

Research Summary

Yuan-Hao Chang

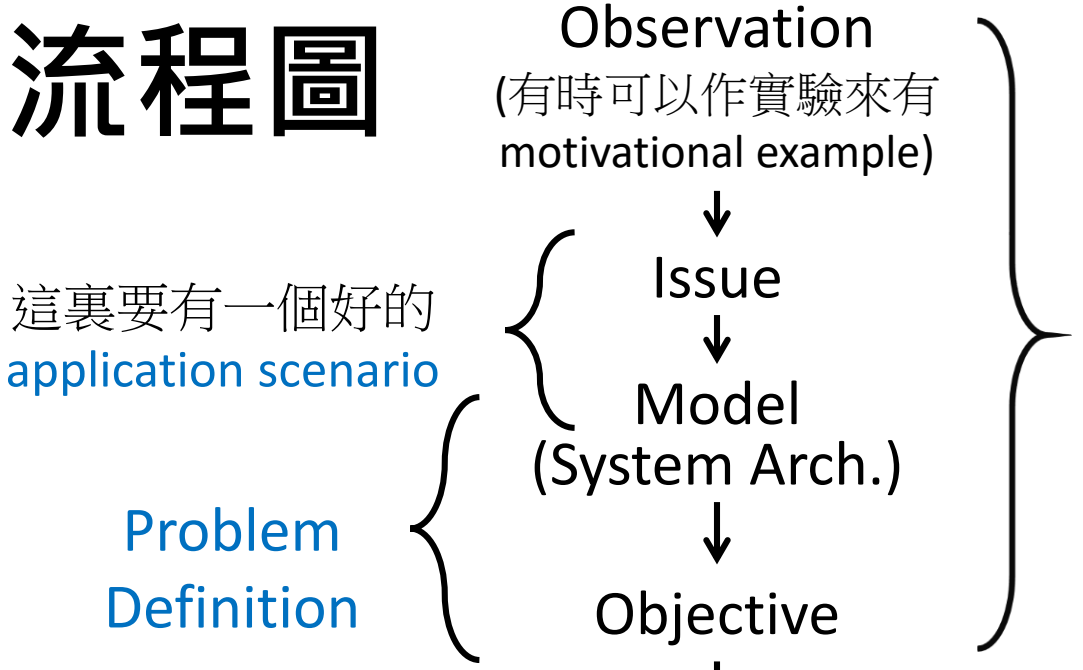
Deputy Director / Research Fellow / Professor

Institute of Information Science,
Academia Sinica

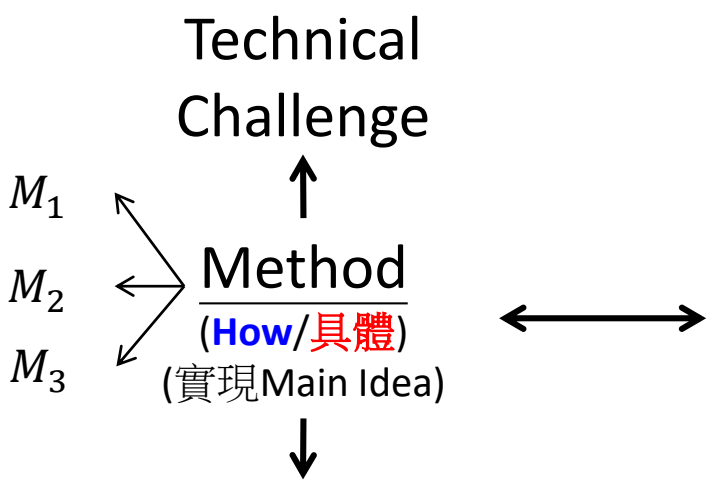
Research Interests

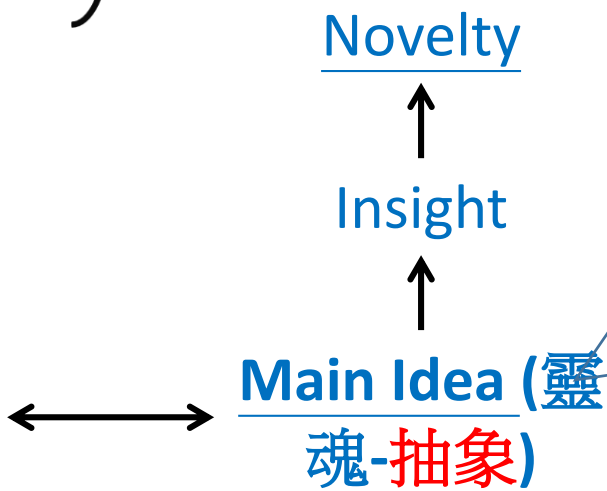
- In/Near Memory Computing
- In-Storage Computing
- Emerging Memory Technologies
- Non-volatile Memories
- Memory/Storage Systems
- Embedded Systems
- Operating Systems

新手論文流程圖



Method (具體)	↔	Main Idea (抽象)
↓		↓
Technical Contribution		Novelty





沒有Main Idea的Method就如行屍走肉，沒有Insight也沒有Novelty

最後用實驗 evaluate 所提 Method，從可檢視 objective 的metrics 開始，再到方法細節 insight的 metrics

↑

可以從中做 analysis，並且 identify technical contribution

Guideline of One-Page Summary

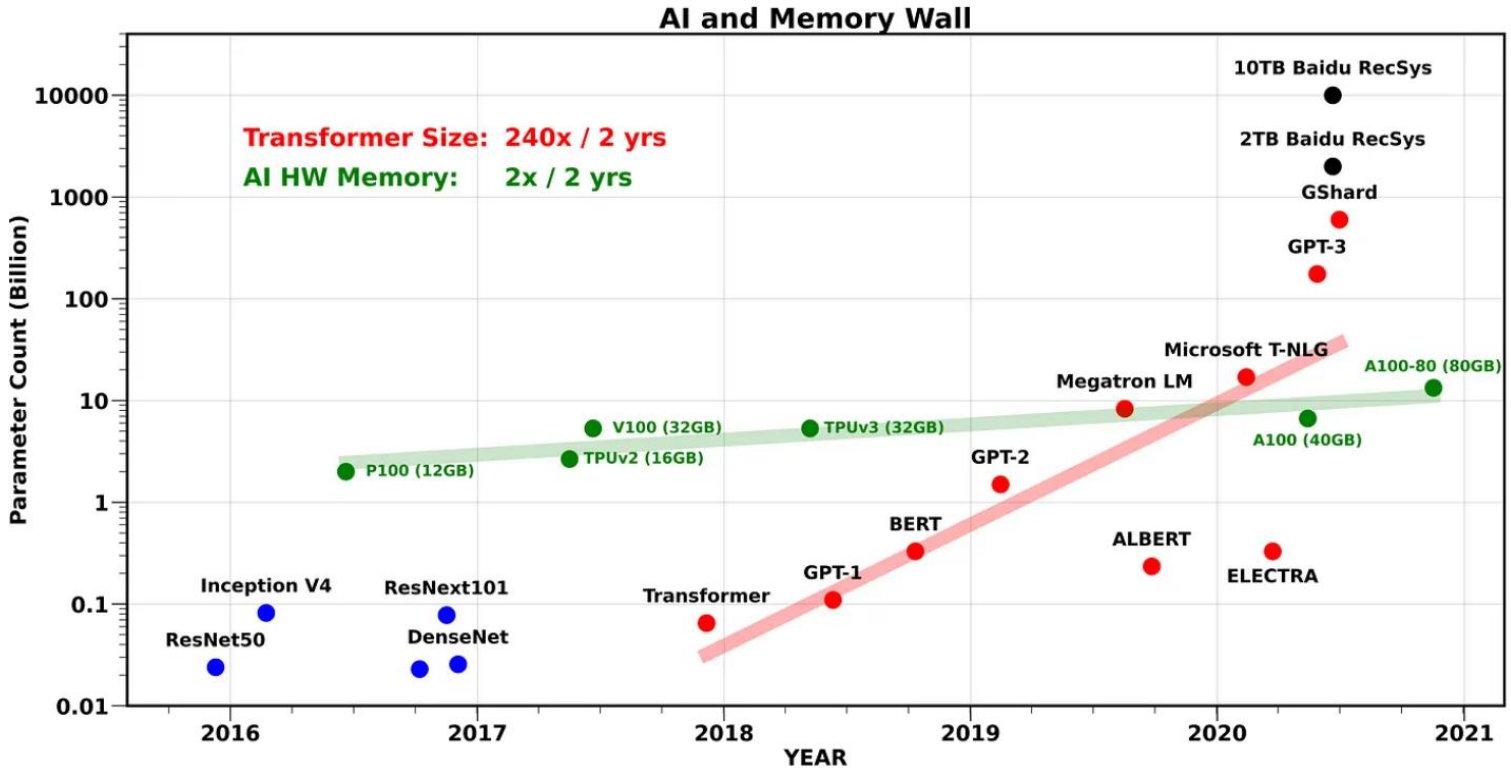
- An easy way to summarize a paper and point out its technical contribution and novelty is to prepare one-page slide for each paper. This slide is called “One-Page Summary.”
- One-page Summary that includes **OOCM-R**:
  - **Observation**: The issues or trends observed
  - **Objective (Goal):** The objectives after resolving the observed issues
  - **Challenge:** The challenges to resolve the issues and achieve the goal
  - **Main Idea (Proposed Method):** The solution to resolve the challenges
    - Main idea leads to the novelty of a paper
    - The proposed method leads to the technical contribution of a paper
  - **Result**:
    - Experiment setups, platform/environment, simulation/implementation, compared methods, workloads/benchmarks, metrics, results, etc.

Outline

- Introduction to In-Memory Computing
- One-Page Summaries

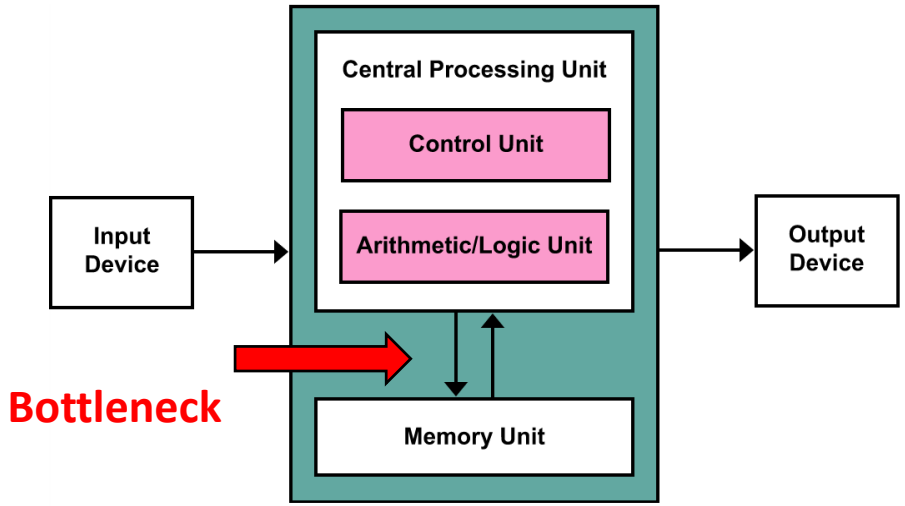
Tremendous Demands and Opportunities in the Era of Artificial Intelligence and Big Data

- To enhance the performance of AI model, more parameters are needed
  - Hardware needs higher memory bandwidth to support novel AI models
- Memory wall issue in AI
  - Growth of AI HW memory << Growth of model size (# of parameters)

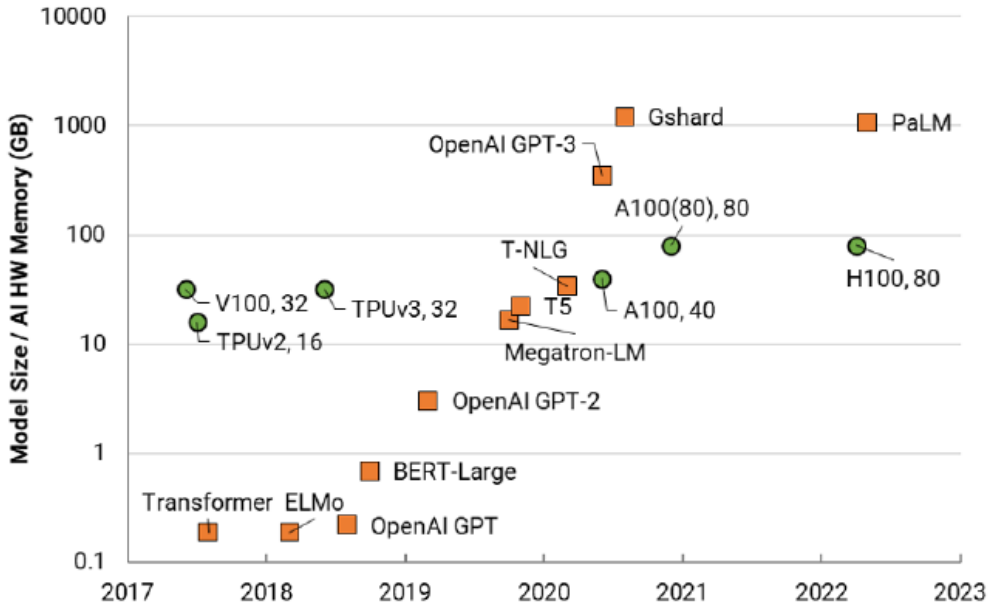


Fundamental Problems in Running AIs

- Bottleneck of von Neumann architectures
  - Memory wall: Lacks of bandwidth (for the growth of AI models)
  - Tremendous data movement (Processing Unit ↔ Memory unit)
- Growth of model size >> Growth of GPU memory
  - GPT-3 model(2020) >> H100 GPU(2022)
  - Need multiple GPUs => **High Cost**



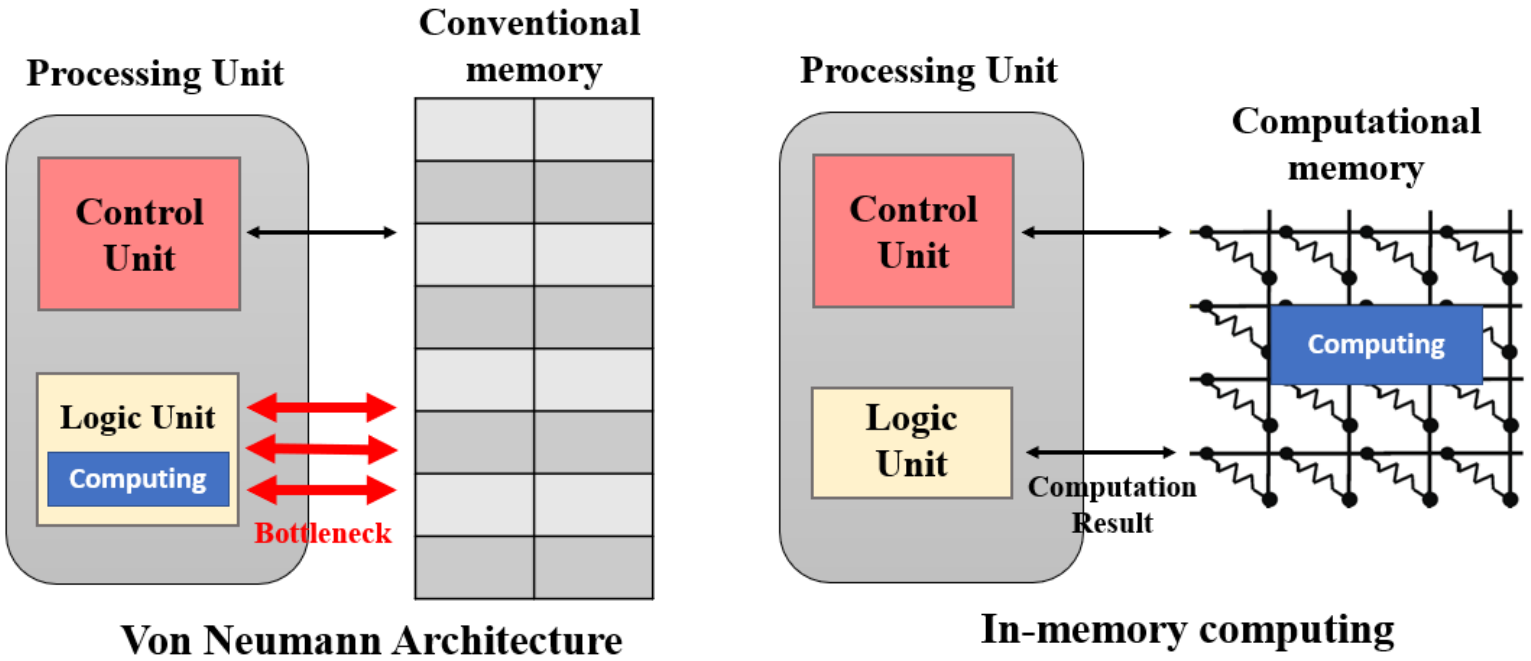
Ref: https://towardsdatascience.com/machine-learning-fundamentals-ii-neural-networks-f1e7b2cb3eef



Ref: Kim, J. et,al. OptimStore: In-Storage Optimization of Large Scale DNNs with On-Die Processing. In 2023 HPCA IEEE.

A Potential Solution

- In-Memory Computing
  - Offload the computation into the memory unit (and even **the storage unit**)
  - Compute the computation during data access
  - Resolve the bottleneck of von Neumann architecture
  - Computational memory – Crossbar array (analog MAC)



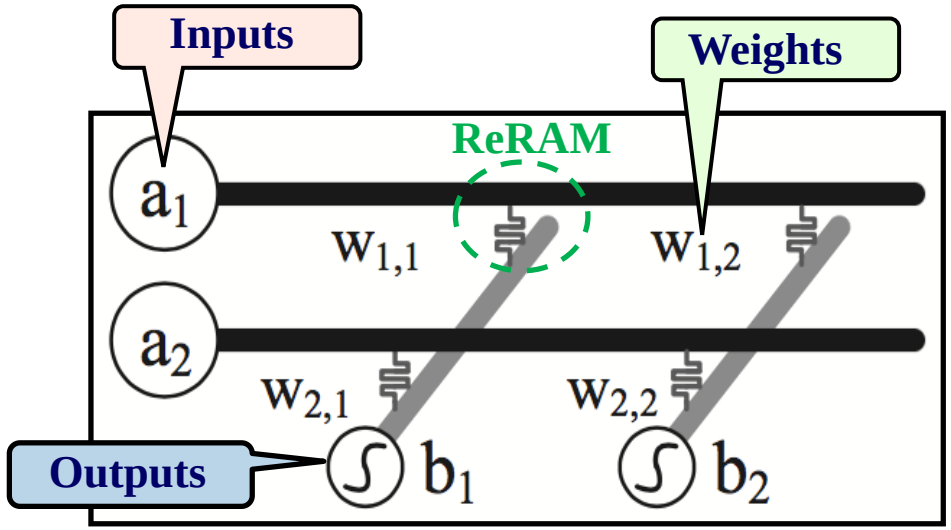
NVM-based Crossbar

- ReRAM-Based Crossbar for In-Memory Computing
  - **Crossbar:** Wordlines and bitlines are orthogonal in the 3-dimensional (3D) space, where ReRAM is used to joint wordlines and bitlines.
  - **ReRAM (or called Memristor):** It works by changing the resistance of the memory cell to represent different data states (e.g., 1 or 0).

Matrix Multiplication:

$$\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} a_1 , a_2 \end{bmatrix} X \begin{bmatrix} w_{1,1} & w_{1,2} \\ w_{2,1} & w_{2,2} \end{bmatrix}$$

Outputs Inputs Weights

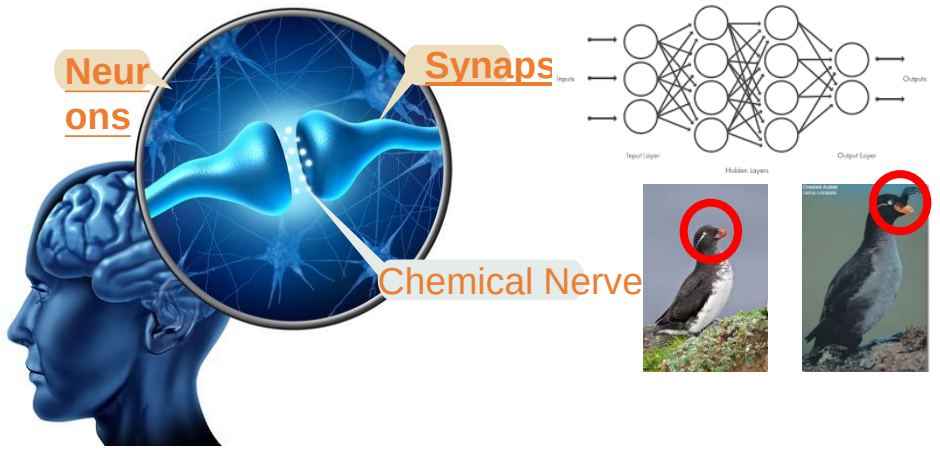


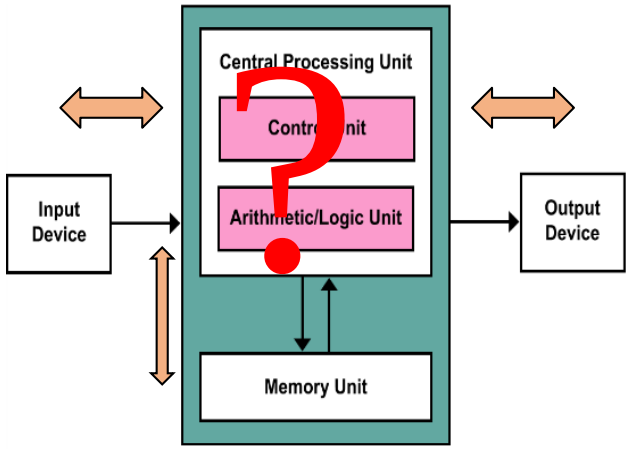
Crossbar Accelerators

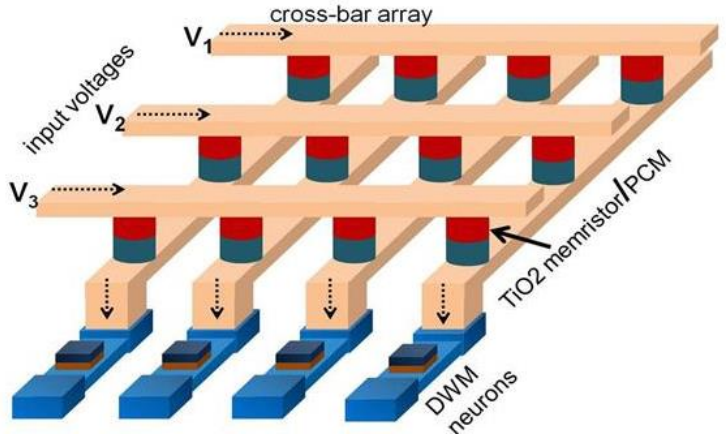
Source: Shafiee, Ali, et al. "**ISAAC: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars.**" Proceedings of the 43rd International Symposium on Computer Architecture. IEEE Press, 2016.

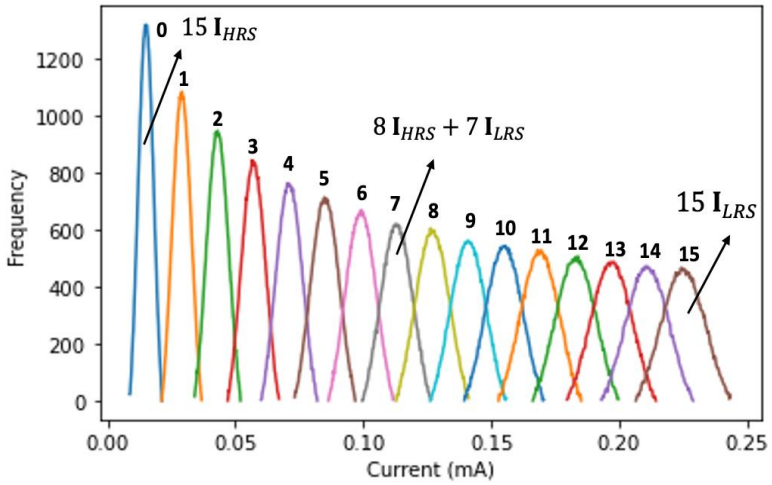
Rethinking of Computing, Memory, and Storage

- Challenges in (Crossbar) In-Memory Computing
  - Reliability (Error rate)
  - Scalability (Space utilization)
  - Functionality (MAC, TCAM, Range)
  - Capacity (ReRAM vs Flash)







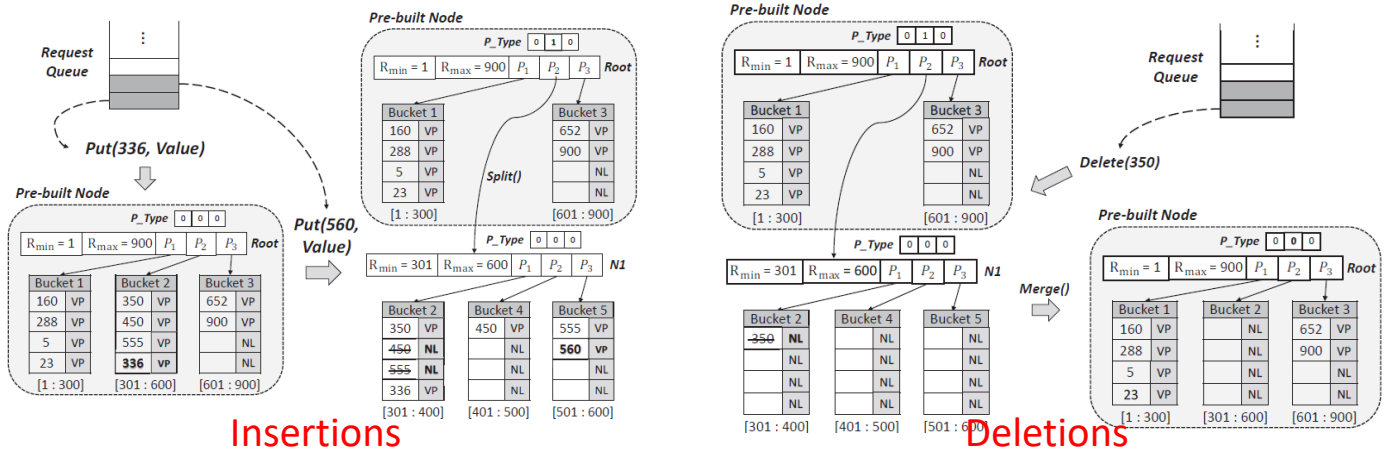


To Appear

APB-tree: An Adaptive Pre-built Tree Indexing Scheme for NVM-based IoT Systems

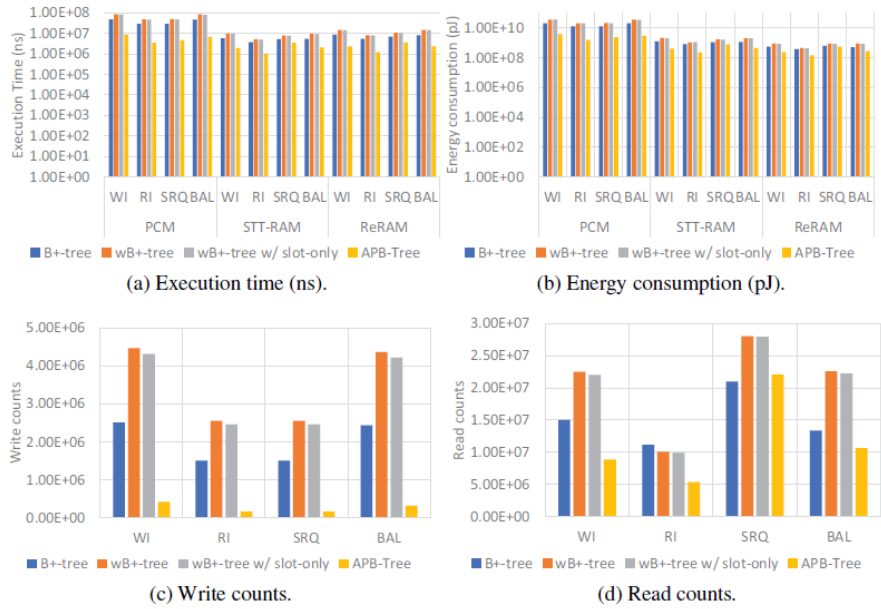
[ACM TECS]

- **Motivation**
  - Traditional B⁺-tree indexing schemes suffer from high write overheads in IoT systems
  - NVM technologies have asymmetric read/write latency and energy consumption, making writes especially costly
- **Goal**
  - Design a new indexing scheme for NVM-based IoT systems to:
    - Minimize dynamic operations triggered by insertions/deletions
    - Adapt to IoT data patterns
    - Leverage NVM characteristics
- **Main Idea**
  - Pre-build initial index structure offline using known hot keys in IoT system
  - Store unsorted keys in fixed-size buckets with sub-ranges
  - Adapt tree structure dynamically as needed at runtime



Execution time: 47% to 72% reduction

Energy consumption: 11% to 72% reduction

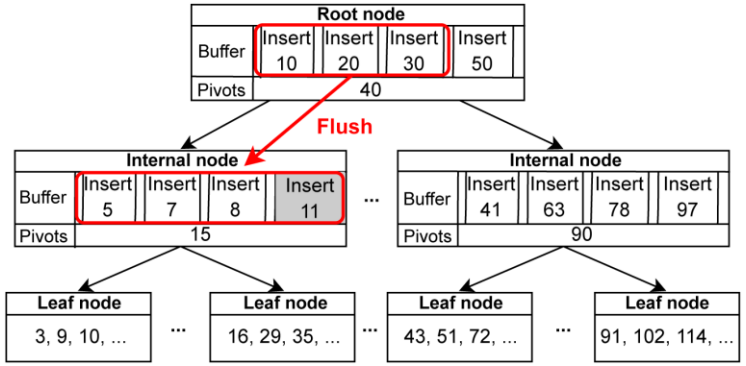


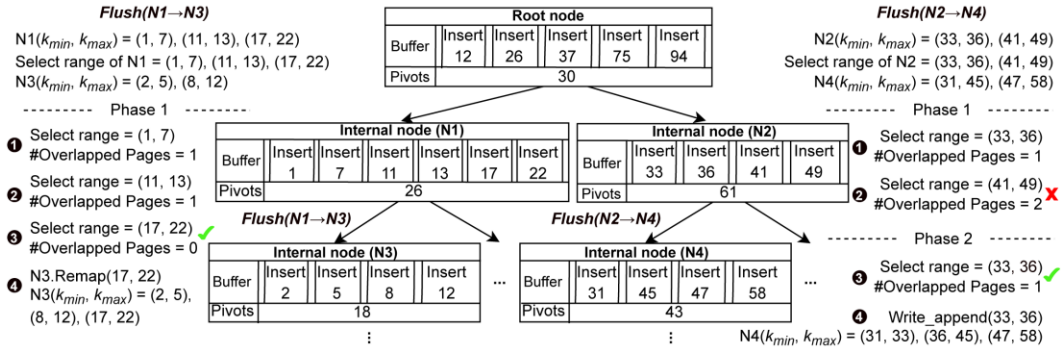
- Shih-Wen Hsu, Yen-Ting Chen, Yuan-Hao Chang, Kam-Yiu Lam, Wei-Kuan Shih, and Han-Chieh Chao, "APB-tree: An Adaptive Pre-built Tree Indexing Scheme for NVM-based IoT Systems," accepted and to appear in ACM Transactions on Embedded Computing Systems (TECS).

PULSE: Progressive Utilization of Log-Structured Techniques to Ease SSD Write Amplification in B-epsilon-tree

[ASP-DAC'25]

- Observation
  - The Bε-tree indexing scheme suffers from severe **write amplification issues**, which significantly affect SSD **endurance** and **performance**—a critical challenge in modern storage systems.
  - The primary causes of write amplification in the Bε-tree indexing scheme are the **buffer flushing mechanism** and the **misalignment of flushed data** with SSD page boundaries.
- Goal
  - We proposed PULSE to **minimize write amplification** during key-value pair insertions and deletions while maintaining the consistency of the indexing scheme.
- Main Idea
  - Tracks node information, enabling precise **alignment of flushed data with SSD page boundaries** and reducing redundant writes.
  - Chooses message subsets for flushing, **prioritizing minimal overlap and maximizing page utilization** to reduce **unnecessary write** operations.
- Contribution
  - The proposed solution, PULSE, significantly **reduces write amplification** by **over 62.6%**, addressing a critical barrier to SSD efficiency and longevity.





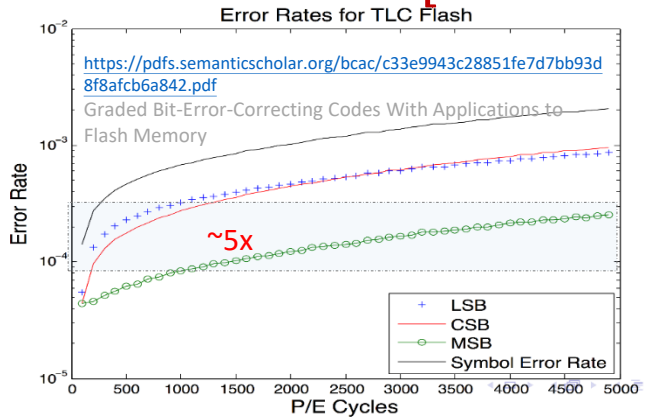
- Huai-De Peng, Yi-Shen Chen, Tseng-Yi Chen, and Yuan-Hao Chang, "PULSE: Progressive Utilization of Log-Structured Techniques to Ease SSD Write Amplification in B-epsilon-tree," ACM/IEEE Asia and South Pacific Design Automation Conference (ASP-DAC), Tokyo, Japan, Jan. 20-23, 2025. (Acceptance rate: 29%)

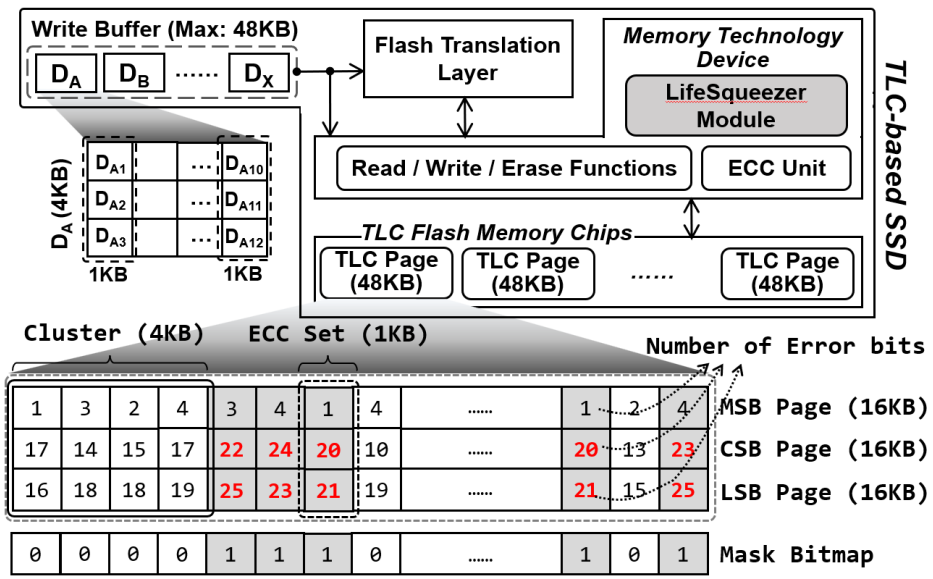
Research Summary 2024

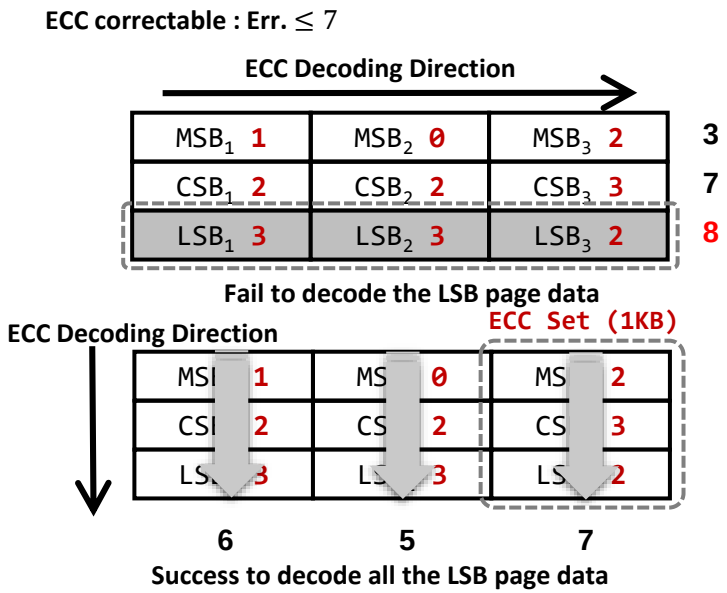
1. Storage Systems - Flash Drives and SMR Disks

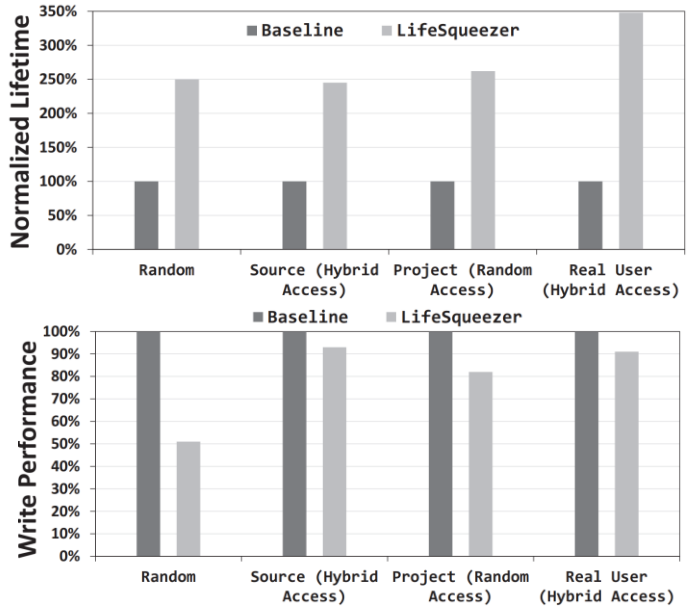
LifeSqueezer: Increase the Tolerability of Weak Pages for Lifetime Improvement on TLC-based SSDs [RACS'24]

- Observation and motivation:
  - CSB and LSB: high error rate, MSB: low error rate
  - The **CSB and LSB pages are also called weak page**, and the **MSB page is called strong page**
  - Considering the asymmetric BER status between weak and strong pages
- Design:
  - Error Locality (errors concentrate in the weak pages)
  - Errors over-concentration because of BER difference on TLC flash memory could be mitigated via **vertical coding**
  - By **distributing errors more evenly, it allows ECC to share its error-correcting capabilities** across all pages on the word-line









- Liang-Chi Chen, Kun-Chi Chiang, Chien-Chung Ho, Yu-Ming Chang, Chin-Chiang Pan, and Yuan-Hao Chang, "LifeSqueezer: Increase the Tolerability of Weak Pages for Lifetime Improvement on TLC-based SSDs with the Off-the-shelf ECC," ACM International Conference on Research in Adaptive and Convergent Systems (RACS), Pompei, Italy, Nov. 5-8, 2024.

