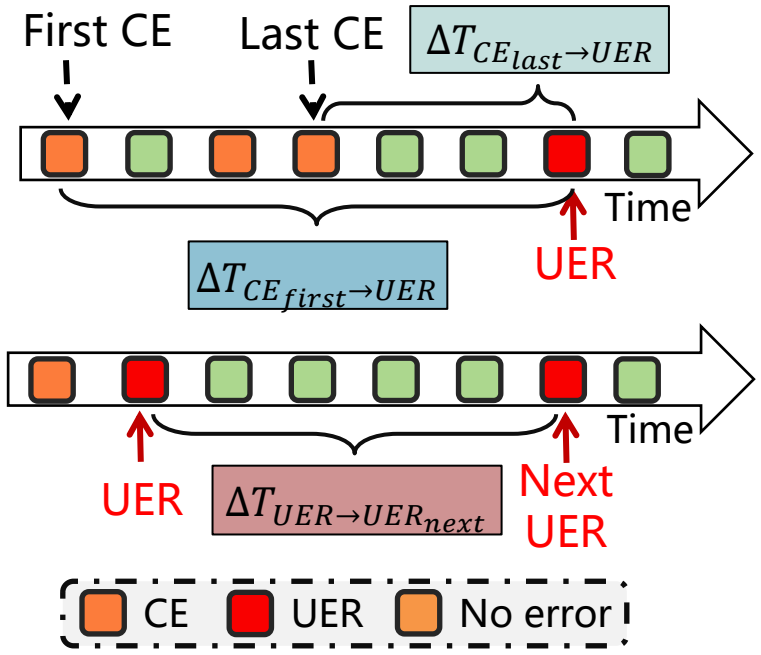
Column-related modes are more commonly observed in HBM, as opposed to the prevalent row-related modes in DRAM

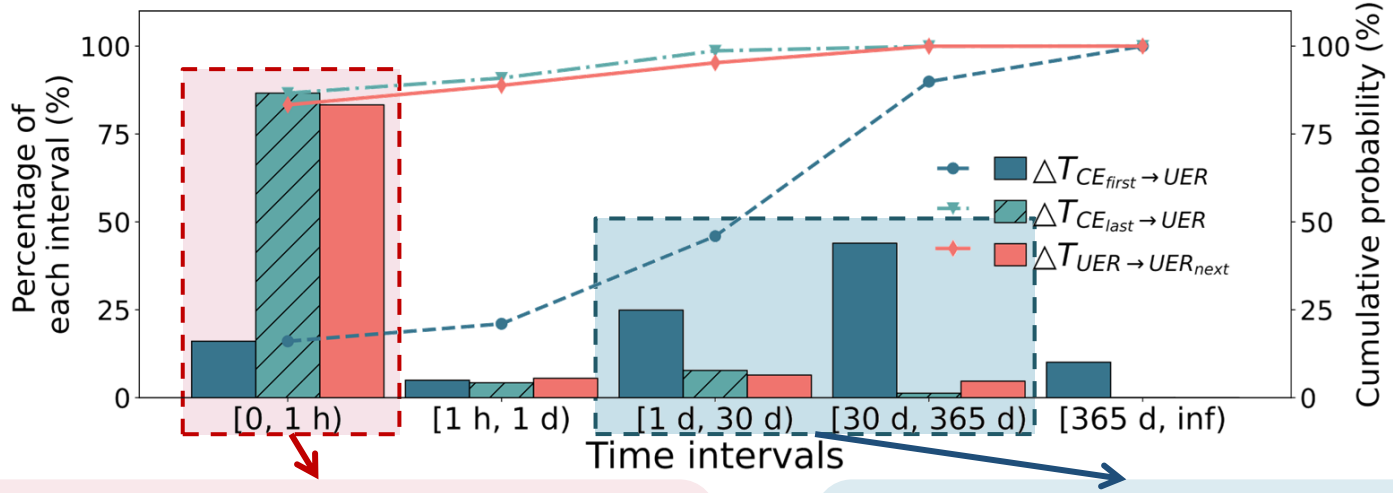
Col-related modes (29.22%)

Analyses: Time between Errors

Time interval

- $\Delta T_{CE_{first} \rightarrow UER}$: how soon will the first CE evolve to the first UER?
- $\Delta T_{CE_{last} \rightarrow UER}$: how much time remains to prevent crashes?
- $\Delta T_{UER \rightarrow UER_{next}}$: how long will the next UER occur once appearing a UER?





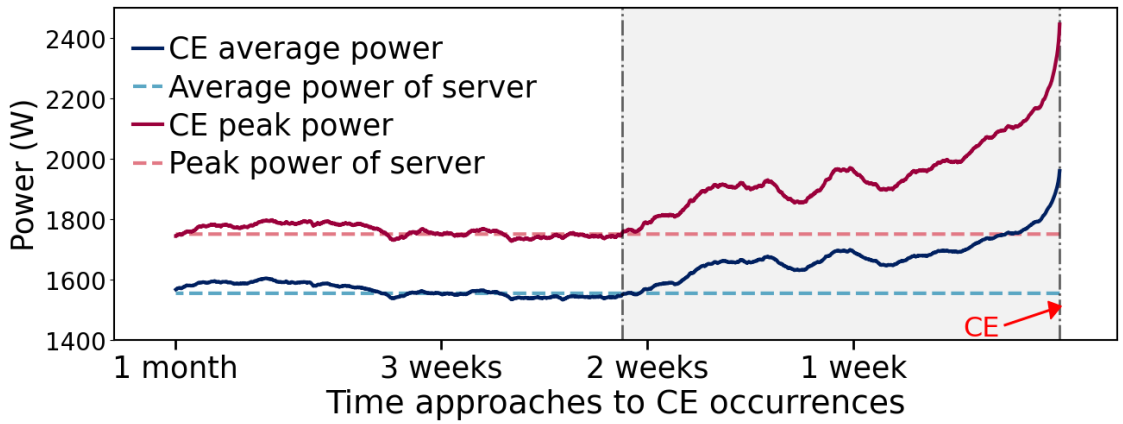
A significant probability that two successive UERs occur within one hour

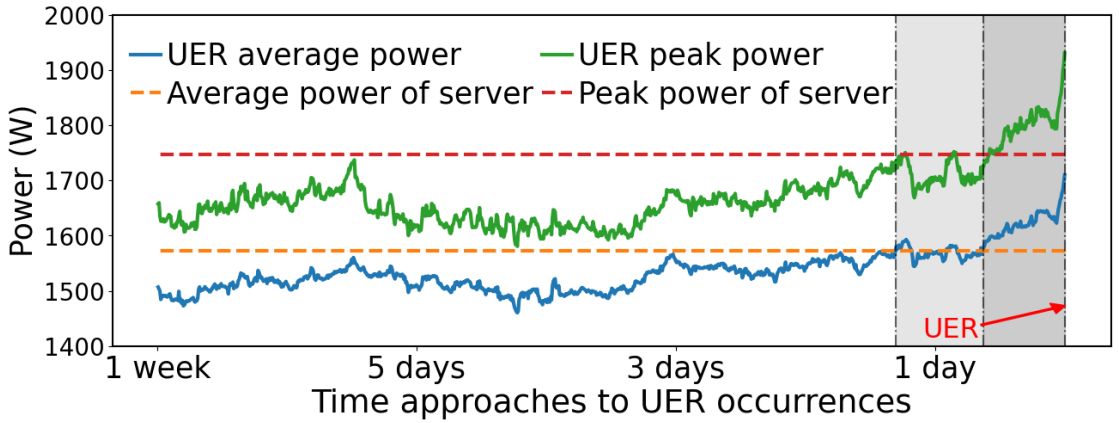
Bank may have encountered CEs long time ago

Analyses: Impact of Power

Power trend before errors

- UER/CE peak/average power: the peak and average power of last ten minutes
- Average and peak power of server: the average of all recorded peak and average power





The average and peak power both exhibit a rapid increase when approaching the error occurrence.

