Unlocking the Full Potential of 3D NAND Flash

A Cross-Layer Approach to Reliability, Performance, and Lifetime

Chun Jason Xue

CHEOPS Workshop @ EuroSys 2026

Talk Outline

Part I: Understanding Flash Reliability
Physical characterization & key insights

Part II: Boosting Read Performance
From LDPC optimization to near-zero read retry

Part III: Smart Data Encoding
Encoding strategies for reliability & lifetime

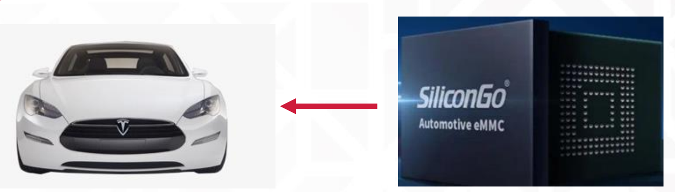
Part IV: The Reprogram Revolution
Repurposing program operations for SSD optimization

Part V: Future Challenges & Opportunities
Where do we go from here?

NAND Flash: The Backbone of Modern Storage

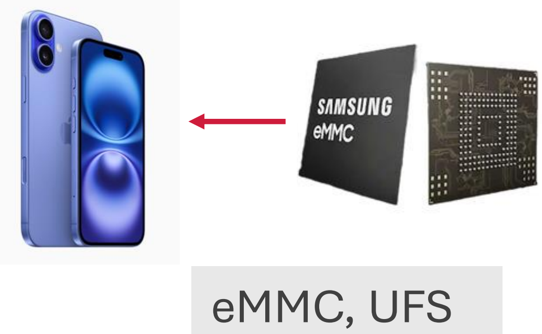
SATA eSSD, PCIe eSSD

Atomotive-grade Storage Products



eMMC

Personal Mobile Phone



eMMC, UFS

Computing Device

SATA cSSD, PCIe cSSD

Consumer Mobile Storage Device

eMMC, UFS

Flash storage is everywhere: from mobile devices to data centers.

NAND Flash: The Backbone of Modern Storage

$100B+
Global NAND market (2025)

60%+
Data center storage is flash

80%+
Data is cold / rarely accessed

Understanding and optimizing NAND flash at every layer is critical to sustaining the storage industry's growth.

Consumer SSDs

Enterprise SSDs

Cloud Storage

AI/ML Training Data

NAND Flash in ISSCC

ISSCC 2024

13.3 A 280-Layer 1Tb 4b/cell 3D...
Density and a 3.2GB/s High-...

H. Kim, W. Jung, D-B. Kim, T-H. K...
T. Kim, H. Kim, G. Lee, J. Lee, J. J...
T. Lee, B-K. Chun, T. Kim, Y. G. Le...
Y. Lee, S. Kim, J. Hwang, R. Son...
M. Lee, S. Jo, C. H. Kim, J. C. Par...
Samsung Electronics, Hwaseong, ...

13.7 A 1Tb Density 3b/Cell 3D-NA...
Write Throughput

K. Kawai[1], Y. Einaga[1], Y. Oikawa...
A. D'alessandro[3], E. Yu[2], A. Mu...
T. Ichikawa[1], J. Yu[2], G. Wang[2], K. ...
S. Bhushan[2], D. Srinivasan[2], H. Ku...
C. Siau[2], R. Ghodsi[2], [1]Micron Tech...
[3]Micron Technology, Avezzano, Ita...
[5]Micron Technology, Hyderabad, ...

W. Cho, J. Jung, J. Kim, J. Ham, S. Lee, Y. N...
H. Cho, J-S. Kim, C. Kwon, C. Park, H. Nam...
T. Shin, J. Jang, J. Mun, J. Choi, H. Choi, S-W...
H. Oh, H. Park, S. Shim, H. Huh, H. Choi, S. ...
SK hynix, Icheon, Korea

7.4 A 1Tb 3b/Cell 8th-Generation 3D-...
Throughput and a 2.4Gb/s Interface

M. Kim, S. W. Yun, J. Park, H. K. Park, J. L...
K. Lee, B. Jeong, S. Kim, J. Park, C. A. Lee, ...
M. Choi, W. Kim, J. Kim, S. Yoon, P. Kwak, M. Lee, R. Song, S. Kim, C. Yoon, D. Kang, J...
J. Song
Samsung Electronics, Hwaseong, Korea

ISSCC 2026

15.1 A 2Tb 4b/Cell 6-Plane 3D-Flash Memory with 37.6Gb/mm² Bit Density and >85MB/s Write Throughput

J. M. Thimmaiah[1], R. Yamashita[2], I-S. Yoon[3], J. Li[3], C. Hsu[3], T. Ariki[2], N. Ookuma[2], Y. Kato[2], K. Hayashi[2], K. Yamauchi[2], I. K V[1], M. Kano[2], S. Bhamidipati[1], S. Bhatia[1], S. Malhotra[1], N. Ojima[2], E. Wu[3], Z. Yang[3], F. W. Tsai[3], M. Bayle[2], N. Minami[2], Y. Fujihara[2], K. Kitamura[2], T. Kitani[2], T. Kodama[4], T. Handa[4], N. Kanagawa[4], Y. Ishizaki[4], S. Fujimura[4], Y. Suzuki[4], M. Sako[4], Y. Higashi[4], Y. Watanabe[4], T. Kouchi[4], A. V[1], C-Y. Chen[3], X. Yang[3], G. Liang[3], J. Wang[3]
[1]Sandisk, Bengaluru, India; [2]Sandisk, Yokohama, Japan
[3]Sandisk, Milpitas, CA; [4]KIOXIA, Yokohama, Japan