CellRejuvo: Rescuing the Aging of 3D NAND Flash Cells with Dense-Sparse Cell Reprogramming

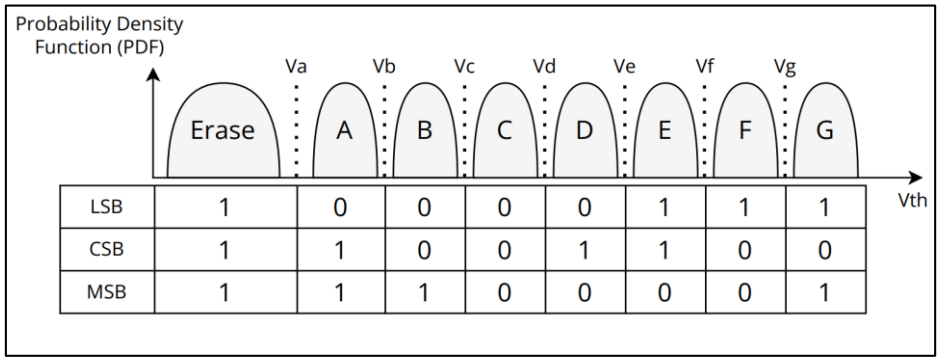
[ICCAD'24]

- **Introduction**
  - The shortened margins make the NAND flash become less tolerant of the error effects caused by charge loss in NAND flash cells.

- **Method**
  - Since G state is the most error prone state in 3D TLC NAND flash, our proposed method will only reprogram cells that store G state in the original user data.

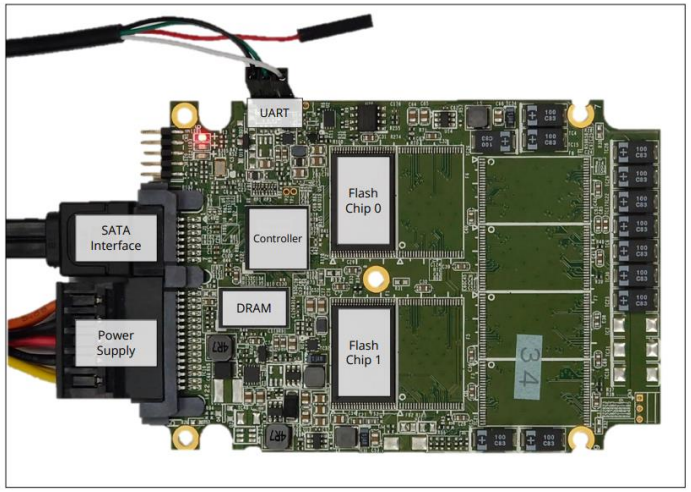
- **Experiment** (Real SSD Platform)
  - We implemented CellRejuvo on a real SSD device, and the experiments show that the method proposed in this article can reduce errors by 38.28% compared to the Baseline on average, and significantly improve read performance by 21.86% in the late stage.

States and Coding Rule of 3D NAND TLC

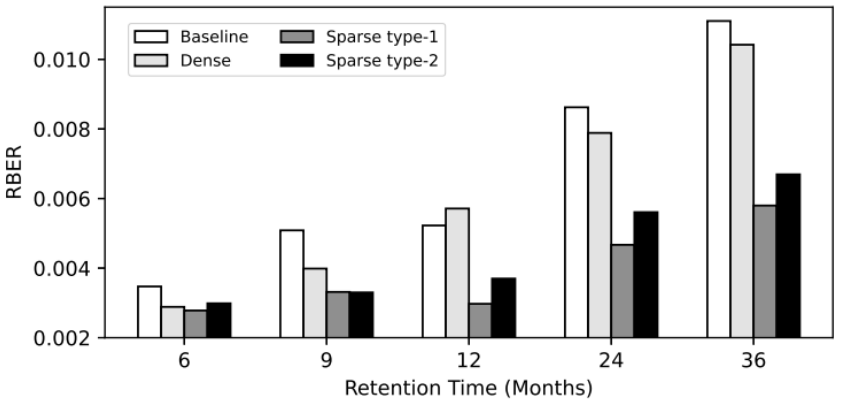


LSB	1	0	0	0	0	1	1	1
CSB	1	1	0	0	1	1	0	0
MSB	1	1	1	0	0	0	0	1

SSD Development Platform



Error Reduction

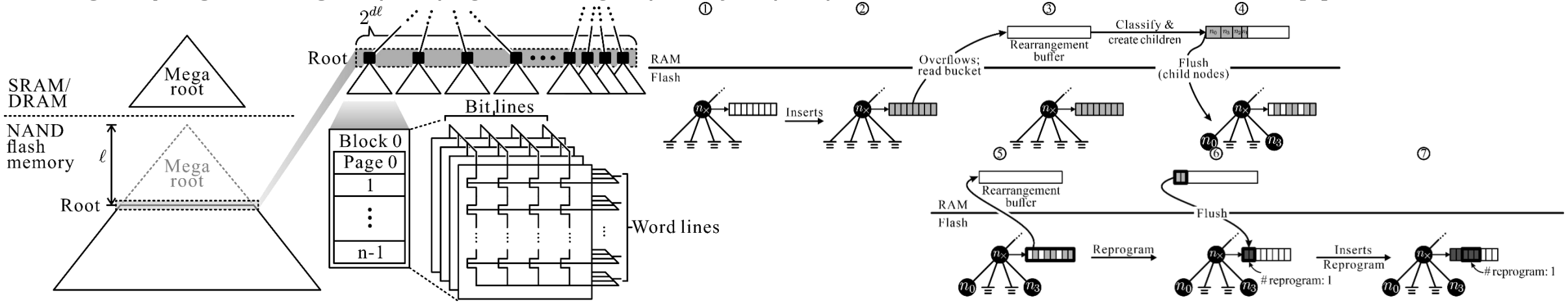


- Han-Yu Liao, Yi-Shen Chen, Jen-Wei Hsieh, Yuan-Hao Chang, Hung-Pin Chen, "CellRejuvo: Rescuing the Aging of 3D NAND Flash Cells with Dense-Sparse Cell Reprogramming," ACM/IEEE International Conference on Computer-Aided Design (ICCAD), Newark, New Jersey, USA, Oct. 27-31, 2024. **(Top Conference)**

FIRM-tree: a Multidimensional Index Structure for Reprogrammable Flash Memory

[CODES'24, IEEE TCAD'24]

- **Observation**
  - Existing multidimensional index data structures often face a management trilemma on flash memory, among access performance, space utilization, and maintenance overheads. This challenge can potentially be addressed by leveraging the unique features of modern flash memory, such as page reprogramming.
- **Goal**
  - Our proposed FIRM-tree, a multidimensional tree designed for NAND flash, significantly reduces tree maintenance overheads on flash storage.
- **Main idea**
  - **Selective data migration**: Flush data points only when there are enough of them to guarantee the space utilization of flash.
  - **Mega root**: Enlarge the root node size of the flash-part tree to maximally utilize the buffer for write amplification alleviation.
  - **Page reprogramming**: Exploit page rewriting capability to postpone block erases and GC overheads [1].



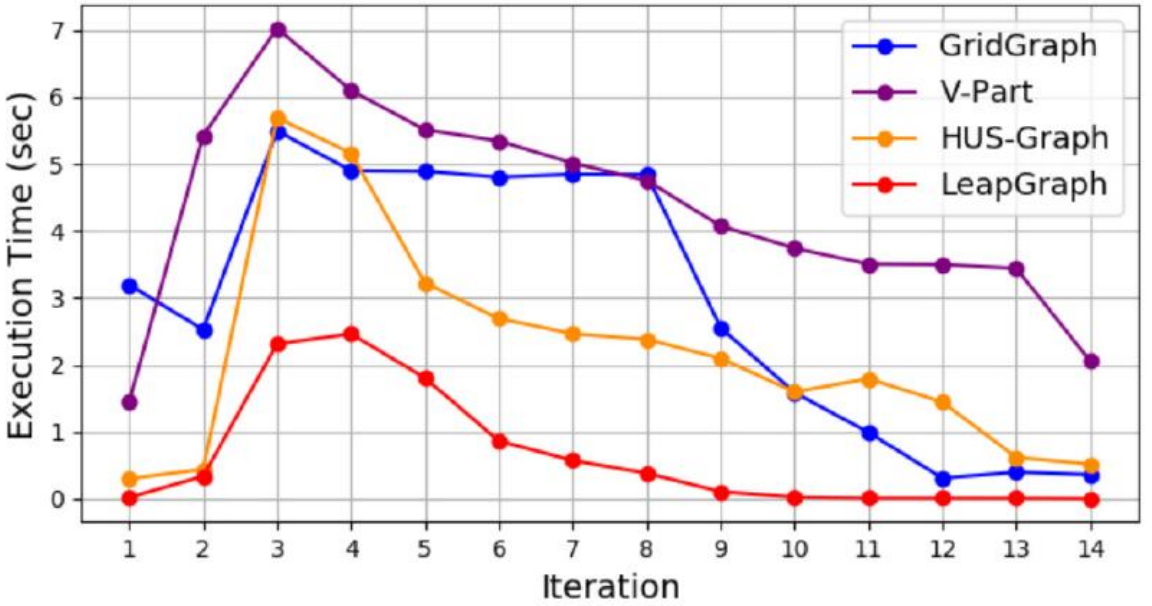
[1] Congming Gao, Min Ye, Chun Jason Xue, Youtao Zhang, Liang Shi, Jiwu Shu, and Jun Yang. 2022. Reprogramming 3D TLC Flash Memory based Solid State Drives. ACM Trans. Storage 18, 1, Article 9 (February 2022), 33 pages. https://doi.org/10.1145/3487064

- Shin-Ting Wu, Pin-Jung Chen, Po-Chun Huang, Wei-Kuan Shih, and Yuan-Hao Chang, "FIRM-tree: a Multidimensional Index Structure for Reprogrammable Flash Memory," ACM/IEEE International Conference on Hardware/Software Codesign and System Synthesis (CODES+ISSS), Raleigh, NC, USA, Sep. 29 – Oct. 4, 2024. (Journal Track, Integrated with IEEE TCAD) **(Top Conference)**
- Shin-Ting Wu, Pin-Jung Chen, Po-Chun Huang, Wei-Kuan Shih, and Yuan-Hao Chang, "FIRM-tree: a Multidimensional Index Structure for Reprogrammable Flash Memory," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD), vol. 43, no. 11, pp. 3600-3613, Nov. 2024. (Integrated with ACM/IEEE CODES+ISSS'24)

LeapGraph: A Fully External Graph Processing System on High-Speed SSD

[NVMSA'24]

- **Observation**
  - Fully external graph systems offer an appealing feature of processing very large-scale graphs with only constant memory in a single machine.
  - Meanwhile, SSDs have evolved rapidly in recent years in terms of access speed and price. Nevertheless, such drastic advancements in storage technology has changes the bottleneck for fully external graph processing. This shift has rendered the traditional design goal for fully external graph processing outdated.
- **Goal**
  - This work presents LeapGraph, a fully external graph system which can keep up with the evolving storage technology trend by better exploiting high-speed SSDs.
- **Main Idea**
  - *Dual Update Mode* can switch between *push* and *pull* execution mode during graph processing, depending on whether the current iteration is bottlenecked by I/O bandwidth or CPU computation.
  - *Lazy Vertex Write* is proposed to delay the access of vertex attributes and then redistribute them. It effectively enhances the locality of access, not only for memory access but also for I/O access.
  - *Subgraph-based Pull Update Mode* further optimizes the performance by determining whether to use push or pull modes with finer granularity: subgraph.



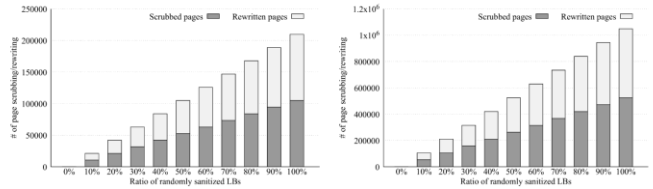
Per-Iteration Execution Time of Running BFS on Twitter.

- Tsun-Yu Yang, Ming-Chang Yang, and Yuan-Hao Chang, "LeapGraph: A Fully External Graph Processing System on High-Speed SSD," IEEE Nonvolatile Memory Systems and Applications Symposium (NVMSA), Sokcho, South Korea,Aug. 21-23, 2024.

PRESS: Persistence Relaxation for Efficient and Secure Data Sanitization on Zoned Namespace Storage

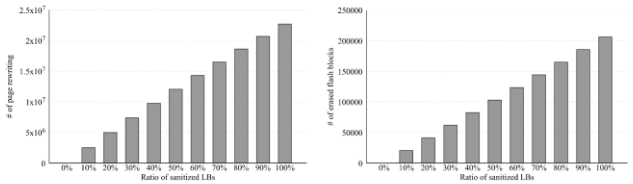
[ASP-DAC'24]

- Motivation
  - **Data sanitization is more challenging on ZNS SSDs** due to:
    - **Large unit of data removal operations**, i.e., zone resetting
      ⇒ time-consuming data sanitization
    - **Asynchronous nature of zone resetting**, like block erasing
      ⇒ unpredictable and insecure data removal
- Goal
  - We present the PRESS data sanitization scheme, which makes use of the limited on-device RAM or NVRAM buffer to postpone the persistence of data and reduce the overheads to sanitize sensitive temporary data.
- Main Idea
  - The more levels of keys have been written from key buffer to the ZNS SSD, the higher overheads to sanitize the data later.
  - **When the key buffer is full, recursively flush low-level keys and replace them with their encryption key in the buffer, making the data "more persistent" and "harder to securely delete."**



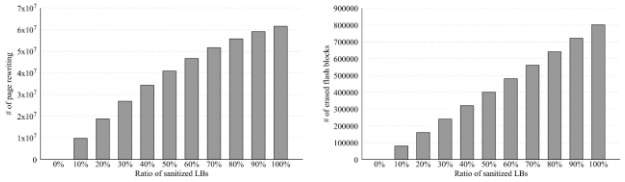
(a) 10% LBs of the zone are sanitized. (b) 50% LBs of the zone are sanitized.

Figure 7: Scrubbing/rewriting overheads of PRESS w.r.t. ratios of sanitized LBs and ratios of random sanitize commands.



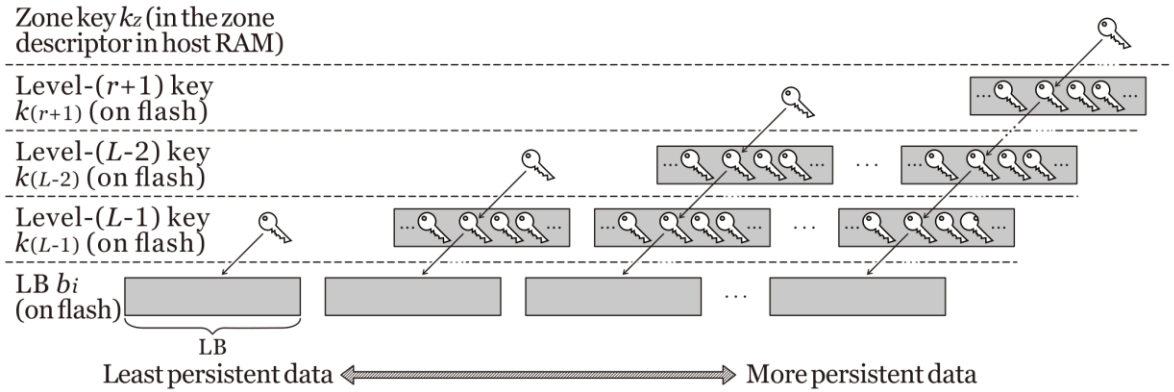
(a) Page rewriting overheads. (b) Block erasing overheads.

Figure 8: Rewriting/erasing overheads of the block erasing approach, w.r.t. different ratios of sanitized LBs. (*80%* of the sanitized LBs are sequentially sanitizes.)



(a) Page rewriting overheads. (b) Block erasing overheads.

Figure 9: Rewriting/erasing overheads of zone resetting, w.r.t. different ratios of sanitized LBs. Among the sanitized LBs. (*20%* of the sanitized LBs are sequentially sanitizes.)



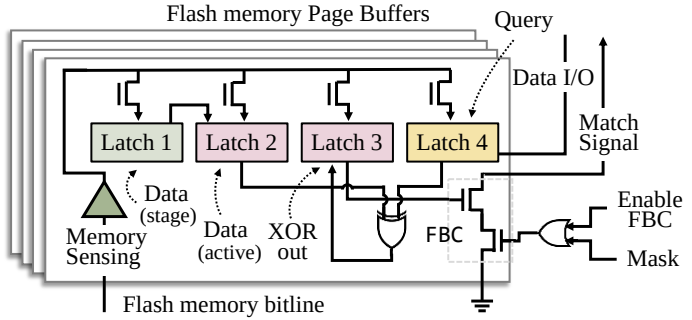
- Yun-Shan Hsieh, Bo-Jun Chen, Po-Chun Huang, and Yuan-Hao Chang, "PRESS: Persistence Relaxation for Efficient and Secure Data Sanitization on Zoned Namespace Storage," ACM/IEEE Asia and South Pacific Design Automation Conference (ASP-DAC), Incheon, South Korea, Jan. 22-25, 2024.

2. NVM Main Memory and Storage

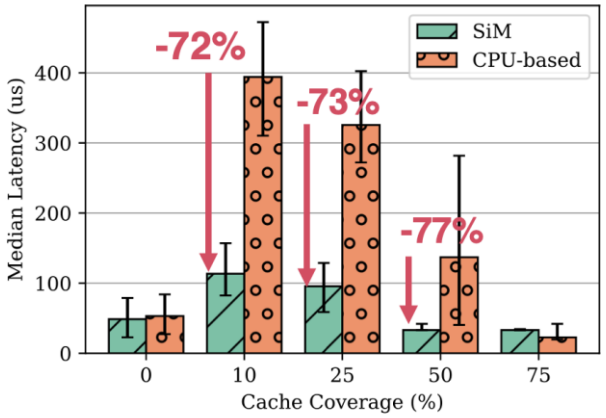
Search-in-Memory (SiM): Conducting Data-Bound Computations on Flash Memory Chip

[DATE'24]IEEE TCAD'24]
[CODES'24 – Best Paper Award]

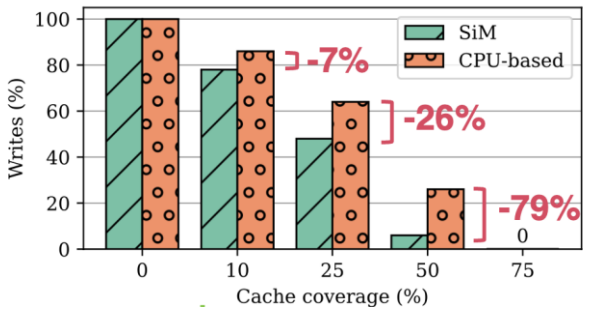
- **Motivation**
  - Data indexing is I/O bound
  - Existing computational memory solutions require intrusive design changes
- **Goal**
  - Good performance even with radically lower bandwidth
  - Improves DRAM's role as write buffer
  - Good performance even with few cache
- **Main Idea**
  - Realize data matching through re-purposing existing circuits
  - Saving I/O by sending query into memory instead of reading page out of memory
  - Generic SIMD command interface useful for wide range of applications



Read delay ↓



Writes ↓



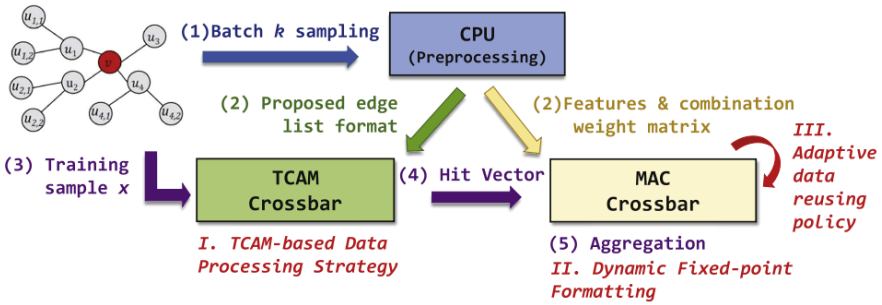
- Yun-Chih Chen, Yuan-Hao Chang, and Tei-Wei Kuo, "Search-in-Memory (SiM): Conducting Data-Bound Computations on Flash Memory urope (DATE), Valencia, Spain, Mar. 25-27, 2024.
- Yun-Chih Chen, Yuan-Hao Chang, and Tei-Wei Kuo, "Reliable, Versatile, and Efficient Data Matching in SSD's NAND Flash Memory Chip for Data Indexing Acceleration," ACM/IEEE International Conference on Hardware/Software Codesign and System Synthesis (CODES+ISSS), Raleigh, NC, USA, Sep. 29 - Oct. 4, 2024. (Journal Track, Integrated with IEEE TCAD) **(Best Paper Award - Top Conference)**
- Yun-Chih Chen, Yuan-Hao Chang, and Tei-Wei Kuo, "Search-in-Memory: Reliable, Versatile, and Efficient Data Matching in SSD's NAND Flash Memory Chip for Data Indexing Acceleration," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD), vol. 43, no. 11, pp. 3864-3875, Nov. 2024. (Integrated with ACM/IEEE CODES+ISSS'24)

3. In/Near-Memory Processing and AI/ML with NVM

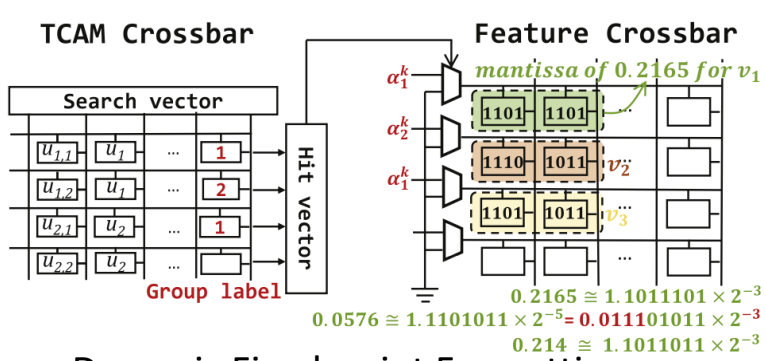
TCAM-GNN: A TCAM-based Data Processing Strategy for GNN over Sparse Graphs

[IEEE TETC'24]

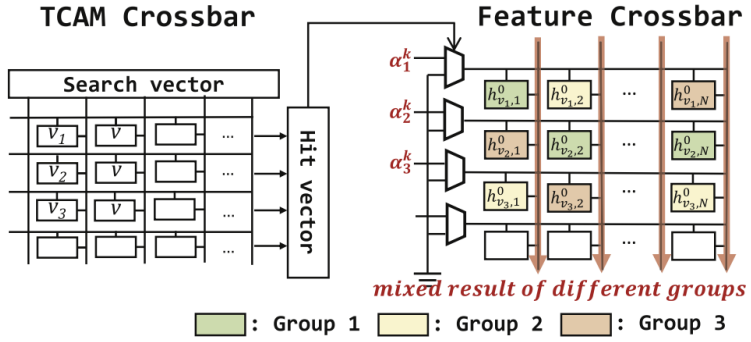
- **Observation**:
  - Smart Utilizing ternary content addressable memory (TCAM) crossbars to enable intensive neighbor vertices sampling operation and efficiently support parallel data processing strategy in training phase of GNN.
- **Goal:** Enhance training/inferencing performance of graph neural network with ReRAM-based PIM accelerator.
- **Main idea**: A high-throughput and energy-efficient *ReRAM-based PIM accelerator* with auxiliary TCAM crossbars is designed for training various graph neural networks over large-scale graphs.
  - A TCAM-based data processing strategy to orchestrate crossbars and TCAMs for handling GNN operations.
  - A dynamic fixed-point formatting approach to improve the resource efficiency of crossbar arrays.
  - An adaptive data reusing policy is designed to enhance the data locality of graph features.



- Overview of TCAM-GNN



- Dynamic Fixed-point Formatting



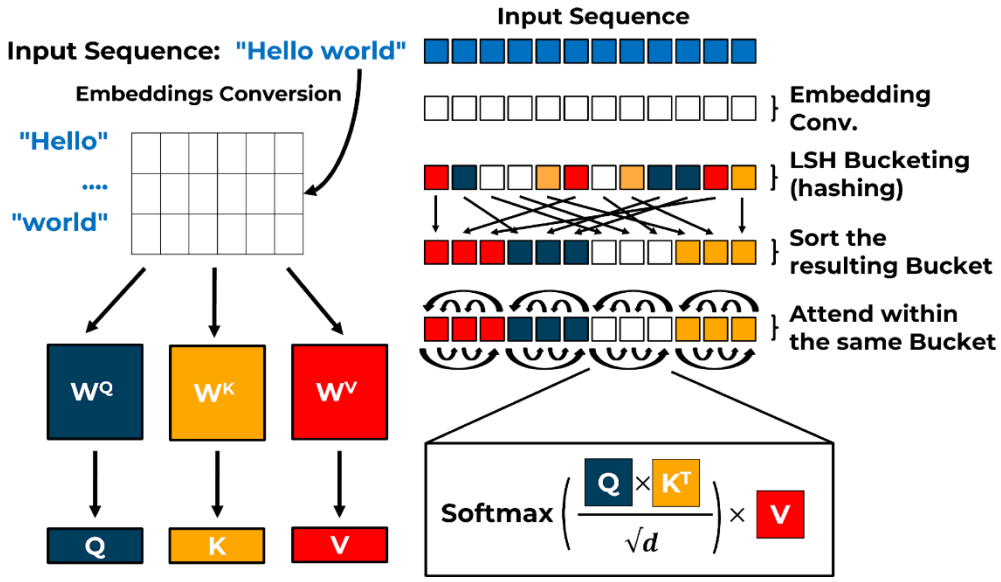
- Adaptive Data Reusing Policy

- Yu-Pang Wang, Wei-Chen Wang, Yuan-Hao Chang, Chieh-Lin Tsai, Tei-Wei Kuo, Chun-Feng Wu, Chien-Chung Ho, and Han-Wen Hu, "TCAM-GNN: A TCAM-based Data Processing Strategy for GNN over Sparse Graphs," IEEE Transactions on Emerging Topics in Computing (TETC), vol. 12, no. 3, pp. 891-904, Jul. 2024.

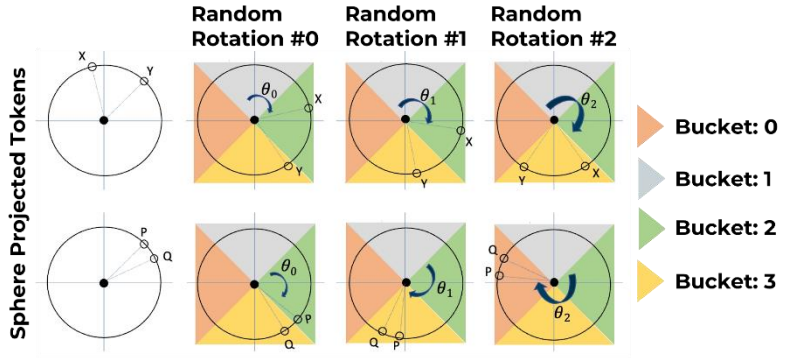
AttentionRC: A Novel Approach to Improve Locality Sensitive Hashing Attention on Dual-addressing Memory

[CODES'24, IEEE TCAD'24]

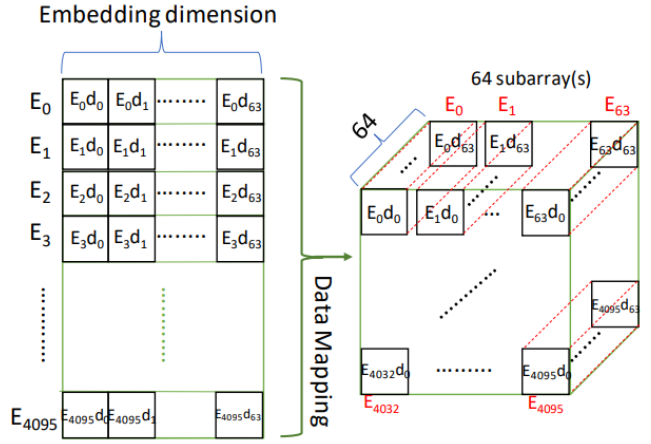
Locality Sensitive Hashing (LSH) in Reformer Model



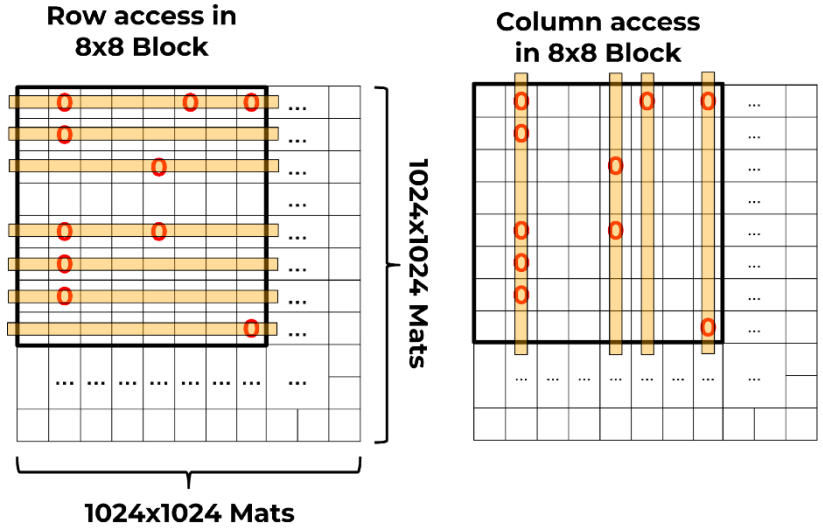
LSH Bucketing



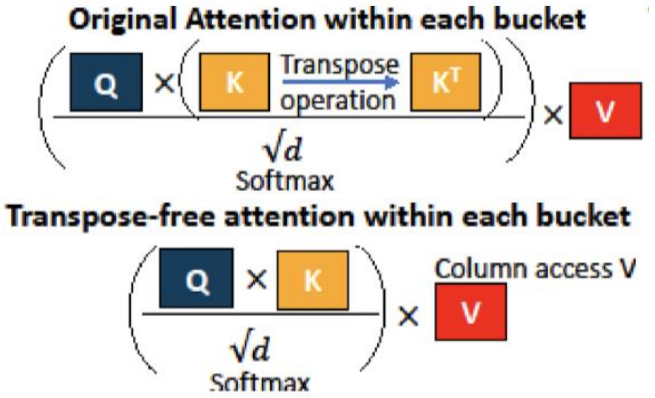
LSH-friendly data mapping



Sort-free bucket access



Transpose-free attention



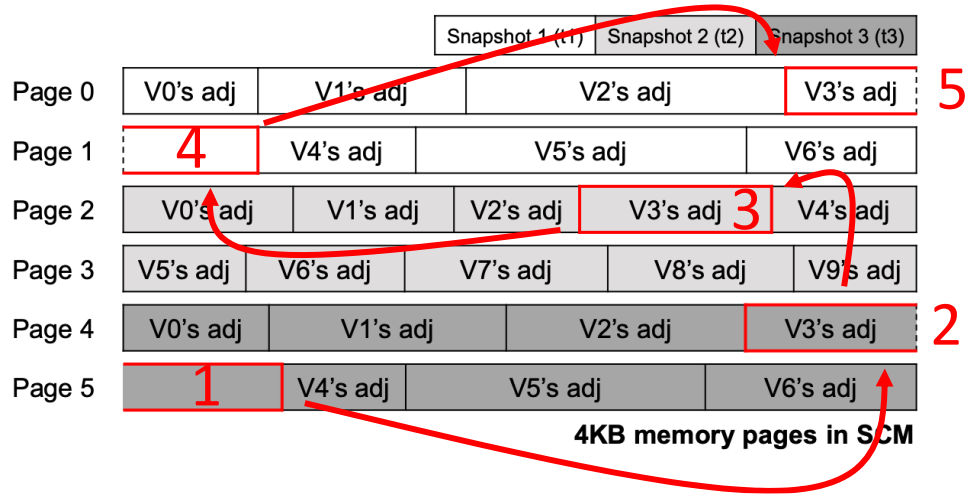
- Input sequences → High-D embeddings
- LSH bucketing groups **similar** embeddings into **the same** buckets
- Sort embeddings in each buckets (easier for in-bucket attention)
- Attend tokens within each bucket
- LSH reduces computational complexity through hashing

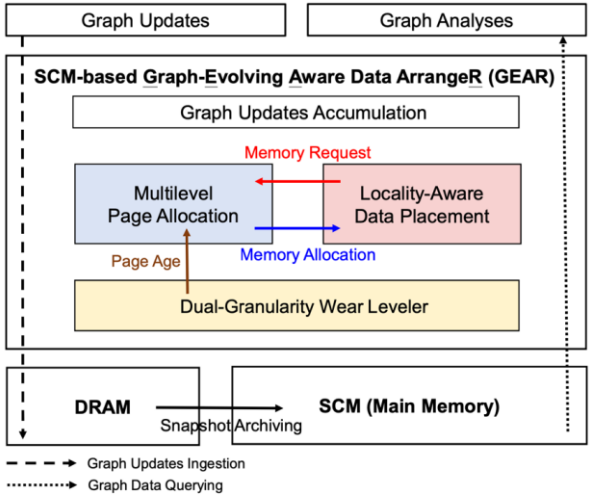
- Chun-Lin Chu, Yun-Chih Chen, Wei Cheng, Ing-Chao Lin, and Yuan-Hao Chang, "AttentionRC: A Novel Approach to Improve Locality Sensitive Hashing Attention on Dual-addressing Memory," ACM/IEEE International Conference on Hardware/Software Codesign and System Synthesis (CODES+ISSS), Raleigh, NC, USA, Sep. 29 - Oct. 4, 2024. (Journal Track, Integrated with IEEE TCAD) **(Top Conference)**
- Chun-Lin Chu, Yun-Chih Chen, Wei Cheng, Ing-Chao Lin, and Yuan-Hao Chang, "AttentionRC: A Novel Approach to Improve Locality Sensitive Hashing Attention on Dual-addressing Memory," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD), vol. 43, no. 11, pp. 3925-3936, Nov. 2024. (Integrated with ACM/IEEE CODES+ISSS'24)

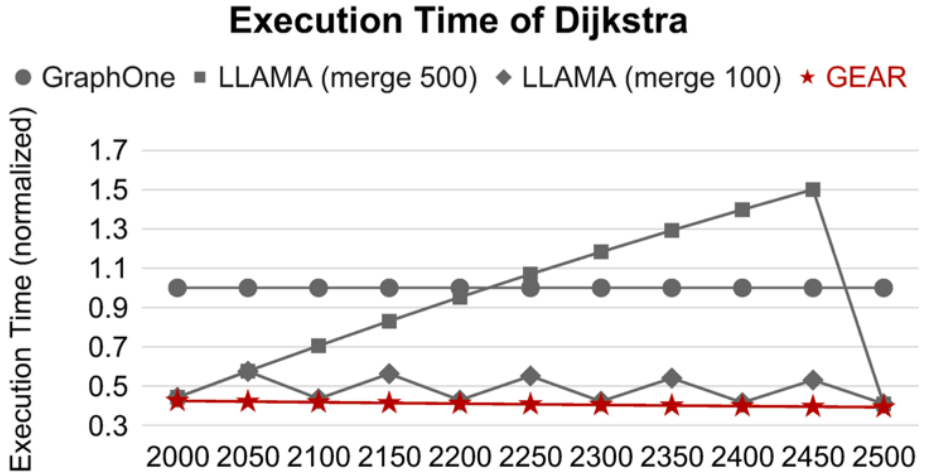
GEAR: Graph-Evolving Aware Data ArrangeR to Accelerate Traversing Evolving Graphs on SCM

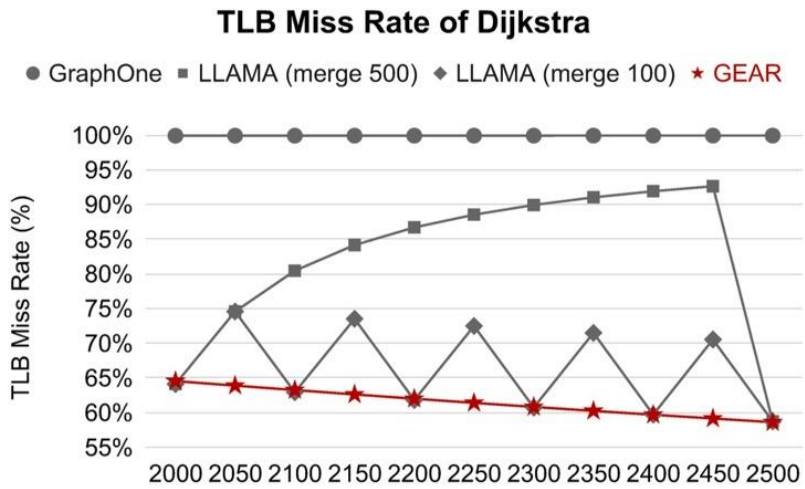
[CODES'24, IEEE TCAD'24]

- **Motivation**
  - Generating delta snapshots for graph evolving breaks locality
  - Running traversal on evolving graph faces high TLB misses
- **Goal**
  - Arrange and write the evolving graph data into SCMs while achieving strong graph spatial locality
- **Main Idea**
  - Allocating subpages based on vertex-neighboring relationships
  - Keeping unused areas for future updates
  - Evenly spreading write operations.