Analyses

- Dataset overview (§3.1)
- Errors in HBM exhibit strong spatial locality. (§3.2)
- Servers that have experienced CE storms may increase the probability of encountering UERs (§3.3)
- The temperature distribution of CEs and UERs exhibits significant differences (§3.4)

More details in the paper

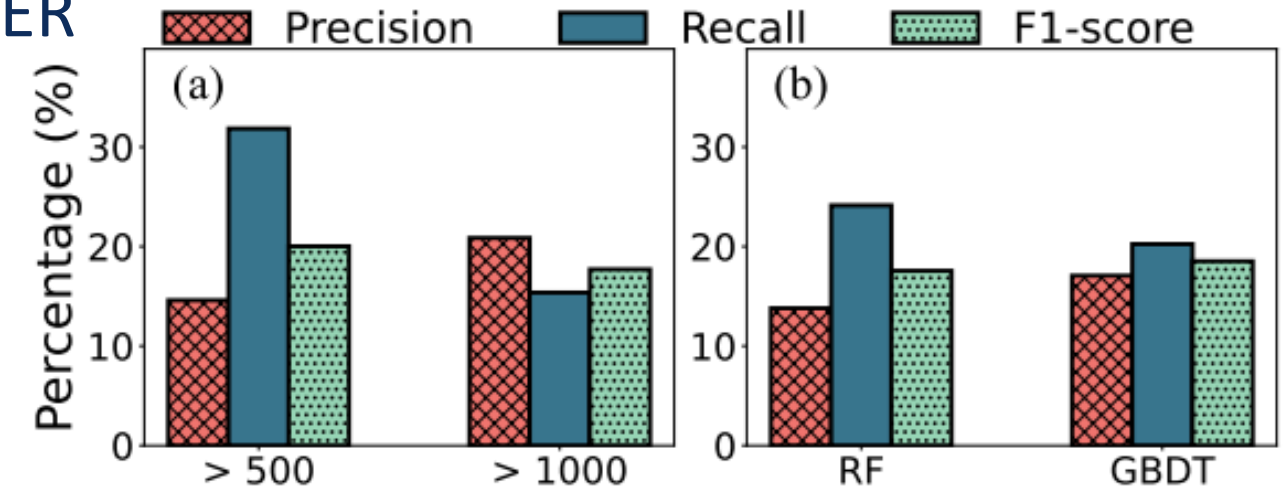
Outline

- Background
- Analyses and Findings
- Calchas: Hierarchical Prediction Framework
- Conclusion

Failed Attempts: CE-based Prediction

Is promise of using CE to predict upcoming UER in HBM?

- **CE rate indicator**[4]: consider a UER may occur when the number of CEs exceed the predefined threshold
- **CE-based predictor**[5]: utilize all the features related to CEs to predict future UER



How to conduct a HBM-specific predictor to improve the prediction performance?

[4] Du et al. Predicting uncorrectable memory errors for proactive replacement: An empirical study on large-scale field data. Proc. of EDCC, 2020.
[5] Boixaderas et al. Cost-aware prediction of uncorrected dram errors in the field. Proc. of SC, 2020.

Calchas: Design Guideline

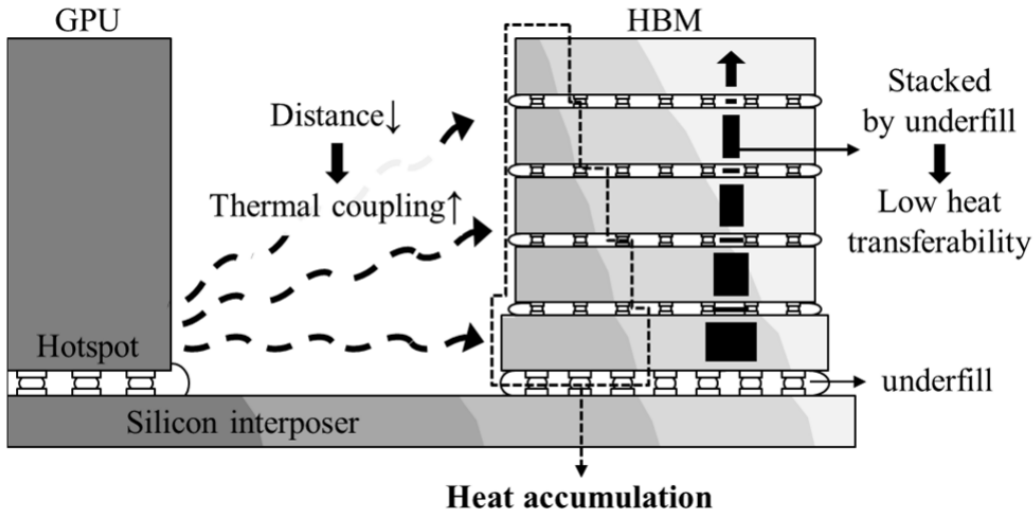
Key factors for designing an efficient predictor

- Which features should be utilized to enhance prediction performance?
  - Component features and statistical features?
- How can predictions be made to fully leverage the correlation between features and failures?
  - Micro-level, HBM-level or server-level?
- When is the appropriate time to make predictions?
  - Period-based or event-driven?

Calchas: Feature Generation

Which features should be utilized to enhance prediction performance?

- Component features: number of components experiencing errors
- Stack features: one-hot code of SID positions
- Sensor features: variation of both power and temperature



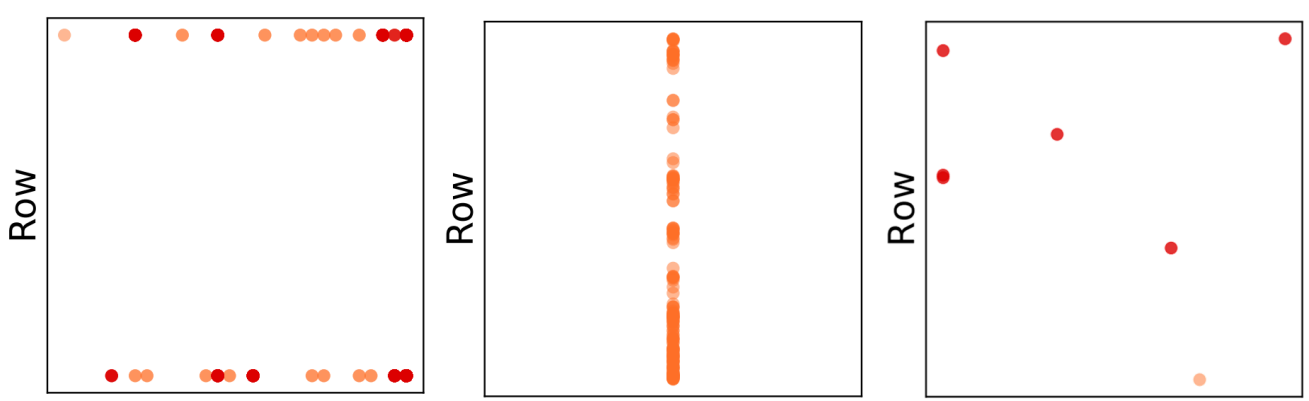
The 3D-stacked structure of HBM cause its characteristic (e.g., poor heat dissipation[6]) different from DRAM.

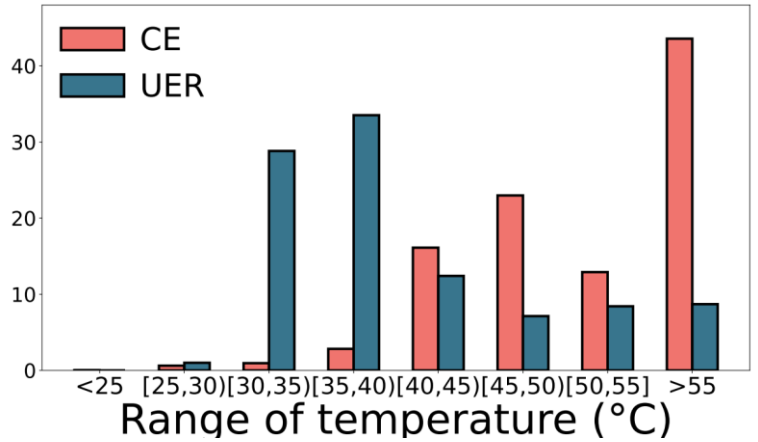
[6] Son et al. Thermal and Signal Integrity Co-Design and Verification of Embedded Cooling Structure With Thermal Transmission Line for High Bandwidth Memory Module. IEEE TCPMT, 1542-1556, 2022.