30.6 A 16Mb 166.8TOPS/W Near-Memory Phase-Domain-Computing Ferroelectric NAND Flash for Approximate Nearest Neighbor Search on Edge Devices

W. Li*[1,2], B. Wang *[1,2], Z. Zhou[1,2], J. Zhu[1,2], Z. Li[1,2], J. Shen[1,2], W. Zha[1,2], Z. Han[1,2], Y. Wang[1,2], L. Wang[3], H. Hu[1,2], Q. Luo[1,2], C. Dou[1,2], M. Liu[1]
[1]Institute of Microelectronics of the Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
[3]Columbia University, New York, NY
*Equally Credited Authors (ECAs)

ISSCC 2025

...cell WF-Bonding 3D-NAND Flash

...g, C-H. Yu, H. Makoto, Y. Kwon, J-H. Park,
... Kim, D-H. Kim, Y. Jo, H. Yoon, J. Park,
...-S. Ku, J. Seo, J. Byun, S-H. Yun, K. Kang,
... Y. Ryu, H. Kim, W. Kim, H. Choi, J. Jeon,
... Lee, M-K. Lee, J-I. Son, J. Cho, M. Kim,
...n, S. Lee, K. Song, S-H. Hur

... with a 29%-Improved-Energy-
...8Gb/s Power-Isolated

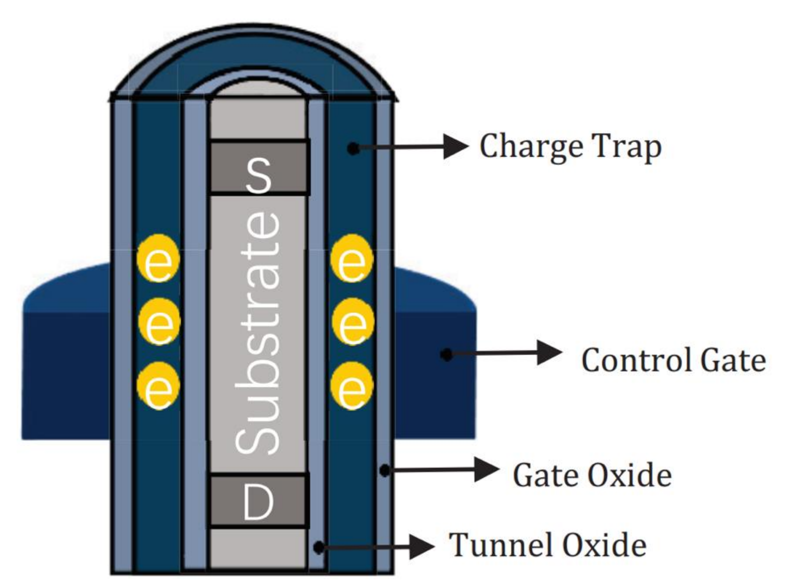
... Matsuno[1], Y. Higashi[1], Y. Shimizu[1],
... Nakano[1], Y. Ochi[1], H. Hoshino[1], T. Hioka[1],
... Y. Kamata[1], H. Chibvongodze[2], N. Ojima[2],
...ukoshi[2], R. Yamashita[2], K. Abe[2],
...uh[2], K. Htoo[2], Y. Kato[2], Y. Watanabe[1],
... Western Digital, Milpitas, CA