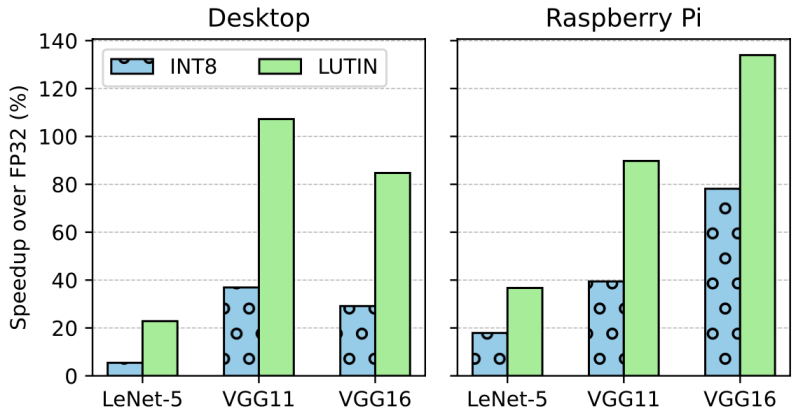


- Wen-Yi Wang, Chun-Feng Wu, Yun-Chih Chen, Tei-Wei Kuo, and Yuan-Hao Chang, "GEAR: Graph-Evolving Aware Data ArrangeR to Accelerate Traversing Evolving Graphs on SCM," ACM/IEEE International Conference on Hardware/Software Codesign and System Synthesis (CODES+ISSS), Raleigh, NC, USA, Sep. 29 - Oct. 4, 2024. (Journal Track, Integrated with IEEE TCAD) **(Top Conference)**
- Wen-Yi Wang, Chun-Feng Wu, Yun-Chih Chen, Tei-Wei Kuo, and Yuan-Hao Chang, "GEAR: Graph-Evolving Aware Data ArrangeR to Accelerate Traversing Evolving Graphs on SCM," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD), vol. 43, no. 11, pp. 3674-3684, Nov. 2024. (Integrated with ACM/IEEE CODES+ISSS'24)

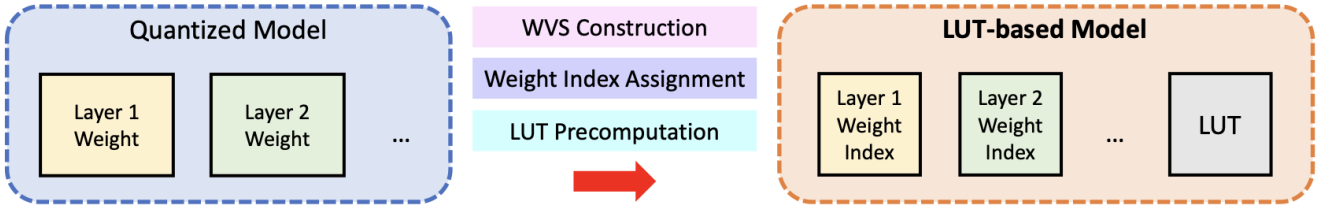
LUTIN: Efficient Neural Network Inference with Table Lookup

[ISLPED'24]

- **Motivation:**
  - DNN can be accelerated with Lookup tables to avoid dot products
  - LUTs with high-dimensional vectors or high bit-widths are expensive
- **Goal:**
  - Improve DNN inference for low-power, resource-constrained hardware
- **Approach (LUTIN):**
  - Reduces matrix multiplication by precomputing and storing into table lookups.
  - hyperparameter optimization: refine the quantization process
  - Vector partitioning: further size reduction



- **Result:**
  - up to a 2.07x speedup in latency
  - 2.04x improvement in energy efficiency over full-precision models.



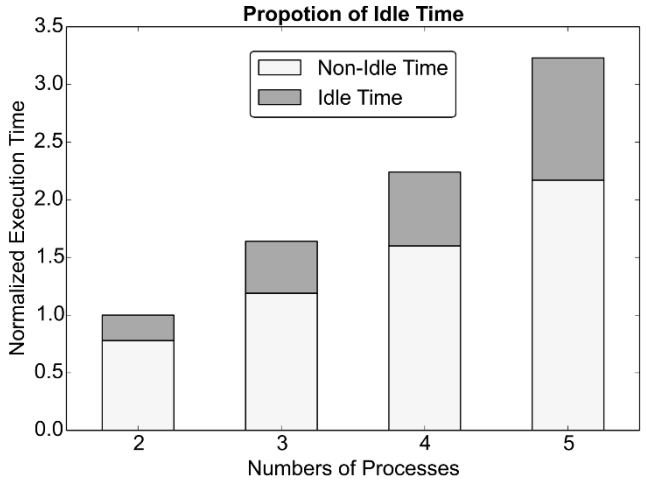
- Shi-Zhe Lin, Yun-Chih Chen, Yuan-Hao Chang, Tei-Wei Kuo, and Hsiang-Pang Li, "LUTIN: Efficient Neural Network Inference with Table Lookup," ACM/IEEE International Symposium on Low Power Electronics and Design (ISLPED), Newport Beach, California, USA, Aug. 5-7, 2024. **(Top Conference)**

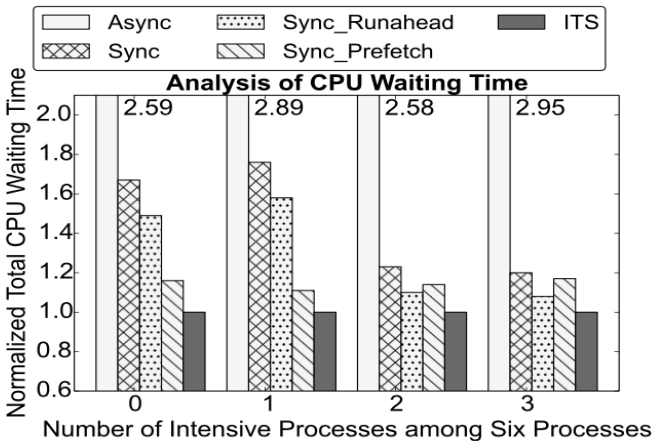
4. Intermittent Systems, Real-time Systems, and Operating Systems

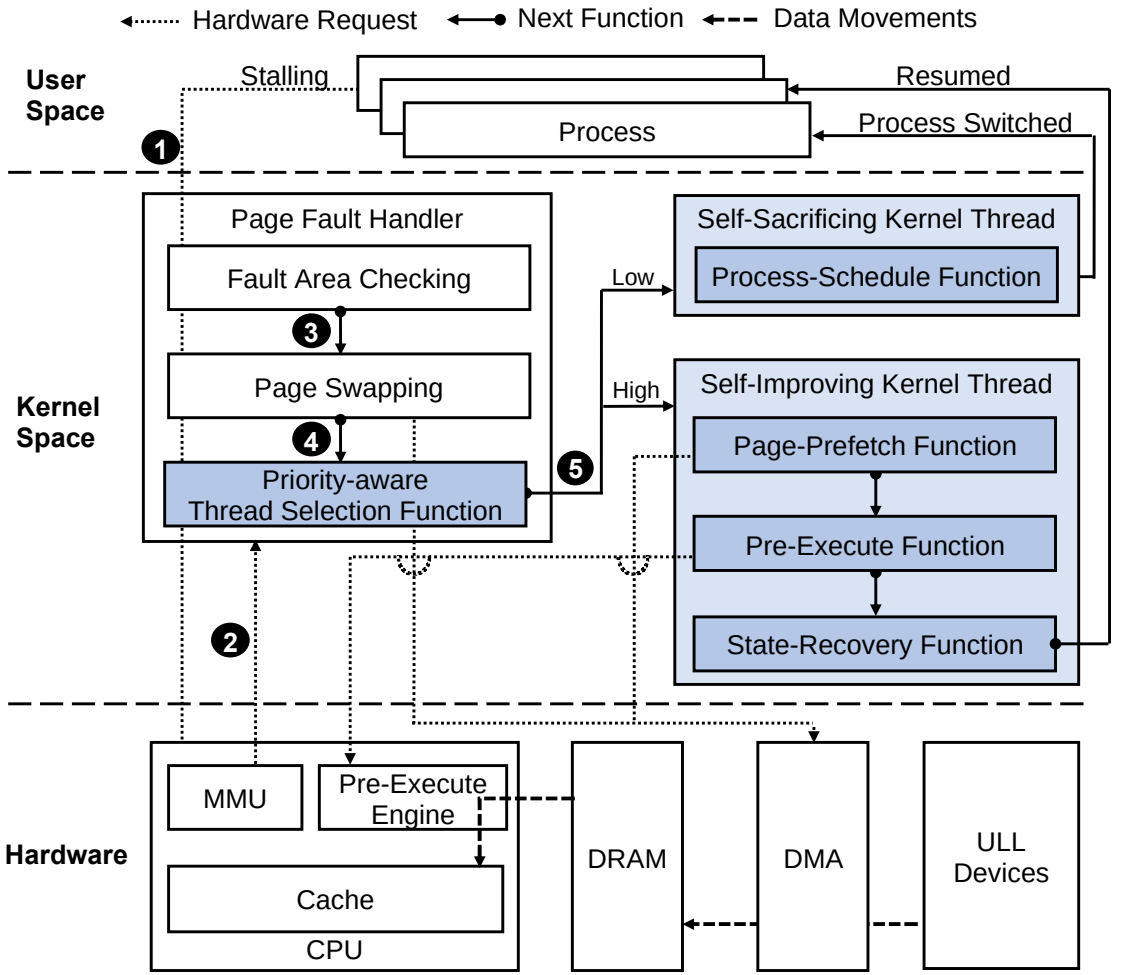
How to Steal CPU Idle Time When Synchronous I/O Mode Becomes Promising

[DAC'24]

- **Motivation**
  - Applying Sync I/O, CPU busy waiting time climbs when the amount of page fault frequency increases.
  - Around 30% of overall system execution time is spent on CPU busy waiting.
- **Goal**
  - How to utilize otherwise-wasted I/O busy time
- **Main Idea**
  - Design different kernel threads to help on the execution progress of process with different priorities.
  - Design speculative optimizations for memory and cache.







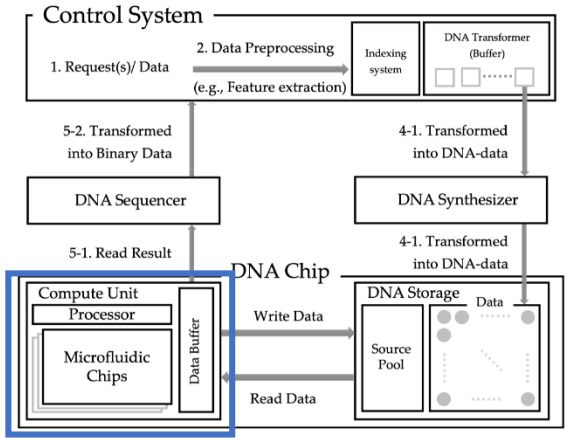
- Chun-Feng Wu, Yuan-Hao Chang, Ming-Chang Yang, and Tei-Wei Kuo, "How to Steal CPU Idle Time When Synchronous I/O Mode Becomes Promising," ACM/IEEE Design Automation Conference (DAC), San Francisco, CA, USA, Jun. 23-27, 2024. (Acceptance rate: 23%) **(Top Conference)**

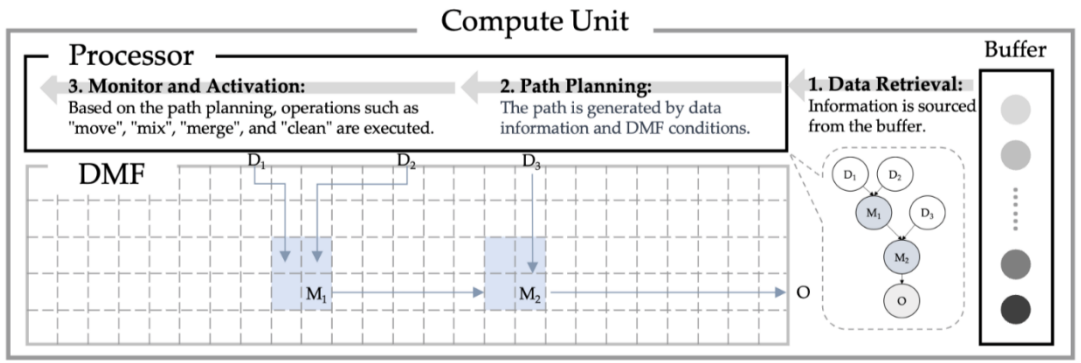
5. Others (including DNA Storage)

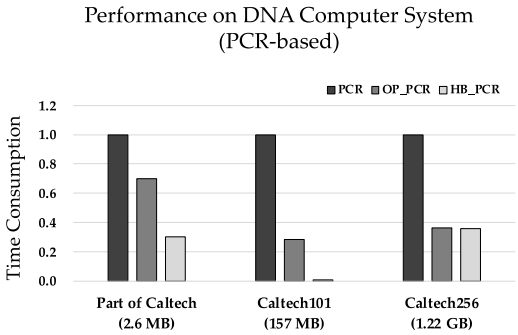
Bridging DNA Storage and Computation: An Integrated Framework for Efficient Biomolecular Data Management

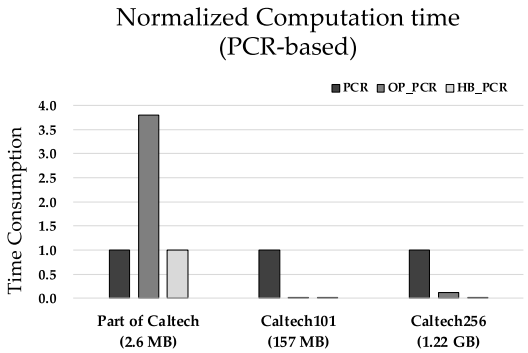
[SAC'24]

- **Observation**
  - DNA computing can achieve massive parallel computation.
  - No existing DNA computer can directly process DNA data.
  - Transferring data between DNA storage and computing units faces challenges like data pollution, leading to high costs.
- **Goal**
  - Develop a DNA computer that seamlessly integrates DNA storage and DNA computing units.
  - Enable efficient DNA data communication and transportation.
- **Main Idea**
  - **DNA Data Indexing System**: Design a unique indexing system tailored to DNA computer characteristics.
  - **DCA (Dual-Phase Clustering Approach)**: Use clustering algorithms to efficiently group data before storage.









- Yi-Syuan Lin, Yu-Pei Liang, Yuan-Hao Chang, and Wei-Kuan Shih, "Bridging DNA Storage and Computation: An Integrated Framework for Efficient Biomolecular Data Management," ACM Symposium on Applied Computing (SAC), Avila, Spain, Apr. 8-12, 2024.

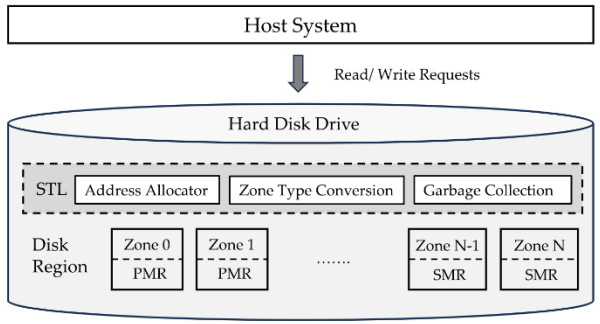
Research Summary 2023

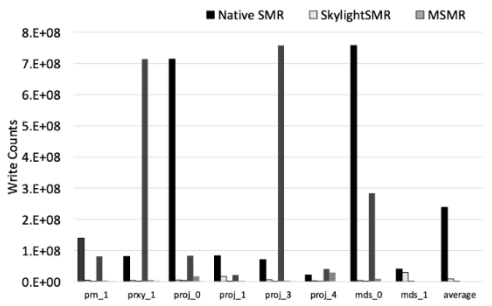
1. Storage Systems - Flash Drives and SMR Disks

Adaptive Mode-Switching for SMR Disks

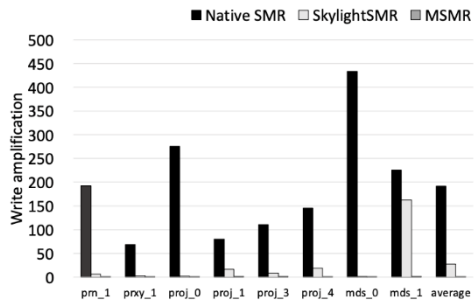
[ISOCC'23]

- **Observation**
  - The growing demand for storage capacity has made Shingled Magnetic Recording (SMR) disks increasingly popular in the storage device market.
  - The overlapping tracks of SMR disks cause write amplification during random writes, degrading performance compared to traditional Perpendicular Magnetic Recording (PMR) disks.
- **Goal**
  - Enable a seamless transition between PMR and SMR with effective management to meet both storage capacity demands and performance efficiency goals.
- **Main Idea**
  - The host system treats SMR disks as PMR, with additional management through the **Shingled Translation Layer (STL)**.
  - STL incorporates three key techniques:
    - **Address Allocator**
    - **Zone Type Conversion**: Dynamically adjusts storage density by transforming SMR and PMR.
      - PMR to SMR
      - SMR to PMR
    - **Garbage Collection**





(a) Write Counts



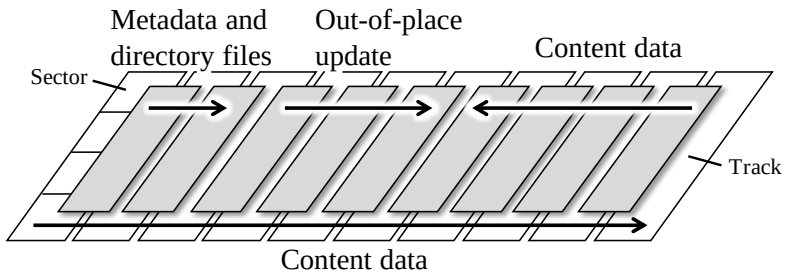
(b) Write amplification

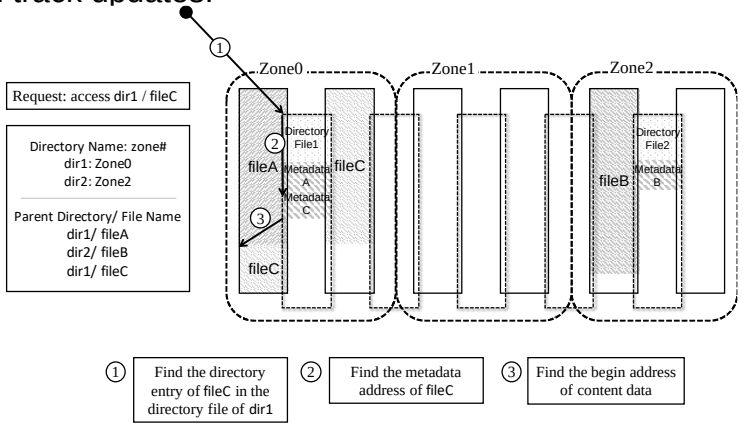
- Yi Syuan Lin, Yu-Pei Liang, Yu-Shan Yen, Yen-Ting Chen, Wei-Kuan Shih, and Yuan-Hao Chang, "Adaptive Mode-Switching for Write-amplification Reduction of SMR Disks," IEEE International SoC Conference (ISOCC), Jeju Island, Korea, Oct. 25-28, 2023.

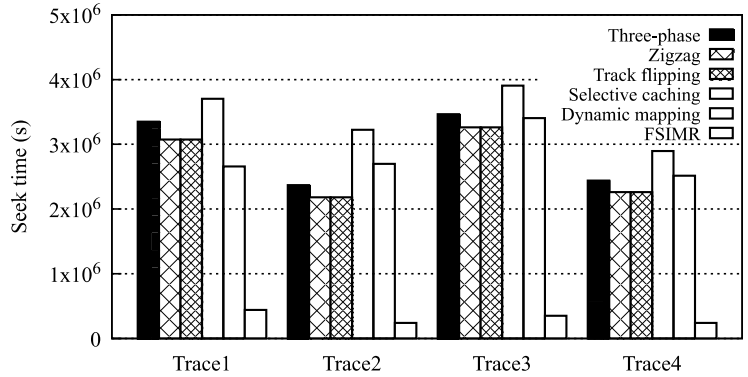
FSIMR: File-system-aware Data Management for Interlaced Magnetic Recording

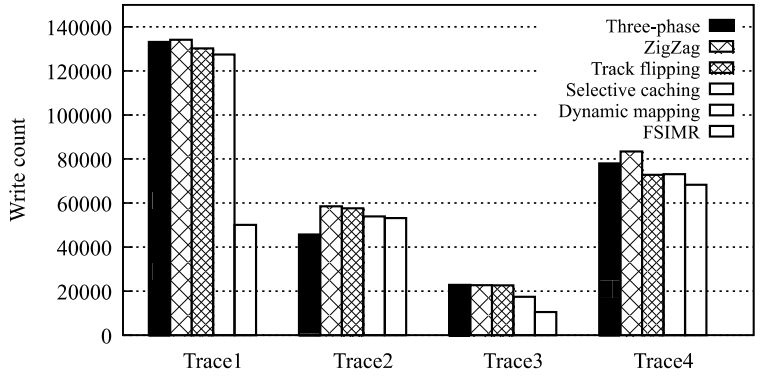
[ACM TECS'23, CODES'23]

- **Observation**
  - Write amplification in bottom tracks updates of IMR is a crucial performance issue.
  - Existing designs are mostly device-level solutions, which are limited due to the unawareness of data semantics of the file system.
- **Goal**: Leverage the data characteristics of file systems in data allocation on IMR drives to improve system performance.
- **Contribution**
  - Reduce access seek time
    - Files under the same directory are frequently updated at the same time.
    - Manage IMR into zones, which are related to each directory in file systems.
    - Data of files in the same directory are allocated to the same zone.
  - Improve write performance
    - Metadata and content data have different access frequencies.
    - Place the hot data (metadata) in top tracks and cold data (content data) in bottom tracks.
    - Adopt out-of-place update in bottom track updates.









- Yi-Han Lien, Yen-Ting Chen, Yuan-Hao Chang, Yu-Pei Liang, and Wei-Kuan Shih, "FSIMR: File-system-aware Data Management for Interlaced Magnetic Recording," ACM Transactions on Embedded Computing Systems (TECS), vol. 22, no. 5s, pp. 128:1-128:18, Sep. 2023. (Integrated with ACM/IEEE CODES+ISSS'23)
- Yi-Han Lien, Yen-Ting Chen, Yuan-Hao Chang, Yu-Pei Liang, and Wei-Kuan Shih, "FSIMR: File-system-aware Data Management for Interlaced Magnetic Recording," ACM/IEEE International Conference on Hardware/Software Codesign and System Synthesis (CODES+ISSS), Germany, Sep. 17-22, 2023. (Journal Track, Integrated with ACM TECS) **(Top Conference)**

HF-Dedupe: Hierarchical Fingerprint Deduplication Scheme for Flash-based Storage Systems

[ICCAD'23]

- Motivation
  - Severe data deduplication overheads
    - **Fingerprint computation & searching overheads**
    - **Fingerprint space overheads**
    - **Byte-by-byte data comparison overheads (on hash collisions)**
- Goal
  - Strike a balance among all different sources of the performance overheads, to optimize the overall storage performance.
- Main Idea
  - Use multi-level fingerprinting schemes to detect duplicate data.
  - Cache high-level fingerprints for fast duplication detection.

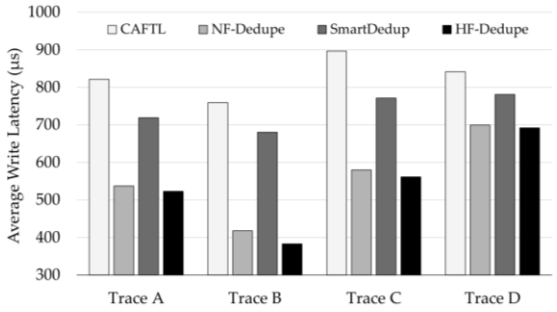


Fig. 5. Average Write Latency

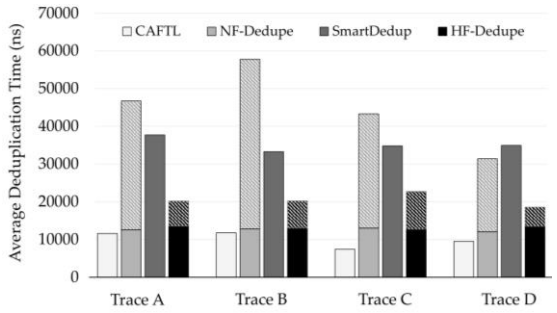
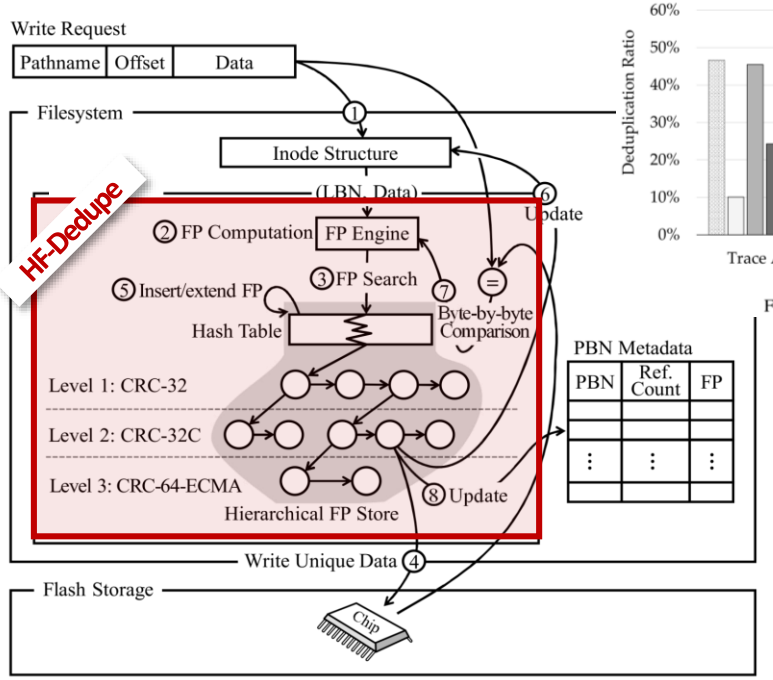


Fig. 6. Average Deduplication Time



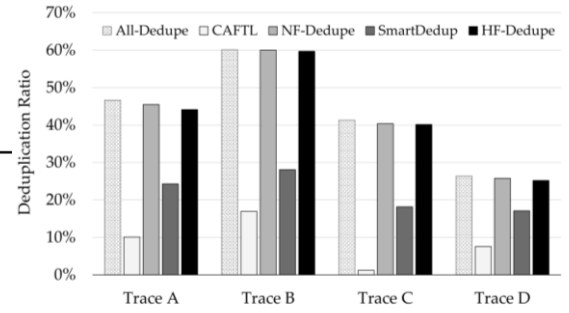


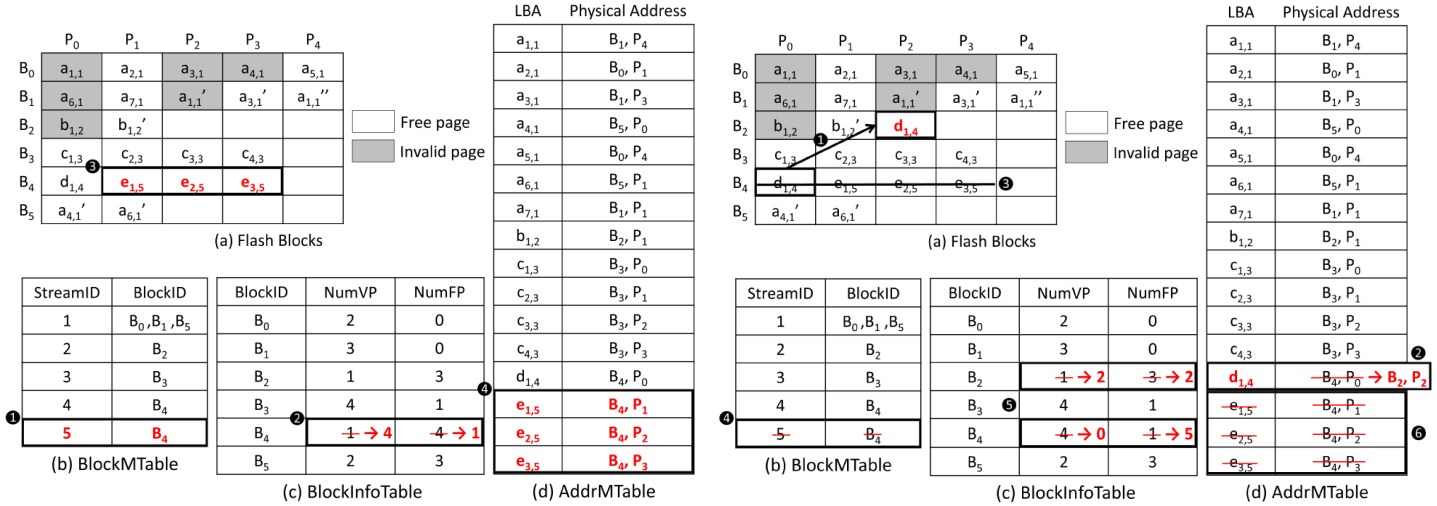
Fig. 7. In-line Deduplication Ratio

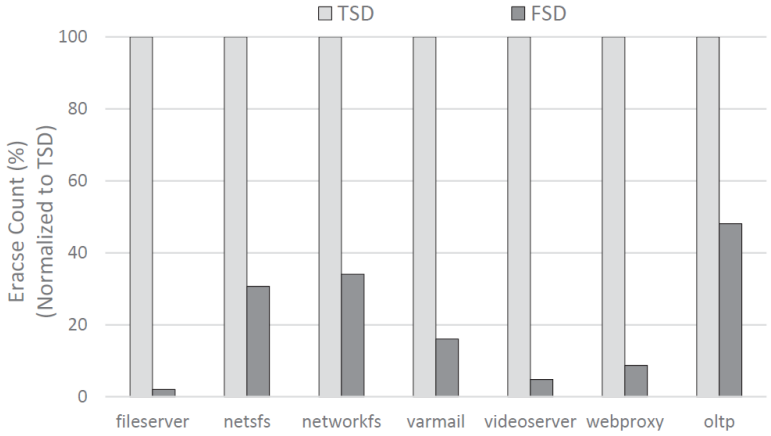
- Kai-Ting Weng, Yun-Shan Hsieh, Yen-Ting Chen, Yu-Pei Liang, Yuan-Hao Chang, Po-Chun Huang, and Wei-Kuan Shih, "HF-Dedupe: Hierarchical Fingerprint Scheme for High Efficiency Data Deduplication on Flash-based Storage Systems," ACM/IEEE International Conference on Computer-Aided Design (ICCAD), San Francisco, California, USA, Oct. 29 - Nov. 2, 2023**. (Top Conference)**

FSD: File-related Secure Deletion for SSD

[NVMSA'23]

- **Observation**
  - SSDs' inherent access characteristics bring in a grand challenge to provide secure deletion to thoroughly remove the sensitive data from the storage devices
  - The existing erase mechanism of SSDs may indirectly reduce the SSD lifetime when performing a secure deletion because of the massive block erases
- **Goal**
  - Propose a file-related secure deletion (FSD) scheme to alleviate the impact of secure deletion for prolonging SSD lifetime
- **Main Idea**
  - Exploit the file information hints to alleviate the potential endurance degradation when performing the secure deletion by optimizing the data allocation of the to-be-written data
  - Implement a file-related secure deletion mechanism to thoroughly remove the related file data from the SSDs with the recorded file information hints





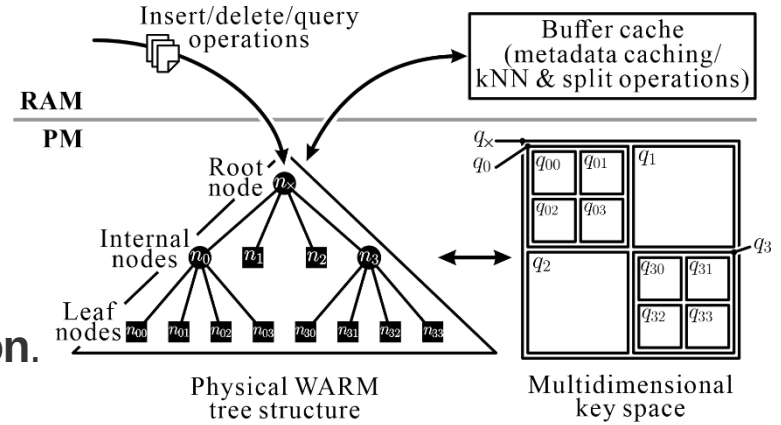
Provide secure deletion on SSD and mitigate about 80% of block erases and extra data movement

- Shih-Chun Chou, Yi-Shen Chen, Ping-Xiang Chen, Yuan-Hao Chang, Ming-Chang Yang, Tei-Wei Kuo, Yu-Fang Chen, and Yu-Ming Chang, "FSD: File-related Secure Deletion to Prolong the Lifetime of Solid-State Drives," IEEE Nonvolatile Memory Systems and Applications Symposium (NVMSA), Niigata, Taiwan, Aug. 30 - Sep. 1, 2023.

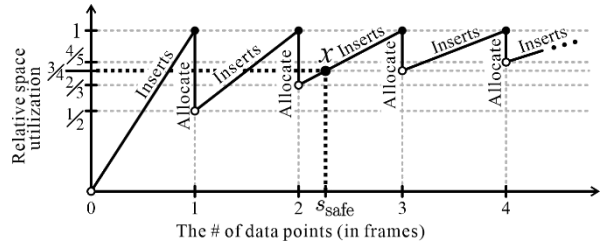
WARM-tree: Making Quadtrees Write-efficient and Space-economic on Persistent Memories

[ACM TECS'23, CODES'23]

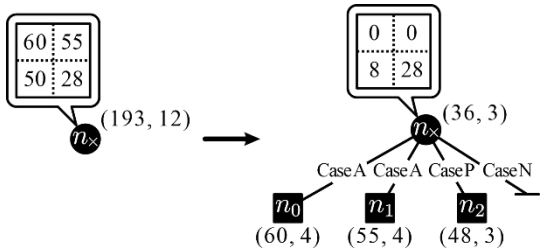
- **Observation**
  - Existing multidimensional data structures such as the kd-tree, R*-tree, and bucket point-region quadtree are not designed for modern PMs, and suffer from the **(1) serious write amplification**, or **(2) inability to satisfy different space utilization requirements.**
- **Goal**
  - Our proposed WARM-tree, a write-amplification-reducing multidimensional tree for point data on PMs, to **suppress write amplification** and **guarantee space utilization**.
- **Main idea**
  - **Incremental space allocation** for space efficiency enhancement
  - **Bucket reusing strategy** for suppressing the write amplification
  - **Providing worst-case space utilization guarantees in the form of $\frac{m-1}{m}$ $(m \in \mathbb{Z}^+)$**
  - **Reducing write traffic of key insertions by up to 48.10% and 85.86%.**



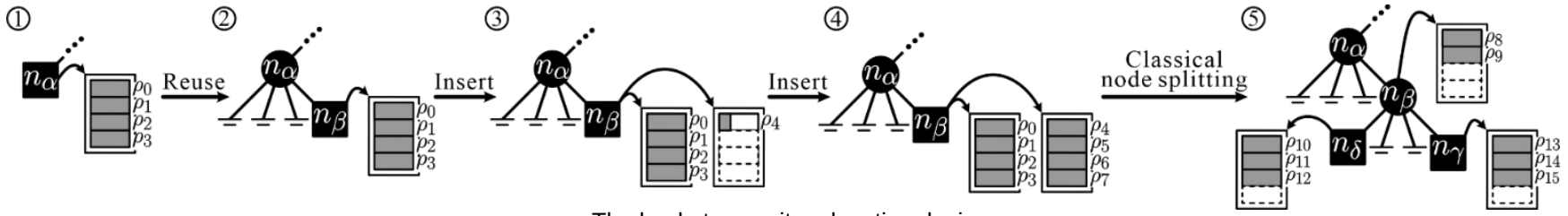
System architecture.