Calchas: Hierarchical Prediction

How can predictions be made to fully leverage the correlation between features and failures?

- Micro-level predictor: consider row-level, col-level, and bank-level prediction, respectively
- Server-level predictor: capturing the symptom of server failure





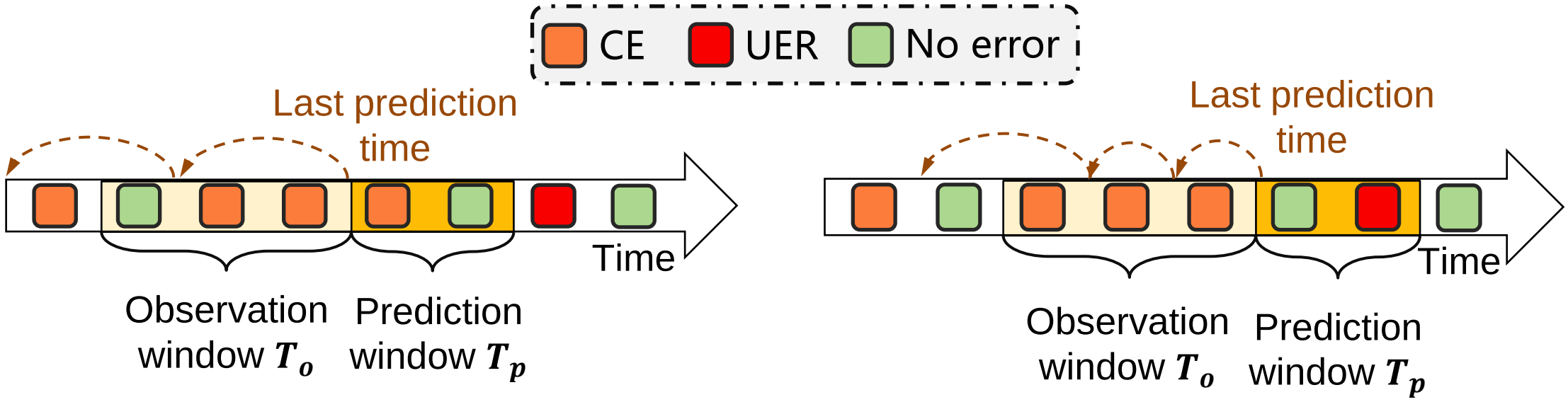
Errors may occur along a row, a column, or across in a bank.

Temperature is different before the UER occurrence.

Calchas: Prediction Timing

When is the appropriate time to make predictions?

- Period-based approach: perform **prediction every cycle**
- Event-driven approach: prediction **triggered by error events**



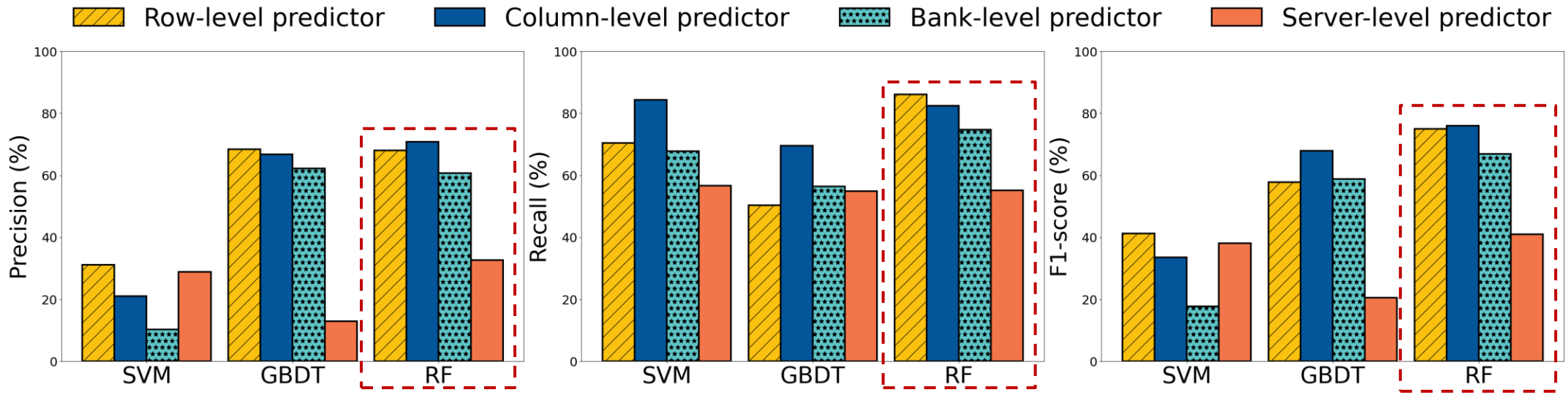
Benificial for regularly capturing the symptom ⇨ Server-level predictor

Provide a rapid response to errors ⇨ Micro-level predictor

Calchas: Prediction Performance

Experimental Setup

- Ratio of train and test datasets: 70% vs 30% based on timestamp
- Metrics: Presion, recall, and F1-score
- Models: SVM, GBDT, and RF (sklearn v0.24.2)



Based on **RF**, Calchas can achieve a highest precision of 58.0%, recall of 74.6%, and F1-score of 64.7% on average.

Outline

- Background
- Analyses and Findings
- Calchas: Hierarchical Prediction Framework
- Conclusion

Conlusion

Key Findings

- Lower SIDs (i.e., SID1) exhibit a higher susceptibility to errors
- Column-related error modes are prevalent in HBM
- Two successive UERs may occur within one hour
- Power increase when approaching the error occurrence
- ……

More details in the paper

Our Solutions: Calchas

- Combines features specific to HBM with those used in traditional DRAM
- Employs both period-based and event-driven methods to realize the elastic and adaptive prediction

Thanks & QA

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

XMU ERAS Research Group: https://xmusys.github.io/
Contact email: Ronglong Wu, rlwoo@stu.xmu.edu.cn

Agenda

Part 1. INTRODUCTION (Min Zhou, 13:10 - 13:30)

1. Reliability Challenges for Huawei Cloud in the AI era
2. Hardware Failure Prediction Progress in Huawei Cloud

Part 2. Memory Failure Prediction (Qiao Yu, 13:30 – 14:15)

1. Background of memory failure
2. Hierarchical memory failure prediction
3. Conclusion and future work

Part 3. HBM Failure Prediction and Reliable Storage System (Zhirong Shen, 14:15 – 15:00)

1. Introduce analysis of HBM errors in the field
2. Introduce HBM failure prediction framework
3. Introduce some techniques for reliable storage