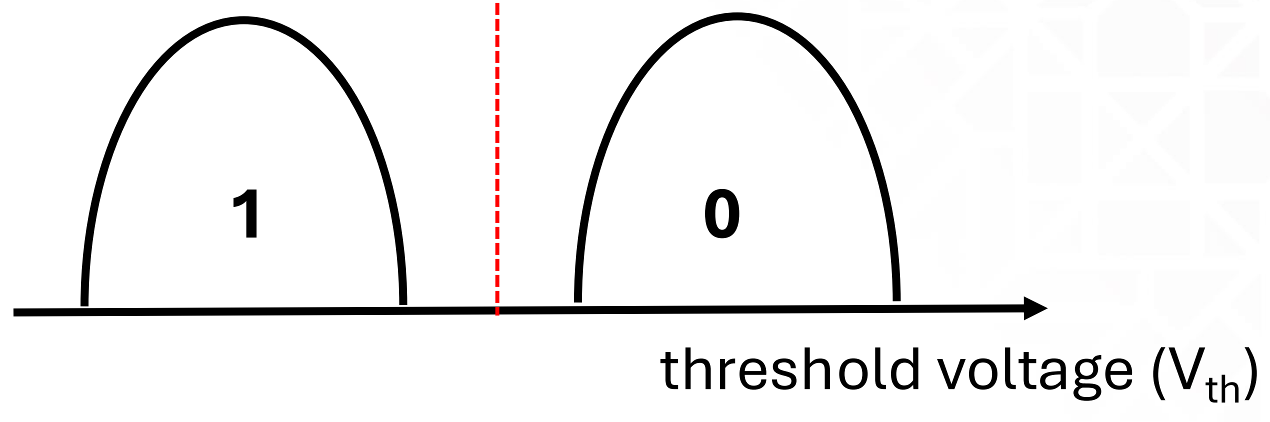
NAND Flash Structure

3D NAND flash package

Flash cell: charge trap based

Storing charges inside gates to represent data





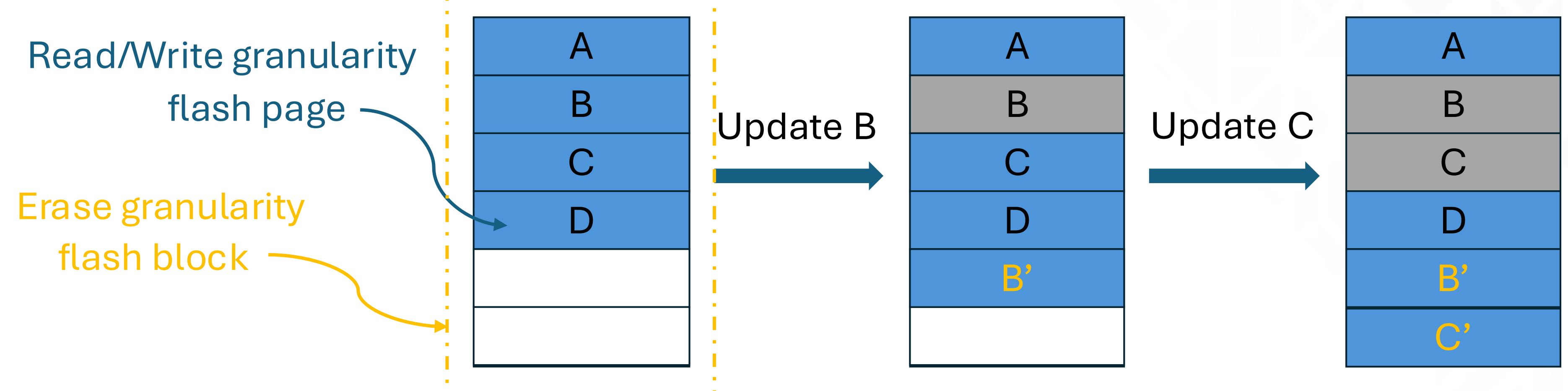
Program (Write): charge cells

Read: identify the voltage levels, low or high

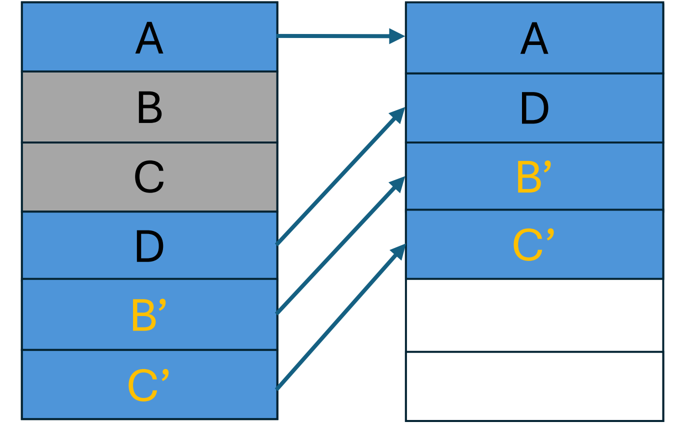
Erase: discharge cells ← Unique

Out-of-Place Update

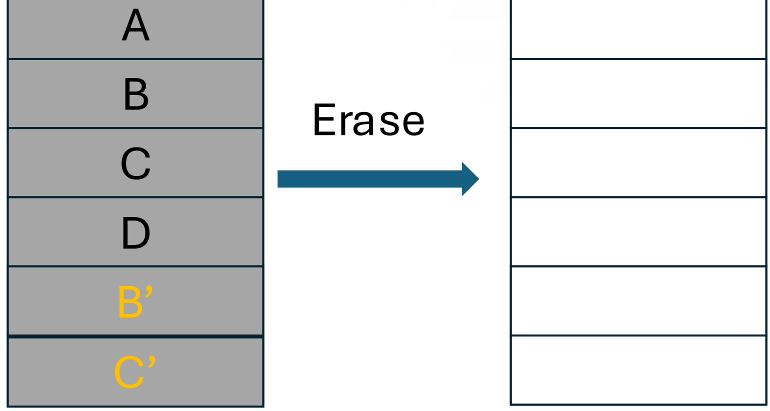
Write updated data to a new free location, mark the original location as invalid.



Garbage collection:
migrate valid data →
erase the block →
release free space.



migrate valid data



Flash Translation Layer

Core firmware layer connecting the host with the NAND physical medium.

Host / File System
LBA (Logical Block Address)

FTL
Flash Translation Layer

Address Mapping | Garbage Collection | Wear Leveling | Bad Block Mgmt

NAND Flash Array
PBA (Physical Block Address)

Address Mapping: Translates host LBAs to NAND PBAs and maintains the mapping table.