The relative space utilization under different node sizes.



Working example of node split.



The bucket capacity relaxation design.

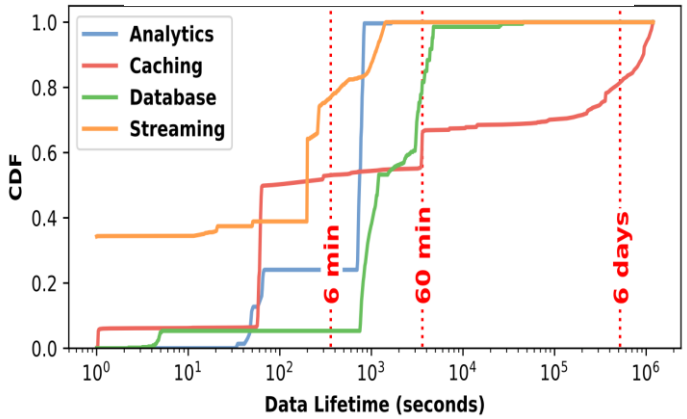
- Shin-Ting Wu, Liang-Chi Chen, Po-Chun Huang, Yuan-Hao Chang, Chien-Chung Ho, and Wei-Kuan Shih, "WARM-tree: Making Quadtrees Write-efficient and Space-economic on Persistent Memories," ACM Transactions on Embedded Computing Systems (TECS), vol. 22, no. 5s, pp. 119:1-119:26, Sep. 2023. (Integrated with ACM/IEEE CODES+ISSS'23)
- Shin-Ting Wu, Liang-Chi Chen, Po-Chun Huang, Yuan-Hao Chang, Chien-Chung Ho, and Wei-Kuan Shih, "WARM-tree: Making Quadtrees Write-efficient and Space-economic on Persistent Memories," ACM/IEEE International Conference on Hardware/Software Codesign and System Synthesis (CODES+ISSS), Germany, Sep. 17-22, 2023. (Journal Track, Integrated with ACM TECS) (Top Conference)

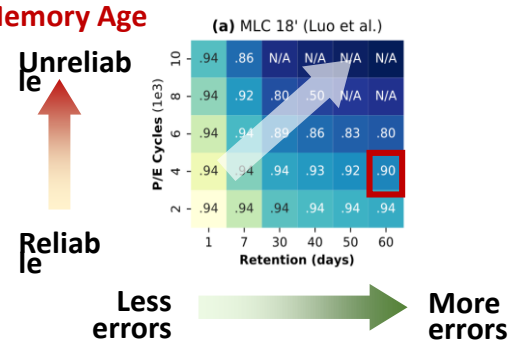
ZoneLife: Using Lifetime Semantics to Make SSDs Smarter

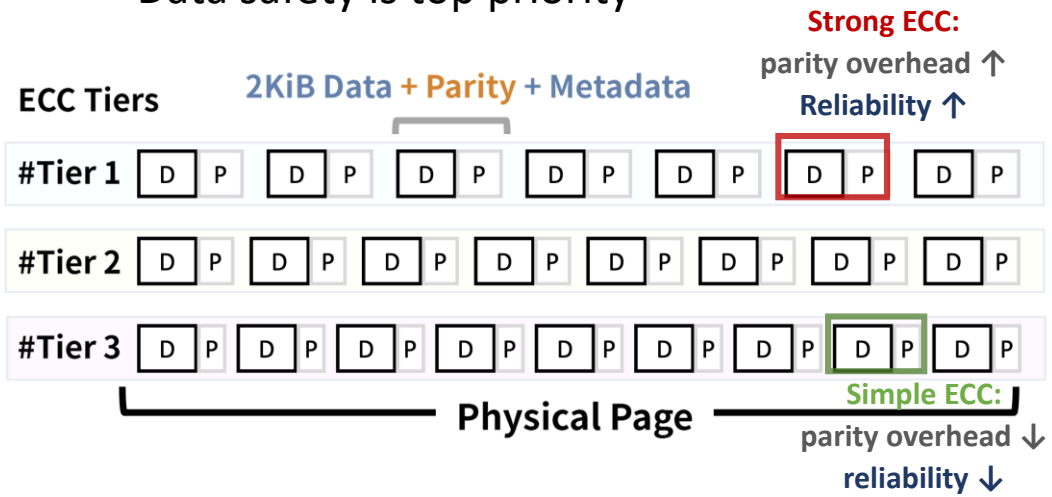
[IEEE TCAD'23]

- **Motivation**
  - Short-lived data are prevalent
  - if data known to retain for < 1 day
    - Write less overhead required for long-term storage protection
    - Write more user data before killing the SSD
- **Goal**
  - Host conveys data lifetime to storage
  - Adaptive error correction (ECC)
- **Main Idea**
  - Protect short-lived data with simple ECC
  - Larger capacity for short-lived data
  - Data safety is top priority

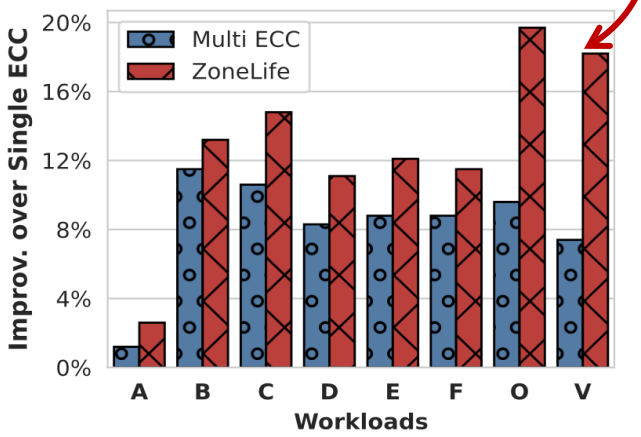
Data Lifetime Cumulative Distribution (CDF)



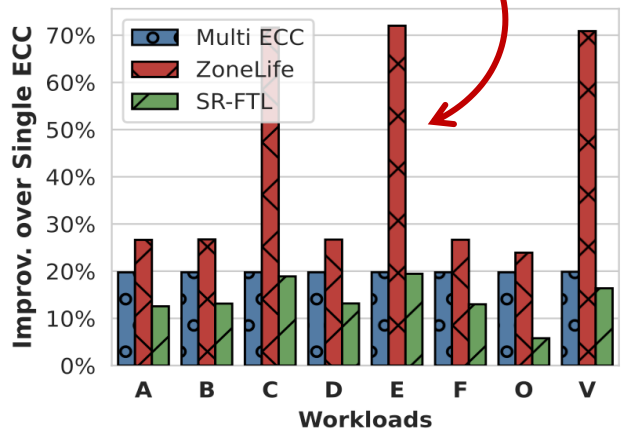




Spend **11 ~ 20%** less memory to complete a workload



Write **21 ~ 71%** more data before the memory wears out.

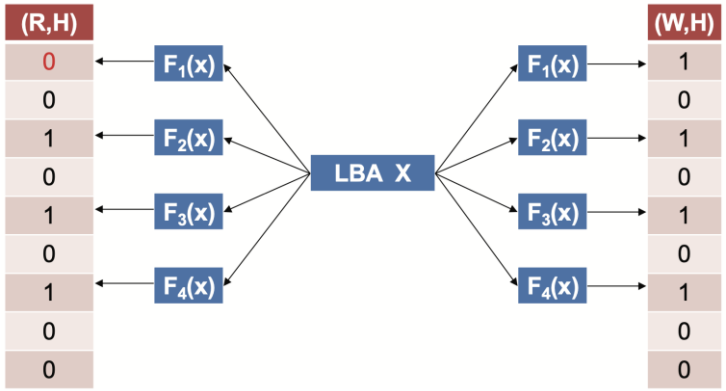


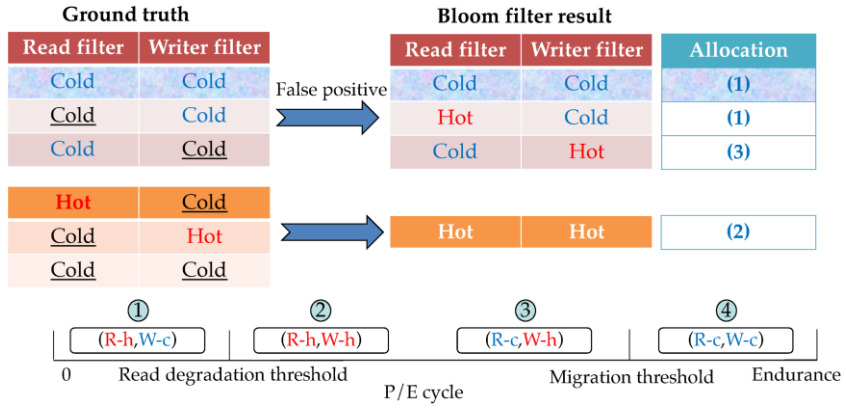
- Yun-Chih Chen, Chun-Feng Wu, Yuan-Hao Chang, and Tei-Wei Kuo, "ZoneLife: How to Utilize Data Lifetime Semantics to Make SSDs Smarter," accepted and to appear in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD).

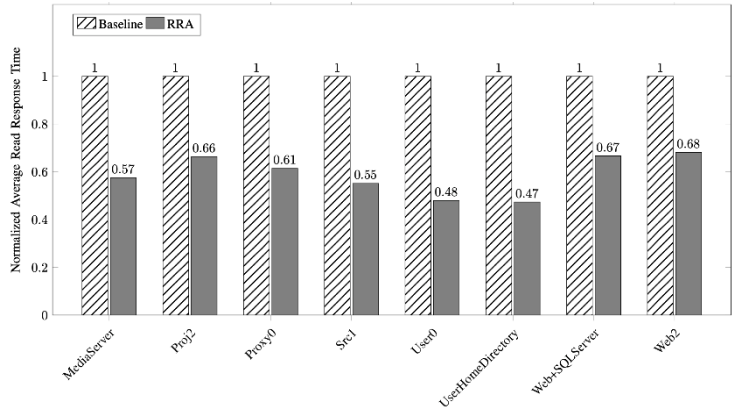
Retention-Aware Read Acceleration for LDPC-based Flash

[IEEE TCAD'23] [US 11,042,308]

- Observation:
  - LDPC-based NAND flash SSD faces the problem of read performance degradation due to the increase in raw bit error rate.
  - The two main factors that affect the raw bit error rate are data retention error and P/E cycle limitation.
- Goal:
  - Aiming to provide a stable and great read performance flash memory system, this work proposes a retention-aware read acceleration design that exploits access patterns to improve read performance.
- Main Idea:
  - Access feature identification efficiently detects and predicts the data lifetime and access behavior.
  - Request-based allocation allocates the suitable blocks for different data (with different data lifetime and access behavior).
  - Migration lazily balances the wearing level among blocks.







- Tse-Yuan Wang, Che-Wei Tsao, Yuan-Hao Chang, and Tei-Wei Kuo, "Retention-Aware Read Acceleration Strategy for LDPC-based NAND Flash Memory," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD), vol. 42, no. 12, pp. 4597-4605, Dec. 2023.
- Wei-Chen Wang, Ping-Hsien Lin, Tse-Yuan Wang, Yuan-Hao Chang, and Tei-Wei Kuo, "Memory Management Apparatus and Memory Management Method," Patent No.: US 11,042,308, Date of Patent: Jun. 22, 2021.

Retention Leveling: Enhancing Flash Reliability with the Awareness of Temperature

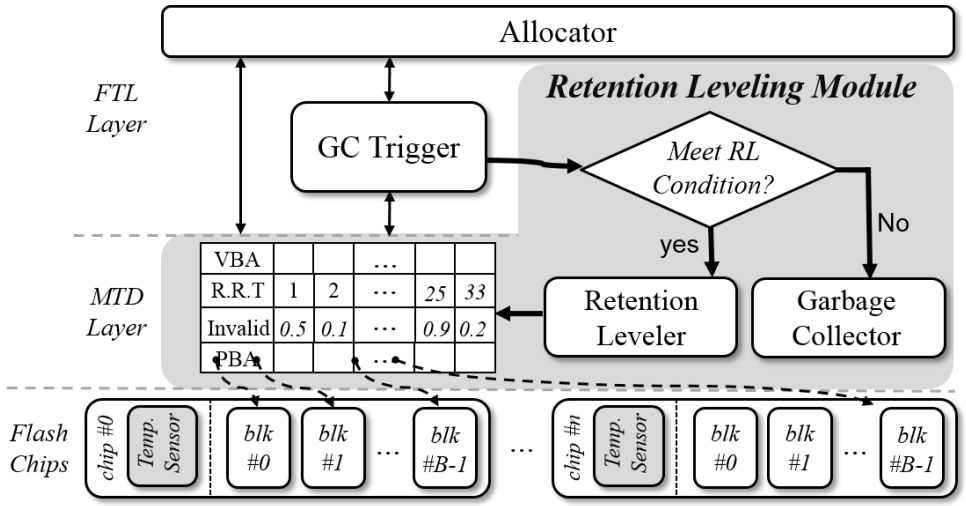
[NVMSA'23]

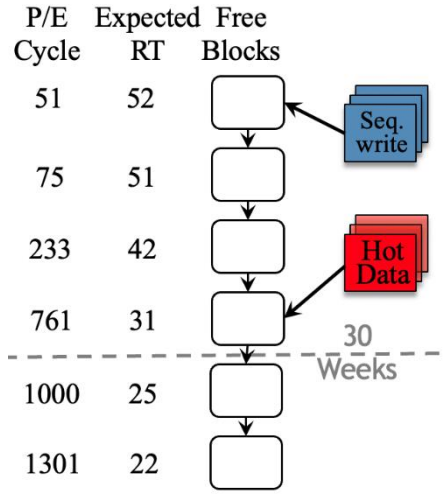
- **Motivation**
  - The error probability is exaggerated as the density of flash manufacture increases
  - Impact of **temperature** on retention time
  - Retention Time Relaxation
- **Goal**
  - Ensure data integrity from being hurt due to the retention error
  - To achieve the objective of "retention leveling"
- **Main Idea**
  - Temperature-aware Write Strategy
  - Retention-aware Block Management

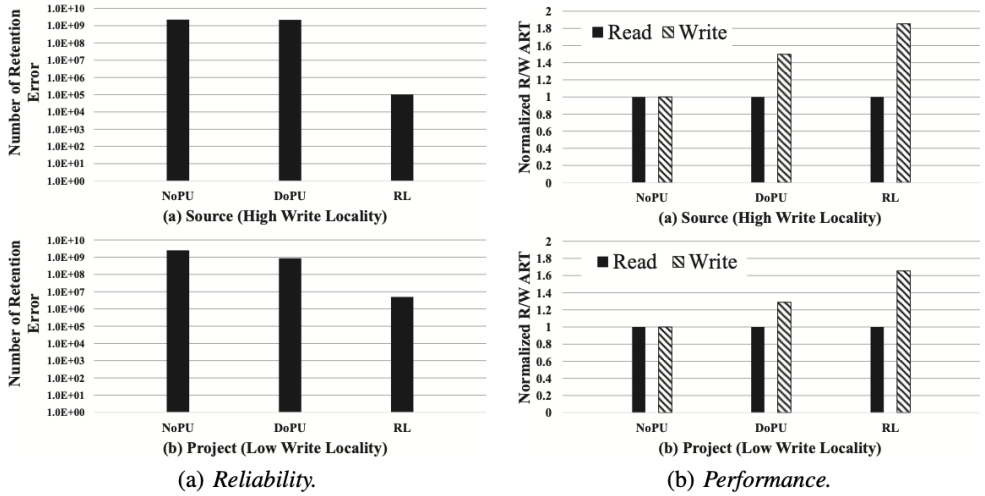
Retention time extension caused by writing on the higher temperature should be considered and exploited

Power Off Temperature \ Active Temperature	25	30	35	40	45	50	55
50						9	15
45					10	17	27
40				14	20	31	52
35			20	26	38	61	101
30		32	39	52	76	120	199
25	58	65	79	105	155	244	404

Source: JEDEC





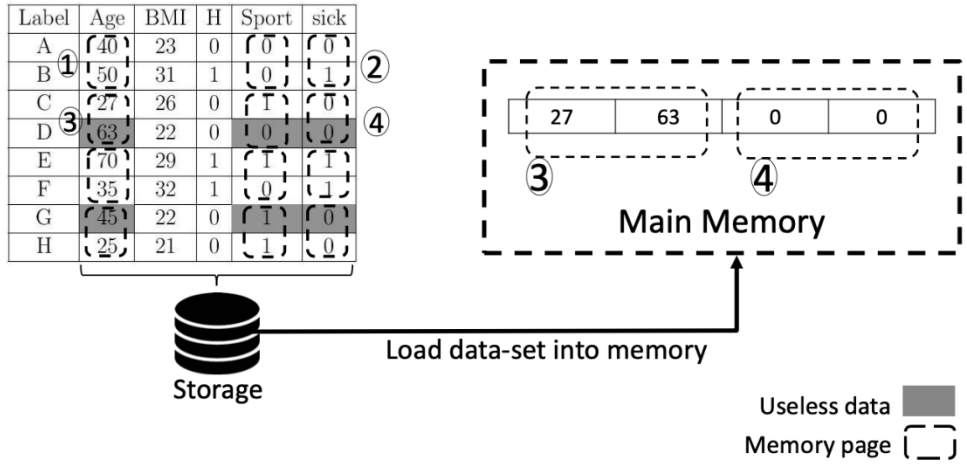


- Wei-Chen Wang, Chien-Chung Ho, Yuan-Hao Chang, Tei-Wei Kuo, and Yu-Ming Chang, "Retention Leveling: Leverage Retention Refreshing and Wear Leveling Techniques to Enhance Flash Reliability with the Awareness of Temperature," IEEE Nonvolatile Memory Systems and Applications Symposium (NVMSA), Niigata, Taiwan, Aug. 30 - Sep. 1, 2023.

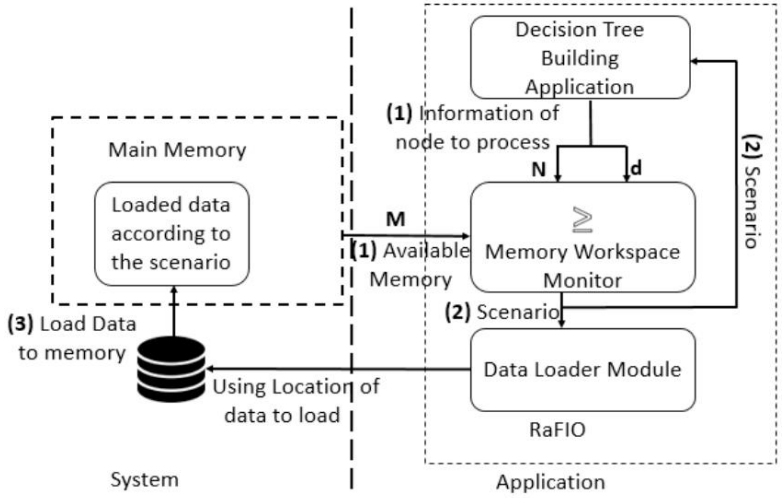
Random Forest I/O-aware Algorithm

[IEEE TC'23, SAC'21]

- **Observation**
  - During training random forest, performance drops significantly when the dataset size is larger than the available memory size.
  - Reasons: Randomly bagging data causes unnecessary data movements.
- **Goal**
  - Reduce unnecessary data movement by avoiding loading useless data and smartly selecting the data according to their reuse pattern in the following tree building steps
- **Main Idea**
  - Decision Tree Building Module: Perform on-demand data loading according to the available memory space.
  - Data Loader Module: Pre-process data to easily locate useful data without reading them multiple times during data loading.



Unnecessary Data Movements



Unnecessary Data Movements

- Camelia Slimani, Chun-Feng Wu, Stephane Rubini, Yuan-Hao Chang, and Jalil Boukhobza, "Accelerating Random Forest on Memory-Constrained Devices through Data Storage Optimization," IEEE Transactions on Computers (TC), vol. 72, no. 6, pp. 1595-1609, Jun. 2023.
- Camelia Slimani, Chun-Feng Wu, Yuan-Hao Chang, Stephane Rubini, and Jalil Boukhobza, "RaFIO: A Random Forest I/O-Aware Algorithm," ACM Symposium on Applied Computing (SAC), Gwangju, South Korea, Mar. 22-26, 2021.

2. NVM Main Memory and Storage

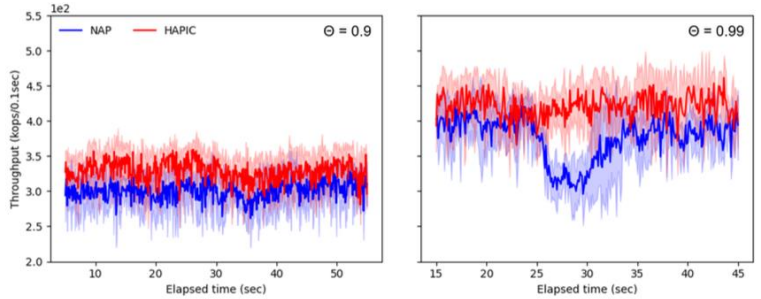
HAPIC: a Scalable, Lightweight and Reactive Cache for Persistent-Memory-based Index

[ICCAD'23]

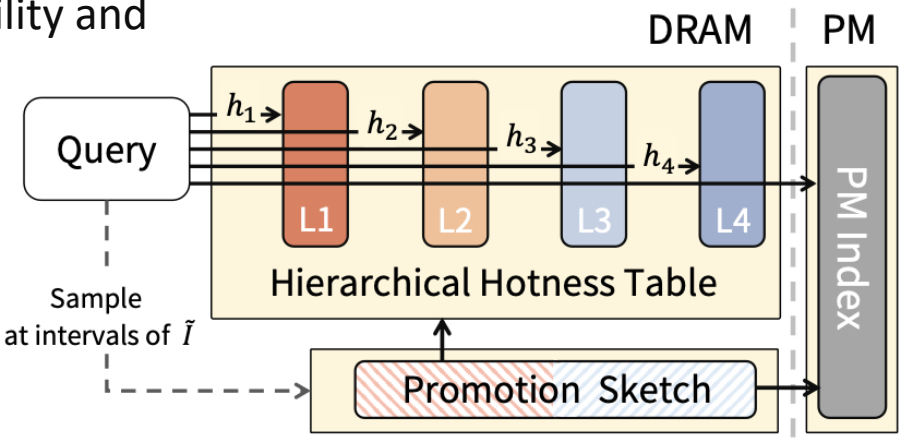
- **Motivation**
  - Persistent memory-based indexes face challenges under read-intensive, skewed, and dynamic workloads
  - Existing strategies like NAP fail to react quickly to shifting query hotspots

- **Approach**
  - **HAPIC**: a scalable index leveraging a hierarchy of hash tables to identify hotspots and adapt to shifting workloads.
  - structural hotness estimation using a multi-level hash table.
  - *Promotion Sketch*: probabilistic, low-overhead hotspot adjustment.
  - *Epoch-based hotness promotion*: prevent overreaction.
  - CPU-aligned hash tables and adaptive sampling to maintain scalability and responsiveness.

- **Results**
  - Up to 14% higher stable read throughput compared to the state-of-the-art approach.
  - Reacts more quickly to workload shifts, minimizing throughput drops and fluctuations.
  - Linear scalability under high concurrency (better than ARC and NAP.)



Prevent throughput drop!

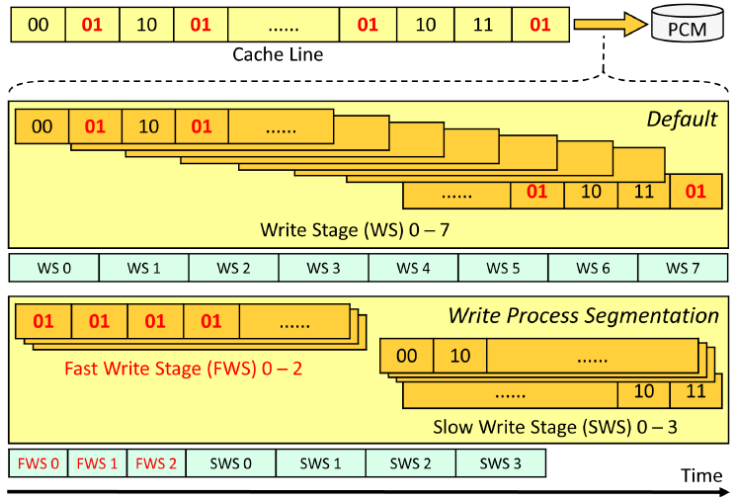


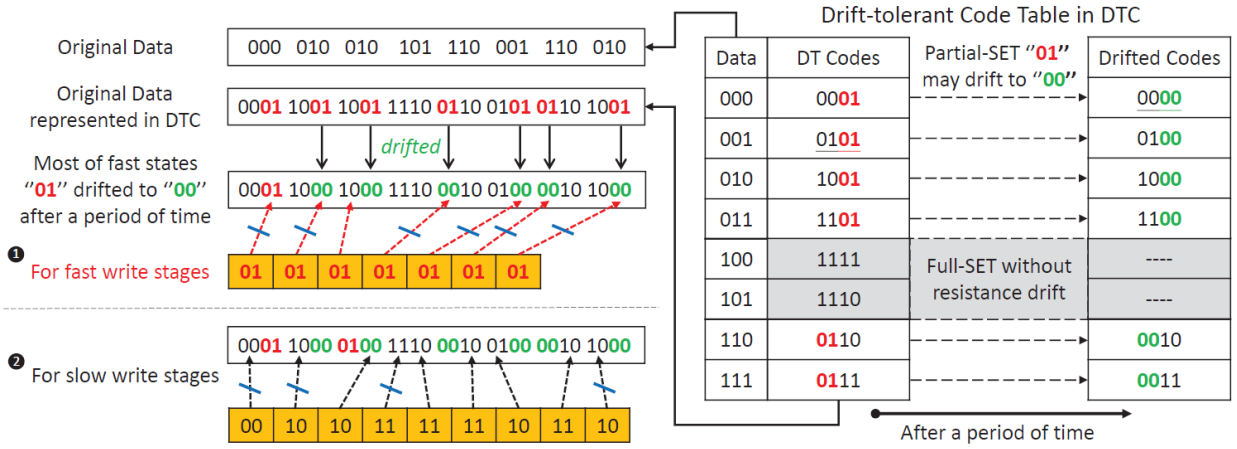
- Chih-Ting Lo, Yun-Chih Chen, Yuan-Hao Chang, and Tei-Wei Kuo, "HAPIC: a Scalable, Lightweight and Reactive Cache for Persistent-Memory-based Index," ACM/IEEE International Conference on Computer-Aided Design (ICCAD), San Francisco, California, USA, Oct. 29 - Nov. 2, 2023. **(Top Conference)**

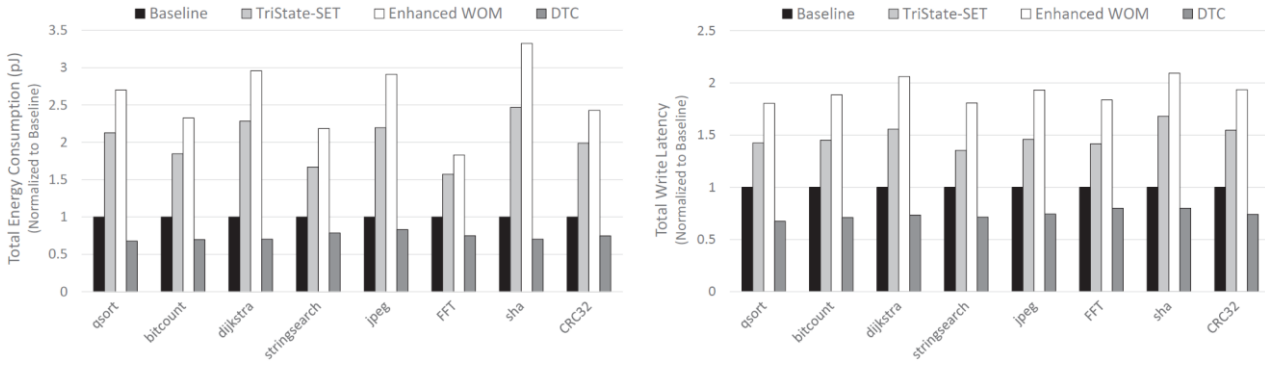
DTC: A Drift-Tolerant Coding for MLC Phase-Change Memory

[IEEE TCAD to appear, ISLPED'22]

- **Observation**
  - MLC PCM suffers from the problems of resistance drift errors and asymmetric writes due to the narrow margins between the multiple states.
- **Goal**
  - Design a drift-tolerant coding (DTC) scheme to efficiently improve the performance and energy efficiency of MLC PCM without sacrificing any data accuracy
- **Main Idea**
  - Propose a two-generation code to tolerate the resistance drift errors with Partial-SET
  - Divide the write process of the cache line into different write stages to improve the write performance
  - Eliminate unnecessary update operations with the read operations to further reduce the write latency and energy consumption







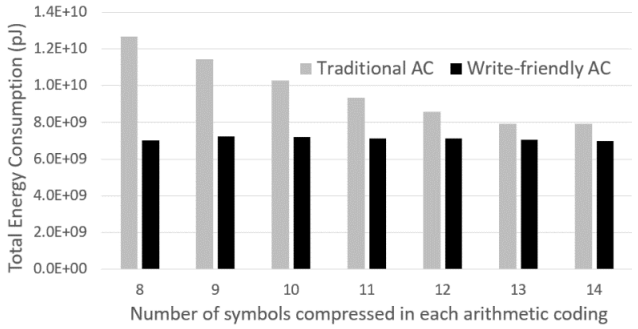
Reduce 16.8-32.1% energy consumption and 20.1-32.6% write latency, compared to the existing well-known schemes

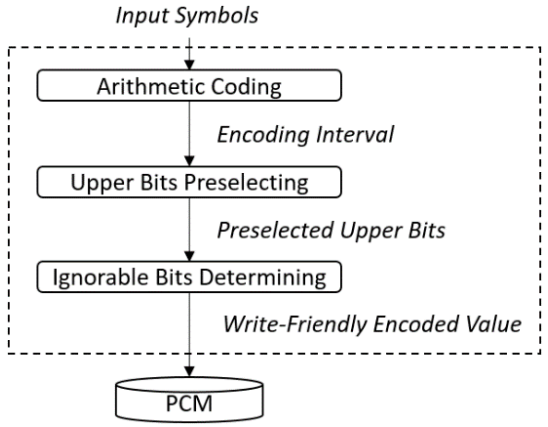
- Yi-Shen Chen, Yuan-Hao Chang, and Tei-Wei Kuo, "DTC: A Drift-Tolerant Coding to Improve the Performance and Energy Efficiency of Multi-Level-Cell Phase-Change Memory," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD), vol. 42, no. 10, pp. 3185-3195, Oct. 2023.
- Yi-Shen Chen, Yuan-Hao Chang, and Tei-Wei Kuo, "Drift-tolerant Coding to Enhance the Energy Efficiency of Multi-Level-Cell Phase-Change Memory" ACM/IEEE International Symposium on Low Power Electronics and Design (ISLPED), Boston, MA, USA, Aug. 1-3, 2022. **(Top Conference)**

Write-friendly Arithmetic Coding for NVM

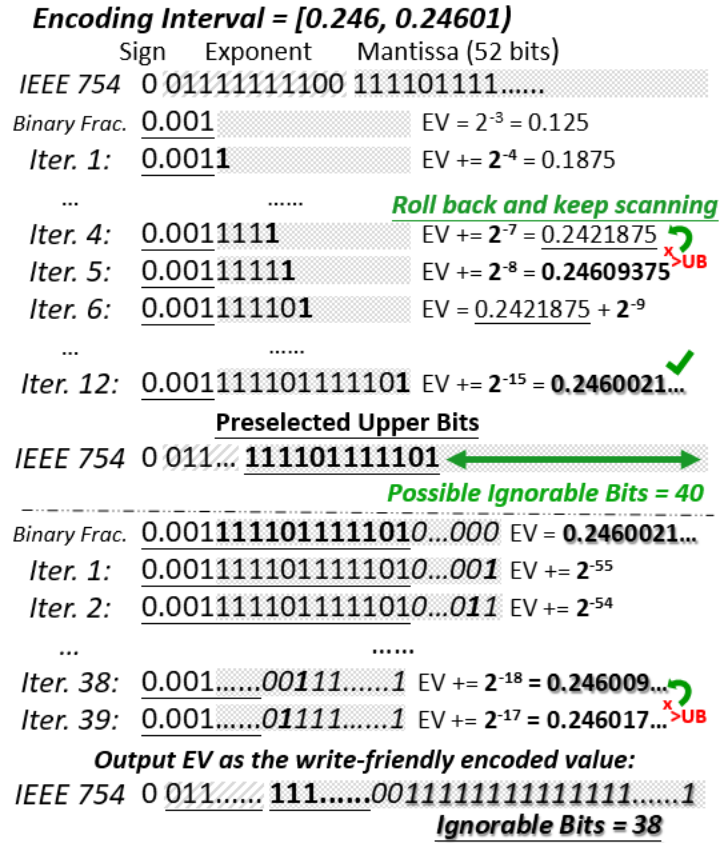
[IEEE TCAD'23, ASP-DAC'21]

- **Observation**
  - Storage-Class Memory technologies and data compression techniques can be used to alleviate the energy consumption of wearable IoT devices
  - However, the information gap between the PCM devices and data compression techniques hinders the cooperation among the two techniques for achieving further performance optimization
- **Goal**
  - Design an energy-aware and write-friendly arithmetic coding (AC) to improve energy efficiency of PCM
- **Main Idea**
  - Exploit the property of encoding interval in arithmetic coding to smartly choose an ideal encoded value consists of most ignorable bits, so as to reduce the number of write operations during the compression
    - Upper Bits Preselecting
    - Ignorable Bits Determining





Reduce 10.6-44.6% energy, compared to the traditional AC

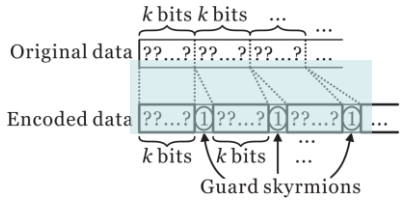


- Yi-Shen Chen, Chun-Feng Wu, Yuan-Hao Chang, and Tei-Wei Kuo, "Energy Efficiency Enhancement of SCM-based Systems: A Write-friendly Coding," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD), vol. 42, no. 5, pp. 1425-1437, May 2023.
- Yi-Shen Chen, Chun-Feng Wu, Yuan-Hao Chang, and Tei-Wei Kuo, "A Write-friendly Arithmetic Coding Scheme for Achieving Energy-Efficient Non-Volatile Memory Systems," ACM/IEEE Asia and South Pacific Design Automation Conference (ASP-DAC), Tokyo, Japan, Jan. 18-21, 2021.

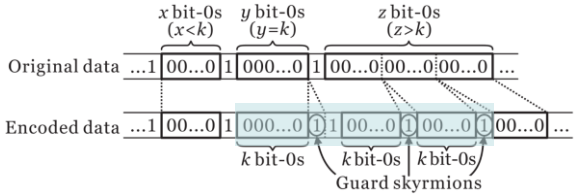
Granularity-driven Management for Reliable Skyrmion Racetrack Memories

[IEEE TETC'23]

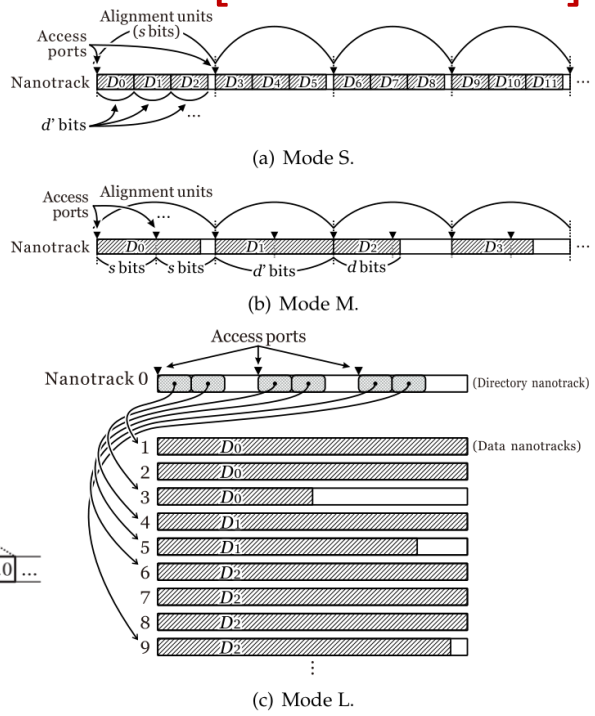
- **Observation**
  - The unique position errors and data representation errors on SK-RM can be solved by the existing encoding schemes. However, they yield variable-length encoded results, which leads to extra complexities of data management, such as data layout and indexing strategies.
- **Goal**
  - A system-level, granularity-driven management scheme for SK-RM is necessary for guaranteeing data reliability and enhancing access performance of SK-RM.
- **Main Idea**
  - Exploit two existing flyweight bit-stuffing approaches, called FGS and GGS in this work, to solve the data representation problem of SK-RM.
  - The design space of different data layout and alignment strategies is explored, with joint consideration over the parallel access capability of SK-RM. ⇒ We propose different management schemes, namely Modes S, M, and L, for data at different degrees of granularity.