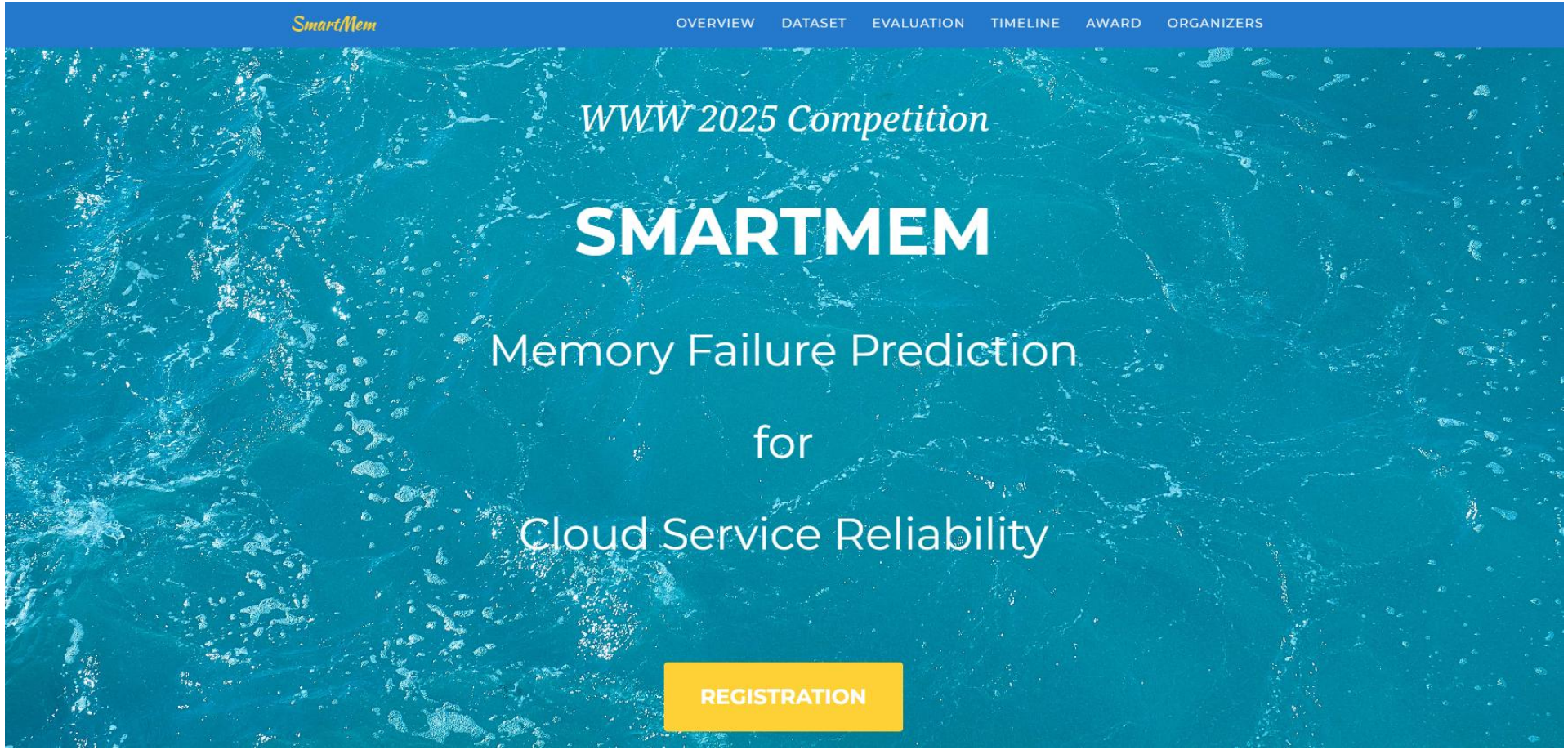
Part 4. Competition Highlights (Min Zhou, 15:00 – 15:30)

1. **Overview of the SmartMem Competition**
2. **Attempts to unified memory prediction solution**
3. **Future work**

Coffee Break (15:00 – 15:30)

Part 5. Hands-on Competition (15:30 – 17:00)

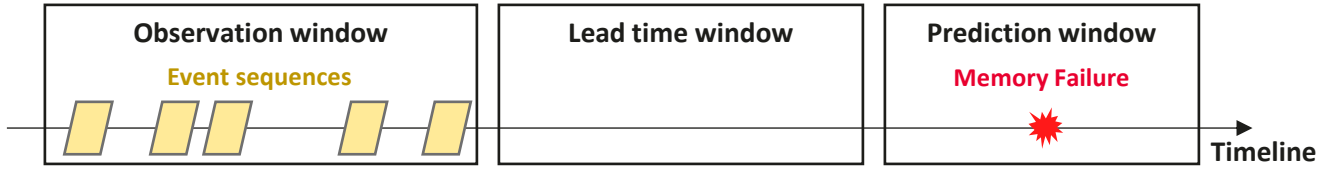
SmartMem Competition



Competition Overview

Task Description

- **Memory Failure Prediction (MFP):** Failure prediction based on **event sequences** requires predicting whether a **failure** will occur within a certain period in the future according to historical event information.



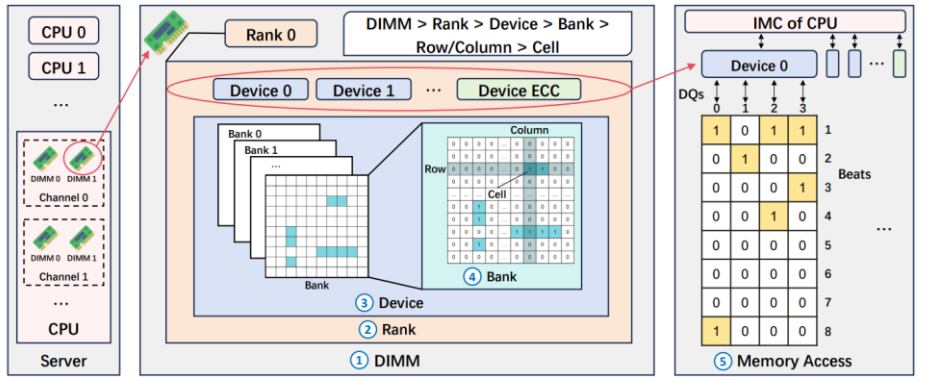
Schematic Diagram of the MFP Task

- **Challenges:**

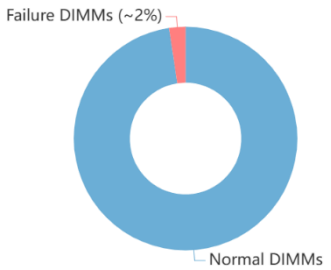
❗ **(1) Complex Event Data: The dataset** includes multi-level event information of the memory.

⚠️ **(2) Extremely Unbalanced Samples**: The proportion of faulty memory modules in the dataset is low.

🔧 **(3) Interference Caused by Hardware Differences**: The dataset contains hardware from different manufacturers and of different models.