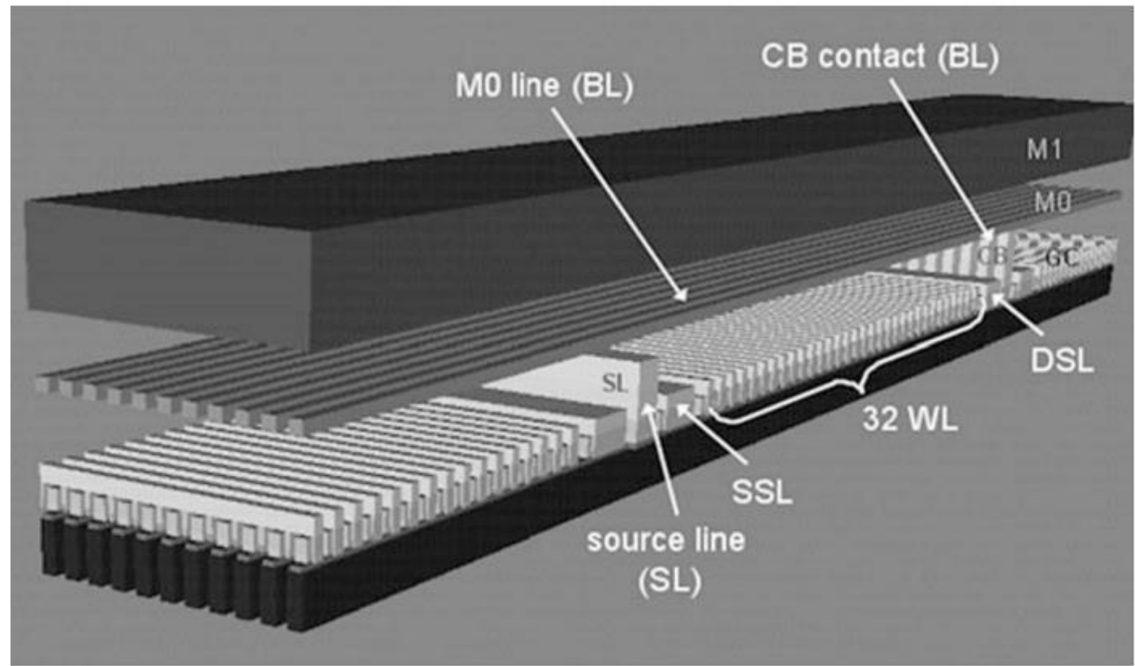
Garbage Collection (GC) : Reclaims blocks containing invalid pages

Wear Leveling : Evenly distributes P/E cycles to prevent premature wear of specific blocks.

Bad Block Management : Detects and isolates factory-bad blocks and runtime-bad blocks, maintaining a bad block table.

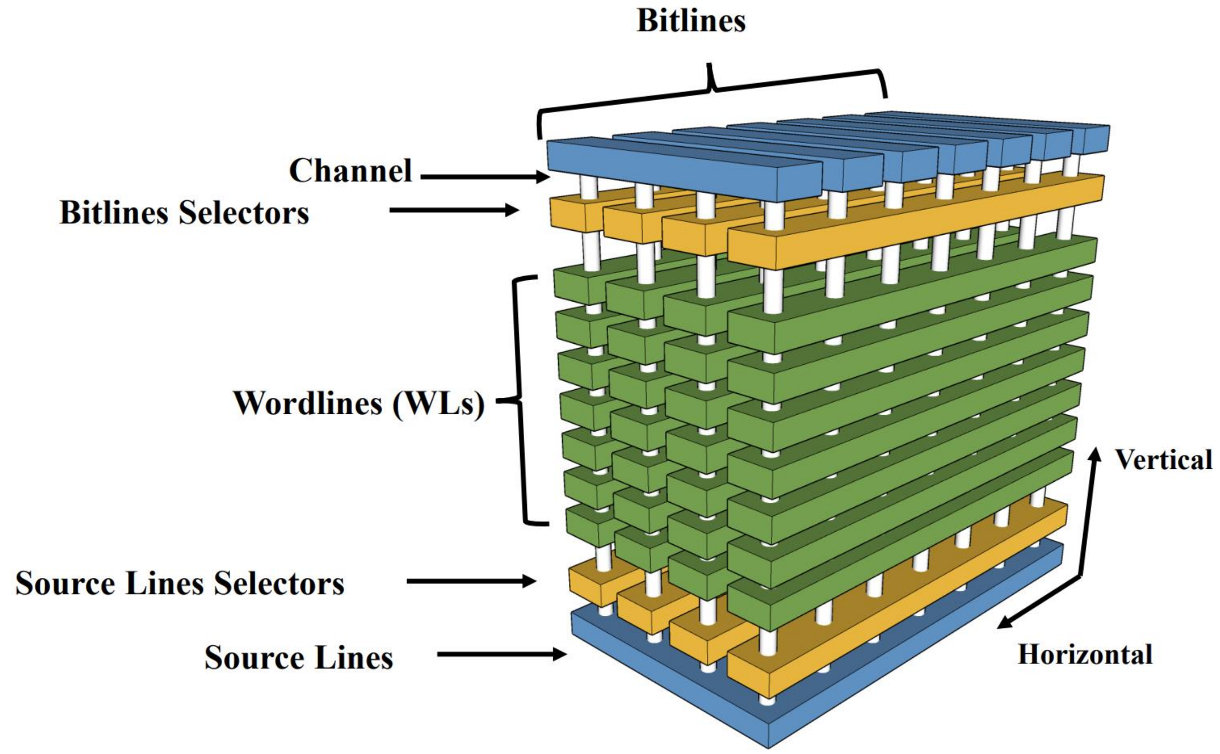
2D VS 3D NAND

2D NAND flash memory



Scaling method: shrinking cell size
Process node: 15–19 nm
Cell-to-cell interference: severe

3D NAND flash memory



Scaling method: vertical layer stacking
Layer count: 128 / 176 / 232+
Cell-to-cell interference: significantly improved