Fixed guard skyrmion (FGS).



Greedy guard skyrmion (GGS)



Data alignment strategies of different granularities.

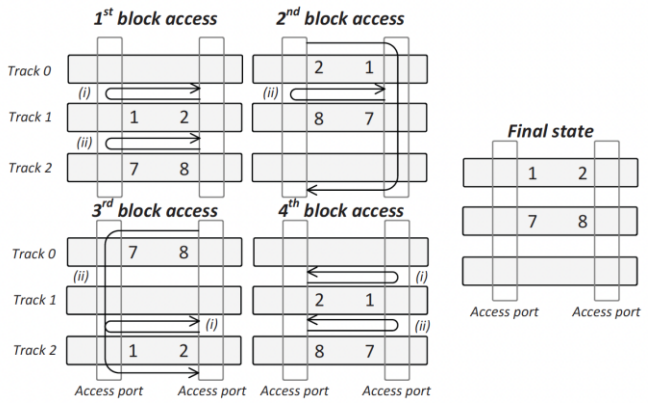
Comparison of Different Modes of the Proposed Management Scheme.

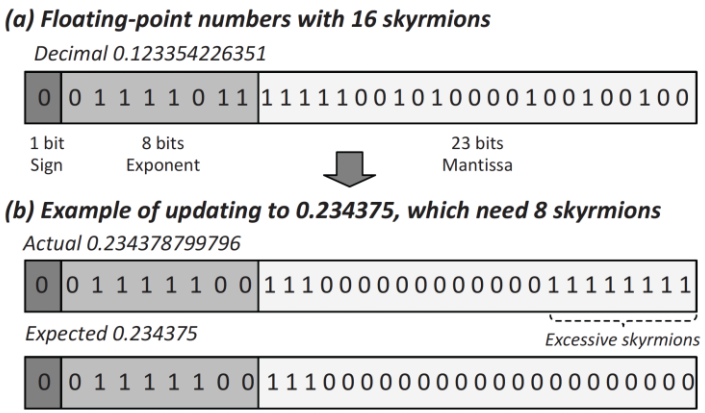
Properties	Mode S	Mode M	Mode L
Encoded data item size	< data segment	≥ data segment, < nanotrack.	≥ nanotrack.
Potential application scenarios	SPM & high-level caches	Low-level caches	Main memory & persistent storage
Variable-sized data item support?	✓ (based on extension of §3.3.2)	✓ (< nanotrack size)	✓ (≥ nanotrack size)
Data encoding	FGS	GGS	GGS
Data alignment	1^+ data items → 1 data segment	1 data item → 1^+ data segments	1 data item → 1^+ nanotracks
Directory needed?	×	×	✓

- Yun-Shan Hsieh, Po-Chun Huang, Yuan-Hao Chang, Bo-Jun Chen, Wang Kang, and Wei-Kuan Shih, "Granularity-driven Management for Reliable and Efficient Skyrmion Racetrack Memories," IEEE Transactions on Emerging Topics in Computing (TETC), vol. 11, no. 1, pp. 95-111, Jan. 2023.

Sky-NN: Enabling Efficient Neural Network Data Processing with Skyrmion Racetrack Memory

- **Observation** [ISLPED'23]
  - Skyrmion racetrack memory (SK-RM) is regarded as a promising NVRAM. However, directly applying existing data process methods of neural networks on SK-RM hinders the benefits and performance.
- **Goal**
  - Reconsider NN computations with the awareness of SK-RM characteristics
- **Main Idea**
  - Enable efficient NN data processing methods on SK-RM by utilizing the distinct shift and re-assemblability capability of skyrmions.
  - Completely remove the need of skyrmion injections and deletions after the first write of NN models on SK-RM





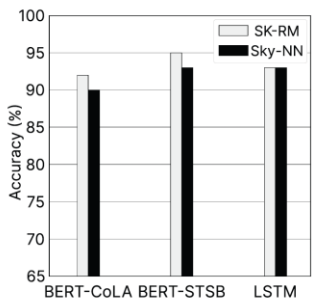


Fig. 12. Accuracy comparison.

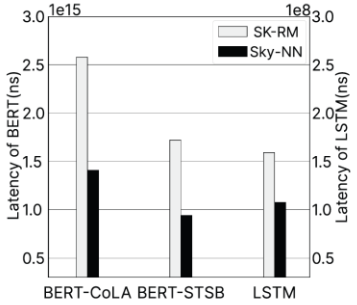


Fig. 13. Latency comparison.

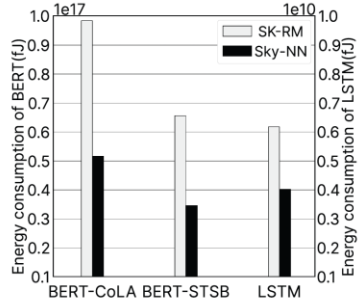


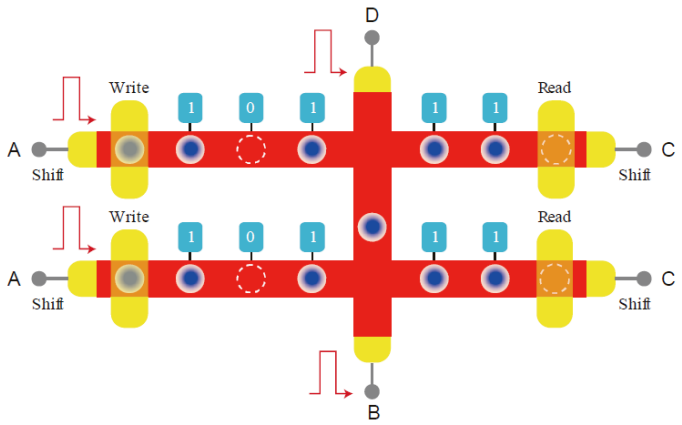
Fig. 14. Energy comparison.

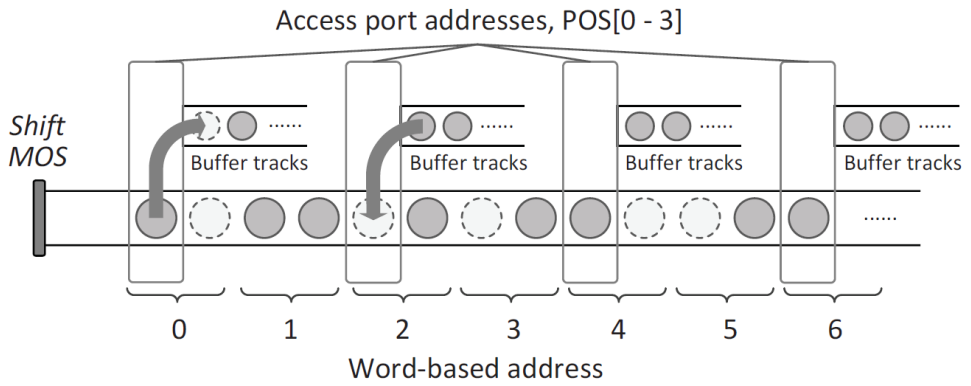
Reduce 41.06% energy and 43.39% latency with 0.66% precision difference

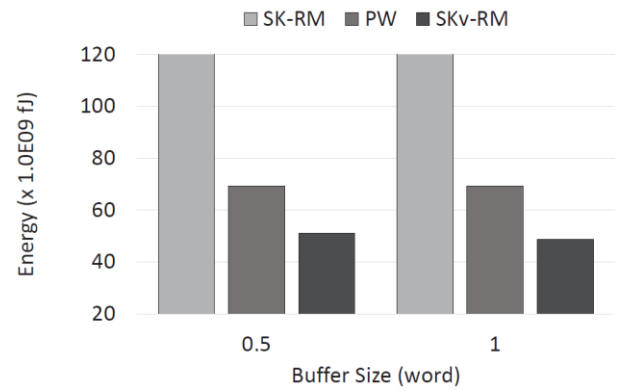
- Yong-Cheng Liao, Shuo-Han Chen, Yuan-Hao Chang, and Yu-Pei Liang, "Sky-NN: Enabling Efficient Neural Network Data Processing with Skyrmion Racetrack Memory," ACM/IEEE International Symposium on Low Power Electronics and Design (ISLPED), Vienna, Austria, Aug. 7-8, 2023. **(Top Conference)**

Skyrmion Vault: Maximizing Skyrmion Lifetime

- **Observation** [ADP-DAC'23]
  - Skyrmion racetrack memory (SK-RM) is regarded as a promising NVRAM. However, it could lead to excessive energy consumption due to ***shift*** and ***insert*** operations
- **Goal**
  - Lowering the number of skyrmion injections for energy conservation by utilizing ***vertical shift*** and ***buffer tracks*** features
- **Main Idea**
  - Preserving skyrmions over multiple write requests while alleviating both shift and injection overhead
  - Introduce SKv-RM to utilize buffer tracks for extending skyrmion lifespan







Reduce the energy consumption up to 56.8% & Prolong the lifespan of skyrmions up to 57.3x

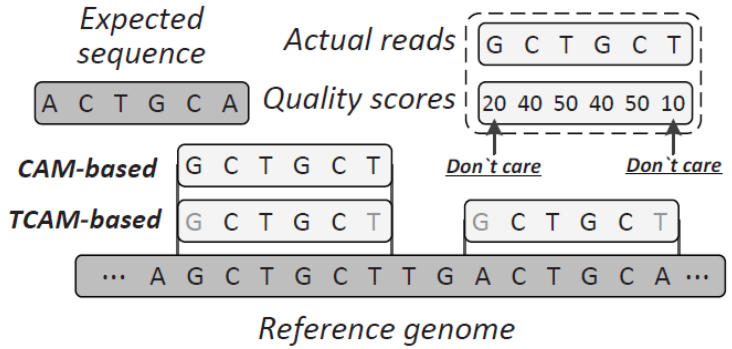
- Syue-Wei Lu, Shuo-Han Chen, Yu-Pei Liang, Yuan-Hao Chang, Kang Wang, Tseng-Yi Chen, and Wei-Kuan Shih, "Skyrmion Vault: Maximizing Skyrmion Lifespan for Enabling Low-Power Skyrmion Racetrack Memory," ACM/IEEE Asia and South Pacific Design Automation Conference (ASP-DAC), Tokyo, Japan, Jan. 16-19, 2023.

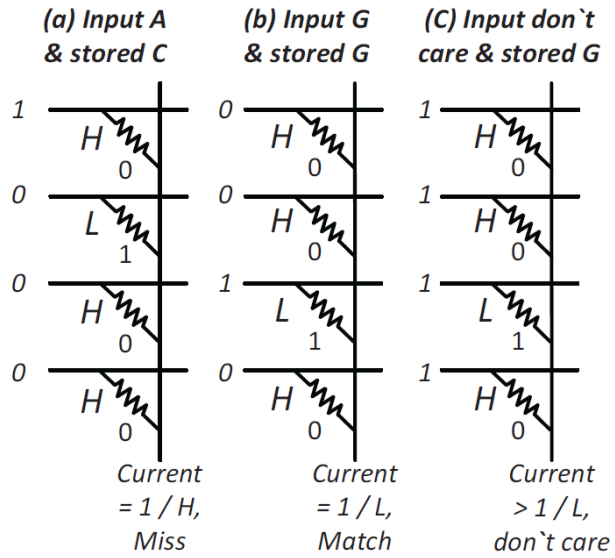
3. In/Near-Memory Processing with NVM

Enabling Highly-Efficient DNA Sequence Mapping via ReRAM-based TCAM

[ISLPED'23]

- **Observation**
  - In the post-pandemic era, third-generation DNA sequencing (TGS) has received increasing attention. However, much less effort has been devoted to DNA sequence mapping acceleration while considering both the memory wall issue and the challenges of TGS technologies.
- **Goal**
  - Propose a novel resistive random-access memory (ReRAM)-based ternary content-addressable memory (TCAM)
  - Exploit the intrinsic parallelity of ReRAM crossbar for mapping acceleration.
- **Main Idea**
  - Exploit the don't care feature of TCAM to mark nucleotides based on quality scores provided by TGS technologies.
  - Implement the functionality of TCAM within ReRAM crossbar circuitry without including new transistors and resistors





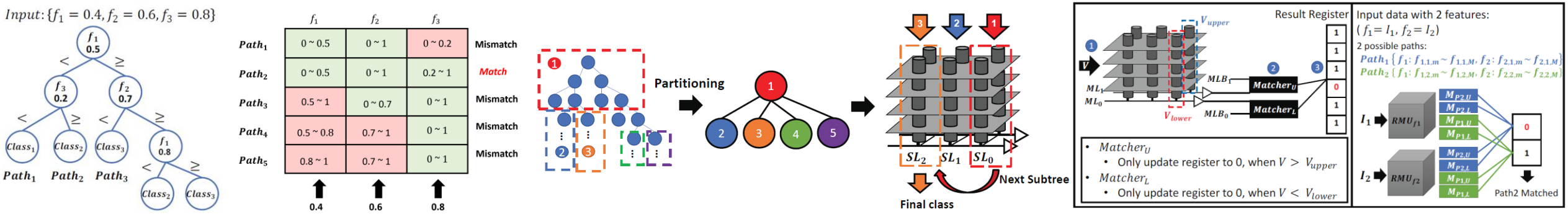
Reduce 99.72% energy and 99.76% latency, compared to the conventional CPU-based Minimap tool

- Yu-Shao Lai, Shuo-Han Chen, and Yuan-Hao Chang, "Enabling Highly-Efficient DNA Sequence Mapping via ReRAM-based TCAM," ACM/IEEE International Symposium on Low Power Electronics and Design (ISLPED), Vienna, Austria, Aug. 7-8, 2023. **(Top Conference)**

A Digital 3D TCAM Accelerator for the Inference Phase of Random Forest

[DAC'23]

- **Observation**
  - Ternary content addressable memory (TCAM) that utilize processing-in memory and high parallelism of crossbar memory and thus is suitable for the memory-bound inference of random forests.
  - However, the reliability and explosive growth of paths become critical issues on applying TCAM to inference phase
- **Objective**
  - A digital 3D TCAM-based accelerator for the inference phase of random forests is proposed with higher reliability than the previous analog based approach.
- **Main Idea**
  - The proposed architecture can check if input values match a specific range in parallel while providing a high density based on the 3D ReRAM TCAM architecture.
  - A subtree-partitioning algorithm spits each decision tree into multiple subtrees to reduce the search complexity and a data placement strategy is designed for the 3D ReRAM TCAM accelerator.
- **Result**: Achieve an average of **3.13x** higher throughput with **22x** more energy saving than the GPU approach

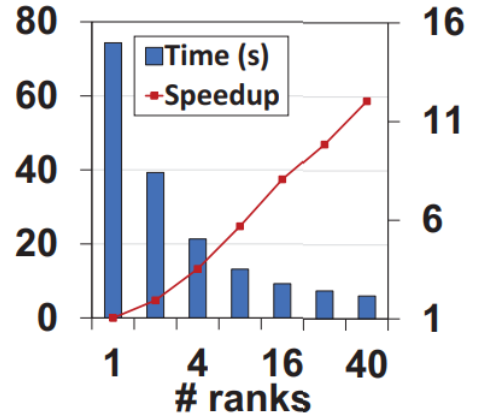


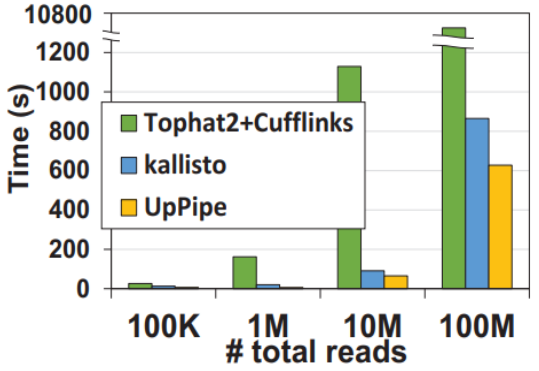
- Chieh-Lin Tsai, Chun-Feng Wu, Yuan-Hao Chang, Han-Wen Hu, Yung-Chun Lee, Hsiang-Pang Li, and Tei-Wei Kuo, "A Digital 3D TCAM Accelerator for the Inference Phase of Random Forest," ACM/IEEE Design Automation Conference (DAC), San Francisco, CA, USA, Jul. 9-13, 2023. (Acceptance rate: 23%) (Top Conference)

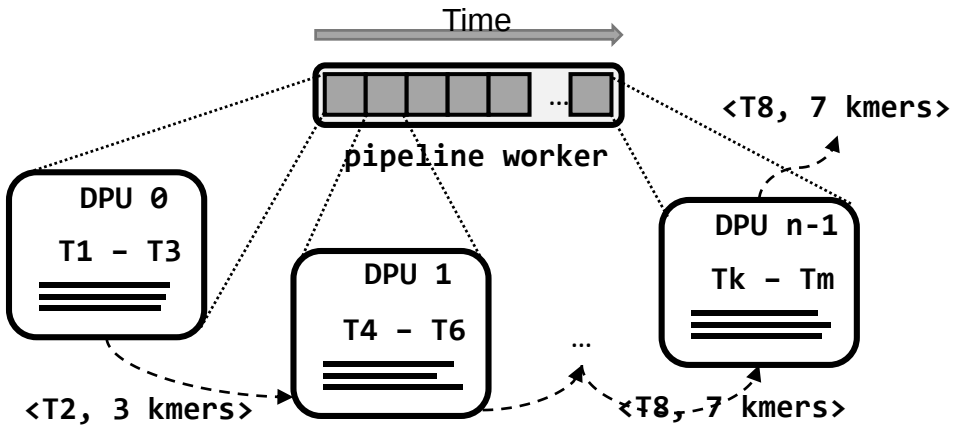
UpPipe: In-Memory Processors for RNA-seq Quantification

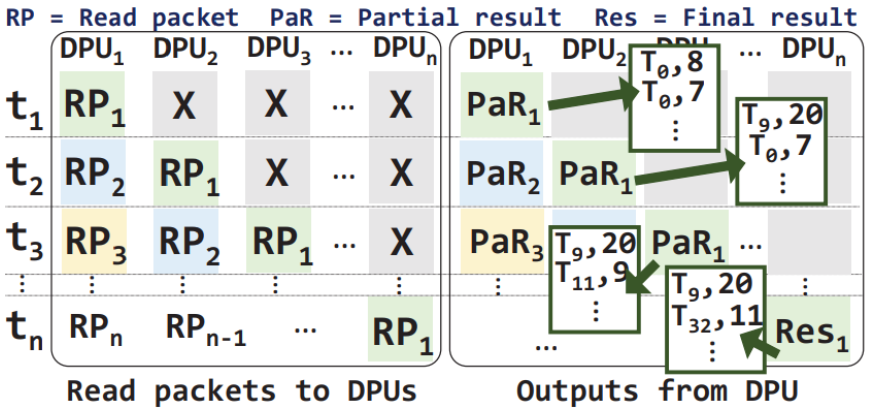
[DAC'23]

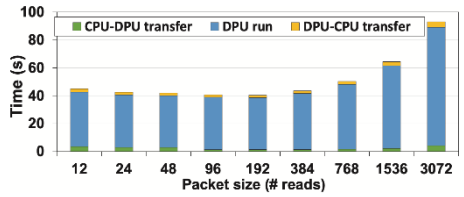
- **Motivation**
  - Limited MRAM capacity means it may not be possible to store a complete hash table on it
  - Data needs to be shared between DPUs, which may incur heavy data transfers
  - Heavy data transfers: additional overheads and more DPU idle time
- **Goal**
  - How to address the problem that **insufficient memory size for storing the hash table**
  - How dose the RNA read **align in each DPU system**
- **Main Idea**
  - We group serval DPUs into "pipeline worker" to hold the transcriptome
  - Using DPU-friendly transcriptome allocation to place the hash table into pipeline worker
  - We propose the DPU-aware pipeline management to finish alignment











- Liang-Chi Chen, Chien-Chung Ho, and Yuan-Hao Chang, "UpPipe: A Novel Pipeline Management on In-Memory Processors for RNA-seq Quantification," ACM/IEEE Design Automation Conference (DAC), San Francisco, CA, USA, Jul. 9-13, 2023. (Acceptance rate: 23%) **(Top Conference)**

4. Intermittent Systems, Real-time Systems, and Operating Systems

TRAIN: Time-Aware Neural Inference on Intermittent Systems [ICCAD'23]

- **Background on Multi-Exit Network:**
  - *Multi-exit networks* is recently proposed to provide a proper balance between energy-accuracy tradeoff.
- **Motivation:** Existing multi-exits network do not take the inference time into account.
- **Goal:** To deploy the neural network models on the intermittent systems by considering energy, time constraint, and delivered model accuracy.
- **Contribution:**
  - TRAIN presents a larger solution space by offering different choices to be made when encountering a to-be-executed model layer
  - A reinforcement learning based method (RL-IIS) is developed to help select a good decision point among the enlarged solution space.
  - A performance metric *inference efficiency* is developed to quantify the actual delivered inference accuracy at runtime.

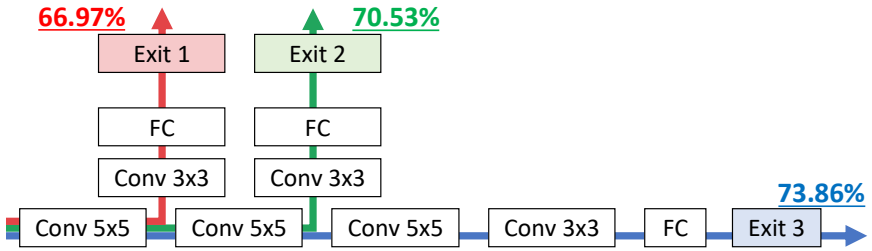


Fig. An example of a multi-exit NN model.

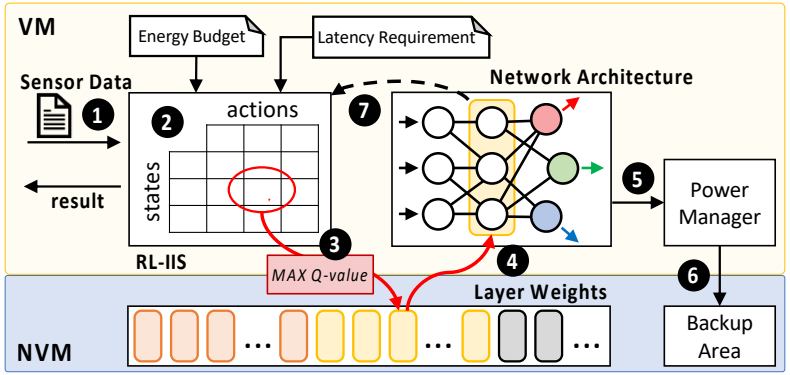
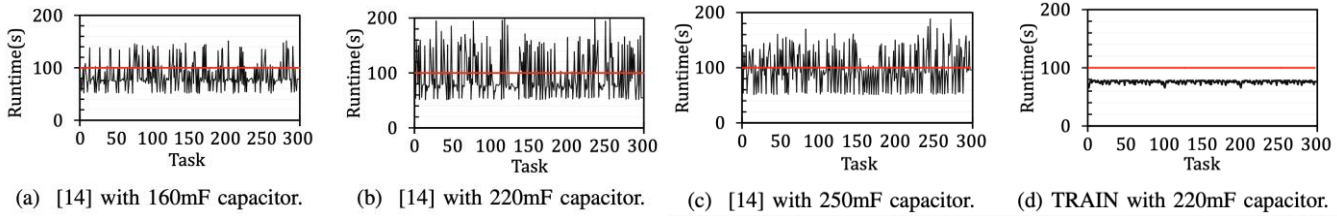
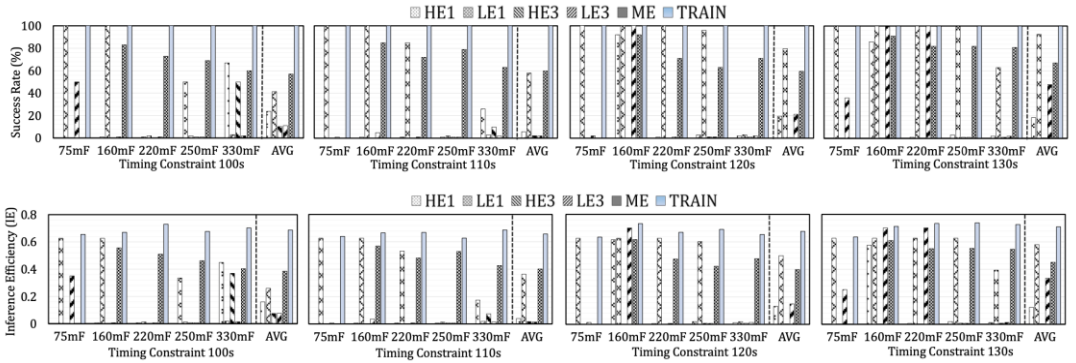


Fig. Workflow of the TRAIN framework.

miss deadline occur!



(a) [14] with 160mF capacitor. (b) [14] with 220mF capacitor. (c) [14] with 250mF capacitor. (d) TRAIN with 220mF capacitor.

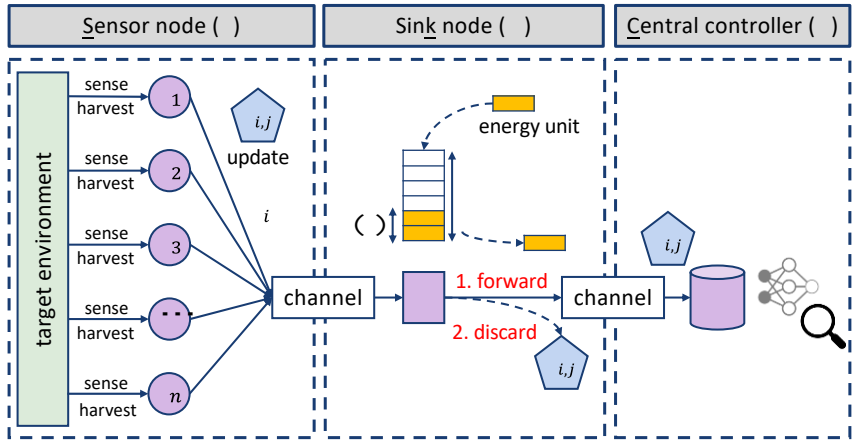


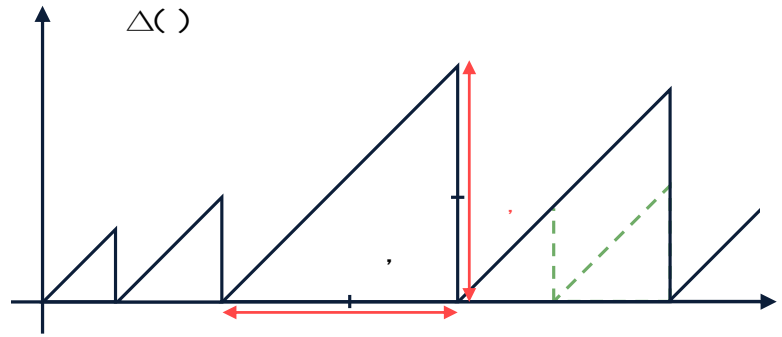
- Shu-Ting Cheng, Wen Sheng Lim, Chia-Heng Tu and Yuan-Hao Chang, "TRAIN: A Reinforcement Learning Based Timing-Aware Neural Inference on Intermittent Systems," ACM/IEEE International Conference on Computer-Aided Design (ICCAD), San Francisco, California, USA, Oct. 29 - Nov. 2, 2023. **(Top Conference)**

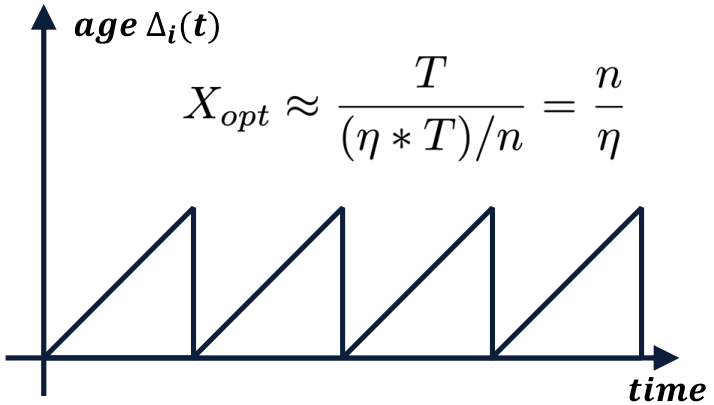
Data Freshness Optimization on Intermittent Systems

[DATE'23]

- **Background on NISs and Data Freshness**
  - **Networked Intermittent Systems (NISs)** use ambient energy to power both the sensor and sink node to to track real-time physical conditions for various purposes.
  - **Data freshness** (i.e., the end-to-end latency between source and destination) is an important metrics to measure the performance of environmental monitoring systems.
- **Motivation on Buffer-Less Design in NISs**
  - The forwarding strategy for the sink node (without requiring data buffer) within a NISs should immediately decide to forward or discard the received updates by considering the *energy causality constraints* (i.e., limited energy buffer and non-deterministic energy harvesting rate).
  - The optimal solution has exponential time complexity to the number of updates.
- **Goal**
  - To minimize the data freshness of the status updates sent by sensors in the NISs to provide the freshest data (i.e., A^3oI) of a monitoring application.
- **Contribution**
  - **AoI-aware Branch-and-Bound Algorithm:** An offline algorithm to find the correct optimal solution by considering energy *causality constraints.*
  - **AoI-aware Update Forwarding Algorithm:** An online algorithm to make constant time decision for an approximate solution by evenly distributed the energy among sensor nodes.





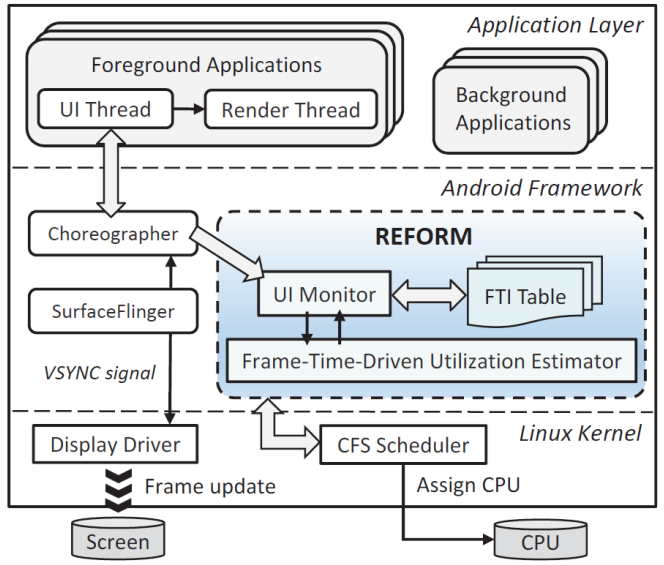


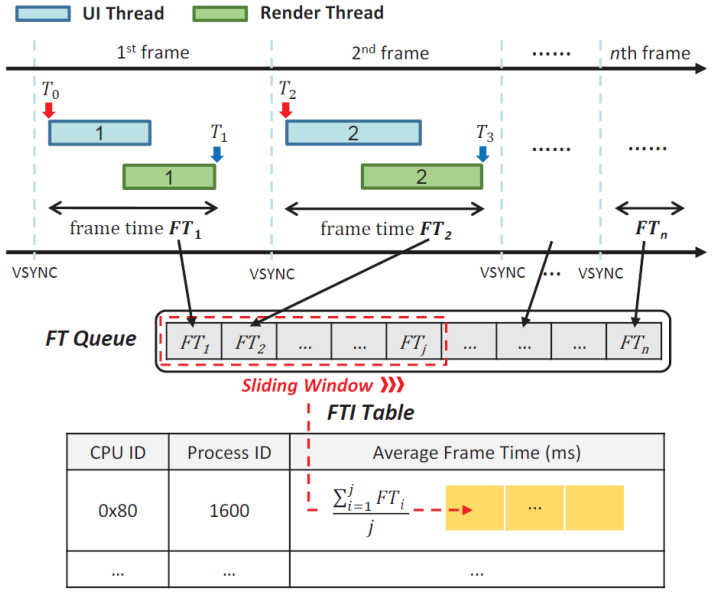
- Hao-Jan Huang, Wen Sheng Lim, Chia-Heng Tu, Chun-Feng Wu, and Yuan-Hao Chang, "Data Freshness Optimization on Networked Intermittent Systems," ACM/IEEE Design, Automation and Test in Europe (DATE), Antwerp, Belgium, Apr. 17-19, 2023.

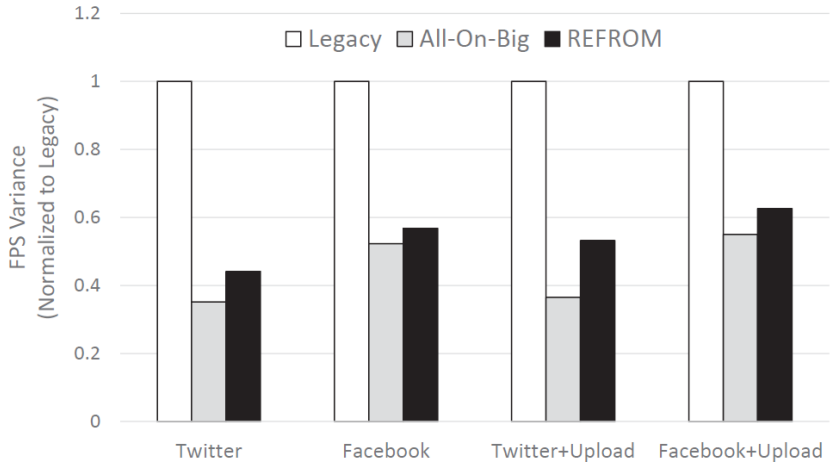
REFROM: Responsive, Energy-efficient Frame Rendering for Mobile Devices

[ISLPED'23, Best Paper Nomination]

- **Observation**
  - The increasing demand for high-quality graphics on mobile devices necessitates a high frame rate for display refresh
  - However, current process scheduling and memory management policies fail to consider the computation demands of frame rendering because they are optimized for saving energy and resource utilization
- **Goal**
  - Develop a new framework, REFROM, to reserve sufficient CPU resources for the upcoming render threads while avoiding unnecessary energy consumption
- **Main Idea**
  - Utilize a history-based frame time estimator to analyze frame time samples from UI threads and predict the computation requirements of upcoming frames





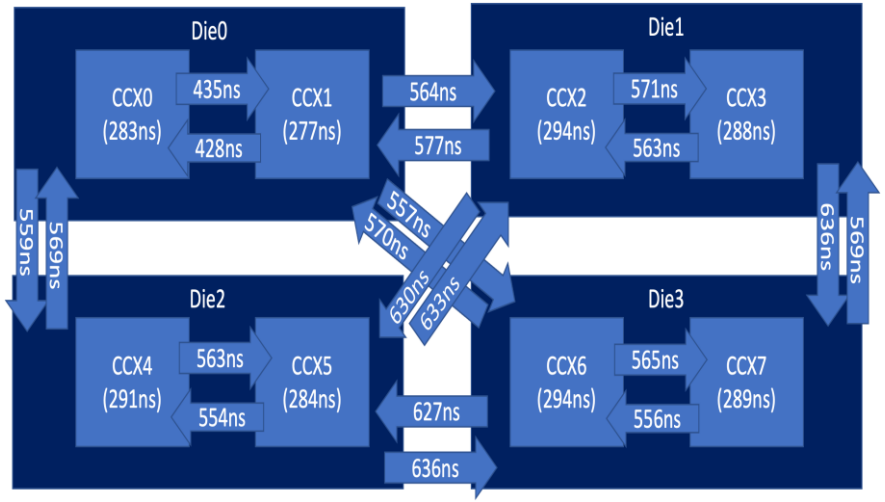


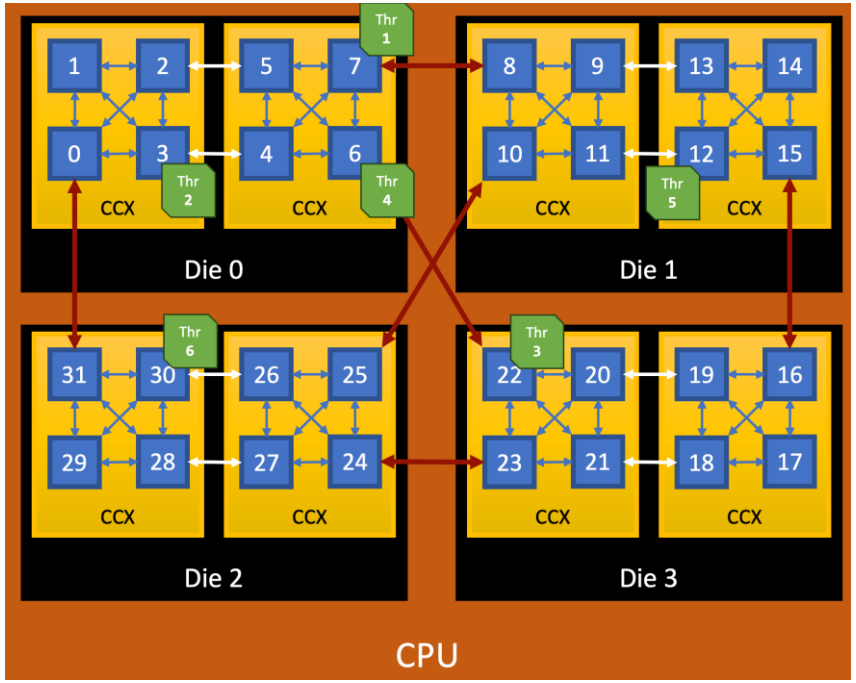
Reduces the number of delayed frames by up to 40% and improves energy efficiency by up to 4%, compared to existing approaches

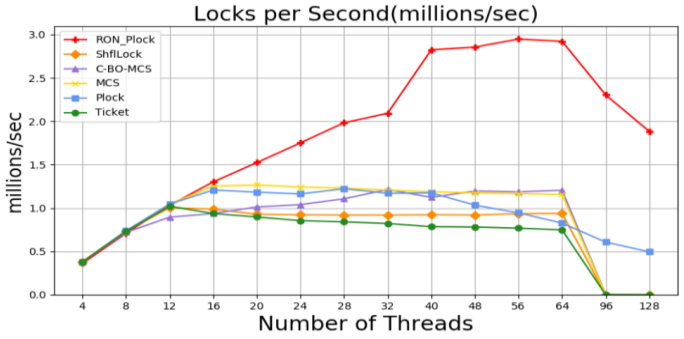
- Tsung-Yen Hsu, Yi-Shen Chen, Yun-Chih Chen, Yuan-Hao Chang, and Tei-Wei Kuo, "REFROM: Responsive, Energy-efficient Frame Rendering for Mobile Devices," ACM/IEEE International Symposium on Low Power Electronics and Design (ISLPED), Vienna, Austria, Aug. 7-8, 2023. **(Top Conference)**

RON: One-Way Circular Shortest Routing to Achieve Efficient and Bounded-waiting Spinlocks [OSDI'23]

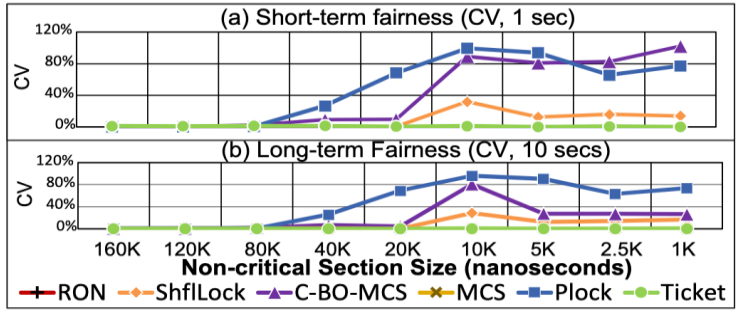
- **Observation**
  - Compared to NUMA-aware spinlock in multi-processor systems, performance optimization becomes more complex in many-core systems due to the increased diversity in core-to-core communication.
  - Certain cores may have a higher likelihood of acquiring the lock due to their proximity to the core holding the lock or their higher execution frequency. Specific mechanisms are necessary to ensure fairness among the cores.
- **Goal**
  - Design an online algorithm to solve the Traveling Salesman Problem. The method must be simple enough to be implemented within a spinlock and must satisfy bounded-waiting to ensure fairness.
- **Main Idea**
  - Precompute the shortest circular path passing through all cores. During runtime, allow the cores requesting to acquire the lock to enter the critical section in the order specified by this path.
    - Assume Thr1 holds the lock. When Thr1 leaves the CS, a spinlock algorithm should "find a core to enter CS."
    - "Find a core to enter CS" is the "traveling salesman problem" (TSP), in terms of minimizing handover costs.