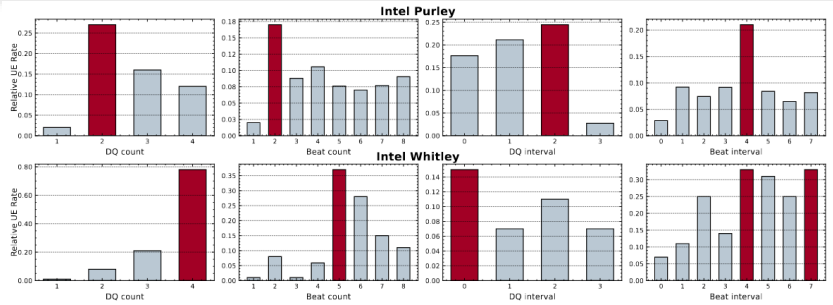


(1) The complex hierarchical organizational structure of the memory



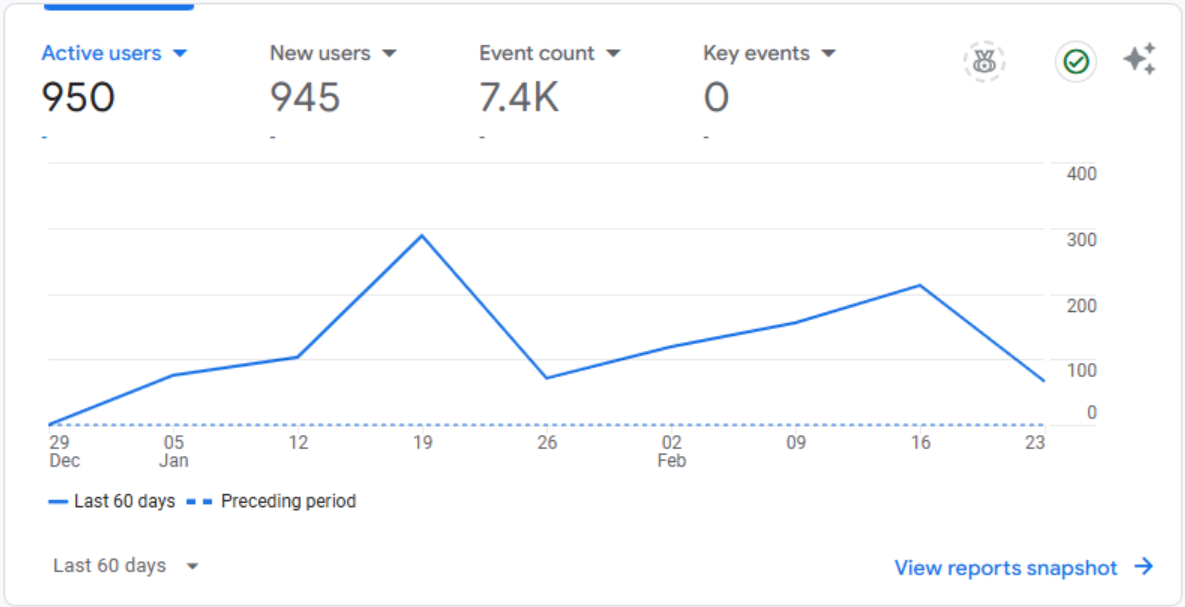
	DIMM count	Failure count	Event count
Intel Purley	65k	1.5k	1,120m
Intel Whitley	7k	0.2k	20m
All	72k	1.7k	1,140m

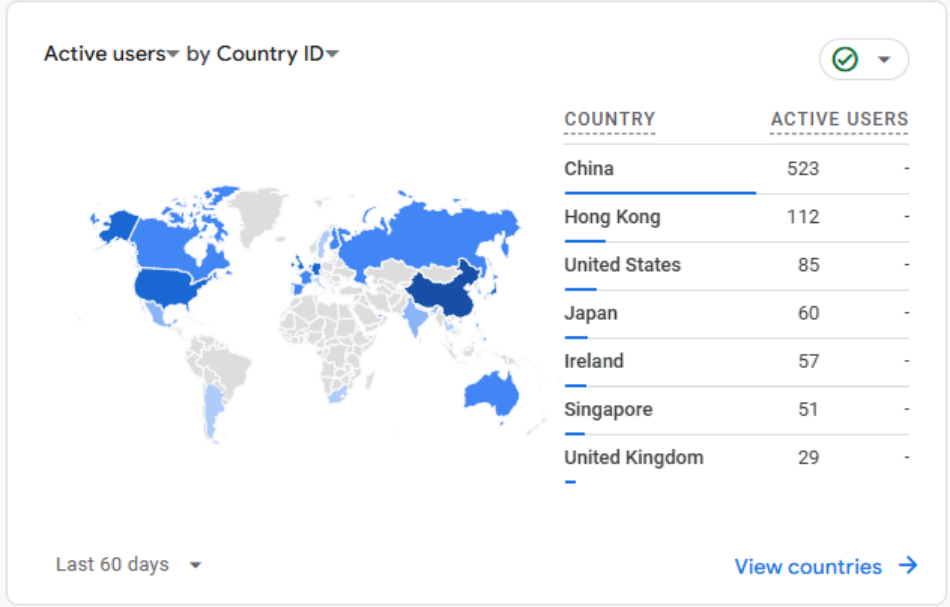
(2) Statistics of DIMM and Event counts



(3) Varied failure modes across different CPU architectures

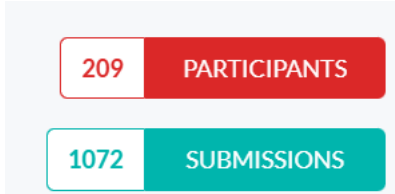
Competition Overview





Competition statistics

- SmartMem has attracted the attention of nearly 1,000 people from 30 countries and regions.
- Over 200 participants have signed up for SmartMem, with 1000+ submissions.



SmartMem Dataset

Large-Scale dataset from **Huawei Cloud datacenters**

- We collected datasets across CPU architectures including Intel Skylake/Cascade Lake and Icelake architectures servers**,** spanning for 9 month
  - Error logs w/ **static configuration, MCE log and memory events.**
  - Over 6,0000 DDR4 DIMMs and 1700 failures are collected.
  - First datasets that including the error correct information during memory access

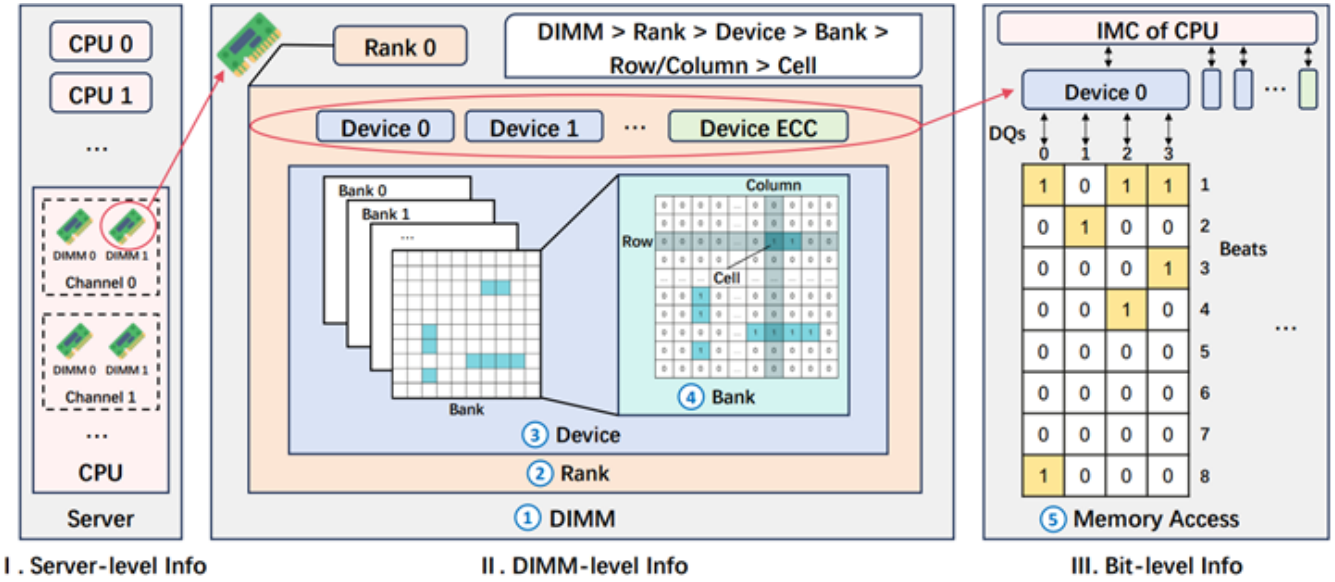


Table 6: Summary of DIMM and CE counts

Month	Intel Purley			Intel Whitley			all		
	DIMM count	Fault DIMM count	CE count	DIMM count	Fault DIMM count	CE count	DIMM count	Fault DIMM count	CE count
2024-01	26450	163	102423615	1466	15	1843056	27916	178	104266671
2024-02	24682	133	97336381	1682	8	1763660	26364	141	99100041
2024-03	26919	174	106745981	1948	16	1682764	28867	190	108428745
2024-04	29776	144	107168332	2067	17	1947564	31843	161	109115896
2024-05	29046	155	119407409	2191	12	2494964	31237	167	121902373
2024-06	32610	169	118064609	2270	16	2580721	34880	185	120645330
2024-07	37853	250	148794345	2564	19	3604404	40417	269	152398749
2024-08	37531	214	165111964	2631	18	3564878	40162	232	168676842
2024-09	37613	160	158444298	2668	21	3296809	40281	181	161741107
Total	64794	1562	1123496934	7175	142	22778820	71969	1704	1146275754