3D NAND Scaling

Cell String
(Poly-Si Channel)
Charge Trap Layer
Layer (Nitride)
Tunnel Oxide
Gate Oxide
Word Line
(Gate)
Vertical Pillar
(Channel Hole)
Bit Lines
32L
32L
64L
128L
232L+
Insulator
Silicon
Substrate

(1) 32-Layer 3D NAND
(approx. 2013)

(2) 64-Layer 3D NAND
(approx. 2016)

(3) 128-Layer 3D NAND
(approx. 2019)

(4) 232-Layer / 256+ Layer 3D NAND
(Current/Advanced)

Number of layers inside 3D flash block in increasing

3D NAND Scaling

3D NAND flash tend to store more bits per cell

SLC
1b/cell
2 states

ER: 1 | P_1: 0 | V_{th}

MLC
2b/cell
4 states

ER: 11 | P_1: 10 | P_2: 01 | P_3: 00 | V_{th}

TLC
3b/cell
8 states

ER: 111 | P_1: 110 | P_2: 100 | P_3: 101 | P_4: 001 | P_5: 000 | P_6: 010 | P_7: 011 | V_{th}

SLC (1b/cell) → MLC (2b/cell) → TLC (3b/cell) → QLC (4b/cell)

3D NAND Scaling

Suffer from multiple types of errors

More errors are introduced because of tighter voltage margins

ER V_{ref1} P_1 V_{ref2} P_2 V_{ref3} P_3 V_{ref4} P_4 V_{ref5} P_5 V_{ref6} P_6 V_{ref7} P_7 V_{Pass}

Shift to the right

111 110 100 101 001 000 010 011

V_{th}

ER V_{ref1} P_1

Error

111 110

⚠ Higher density → tighter voltage margins → more errors → worse performance & shorter lifetime

The Flash Scaling Trilemma

Bitlines
Channel
Bitlines Selectors
Wordlines (WLs)
Vertical
Source Lines Selectors
Source Lines
Horizontal

Reliability

Trade-offs

Performance

Lifetime

As 3D NAND scales to higher density, all three attributes degrade simultaneously.

Our approach: Cross-layer understanding and optimization to push the Pareto frontier on all three dimensions simultaneously.

A Decade of Flash Research (2015–2026)

- **2015–16** FAST '16: Access-guided R/W regulation
- **2017–18** ASP-DAC '17 (Best Paper Nominee) · ICCD '18: Selective compression
- **2019** DAC '19 · MICRO '19: Reprogramming TLC · TC '19: Asymmetric LDPC
- **2020** MICRO '20: Sentinel Cells · DAC '20: Partition for read perf.
- **2021–22** HotStorage '21 ★ **Best Paper** · TCAD '22: Open blocks
- **2023** HPCA '23: MGC · NVMSA '23 ★ **Best Paper** · TCAD '23: LDPC prediction
- **2024** ASPLOS '24: Near-zero retry · HPCA '24: Midas Touch · TACO '24
- **2026** ISCA '26: LOONG · EuroSys '26: ColdCode

Venues: ISCA, MICRO, HPCA, ASPLOS, EuroSys, FAST, DAC

Part I

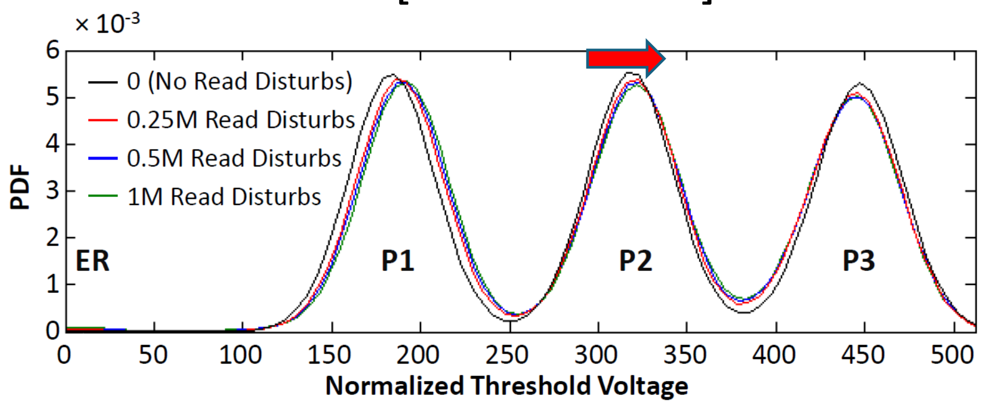
Understanding Flash Reliability

Deep physical characterization as the foundation for all optimizations

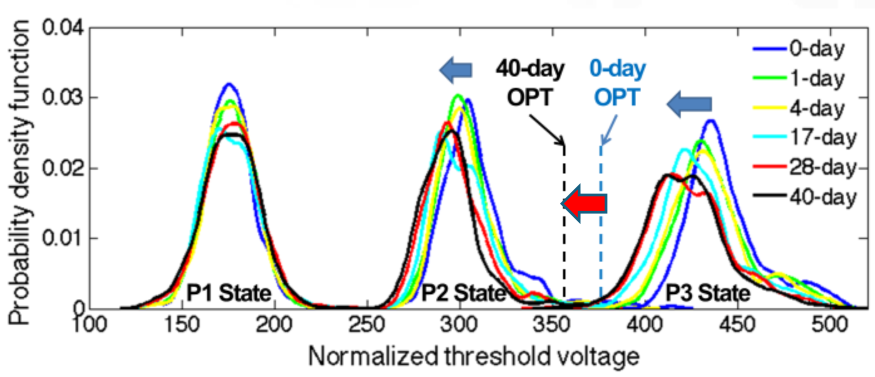
NAND Flash Reliability

Different types of error sources impact flash reliability.