Performance

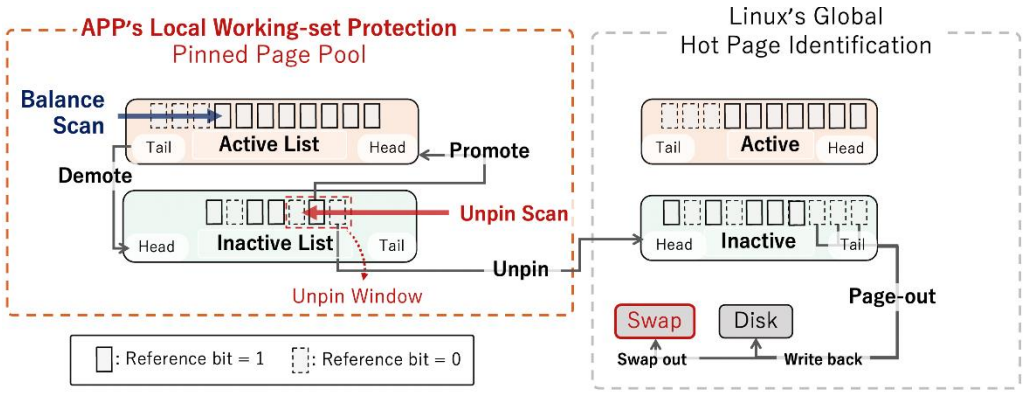


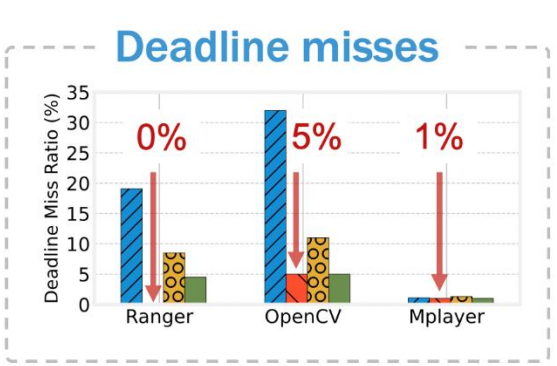
Fairness

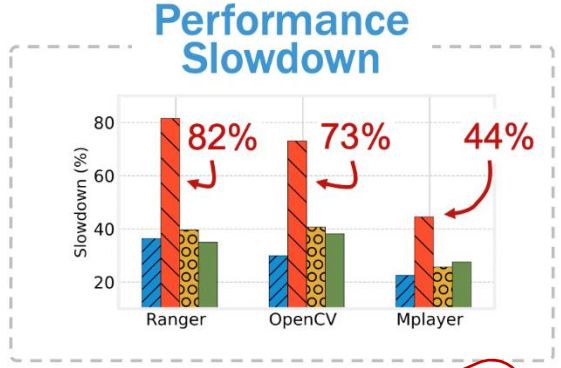
- Shiwu Lo, Han-Ting Lin, Yaohong Xie, Zhaoting Lin, Yu-Hsueh Fang, Jingshen Lin, Jim Huang, Kam Yiu Lam, and Yuan-Hao Chang, "RON: One-Way Circular Shortest Routing to Achieve Efficient and Bounded-waiting Spinlocks," USENIX Symposium on Operating Systems Design and Implementation (OSDI), Boston, MA, USA, Jul. 10-12, 2023. (Acceptance rate: 19.6% (50/255)) **(Top Conference)**

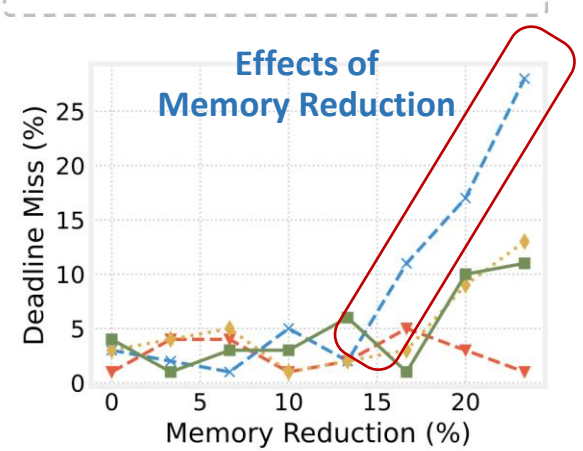
APP: Enabling Soft Real-time Execution on Densely-populated Hybrid Memory Systems [DAC'23]

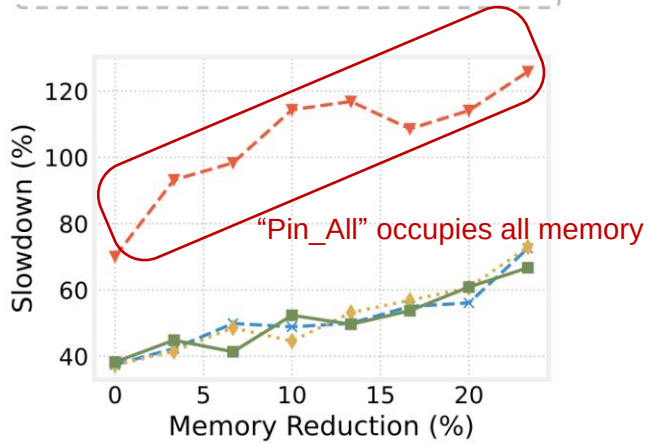
- **Motivation**
  - Virtual memory, which combines DRAM with low-latency SSD swap, is widely adopted in multi-tenant data center servers
  - Memory swapping has overhead. It is possible that the overhead would delay real-time applications
  - When the soft real-time task is scheduled in the next period, it needs to swap in a larger number of pages, introducing excessive swap-in overh
- **Main Idea**
  - **APP (Adaptive Page Pinning)**
    - Protect just enough memory pages of real-time task
  - **Pinned Page Pool**
    - Tracking the memory access frequency of soft real-time process in isolation
    - Linux's global working set tracking











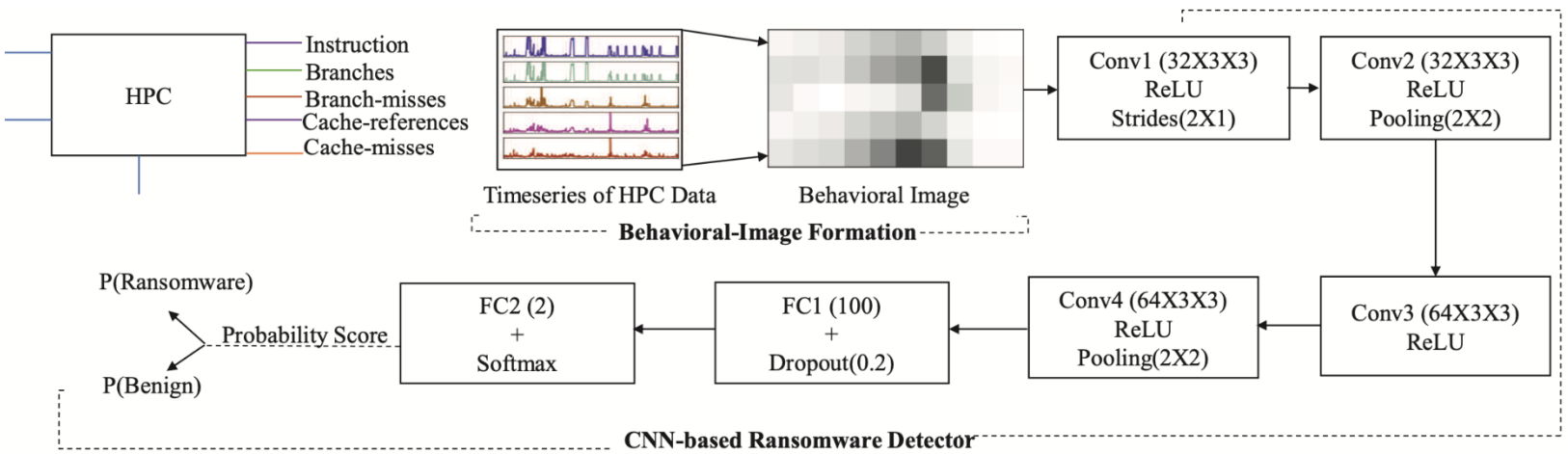
Linux suffers from memory thrashing

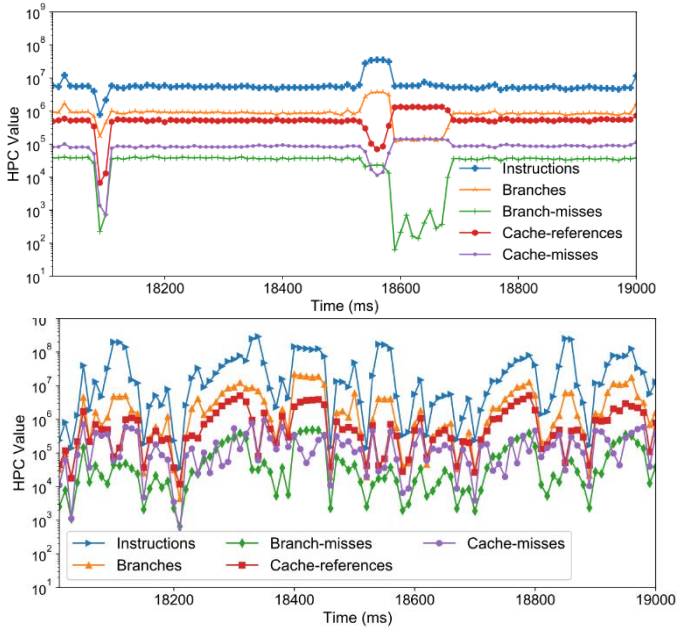
- Zheng-Wei Wu, Yun-Chih Chen, Yuan-Hao Chang, and Tei-Wei Kuo, "APP: Enabling Soft Real-time Execution on Densely-populated Hybrid Memory Systems," ACM/IEEE Design Automation Conference (DAC), San Francisco, CA, USA, Jul. 9-13, 2023. (Acceptance rate: 23%) **(Top Conference)**

5. Others (including Ransomware)

DeepWare: Imaging Performance Counter with Deep Learning to Detect Ransomeware

- **Observation** [IEEE TC'23]
  - In contrast to normal processes, executing ransomware seriously fluctuates the trend of hardware performance counters (HPC).
- **Goal**
  - Detect ransomware by capturing the feature of ransomware, and this approach should be able to detect unseen classes of ransomware.
- **Main Idea**
  - Imaging hardware performance counters with deep learning to detect ransomware.

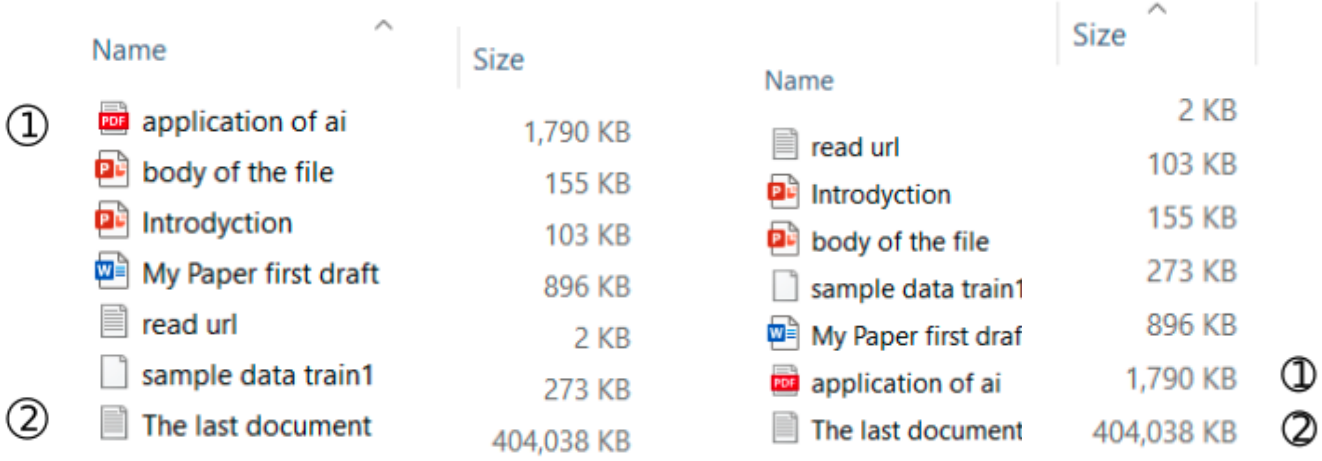




- Gaddisa Olani Ganfure, Chun-Feng Wu, Yuan-Hao Chang, and Wei-Kuan Shih, "DeepWare: Imaging Performance Counters with Deep Learning to Detect Ransomware," IEEE Transactions on Computers (TC), vol. 72, no. 3, pp. 600-613, Mar. 2023.

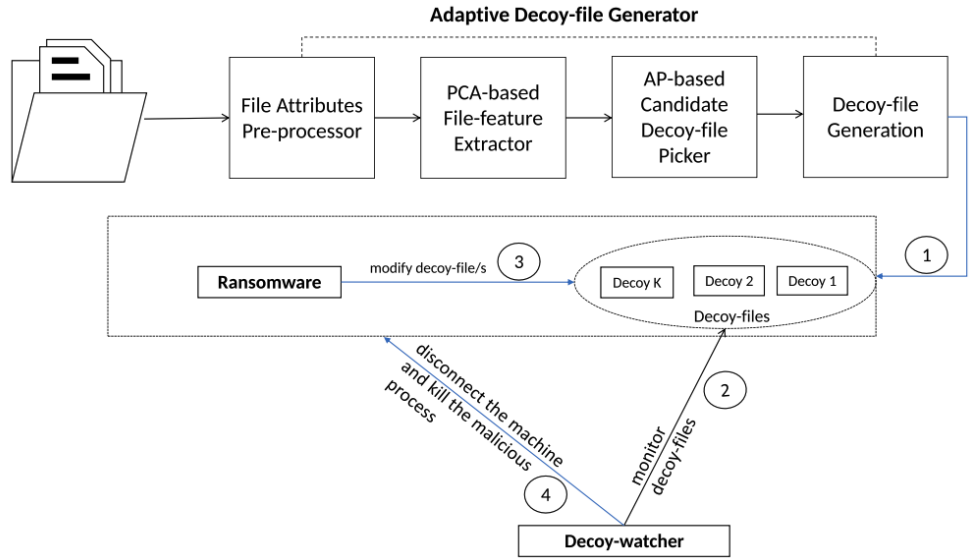
RTrap: Trapping Ransomware with Machine Learning

- **Observation** [IEEE TIFS'23]
  - As the ransomware file-ordering/prioritization differs for each family or variant, relying solely on specific file attributes (e.g., file name or size) is not effective in detecting a variant class of ransomware.
- **Goal**
  - Propose a systematic framework to detect ransomware efficiently and effectively via machine learning-generated deceptive files.
- **Main Idea**
  - Using a data-driven decoy file selection and generation strategy, RTrap plants deceptive decoy files across the directory to lure the ransomware to access it.



Defender's Generation Order

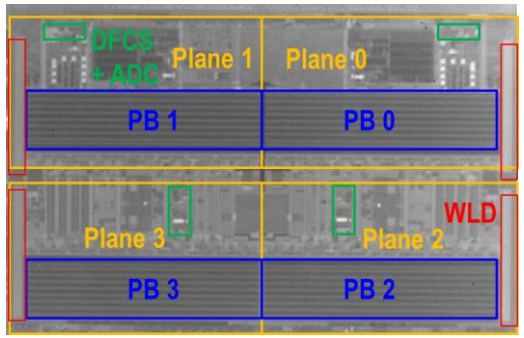
Attacker Sorts Files

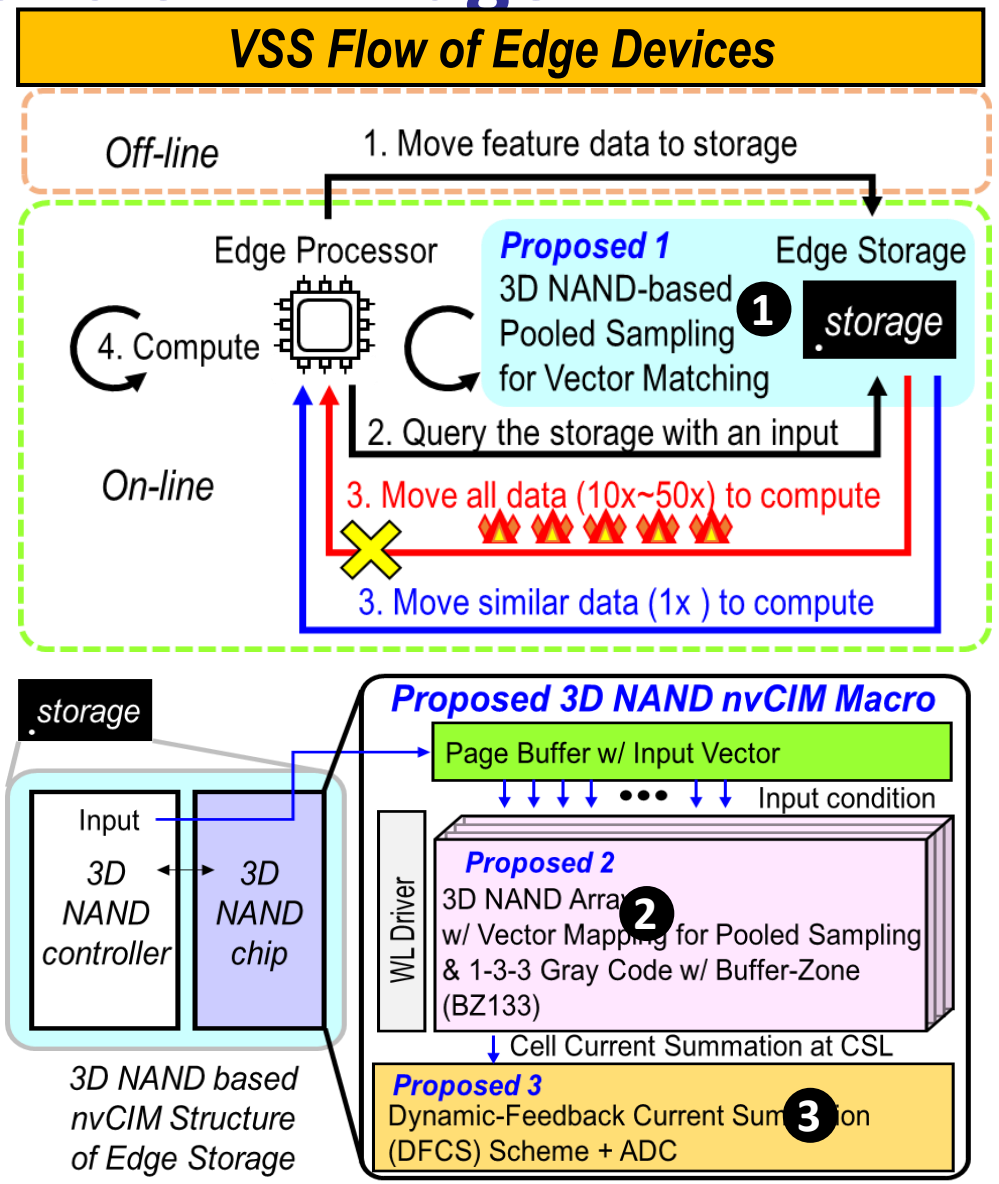


- Gaddisa Olani Ganfure, Chun-Feng Wu, Yuan-Hao Chang, and Wei-Kuan Shih, "RTrap: Trapping and Containing Ransomware with Machine Learning," IEEE Transactions on Information Forensics and Security (TIFS), vol. 18, pp. 1433-1448, Jan. 2023.

In-Memory-Computing 3D NAND Flash: Supporting Similar Vector Matching Operations on AI Edge [ISSCC'22]

- **Observation**
  - Existing vector similarity search (VSS) on edge devices is not inefficient
    - Long search latency and large search energy due to large invalid data movement
  - Exploiting 3D NAND with in-memory computing (IMC) for VSS will face two major challenges:
    - A low-readout accuracy by using the wide range Vt-level of cells
    - The large-readout power consumption for the possible data-patterns.
- **Goal**
  - Enable 3D NAND-based IMC for similar vector matching to boost the VSS performance
- **Main Idea**
  - Adopted "pool sampling" as the major search algorithm
    - $\vec{V}_{INPUT} \cdot \vec{V}_{INDEX_0} + \cdots + \vec{V}_{INPUT} \cdot \vec{V}_{INDEX_K}$
      $= \vec{V}_{INPUT} \cdot (\vec{V}_{INDEX\_0} + \cdots + \vec{V}_{INDEX\_K})$
  - Reuse the selective-BL read function on page buffer with unary data format [HTLue'19:IEDM]
  - A 1-3-3 Gray code with buffer zone (BZ133) for TLC cells, guarding against a low readout accuracy for VVM operation
  - Dynamic-feedback-based current-summation (DFCS) scheme to guard against the wide summation current range of VVM operations



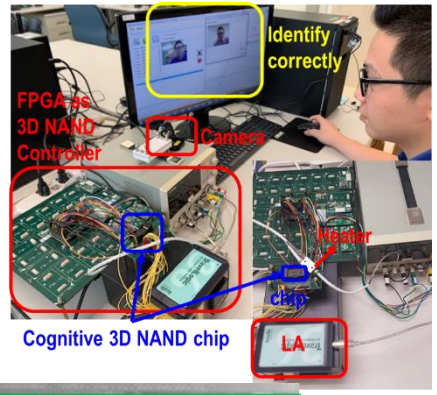


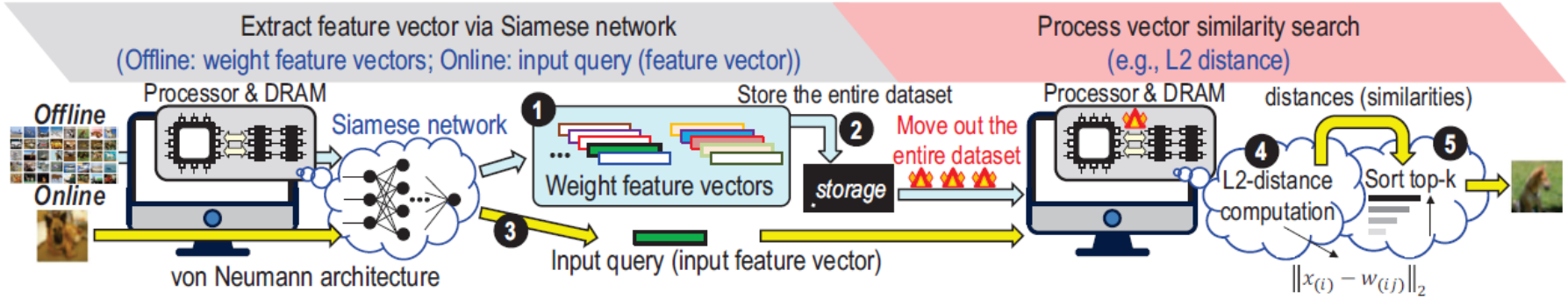
- Han Wen Hu, Wei-Chen Wang, Chung-Kuang Chen, Yung-Chun Lee, Bo-Rong Lin, Huai-Mu Wang, Yen-Bo Lin, Yu-Chao Lin, Chih-Chang Hsieh, Chia-Ming Hu, Yi-Ting Lai, Yuan-Hao Chang, Hsiang-Pang Li, Han-Sung Chen, Tei-Wei Kuo, Keh-Chung Wang, Meng-Fan Chang, Chun-Hsiung Hung, and Chih-Yuan Lu, "A 512Gb In-Memory-Computing 3D NAND Flash Supporting Similar Vector Matching Operations on AI Edge Devices," IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, California, USA, Feb. 20-24, 2022. **(Top Conference)**

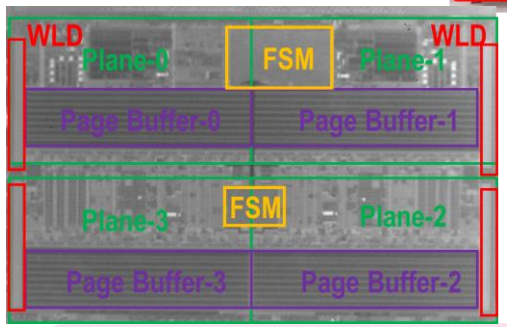
ICE: Intelligent Cognition Engine with NAND In-Memory Computing for Vector Similarity Search

[MICRO'22]

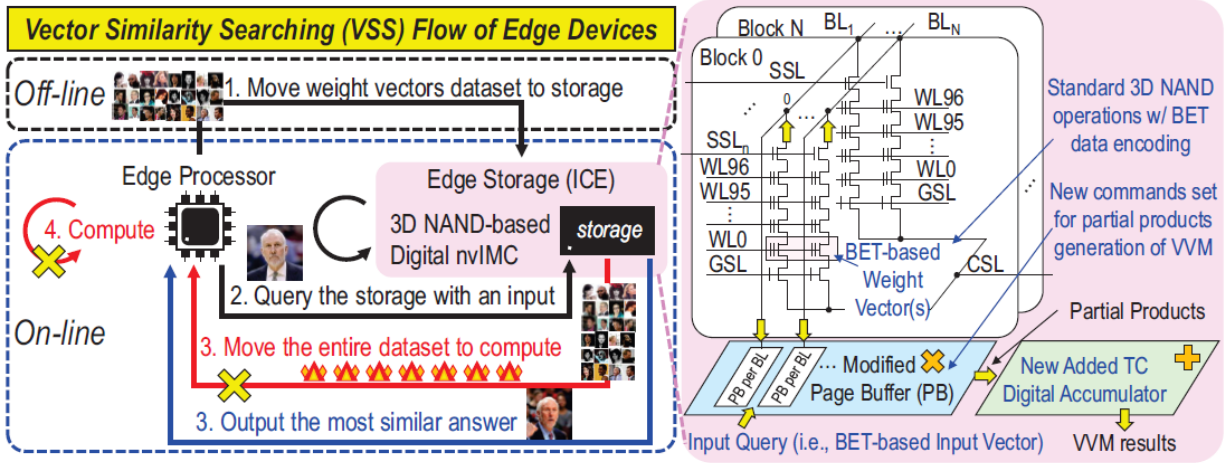
- **Observation**
  - Existing vector similarity search (VSS) on edge devices is inefficient
    - Long search latency and large search energy due to large invalid data movement
  - Exploiting 3D NAND with nonvolatile IMC (nvIMC) for VSS will face two major challenges:
    - Digital-based solution: ECC is critical to the nvIMC design for VSS app. since it guarantees data reliability
    - Analog-based solution: numerous ADCs and DACs increases the chip size







- **Goal**: Enable 3D NAND-based digital nvIMC to accelerate the VSS applications
- **Main Idea**
  - Exploit bit-error tolerant data encoding to mitigate the bit-error influence
  - Adopt modified page buffer to achieve single bit multiplication after computation unfolding
  - Add a new two's complement accumulator to achieve sign-bit computations in accumulation state
  - Propose a hierarchical top-n search to filter invalid data and output the most similar answer during conducting VSS applications



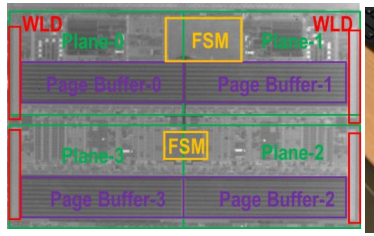
- Han-Wen Hu, Wei-Chen Wang, Yuan-Hao Chang, Yung-Chun Lee, Bo-Rong Lin, Huai-Mu Wang, Chong-Ying Lee, Yu-Ming Huang, Yen-Po Lin, Tzu-Hsiang Su, Chih-Chang Hsieh, Chia-Ming Hu, Yi-Ting Lai, Chung-Kuang Chen, Han-Sung Chen, Hsiang-Pang Li, Tei-Wei Kuo, Keh-Chung Wang, Meng-Fan Chang, Chun-Hsiung Hung, and Chih-Yuan Lu, "ICE: An Intelligent Cognition Engine with 3D NAND-based In-Memory Computing for Vector Similarity Search Acceleration," ACM/IEEE International Symposium on Microarchitecture (MICRO), Chiago, Illinois, USA, Oct. 1-5, 2022. **(Top Conference)**

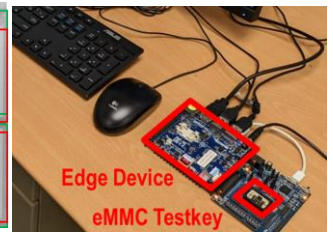
ICE: Intelligent Cognition Engine with NAND In-Memory Computing for Vector Similarity Search

[MICRO'22]

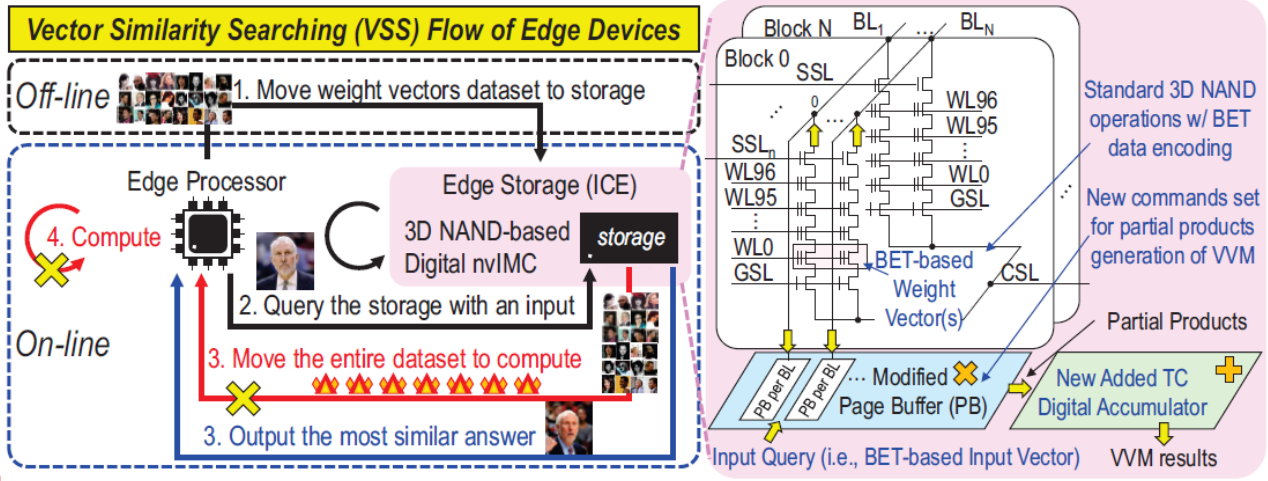
with Macronix

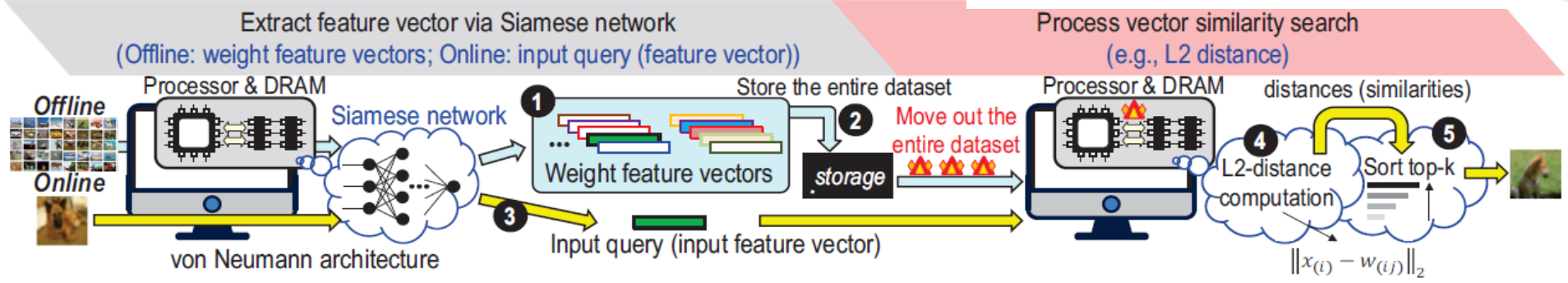
- **Observation**
  - Existing vector similarity search (VSS) on edge devices is inefficient
    - Long search latency and large search energy due to **large unnecessary data movement**
  - Exploiting 3D NAND with nonvolatile IMC (nvIMC) for VSS will face two major challenges:
    - Digital-based solution: ECC is critical to the nvIMC design for VSS app., since it guarantees data reliability
    - Analog-based solution: Numerous ADCs and DACs increases the chip size

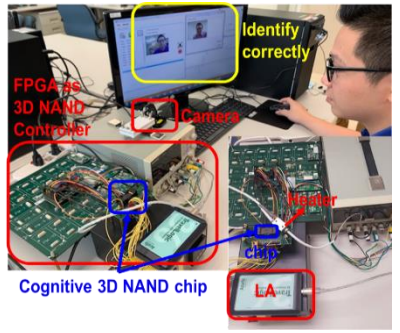




- **Goal**: Enable 3D NAND-based **digital** nvIMC to accelerate the VSS applications
- **Method**: We enable digital flash-based IMC accelerator(**ICE**) supporting VSS **on existing flash cards (e.g., eMMC)**
  - Enable digital IMC to mitigate the bit-error influence
  - Propose a hierarchical top-n search to filter out unneeded data
  - Remove ADC/DAC to resolve the energy issue
- **Result**:
  - ICE enhances the system execution time by **17× to 95×** and energy efficiency by **11× to 140×.**