SmartMem-baseline

- With features that commonly used for memory failure prediction
- The relatively low performance stems from a combination of noise inherent in the production environment, the intricate mechanism underlying the UE and the severe class imbalance
- Stage 2's slightly lower F1-score than stage 1 demonstrates performance degradation in online application

Feature Categories

Category	Feature Name	Type
Temporal	Reported CE count	Numerical
	Error CE count	Numerical
	CE time interval	Numerical
	CE storm count	Numerical
Spatial (Macro-level)	Cell fault count	Numerical
	Row fault count	Numerical
	Column fault count	Numerical
	Bank fault count	Numerical
	Device fault count	Numerical
Spatial (Micro-level)	Error DQ/Beat count	Numerical
	Error DQ/Beat interval	Numerical
	Total error bits count	Numerical

Baseline performance

	Precision	Recall	F_1-score
Stage 1	0.1454	0.3501	0.2055
Stage 2	0.1252	0.3500	0.1845

- Starterkit

https://github.com/hwcloud-RAS/SmartHW/tree/main/competition_starterkit

Competition Link

Agenda

Part 1. INTRODUCTION (Min Zhou, 13:10 - 13:30)

1. Reliability Challenges for Huawei Cloud in the AI era
2. Hardware Failure Prediction Progress in Huawei Cloud

Part 2. Memory Failure Prediction (Qiao Yu, 13:30 – 14:15)

1. Background of memory failure
2. Hierarchical memory failure prediction
3. Conclusion and future work

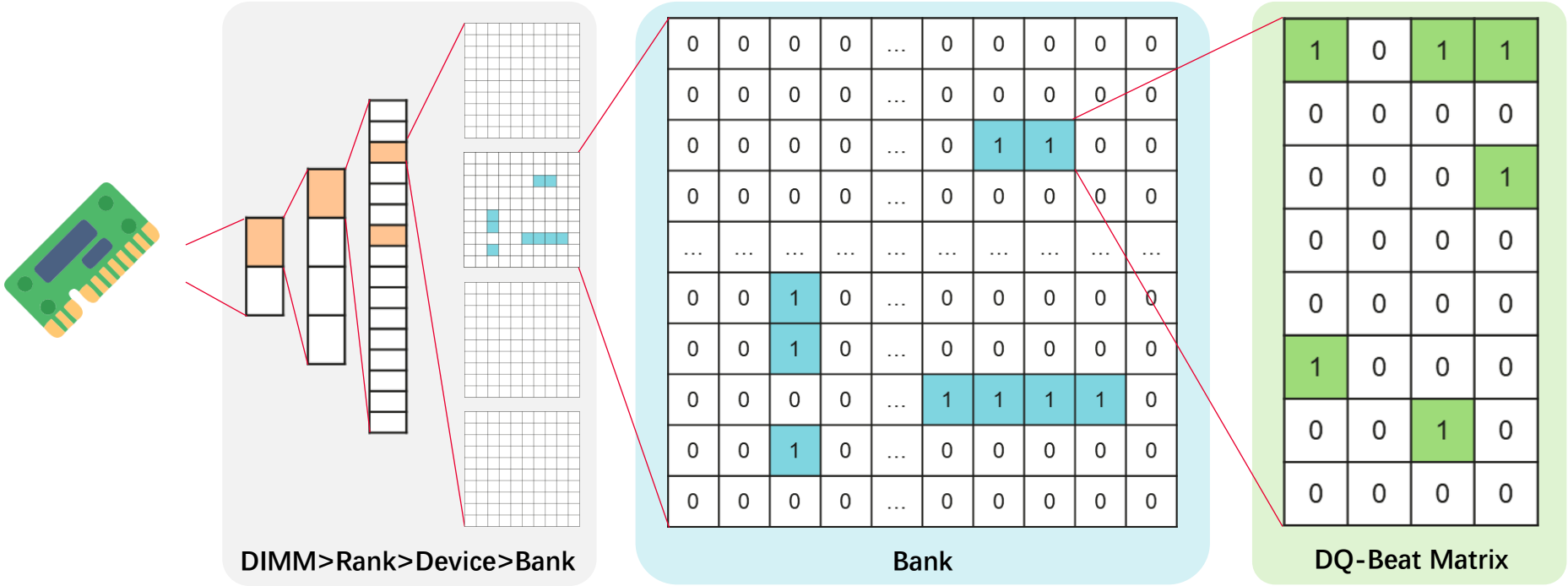
Part 3. HBM Failure Prediction and Reliable Storage System (Zhirong Shen, 14:15 – 15:00)

1. Introduce analysis of HBM errors in the field
2. Introduce HBM failure prediction framework
3. Introduce some techniques for reliable storage

Part 4. Competition Highlights (Min Zhou, 15:00 – 15:30)

1. Overview of the SmartMem Competition
2. Attempts to unified memory prediction solution
3. Future work

Motivation



Multi-level

LogTime	*Cell Position*	*DQ-Beat Info*
1718659391	xxx	xxx
1718659396	xxx	xxx
1718659460	xxx	xxx
1718659465	xxx	xxx
1718659481	xxx	xxx
1718659486	xxx	xxx
1718659491	xxx	xxx

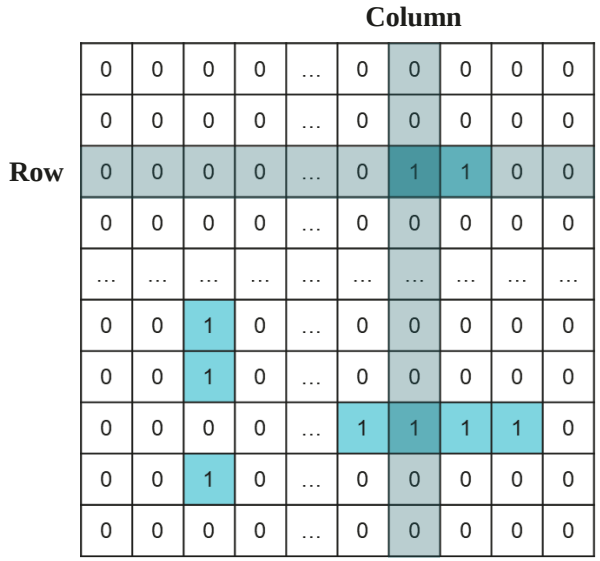
Multi-scale

- A DIMM generates **a large number of event records**, with each event corresponding to one **blue cell** and one **green matrix**.

We need a Multi-scale & Multi-Level MFP Framework!

- **M²-MFP: A Multi-Scale and Multi-Level MFP Framework.** To automatically extracts and organically aggregates multi-level features to establish a MFP model across multiple scales, achieving high performance, high efficiency, universality, and strong generalization ability.