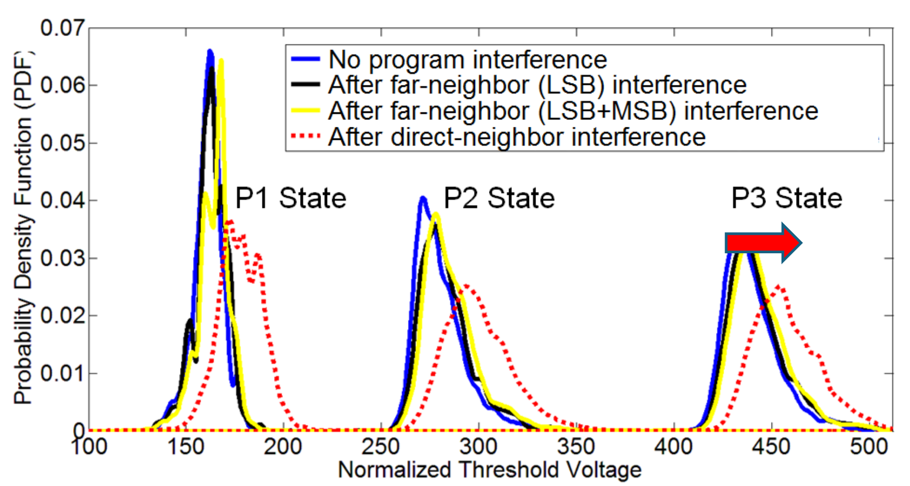
Read disturb [Cai et al. DSN]



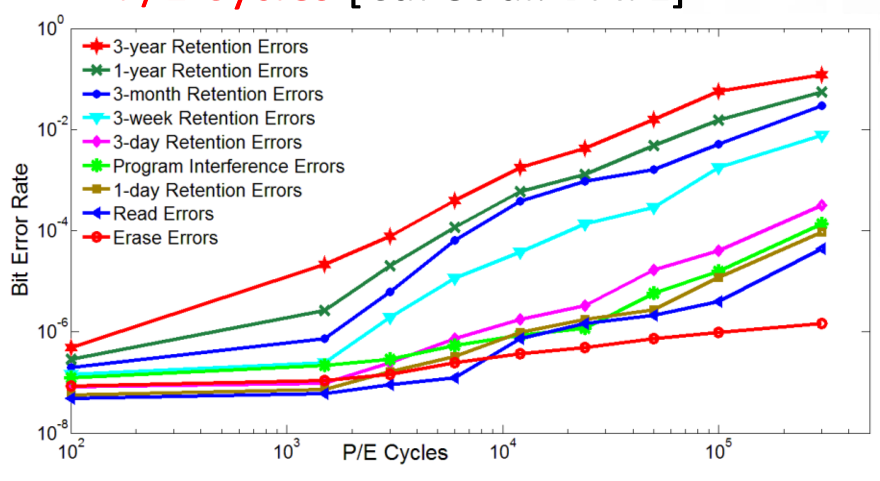
Retention error [Cai et al. HPCA]



Program Interference [Cai et al. ICCD]