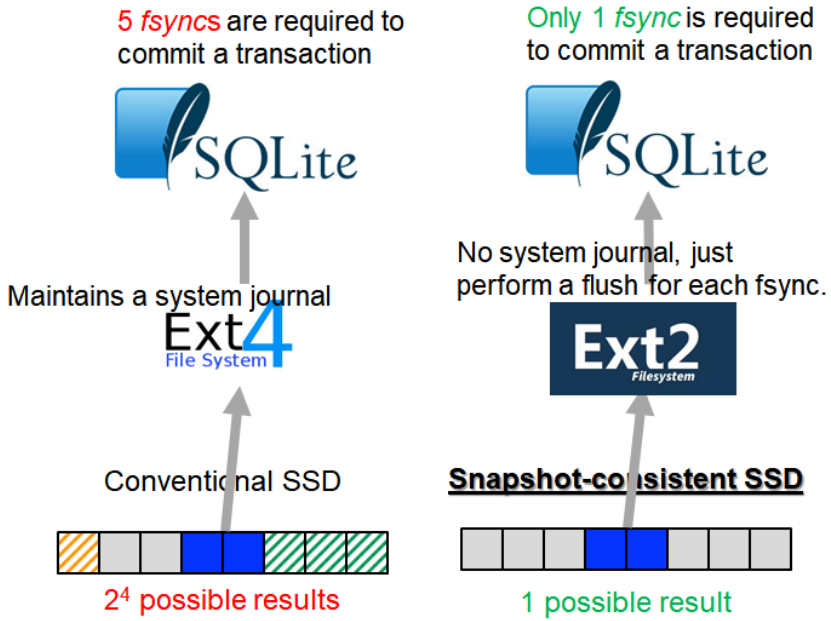


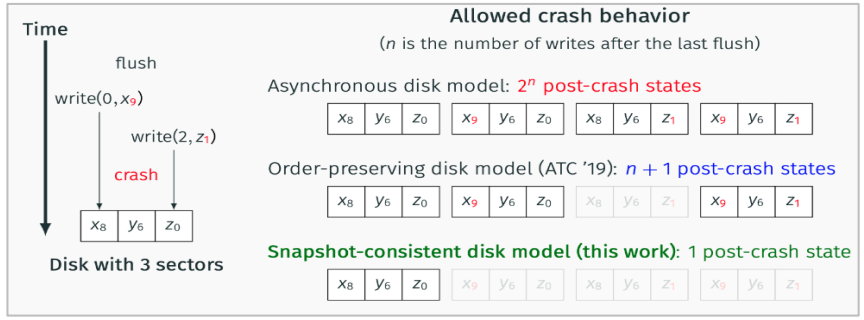
- Han-Wen Hu, Wei-Chen Wang, Yuan-Hao Chang, Yung-Chun Lee, Bo-Rong Lin, Huai-Mu Wang, Chong-Ying Lee, Yu-Ming Huang, Yen-Po Lin, Tzu-Hsiang Su, Chih-Chang Hsieh, Chia-Ming Hu, Yi-Ting Lai, Chung-Kuang Chen, Han-Sung Chen, Hsiang-Pang Li, Tei-Wei Kuo, Keh-Chung Wang, Meng-Fan Chang, Chun-Hsiung Hung, and Chih-Yuan Lu, "ICE: An Intelligent Cognition Engine with 3D NAND-based In-Memory Computing for Vector Similarity Search Acceleration," ACM/IEEE International Symposium on Microarchitecture (MICRO), Chiago, Illinois, USA, Oct. 1-5, 2022. **(Top Conference)**

Crash Recovery Support from the Storage Level

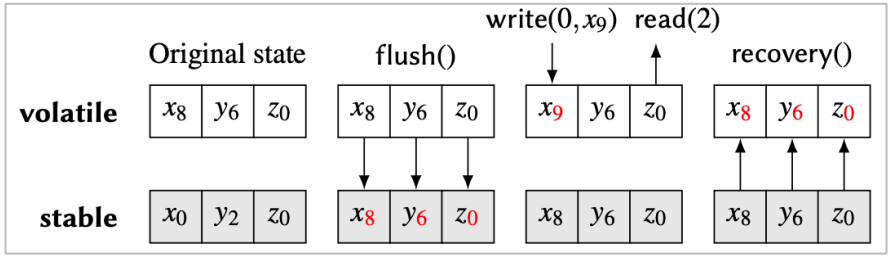
- **Observation**: Existing storages are unreliable so the host (e.g., file system and DB) needs to have a complex crash recovery mechanism that takes time without guaranteed recovery time.
- **Our Method**:
  - This is a verified Snapshot-Consistent Flash Translation Layer (SCFTL) to guarantee determinized time on recovering a flash drive to the state right before the last flush.
  - This is the first attempt to **leverage/apply formal verification** techniques to ensure the correctness of a complex FTL implementation **with guaranteed recovery time**.
  - SCFTL is the first work providing a determinized storage crash recovery mechanism to enable an efficient design of upper layers in the storage stack (e.g., the file system or database system to relax the complexity of crash recovery mechanism).
  - SCFTL is available at: https://github.com/yunshengtw/scftl

[OSDI'20]





Existing non-determinizied work vs. SCFTL



Design Concept of the proposed SCFTL

Achieving Lossless Accuracy with Lossy Programming for Neural Network Training

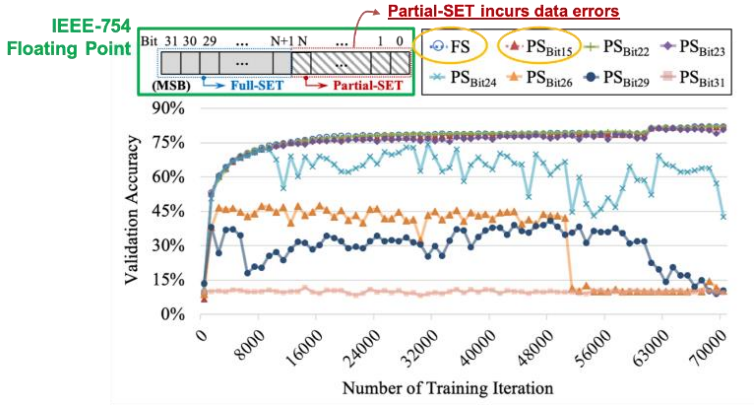
[CODES+ISSS'19, ACM TECS'19] [US 11,550,709, 11,526,285]

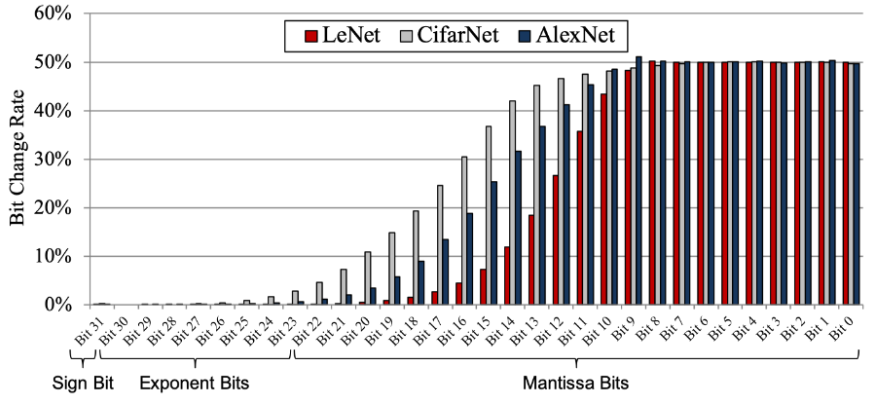
The first ESWEEK best paper award from Taiwan in the past 28 years

- **Observation**:
  - Smart Utilizing of lossy-SET operations of NVM in taking advantage of approximate computing of neural networks (NN). The challenge is on how to consider ***performance, endurance and NN accuracy simultaneously.***
- **Objective**: Enhance training/inferencing performance of neural network with NVM-based systems
- **Challenge**: Compared with DRAM, NVM has a large capacity but has longer write latency and limited write cycles
- **Technical Contributions**: A *Data-Aware Programming Design* *is proposed* to exploit Dual-SET operations to program NN data from the unique viewpoints of data flow and data content.
  - A bit-aware dual-SET policy to efficiently program weights and biases.
  - A layer-aware SET policy to efficiently program intermediate data.
  - A buffered marching-based wear leveling to balance the asymmetric damages of different data c

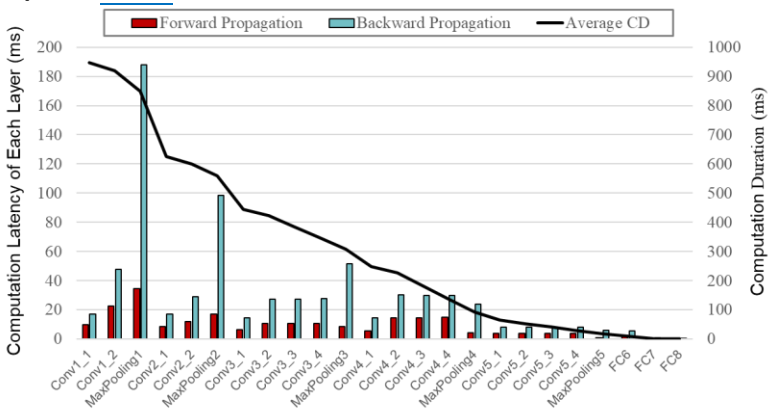
MSB of Weights and Biases (WB_{MSB})	LSB of Weights and Biases (WB_{LSB})	***Buffer Zone***	Shallow Layers of Intermediate Data (ID_{SL})	Deep Layers of Intermediate Data (ID_{DL})

- **Results**: Improve the average memory access latency up to 4.3x and enhance the lifetime up to 3.4x.





Bit Change Rate of Weight and Bias



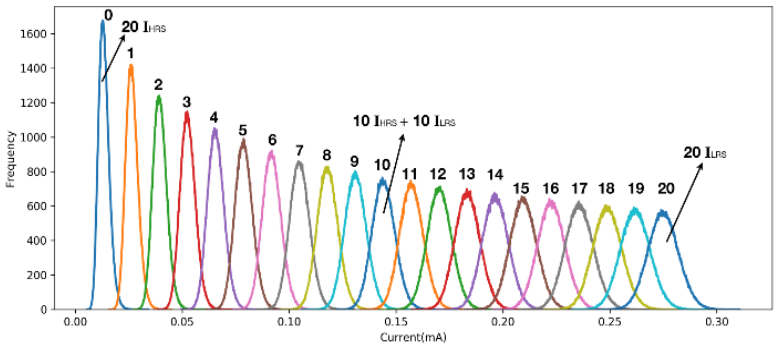
Layer Latency and Round-trip Latency

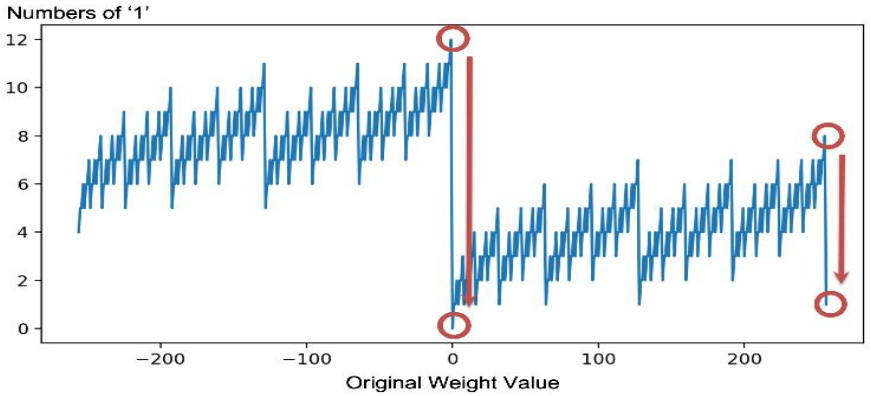
- Wei-Chen Wang, Yuan-Hao Chang, Tei-Wei Kuo, Chien-Chung Ho, Yu-Ming Chang, and Hung-Sheng Chang, "Achieving Lossless Accuracy with Lossy Programming for Efficient Neural-Network Training on NVM-Based Systems," ACM/IEEE International Conference on Hardware/Software Codesign and System Synthesis (CODES+ISSS), New York, NY, USA, Oct. 13-18, 2019. (Journal Track, Integrated with **ACM TECS**) **(Best Paper Award - Top Conference)**
- Wei-Chen Wang, Hung-Sheng Chang, Chien-Chung Ho, Yuan-Hao Chang, and Tei-Wei Kuo, "Memory Device and Wear Leveling Method for the Same," Patent No.: US 11,550,709, Date of Patent: Jan. 10, 2023.
- Wei-Chen Wang, Hung-Sheng Chang, Chien-Chung Ho, Yuan-Hao Chang, and Tei-Wei Kuo, "Memory Device for Neural Networks," Patent No.: US 11,526,285, Date of Patent: Dec. 13, 2022.

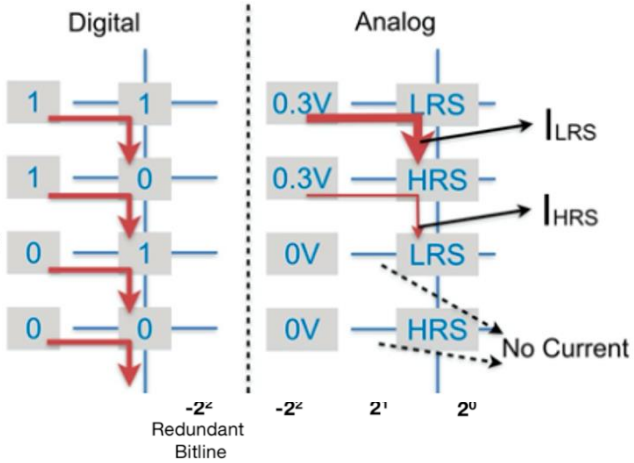
Minimizing Analog Variation Errors of ReRAM Crossbar

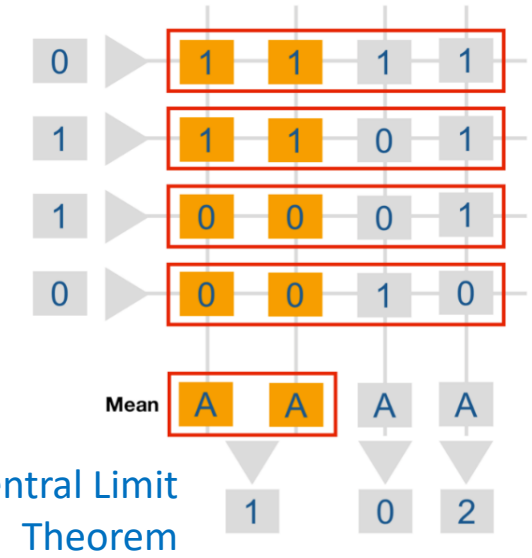
[IEEE TCAD'20 , EMSOFT'20] [US 11,443,797, 11,594,277]

- **Observation:**
  - Variation errors hurt scalability of in-memory computing
- **Objective:** Manage errors to enable in-memory computing for large-scaled inferencing
- **Technical Contributions:** An *Adaptive Data Manipulation Strategy* to significantly reduce the occurrence of the overlapping variation error.
  - Overlapping variation error: Current distributions becomes wider while more ReRAM cells in the LRS state are involved; and wider distribution overlaps with neighbors.
  - Our design is to amortize the sensing results retrieved from redundant bit-lines so that the magnitude of the overlapping variation error can be alleviated
- **Result:**
  - The proposed design can even improve the accuracy in running MNIST and CIFAR-10 by 1.3x and 2.6x respectively.







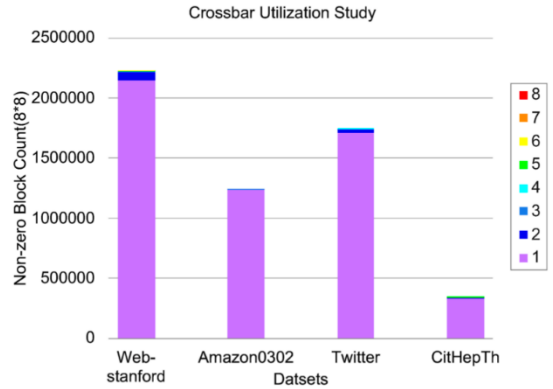


- Yao-Wen Kang, Chun-Feng Wu, Yuan-Hao Chang, Tei-Wei Kuo, and Shu-Yin Ho, "On Minimizing Analog Variation Errors to Resolve the Scalability Issue of ReRAM-based Crossbar Accelerator," accepted and to appear in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD). (Integrated with ACM/IEEE EMSOFT'20)
- Yao-Wen Kang, Chun-Feng Wu, Yuan-Hao Chang, Tei-Wei Kuo, and Shu-Yin Ho, "On Minimizing Analog Variation Errors to Resolve the Scalability Issue of ReRAM-based Crossbar Accelerator," ACM/IEEE International Conference on Embedded Software (EMSOFT), Germany, Sep. 20 - 25, 2020. (Journal Track, Integrated with IEEE TCAD) **(Top Conference)**
- Shu-Yin Ho, Hsiang-Pang Li, Yao-Wen Kang, Chun-Feng Wu, Yuan-Hao Chang, and Tei-Wei Kuo, "Neural Network Computation Met hod and Apparatus Using Adaptive Data Representation," Patent No.: US 11,443,797, Date of Patent: Sep. 13, 2022.
- Shu-Yin Ho, Hsiang-Pang Li, Yao-Wen Kang, Chun-Feng Wu, Yuan-Hao Chang, and Tei-Wei Kuo, "Neural Network Computation Method Using Adaptive Data Representation," Patent No.: US 11,594,277, Date of Patent: Feb. 28, 2023.

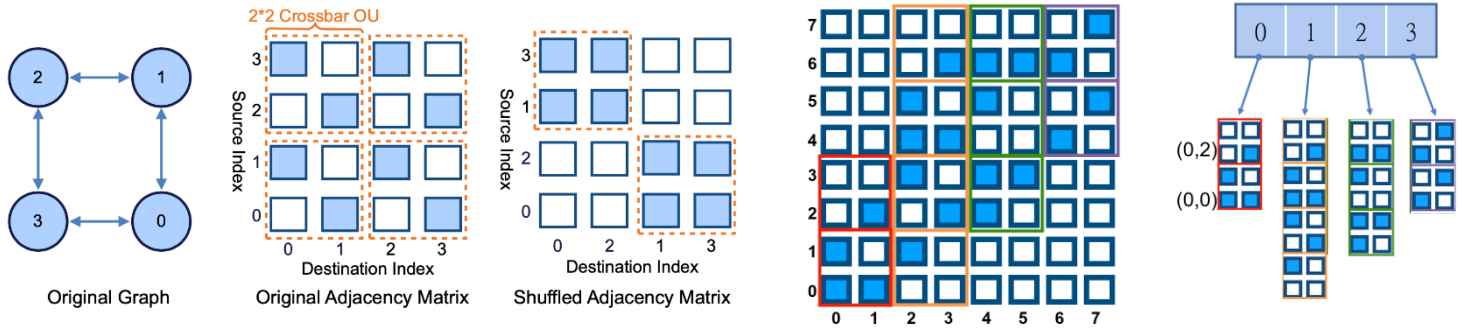
Space Utilization Issue and Allocation Challenge – Example Work in Crossbar Utilization over Irregular Data Structure

[ISLPED'21] [US 11,640,255]

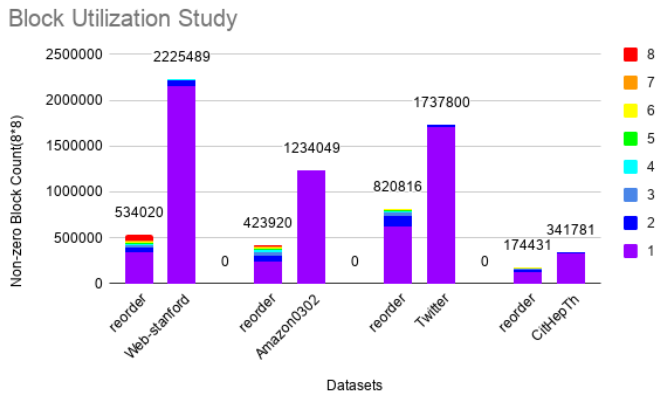
- **Observation**:
  - Placing an adjacency matrix on the crossbar array for accelerating matrix multiplication may lead to unnecessary energy wasting.
  - Reason: Elements in the graph adjacency matrix are usually sparse and discrete, and thus extra crossbar Operation Units (OUs) are required for processing because of the low-utilization.
- **Objective**:
  - Proposing a hardware/software co-design solution to solve the sparse and discrete issues by clustering graph nodes on the crossbar accelerators.
- **Technical Contribution**:
  - Remap and shuffle the original adjacency matrix with the awareness of the graph localities.
- **Result**:
  - The proposed strategy could save up to *2.79x* of the crossbar memory usage and reduce *2.1x* of the energy



Low Utilization on Crossbar Accelerator



Design Concept: Remapping and Reshuffling

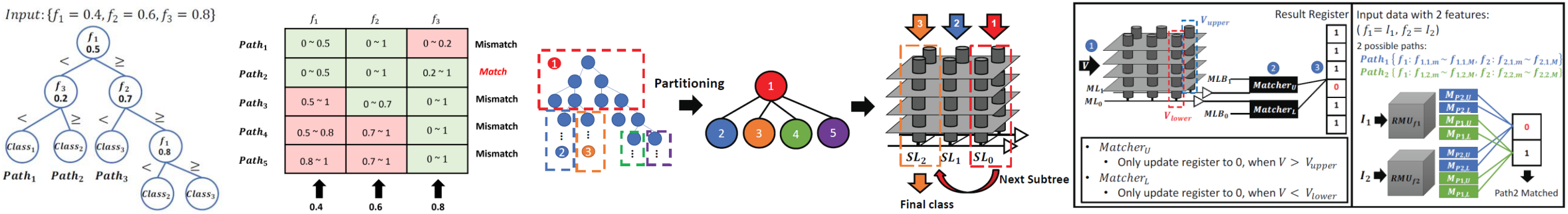


- Ting-Hsuan Lo, Chun-Feng Wu, Yuan-Hao Chang, Tei-Wei Kuo, and Wei-Chen Wang, "Space-efficient Graph Data Placement to Save Energy of ReRAM Crossbar," ACM/IEEE International Symposium on Low Power Electronics and Design (ISLPED), Virtual Conference, Jul. 26-28, 2021. (Top Conference)
- Wei-Chen Wang, Ting-Hsuan Lo, Chun-Feng Wu, Yuan-Hao Chang, and Tei-Wei Kuo, "Memory Device and Operation Method Thereof," Patent No.: US 11,640,255, Date of Patent: May 2, 2023.

A Digital 3D TCAM Accelerator for the Inference Phase of Random Forest

[ACM/IEEE DAC 2023]

- **Observation**
  - Ternary content addressable memory (TCAM) that utilize processing-in memory and high parallelism of crossbar memory and thus is suitable for the memory-bound inference of random forests.
  - However, the reliability and explosive growth of paths become critical issues on applying TCAM to inference phase
- **Objective**
  - A digital 3D TCAM-based accelerator for the inference phase of random forests is proposed with higher reliability than the previous analog based approach.
- **Main Idea**
  - The proposed architecture can check if input values match a specific range in parallel while providing a high density based on the 3D ReRAM TCAM architecture.
  - A subtree-partitioning algorithm spits each decision tree into multiple subtrees to reduce the search complexity and a data placement strategy is designed for the 3D ReRAM TCAM accelerator.
- **Result**: Achieve an average of **3.13x** higher throughput with **22x** more energy saving than the GPU approach

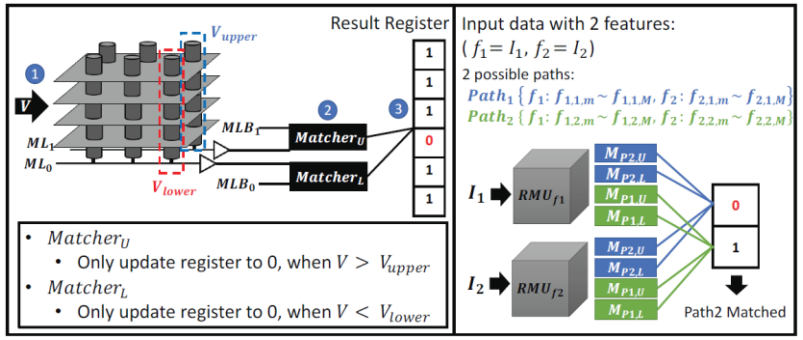


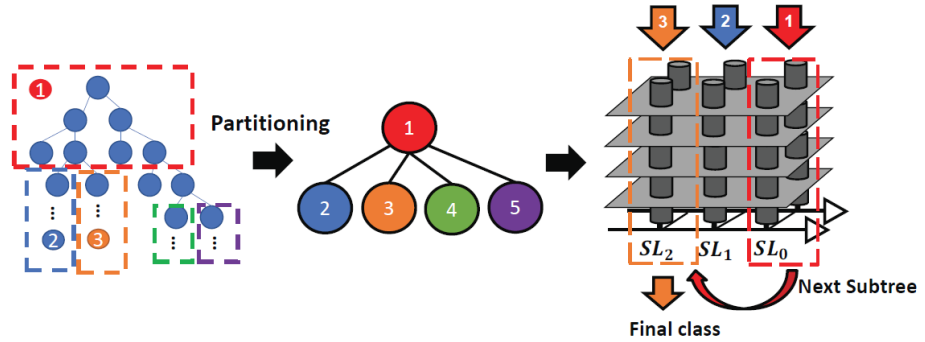
- Chieh-Lin Tsai, Chun-Feng Wu, Yuan-Hao Chang, Han-Wen Hu, Yung-Chun Lee, Hsiang-Pang Li, and Tei-Wei Kuo, "A Digital 3D TCAM Accelerator for the Inference Phase of Random Forest," ACM/IEEE Design Automation Conference (DAC), San Francisco, CA, USA, Jul. 9-13, 2023. (Acceptance rate: 23%) (Top Conference)

A Digital 3D TCAM Accelerator for the Inference Phase of Random Forest

[DAC'23]

- **Observation**
  - Ternary content addressable memory (TCAM) that utilize processing-in memory and high parallelism of crossbar memory and thus is suitable for the memory-bound inference of random forests.
  - However, the reliability and explosive growth of paths become critical issues on applying the TCAM to the inference phase
- **Goal**
  - A digital 3D TCAM-based accelerator for the inference phase of random forests is proposed with higher reliability than the previous analog based approach.
- **Main Idea**
  - The proposed architecture can check if input values match a specific range in parallel while providing a high density based on the 3D ReRAM TCAM architecture.
  - A subtree-partitioning algorithm spits each decision tree into multiple subtrees to reduce the search complexity and a data placement strategy is designed for the 3D ReRAM TCAM accelerator.



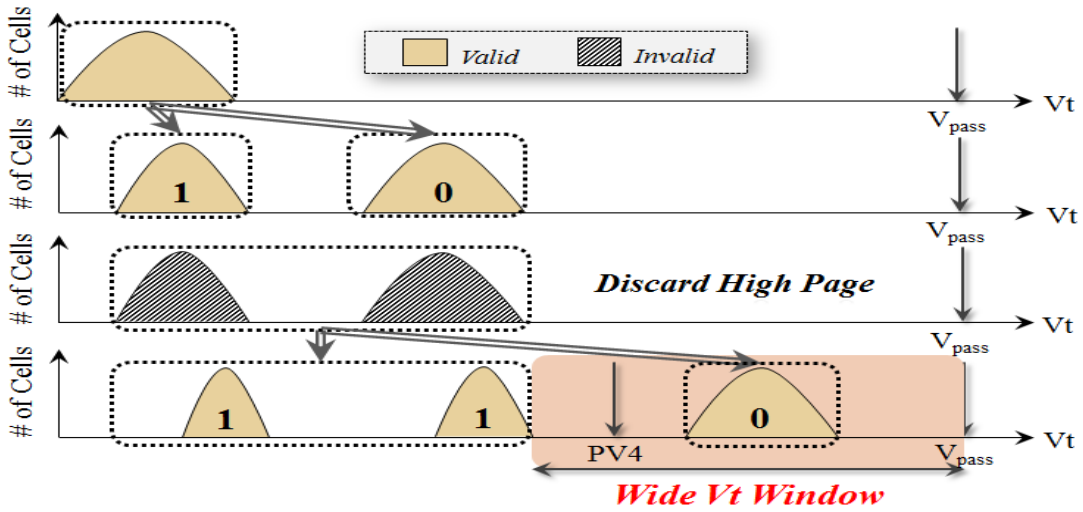


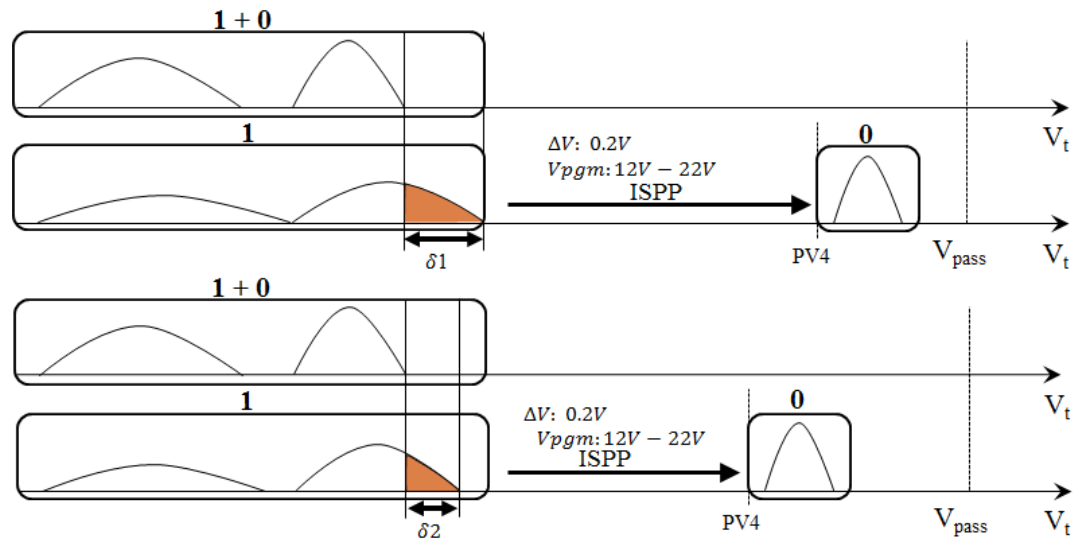
- Chieh-Lin Tsai, Chun-Feng Wu, Yuan-Hao Chang, Han-Wen Hu, Yung-Chun Lee, Hsiang-Pang Li, and Tei-Wei Kuo, "A Digital 3D TCAM Accelerator for the Inference Phase of Random Forest," ACM/IEEE Design Automation Conference (DAC), San Francisco, CA, USA, Jul. 9-13, 2023. (Acceptance rate: 23%) (Top Conference)

Achieving SLC Perf. with MLC Flash

[DAC'15, ACM TOS'18]
[US 9,740,602, 9,627,072]

- **Motivation**:
  - MLC flash has high density but has very write performance.
- **Main Idea**:
  - Develop a trim-like programming scheme to intelligently utilize the knowledge of the ***data validity*** so as to program low page with the speed of SLC flash.
  - Resolve the fundamental issue of ISPP in programming MLC flash.
- **Results:**
  - The trim-like programming scheme could accelerate the programming speed up to 742% and even reduce the bit error rate up to 471% for MLC pages.



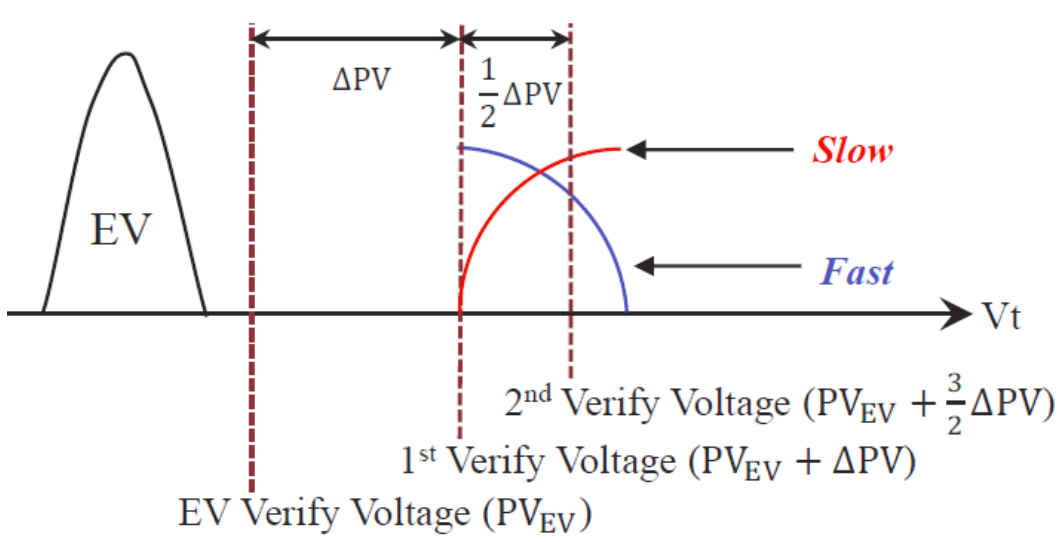


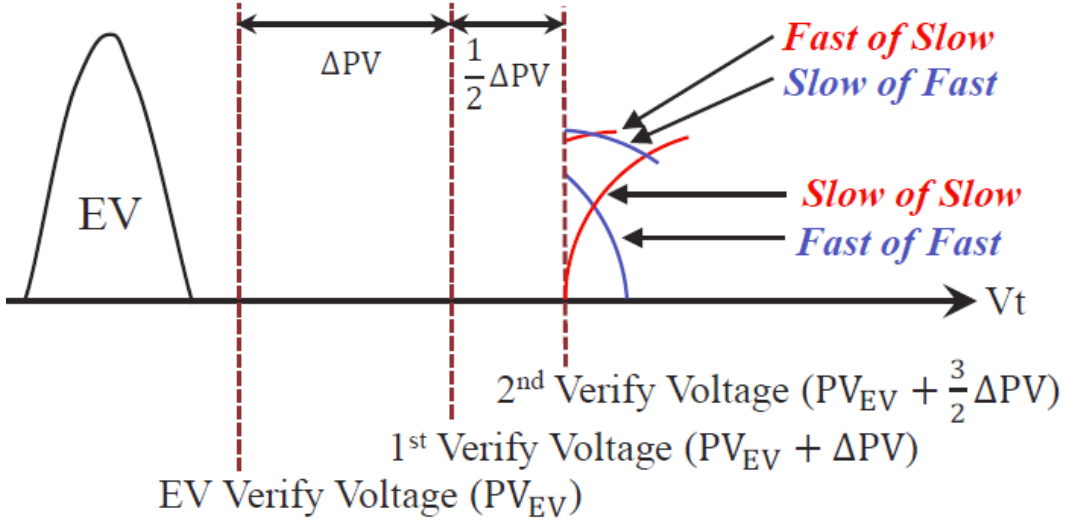
Chien-Chung Ho, Yu-Ming Chang, Yuan-Hao Chang, and Tei-Wei Kuo, "SLC-Like Programming Scheme for MLC Flash Memory," ACM Transactions on Storage (TOS), vol. 14, no. 1, pp. 11:1-11:26, Mar. 2018.
Yu-Ming Chang, Yuan-Hao Chang, Tei-Wei Kuo, Yung-Chun Li, and Hsiang-Pang Li, "Achieving SLC Performance with MLC Flash Memory," ACM/IEEE Design Automation Conference (DAC), San Francisco, California, USA, Jun. 7-11, 2015. (Top Conference)
Yu-Ming Chang, Yung-Chun Li, Hsing-Pang Li, Yuan-Hao Chang, and Tei-Wei Kuo, "Memory Orprating Method and Memory Device Using the Same," Patent No.: US 9,740,602, Date of Patent: August 22, 2017.
Yu-Ming Chang, Yung-Chun Li, Hsiang-Pang Li, Yuan-Hao Chang, and Tei-Wei Kuo, "Variant Operation Sequences for Multibit Memory," Patent No.: US 9,627,072, Date of Patent: April 18, 2017.

Relaxing Program Disturbance

[ICCAD'15]

- **Motivation**:
  - 3D NAND flash memory has serious disturbance issues due to the high cell density and the program speed differences.
- **Main Idea**:
  - A bi-group programming method is proposed resolve the *slow cell effects* (in ISPP).
  - The proposed method is orthogonal to ***wear leveling*** and ***ECC***.
  - Main idea: Assign proper program voltages for cells with different program speeds by means of classifying and programming the cells simultaneously and in a progressive way.
- **Results:**
  - Reduce more than 93% of bit errors.