A Multi-Scale and Multi-Level MFP Framework



1d-BSFE for one-dimensional matrix

Guided by 3 principles

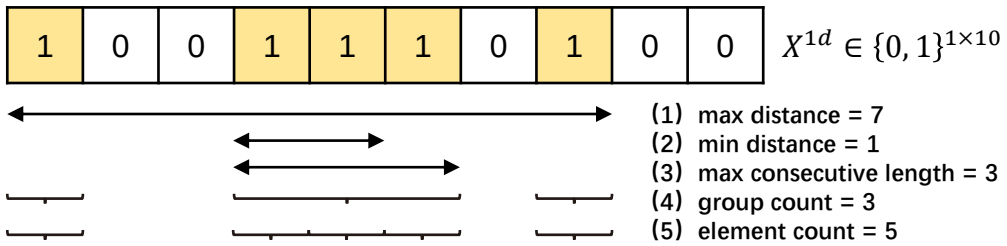
Symmetry Generality Sensitivity

Composed by 2 paths

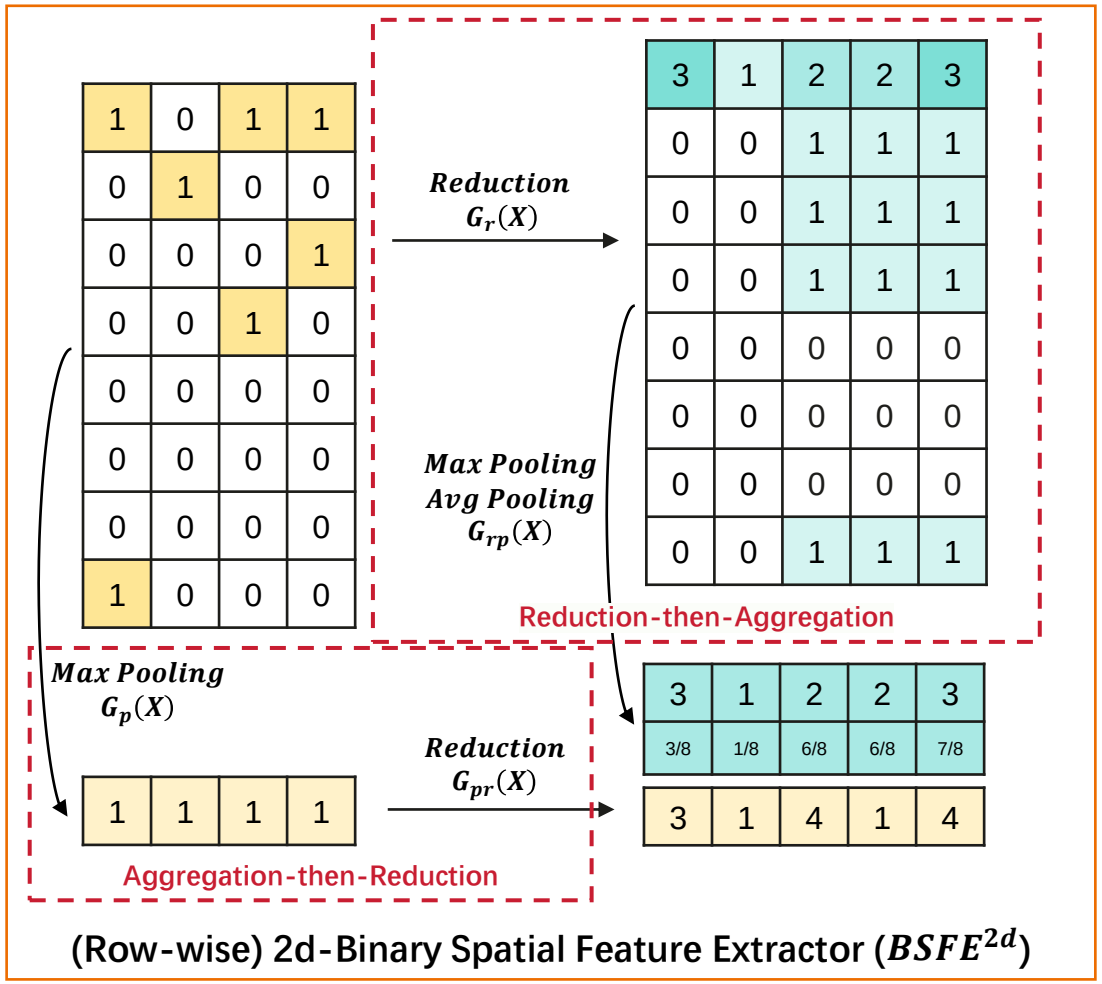
Reduction-then-Aggregation

Aggregation-then-Reduction

2d-BSFE for two-dimensional matrix



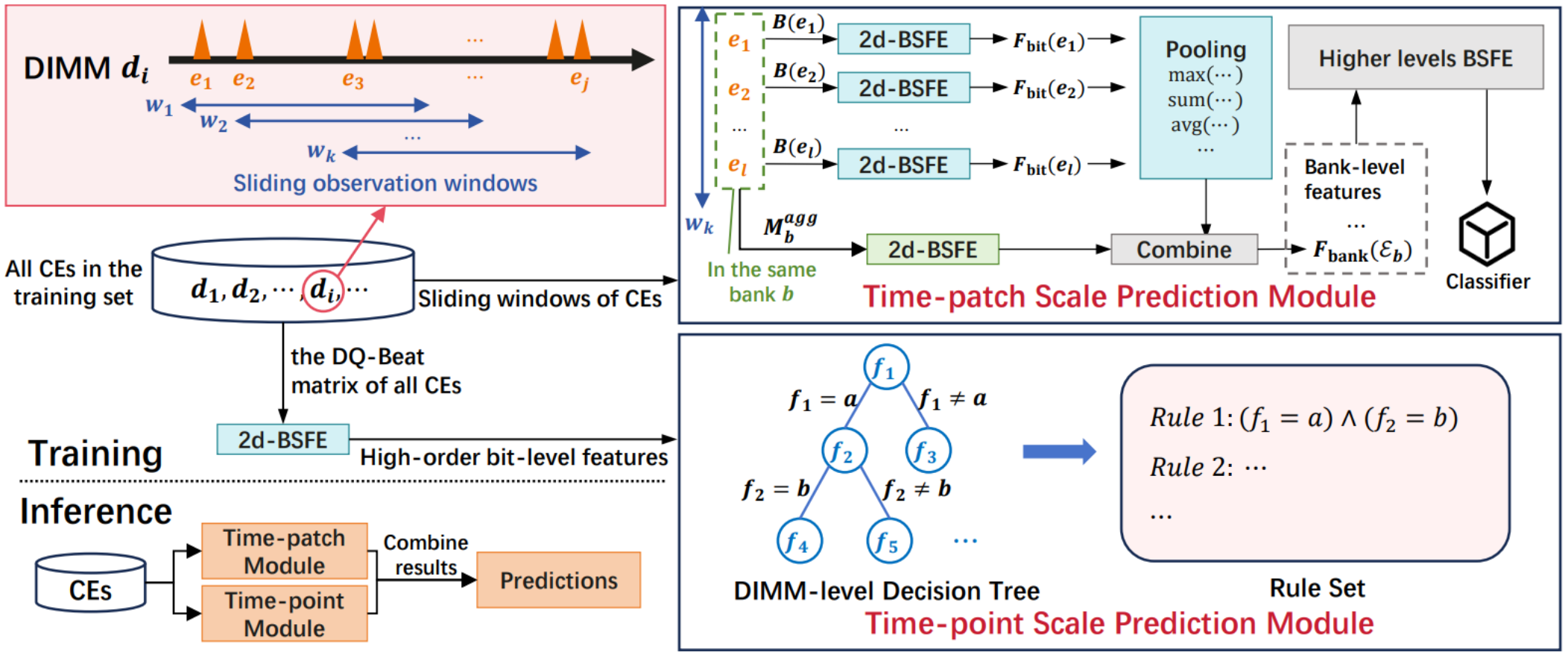
1d-Binary Spatial Feature Extractor ($BSFE^{1d}$)



(Row-wise) 2d-Binary Spatial Feature Extractor ($BSFE^{2d}$)

- BSFE provides a solid foundation for memory fault analysis and prediction.
  - **1D-BSFE** effectively extracts potential fault information from events.
  - **2D-BSFE** efficiently accomplishes feature aggregation.

A Multi-Scale and Multi-Level MFP Framework



- Based on BSFE, we propose a two-scale fault prediction model.
  - **Time-patch Scale Prediction Module:** Utilizes multi-level BSFE to process and extract features from hierarchical event information.
  - **Time-point Scale Prediction Module**: Leverages bit-level features extracted by BSFE to train a custom DIMM-level decision tree.

Performance

- **M²-MFP** achieves sota performance by effectively leveraging its multi-scale hierarchical architecture for enhanced predictive capability.