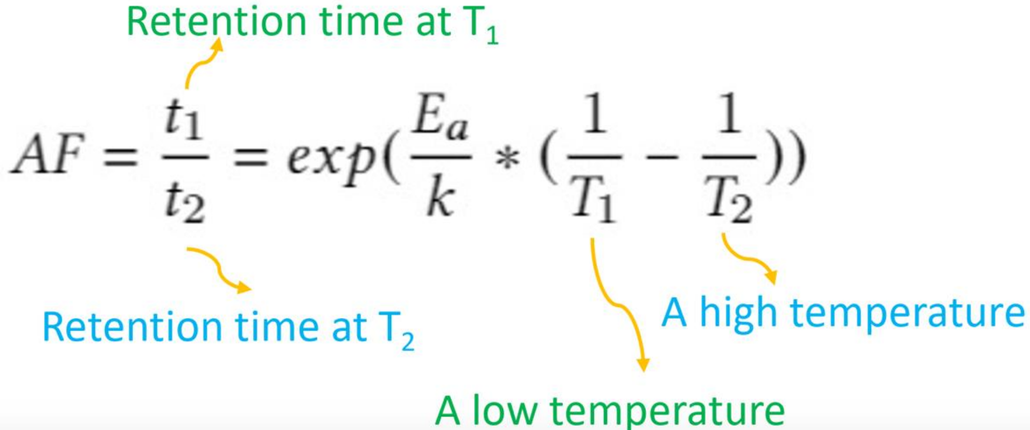


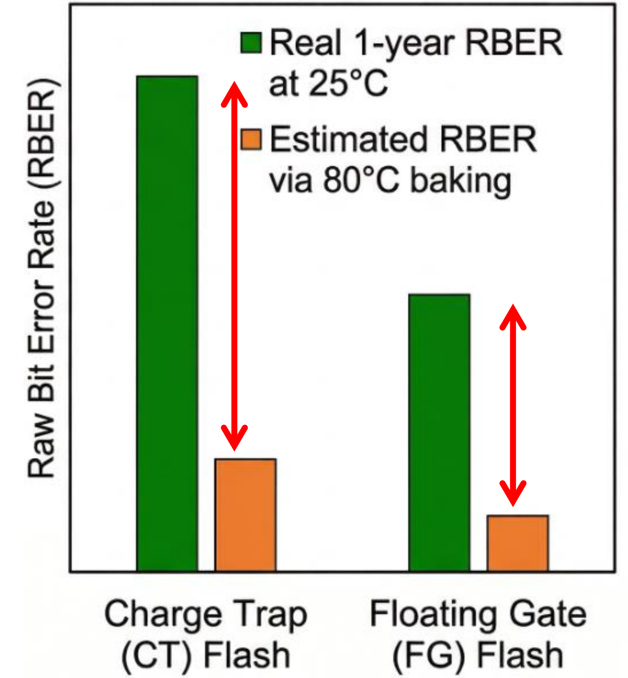
P/E Cycles [Cai et al. DATE]



When Retention Testing Goes Wrong

Industry uses bake acceleration (high-temperature baking) to predict long-term retention degradation. But does this common method actually work for 3D NAND?





Over 10X difference

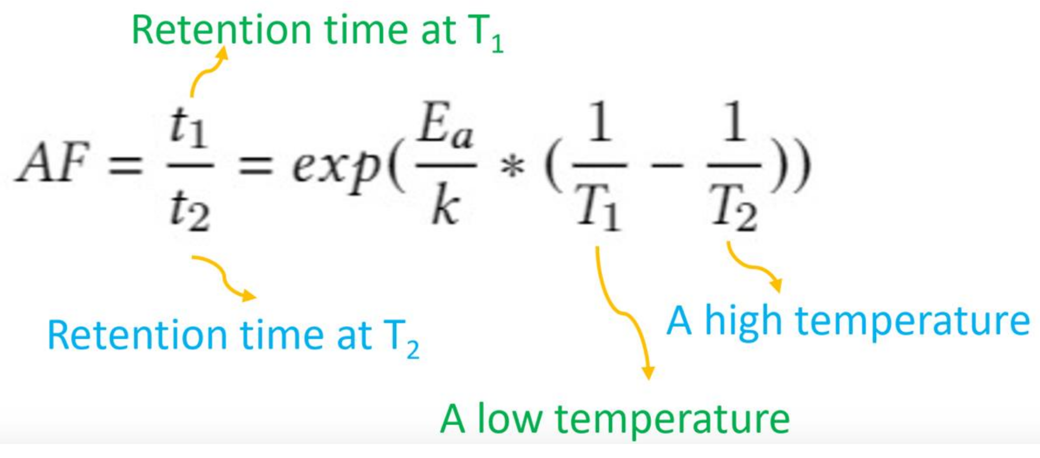
The answer is NO!

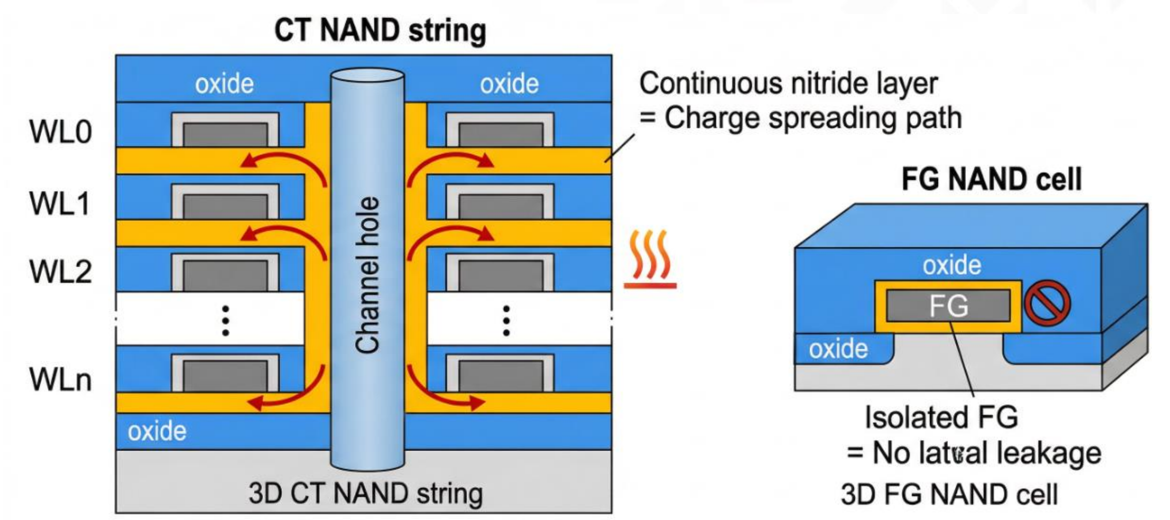
The baking underestimates real long-time errors

Li et al., "How the common retention acceleration method of 3D NAND flash memory goes wrong?" HotStorage 2021

When Retention Testing Goes Wrong

Industry uses bake acceleration (high-temperature baking) to predict long-term retention degradation. But does this common method actually work for 3D NAND?