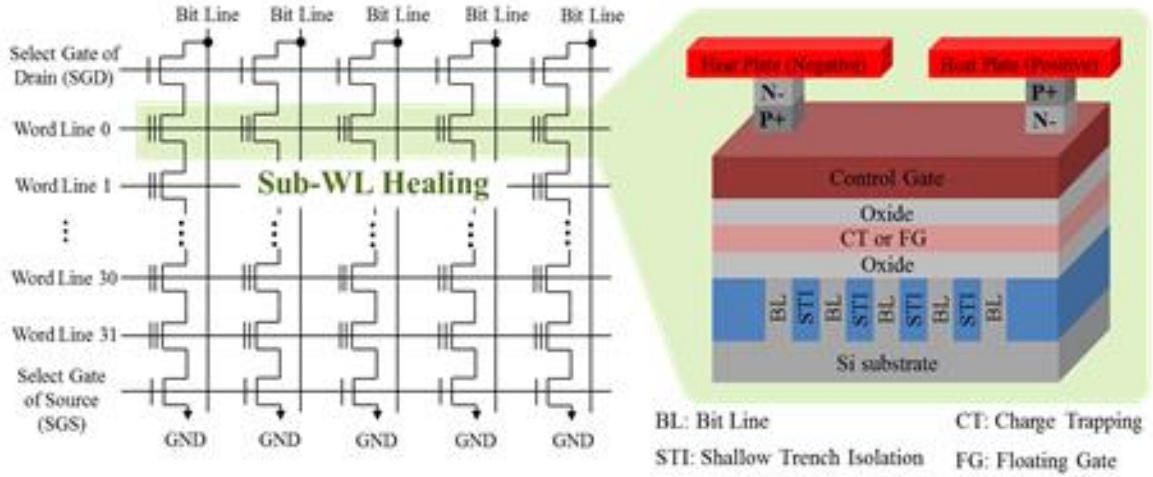
Yu-Ming Chang, Yung-Chun Li, Yuan-Hao Chang, Tei-Wei Kuo, Chih-Chang Hsieh, and Hsiang-Pang Li, "On Relaxing Page Program Disturbance over 3D MLC Flash Memory," ACM/IEEE International Conference on Computer-Aided Design (ICCAD), Austin, Texas, USA, Nov. 2-6, 2015. (Acceptance rate: 24.7%(94/381)) (Top Conference)

Heal Leveling to Replace Wear Leveling for 3D Flash

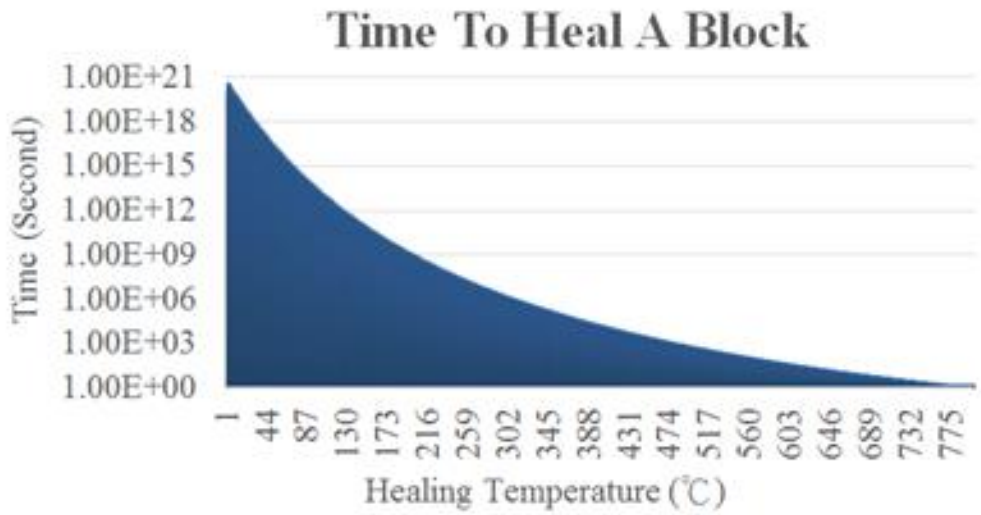
[DAC'14 – Best Paper Nomination]

[US 9,348,748]

- **Motivation**:
  - High-density 3D flash memory has low endurance and PE cycles as the cell-density is increased.
- **Main Idea**:
  - Propose to integrate a **self-healing component** into flash chips, the flash industry will be wholly changed.
  - We are the first team to adopt **heal-leveling** on real 3D flash to significantly enhance 3D flash's lifetime and replace wear leveling.
- **Results:**
  - Achieve almost no lifetime limitation while reducing 47% live-page copyings, compared to traditional wear leveling.



Internal Heating Architecture

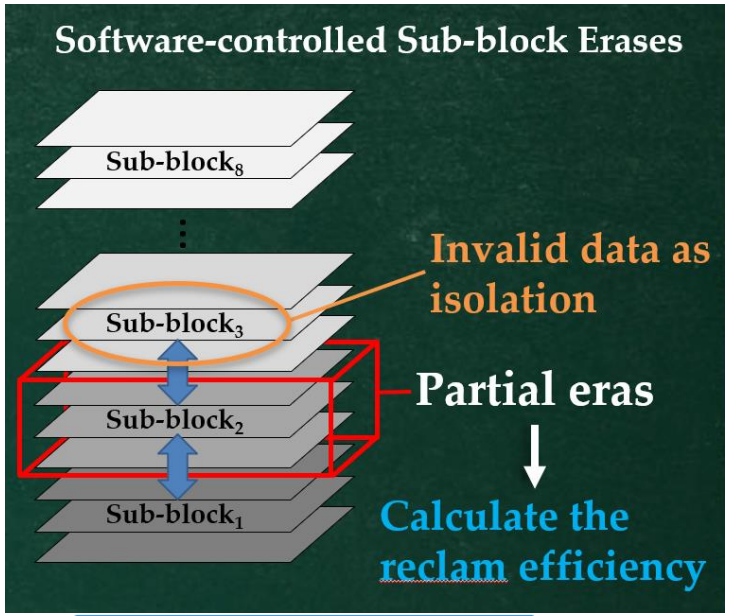


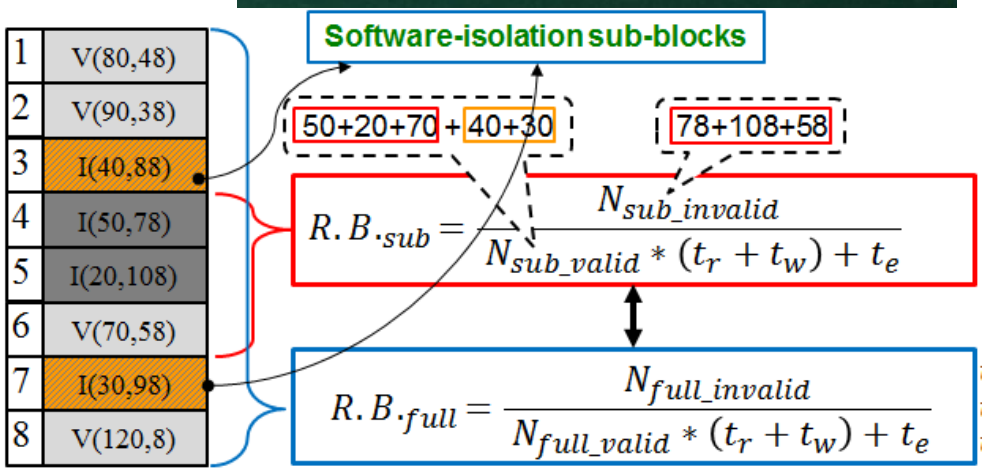
Yu-Ming Chang, Yuan-Hao Chang, Jian-Jia Chen, Tei-Wei Kuo, Hsiang-Pang Li, and Hang-Ting Lue, "On Trading Wear-leveling with Heal-leveling," ACM/IEEE Design Automation Conference (DAC), San Francisco, California, USA, Jun. 1-5, 2014. (Best Paper Nomination (12/787) - Top Conference)
Yu-Ming Chang, Hsiang-Pang Li, Hang-Ting Lue, Yuan-Hao Chang, and Tei-Wei Kuo, "Heal Leveling," Patent No.: US 9,348,748, Date of Patent: May 24, 2016.

Sub-Block Erase

[CODES'16][US 9,754,637]

- **Motivation**:
  - The fast-growing block size in 3D flash, the erase overhead becomes a major performance bottleneck.
  - Sub-block erases are not possible due to the strong disturbance in the adjunct layers of the same block.
- **Main Idea**:
  - This is the first work that *enables sub-block erase with **software isolation** and without hardware cost* to reduce GC overhead of large-block 3D flash.
  - We propose a new evaluate metric called recycle benefit to evaluate whether the area isolated by the ***software-isolation sub-block*** can be erased.
- **Results:**
  - This design reduces at least 20% GC overhead without extra hardware cost.





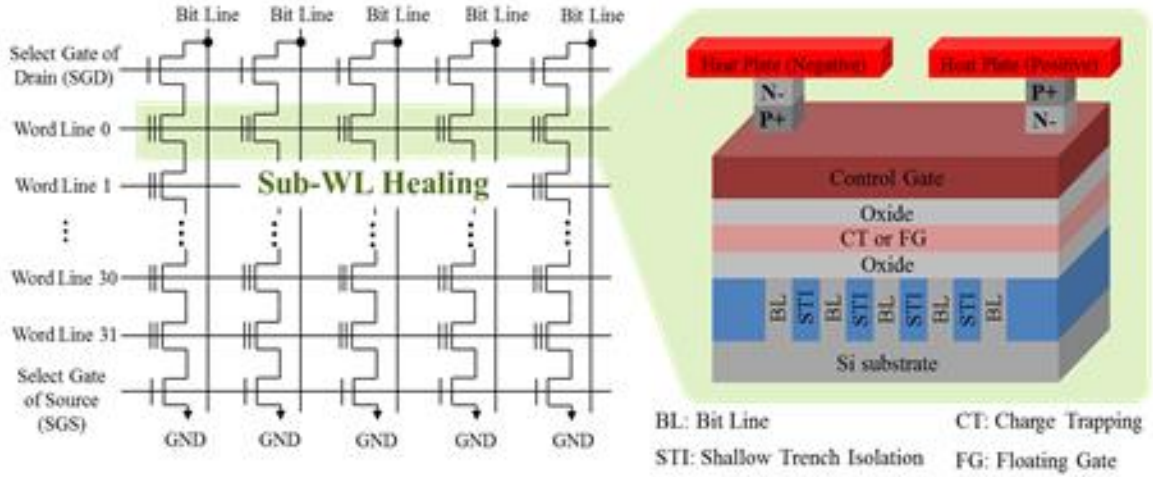
Hsin-Yu Chang, Chien-Chung Ho, Yuan-Hao Chang, Yu-Ming Chang, and Tei-Wei Kuo, "How to Enable Software Isolation and Boost System Performance with Sub-block Erase over 3D Flash Memory," ACM/IEEE International Conference on Hardware/Software Codesign and System Synthesis (CODES+ISSS), Pittsburgh, Pennsylvania, USA, Oct. 2-7, 2016. (Top Conference)
Yu-Ming Chang, Hsing-Pang Li, Hsin-Yu Chang, Chien-Chung Ho, and Yuan-Hao Chang, "Erasing Method and Memory Device Using the Same," Patent No.: US 9,754,637, Date of Patent: Sep. 5, 2017.

Heal Leveling to Replace Wear Leveling for 3D Flash

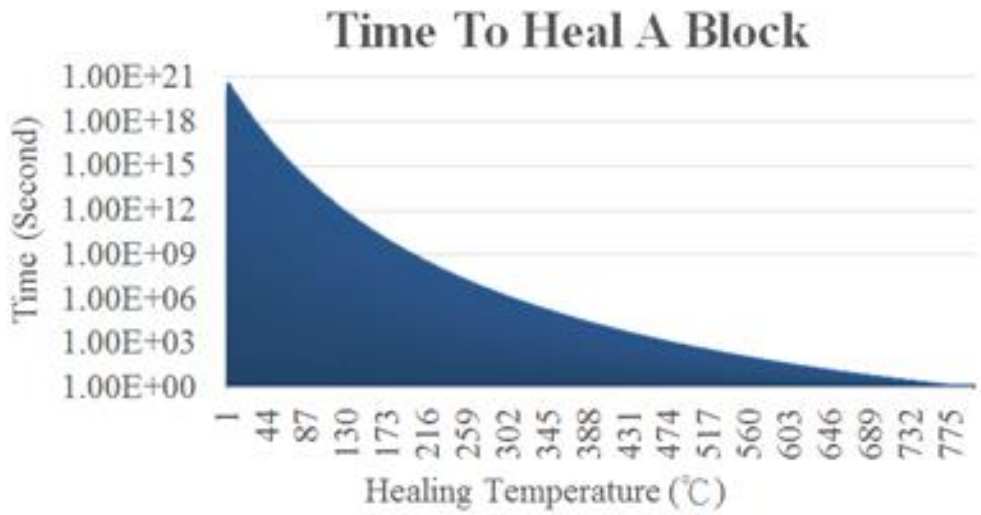
[DAC'14 – Best Paper Nomination]

[US 9,348,748]

- **Motivation**:
  - High-density 3D flash memory has low endurance and PE cycles as the cell-density is increased.
- **Main Idea**:
  - Propose to integrate a **self-healing component** into flash chips, the flash industry will be wholly changed.
  - We are the first team to adopt **heal-leveling** on real 3D flash to significantly enhance 3D flash's lifetime and replace wear leveling.
- **Results:**
  - Achieve almost no lifetime limitation while reducing 47% live-page copyings, compared to traditional wear leveling.



Internal Heating Architecture

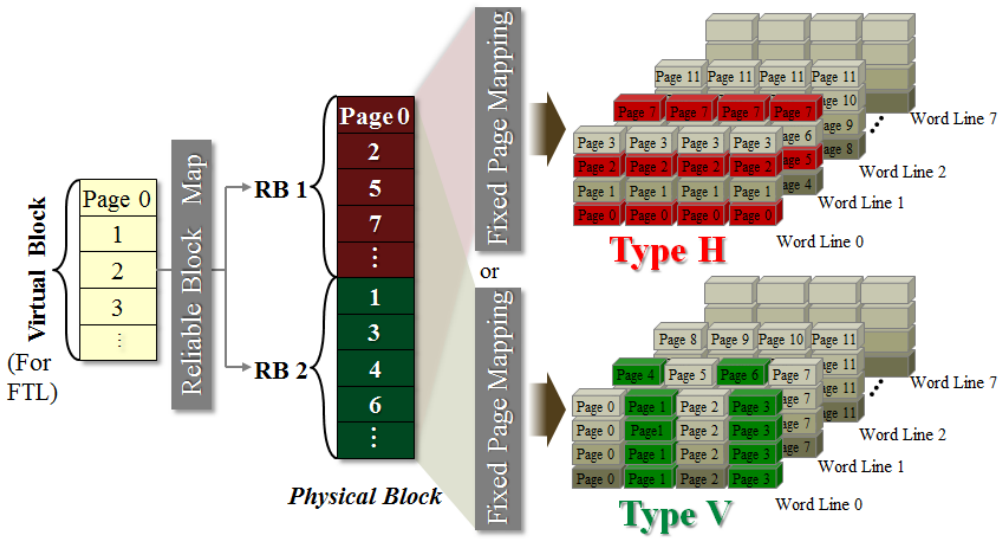


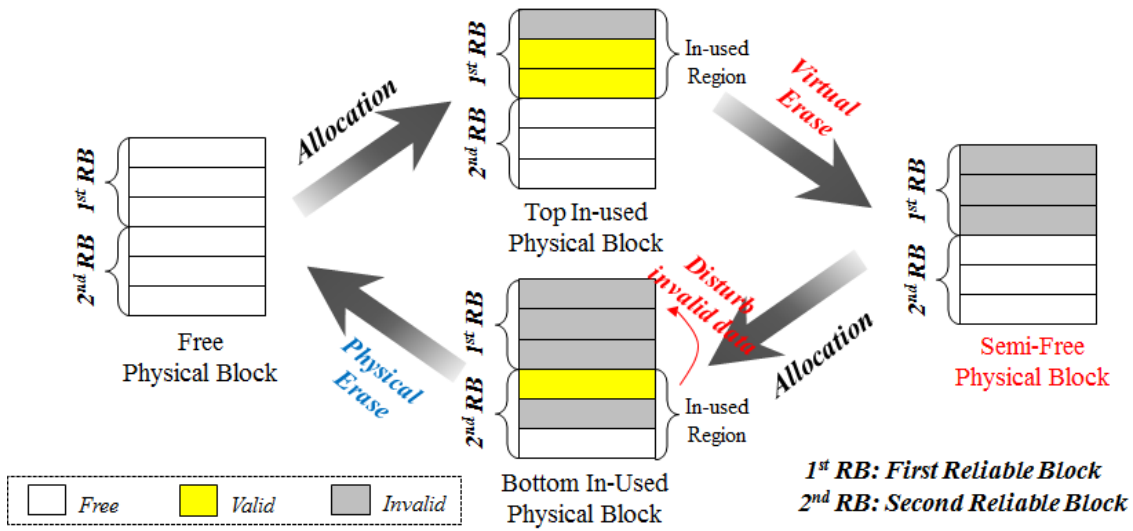
Yu-Ming Chang, Yuan-Hao Chang, Jian-Jia Chen, Tei-Wei Kuo, Hsiang-Pang Li, and Hang-Ting Lue, "On Trading Wear-leveling with Heal-leveling," ACM/IEEE Design Automation Conference (DAC), San Francisco, California, USA, Jun. 1-5, 2014. (Best Paper Nomination (12/787) - Top Conference)
Yu-Ming Chang, Hsiang-Pang Li, Hang-Ting Lue, Yuan-Hao Chang, and Tei-Wei Kuo, "Heal Leveling," Patent No.: US 9,348,748, Date of Patent: May 24, 2016.

Disturbance Alleviation for 3D Flash

[ICCAD'13, IEEE TC'16]
[US 9,558,108, 9,025,375]

- **Motivation**:
  - 3D flash memory presents a grand opportunity for huge-capacity non-volatile memory, it suffers from serious program disturb problems.
- **Main Idea**:
  - The first work that proposes a software solution **with the concept of virtual block and virtual erase** to reduce the disturb bit error rate of **real 3D flash**.
  - We use software solution to *redirect write disturbs to invalid data*.
- **Results:**
  - Experiments conducted on 3D real chips show that the proposed schemed can reduce bit error rate for 71%.



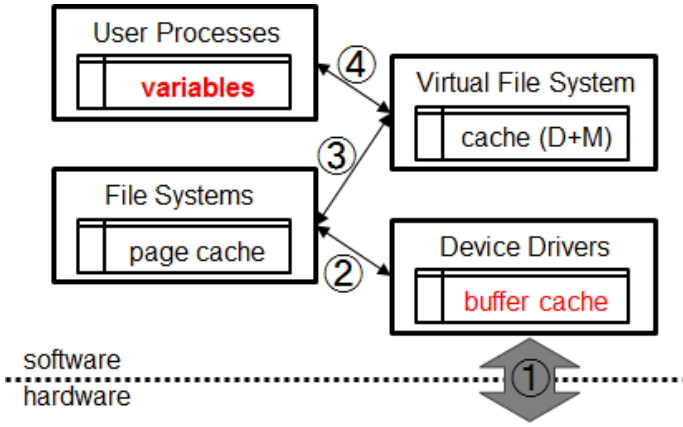


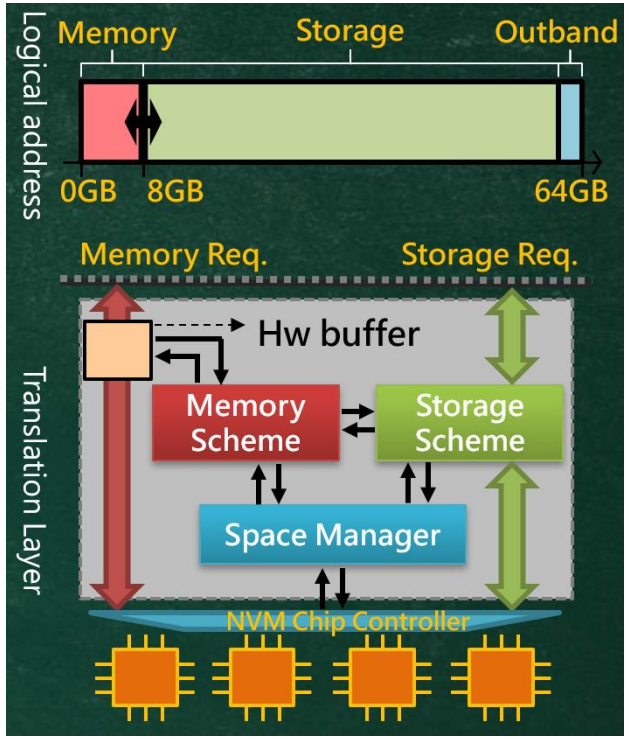
- Yu-Ming Chang, Yuan-Hao Chang, Tei-Wei Kuo, Hsiang-Pang Li, and Yung-Chun Li, "A Disturb-Alleviation Scheme for 3D Flash Memory," ACM/IEEE International Conference on Computer-Aided Design (ICCAD), USA, Nov. 18-21, 2013. (Top Conference)
- Yu-Ming Chang, Yuan-Hao Chang, Tei-Wei Kuo, Yung-Chun Li, and Hsiang-Pang Li, "Disturbance Relaxation for 3D Flash Memory," IEEE Transactions on Computers (TC), vol. 65, no. 5, pp. 1467-1483, May 2016. **[US 9,558,108, 9,025,375]**
- Yu-Ming Chang, Yung-Chun Li, Hsing-Chen Lu, Hsiang-Pang Li, Cheng-Yuan Wang, Yuan-Hao Chang, and Tei-Wei Kuo, "Half Block Management for Flash Storage Devices," Patent No.: US 9,558,108, Date of Patent: January 31, 2017.
- Yu-Ming Chang, Yung-Chun Li, Hsing-Chen Lu, Hsiang-Pang Li, Cheng-Yuan Wang, Yuan-Hao Chang, and Tei-Wei Kuo, "Memory Disturb Reduction for Nonvolatile Memory," Patent No.: US 9,025,375, Date of Patent: May 5, 2015.

One Memory – PCM Translation Layer

[CODES'14 – Best Paper Nomination][US 9,513,815]

- **Motivation**:
  - Phase change memory (PCM) is known for its potentials as main memory and as storage, but its limited write endurance, compared to DRAM, leads to the lifetime issue.
- **Main Idea**:
  - We propose the concept of "one memory" by using NVM as both memory and storage.
  - We are the **first team** to develop joint management of memory and storage
    - To reduce the data movement overheads.
    - To resolve the lifetime issue of NVM.
      - Stealing the lifetime of the large storage space to rescue the lifetime of the small memory space.
- **Results**: Improve the average memory access latency up to 4.3x and enhance the lifetime up to 3.4x.



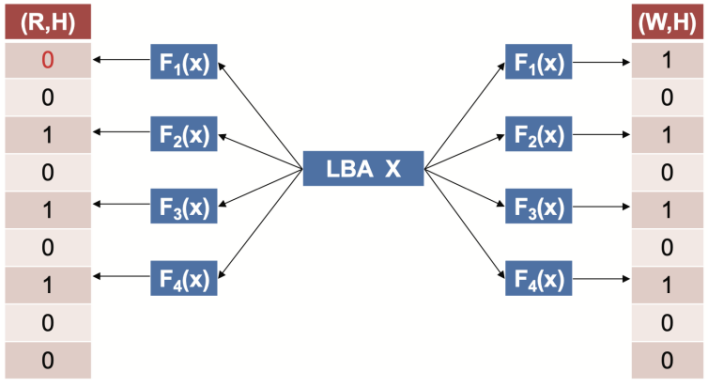


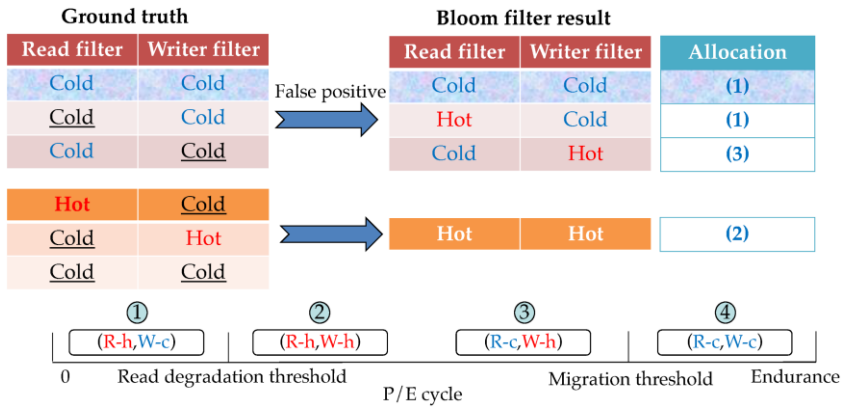
- Bing-Jing Chang, Yuan-Hao Chang, Hung-Sheng Chang, Tei-Wei Kuo, and Hsiang-Pang Li, "A PCM Translation Layer for Integrated Memory and Storage Management," ACM/IEEE International Conference on Hardware/Software Codesign and System Synthesis (CODES+ISSS), New Delhi, India, Oct. 12-17, 2014. **(Best Paper Nomination - Top Conference)**
- Ping-Chun Chang, Yuan-Hao Chang, Hung-Sheng Chang, Tei-Wei Kuo, and Hsiang-Pang Li, "Memory Management Based on Usage Specifications," Patent No.: US 9,513,815, Date of Patent: December 6, 2016.
- Ping-Chun Chang, Yuan-Hao Chang, Hung-Sheng Chang, Tei-Wei Kuo, and Hsiang-Pang Li, "Memory Management Based on Usage Specifications," Patent No.: US 9,513,815, Date of Patent: Dec. 6, 2016.

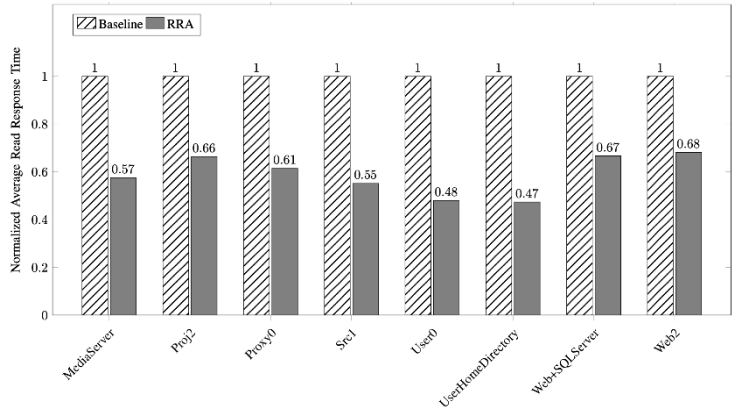
Retention-Aware Read Acceleration for LDPC-based Flash

[IEEE TCAD'23] [US 11,042,308]

- **Motivation**:
  - LDPC-based NAND flash SSD faces the problem of read performance degradation due to the increase in raw bit error rate.
  - The two main factors that affect the raw bit error rate are data retention error and P/E cycle limitation.
- **Main Idea**: We propose a retention-aware read acceleration design that exploits access patterns to improve read performance.
  - Access feature identification efficiently detects and predicts the data lifetime and access behavior.
  - Request-based allocation allocates the suitable blocks for different data (with different data lifetime and access behavior).
  - Migration lazily balances the wearing level among blocks.
- **Results**:
  - The average read response time is improved by at least about 32% and the number of total live-page copying is reduced by at least about 11%.





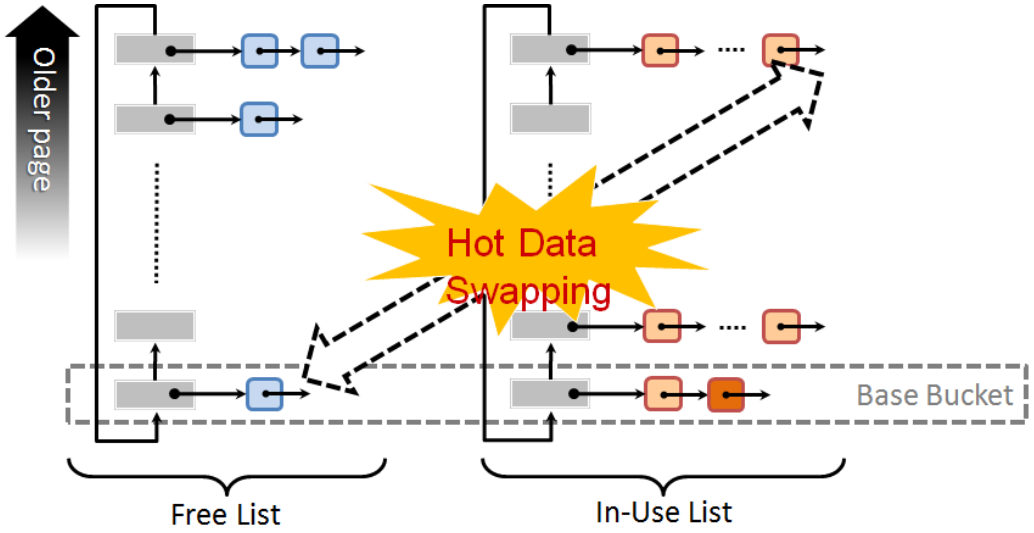


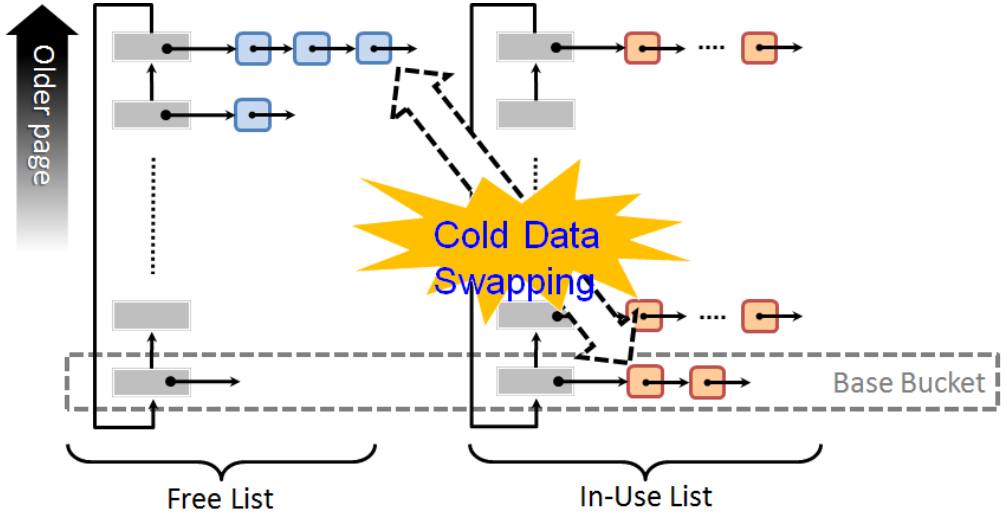
- Tse-Yuan Wang, Che-Wei Tsao, Yuan-Hao Chang, and Tei-Wei Kuo, "Retention-Aware Read Acceleration Strategy for LDPC-based NAND Flash Memory," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD), vol. 42, no. 12, pp. 4597-4605, Dec. 2023.
- Wei-Chen Wang, Ping-Hsien Lin, Tse-Yuan Wang, Yuan-Hao Chang, and Tei-Wei Kuo, "Memory Management Apparatus and Memory Management Method," Patent No.: US 11,042,308, Date of Patent: Jun. 22, 2021.

Age-based PCM Wear Leveling with Nearly Zero Search Cost

[DAC'12][US 9,513,815]

- **Motivation**:
  - Improving PCM endurance is a fundamental issue when it is considered as an alternative to replace DRAM as main memory.
- **Main Idea**:
  - We propose a age-based wear leveling design to achieve WL with nearly-zero search cost by realizing the concept of "placing old pages far away so that they are less likely to be used."
- **Results**: The proposed design was implemented in QEMU, and evaluation results show the proposed design can achieve 80% of the lifetime of the ideal case.



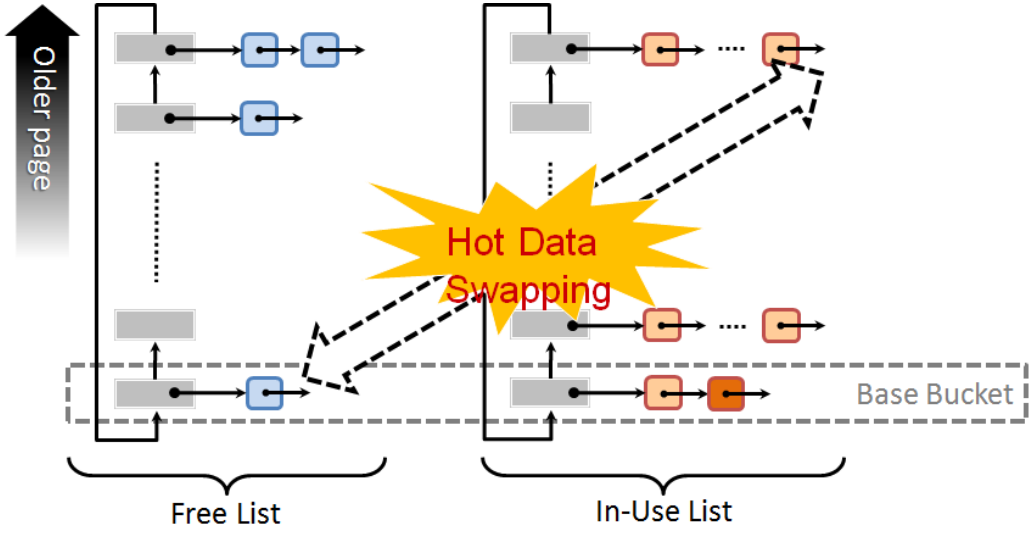


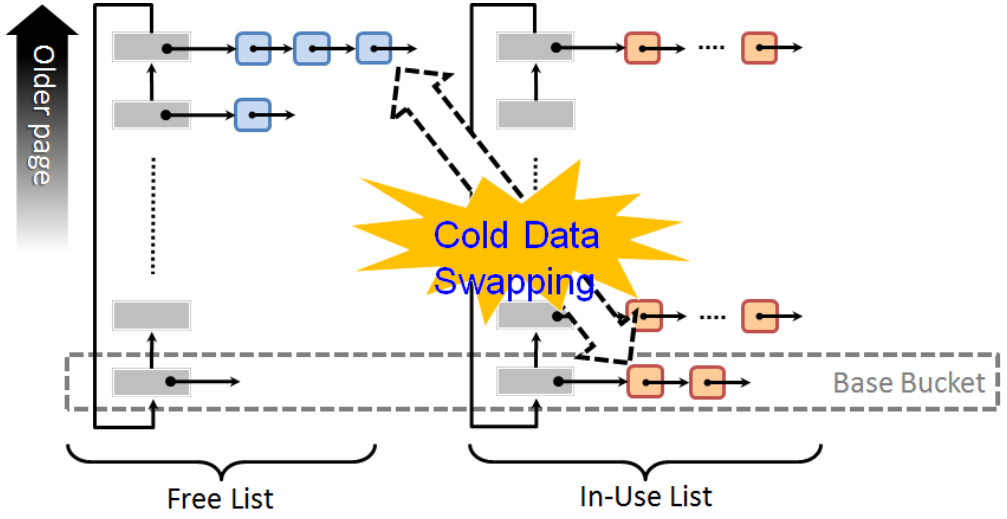
- Yu-Ming Chang, Pi-Cheng Hsiu, Yuan-Hao Chang, Chi-Hao Chen, Tei-Wei Kuo, and Cheng-Yuan Michael Wang, "Improving PCM Endurance with a Constant-cost Wear Leveling Design," ACM Transactions on Design Automation of Electronic Systems (TODAES), vol. 22, no. 1, pp. 9:1-9:27, Jun. 2016.
- Chi-Hao Chen, Pi-Cheng Hsiu, Tei-Wei Kuo, Ciha-Lin Yang, Cheng-Yuan and Michael Wang, "Age-based PCM Wear Leveling with Nearly Zero Search Cost", ACM/IEEE Design Automation Conference (DAC), Jun., 2012

Constant-Cost PCM Wear Leveling with Nearly Zero Search Cost

[DAC'12, ACM TODAES'16]

- **Motivation**:
  - Improving PCM endurance is a fundamental issue when it is considered as an alternative to replace DRAM as main memory.
- **Main Idea**:
  - We propose a age-based wear leveling design to achieve WL with nearly-zero search cost by realizing the concept of "placing old pages far away so that they are less likely to be used."
- **Results**: The proposed design was implemented in QEMU, and evaluation results show the proposed design can achieve 80% of the lifetime of the ideal case.





- Yu-Ming Chang, Pi-Cheng Hsiu, Yuan-Hao Chang, Chi-Hao Chen, Tei-Wei Kuo, and Cheng-Yuan Michael Wang, "Improving PCM Endurance with a Constant-cost Wear Leveling Design," ACM Transactions on Design Automation of Electronic Systems (TODAES), vol. 22, no. 1, pp. 9:1-9:27, Jun. 2016.
- Chi-Hao Chen, Pi-Cheng Hsiu, Tei-Wei Kuo, Ciha-Lin Yang, Cheng-Yuan and Michael Wang, "Age-based PCM Wear Leveling with Nearly Zero Search Cost", ACM/IEEE Design Automation Conference (DAC), Jun., 2012


рахмат
danke
謝謝
ngiyabonga
teşekkür ederim
Баярлалаа
mersi
kia ora
barka
welalin
tack
spas
спасибо
faafetai lava
vinaka
спасибі
blagodaram
dank je
misaotra
matondo
paldies
grazzi
mahalo
tapadh leat
kiitos
dankie
nanni
nandri
dhanyavad
thank you
gracias
хвала
asante
manana
obrigada
tenki
bedankt
bayarlalaa
gracie
hvala
maururu
köszönöm
dankon
aciu
akun
murakoze
djiere dieuf
tau
mochchakkeram
enkosi
dziękuję
дякую
mamnun
chnorakaloutioun
gratias ago
gràcies
sulpáy
go raibh maith agat
chokrane
sobodi
dēkuji
sagolun
sukriya
kop khun krap
taiku
obrigado
mèsi
ありがとう
tanemirt
grazie
arigatô
takk
dakujem
trugarez
didi madloba
najis tuke
rahmet
shukriya
mercé
мерси
kam sah hamnida
rahmat
terima kasih
ευχαριστώ
diolch
dhanyavadagalu
তোমাকে ধন্যবাদ
감사합니다
xièxie
merci