Table 2: Performance comparison among different memory failure prediction models

Category	Method	Intel Purley			Intel Whitley			all		
		Precision	Recall	F_1-score	Precision	Recall	F_1-score	Precision	Recall	F_1-score
Time-point	Naive	0.2085	0.2444	0.2250	0.0698	0.0417	0.0522	0.2024	0.2269	0.2140
	Risky CE	0.0476	0.4875	0.0868	0.1088	0.3750	0.1686	0.0494	0.4778	0.0895
	DQ Beat Predictor	0.0487	0.5059	0.0888	0.1020	0.3750	0.1603	0.0503	0.4946	0.0913
	CNN	0.0795	0.2168	0.1164	0.1681	0.2917	0.2133	0.0836	0.2185	0.1209
	Time-point Ours	0.4047	0.2378	0.2996	0.3529	0.0833	0.1348	0.4029	0.2245	0.2883
Time-patch	CNN	0.0796	0.2838	0.1244	0.1053	0.2778	0.1527	0.0811	0.2689	0.1246
	CNN (1D kernal)	0.0856	0.2799	0.1311	0.1563	0.1389	0.1471	0.0870	0.2665	0.1311
	ViT	0.0596	0.0302	0.0401	0.2308	0.0556	0.0896	0.0516	0.0360	0.0424
	ViT (1D patch)	0.0759	0.3798	0.1265	0.1197	0.1944	0.1481	0.0771	0.3649	0.1273
	STIM	0.0546	0.0644	0.0591	0.0641	0.1389	0.0877	0.0560	0.0708	0.0626
	Himfp	0.1790	0.3246	0.2307	0.1918	0.1944	0.1931	0.1806	0.3097	0.2282
	Time-patch Ours	0.3436	0.3022	0.3216	0.2542	0.2222	**0.2372**	0.3446	0.2893	0.3145
Combined	M²-MFP	0.3344	0.3942	**0.3619**	0.2500	0.2222	0.2353	0.3208	0.3942	**0.3537**

- **Rule-based methods** exhibit high false positive rates in the results.
- **Deep learning methods** struggle to extract features effectively, resulting in poor performance.
- The **Time-point** and **Time-patch** modules outperform their counterparts.
- **M²-MFP** achieves state-of-the-art performance by leveraging its multi-scale hierarchical architecture.

Table 3: Time-patch Module Ablation Results

Method	Precision	Recall	F_1-score
w/o Reduct. then Aggreg.	0.3286	0.2725	0.2979
w/o Aggreg. then Reduct.	0.3563	0.2461	0.2911
w/o DIMM-level features	0.2980	0.2725	0.2847
w/o bit-level features	0.3744	0.1849	0.2475
Time-patch Ours	0.3446	0.2893	**0.3145**

- **Dual-path** feature extraction and **multi-level** feature integration are necessary.

Table 4: Time-point Module Ablation Results

Method	Precision	Recall	F_1-score
LightGBM	0.0916	0.3349	0.1439
XGBoost	0.0951	0.3373	0.1484
FTtransformer	0.0664	0.2749	0.1069
Decision Tree (Gini)	0.1036	0.3433	0.1591
Time-point Ours	0.4029	0.2245	**0.2883**

- **DIMM-level Decision Tree** achieves the best performance compared to these alternatives.

Future Work – Loss Function

- MFP aims to forecast **DIMM failures**, but since a DIMM maps to multiple samples during inference, **sample-level** loss functions fail to accurately represent **DIMM-level** predictions.
- Let the set of DIMMs be S. For $s \in S$, the sample set of s in the training set is M_s.

DIMM-level Loss

$$\mathcal{L} = -\frac{1}{|S|}\sum_{s \in S}[y_s \log(\hat{y_s}^{+}) + (1 - y_s)\log(\hat{y_s}^{-})]$$

Sample-level Loss

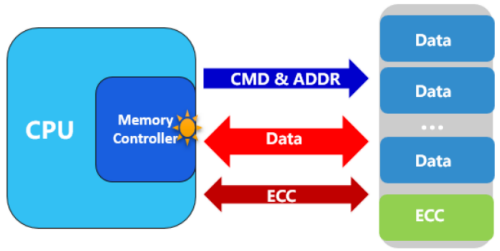
$$\mathcal{L} = -\frac{1}{\sum_{s \in S}|M_s|}\sum_{s \in S}\sum_{m \in M_s}[y_{s,m} \log(\widehat{y_{s,m}}^{+}) + (1 - y_{s,m})\log(\widehat{y_{s,m}}^{-})]$$

Sample Index	1	2	3	4	5
Label	0	0	1	0	1
If Failure	TP = 2, FP = 3				
No Failure	TP = 3, FP = 2				

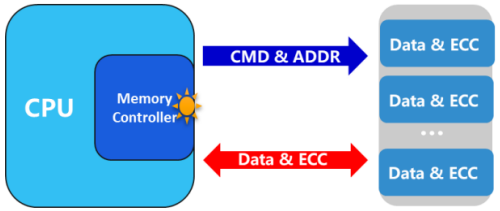
DIMM Index	s
Label	1
If Failure	TP = 1, FP = 0
No Failure	TP = 0, FP = 1

- How to apply **DIMM-level** loss in a **Sample-level** prediction scenario becomes a problem, and Multi-Instance Learning (MIL) may be applicable in this context.

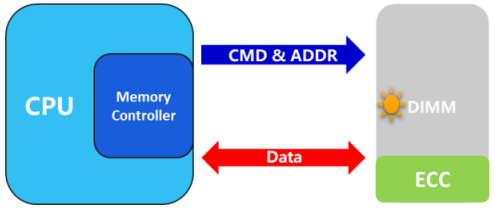
Future Work – Risky Parity Pattern



Side-band ECC



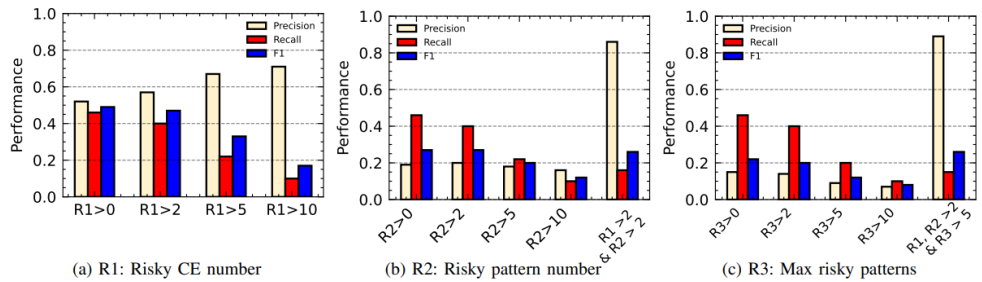
Inline ECC



On-die ECC

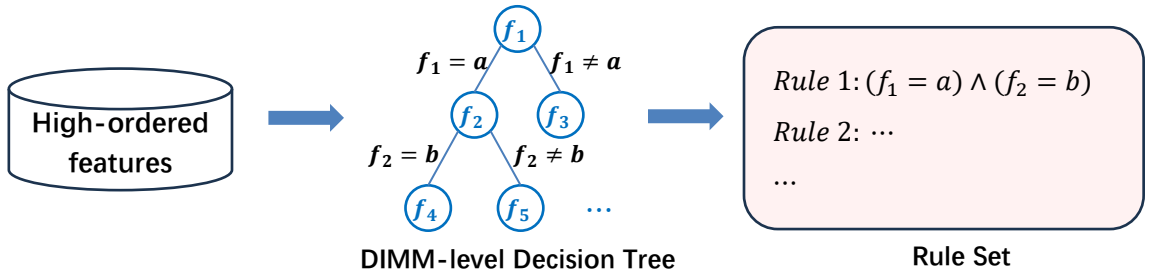
1	0	1	1
0	0	0	0
0	0	0	1
0	0	0	0
0	0	0	0
1	0	0	0
0	0	1	0
0	0	0	0

Is it Risky Parity?



(a) R1: Risky CE number (b) R2: Risky pattern number (c) R3: Max risky patterns

Statistical analysis to find risky CE patterns



DIMM-level Decision Tree

Rule Set

Custom Decision Tree to find risky CE patterns

- **Stage 1:** Design rules based on expert knowledge according to the mechanism of specific ECC.
- **Stage 2:** Design rules based on experience according to statistical laws.
- **Stage 3:** Utilize general models to mine general-purpose models applicable to universal datasets.

Q&A

Hands-on Online Competition