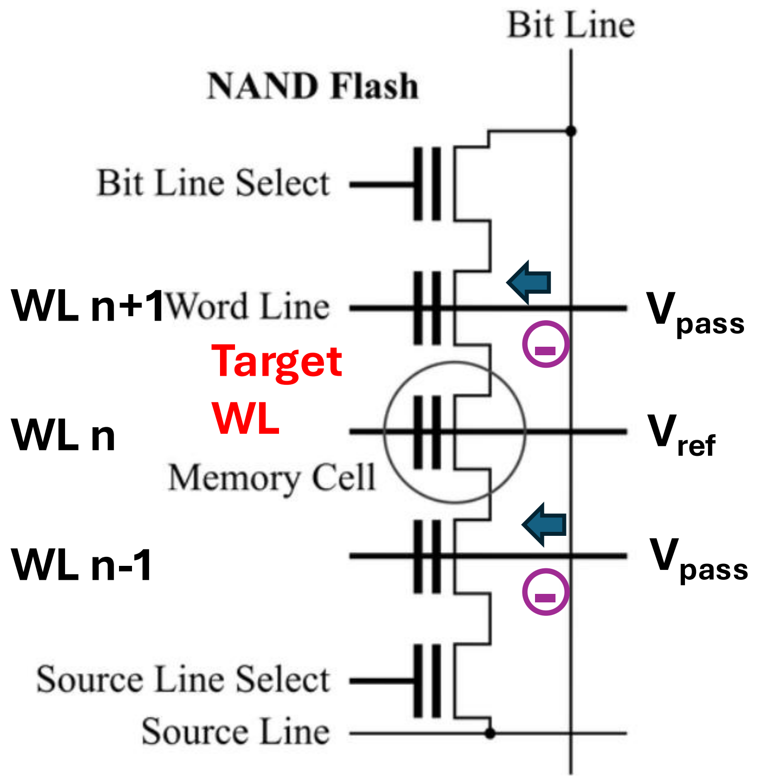


The answer is NO!

Li et al., "How the common retention acceleration method of 3D NAND flash memory goes wrong?" HotStorage 2021

Read Disturb: The Complete Story for 3D CT NAND

Read operations unintentionally disturb stored data in neighboring cells. This is a growing concern as layer counts increase.



New Insights

Read disturb behavior differs fundamentally from 2D NAND. It is location-dependent, asymmetric, and coupled with other error sources in complex ways.

Takeaway: Understanding read disturb's full picture in 3D CT NAND is essential for designing robust read management strategies.

Ren et al., "Read Disturb and Reliability: The Complete Story for 3D CT NAND Flash " NVMSA 2023

Read Disturb: The Complete Story for 3D CT NAND

The read impacts as the compensation can **reduce** the RBERs at first, and than more reads will **increase** RBERs, the impacts of read will turn to be the disturbance.
Transformation from **compensation** to **disturbance**

Bit Line
NAND Flash
Bit Line Select
WL n+1 Word Line
Target WL
WL n
Memory Cell
WL n-1
Source Line Select
Source Line
V_{pass}
V_{ref}
V_{pass}

Flash 1

1 P/E, 1K P/E, 2K P/E, 3K P/E, 4K P/E, 5K P/E

RBERs ($\times 10^{-2}$)

Read Counts ($\times 10^{6}$)

Flash 2

1 P/E, 1K P/E, 2K P/E, 3K P/E, 4K P/E, 5K P/E

RBERs ($\times 10^{-4}$)

Read Counts ($\times 10^{6}$)

Ren et al., "Read Disturb and Reliability: The Complete Story for 3D CT NAND Flash " NVMSA 2023

Data Randomization

All NAND flash chips apply data **randomization** before programming to avoid extreme data patterns that generate worst-case RBER.

(d) PS34

(b) PS2

(e) PS07

(f) PS24

randomize

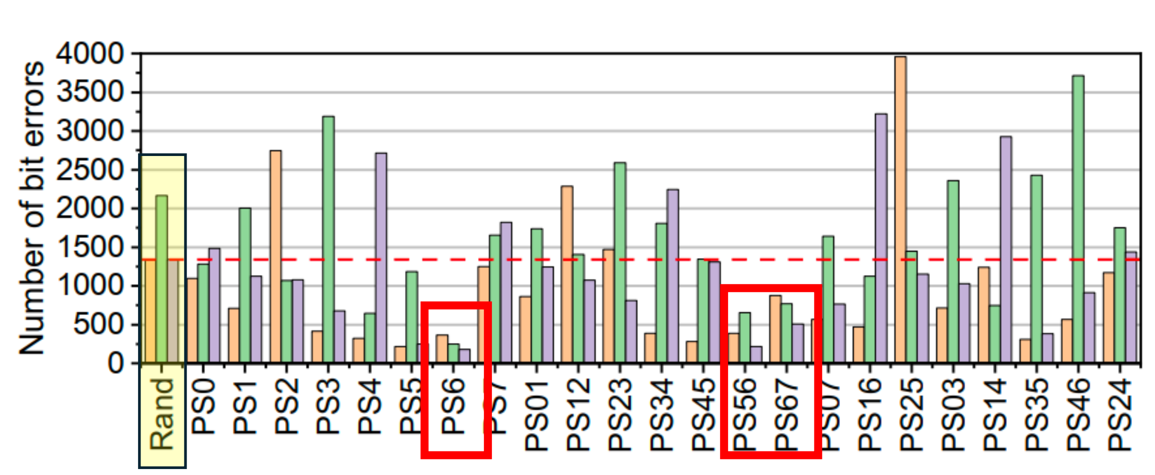
(a) Rand

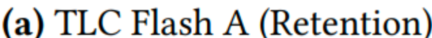
We tested the reliability of 24 different data patterns.

Data Randomization

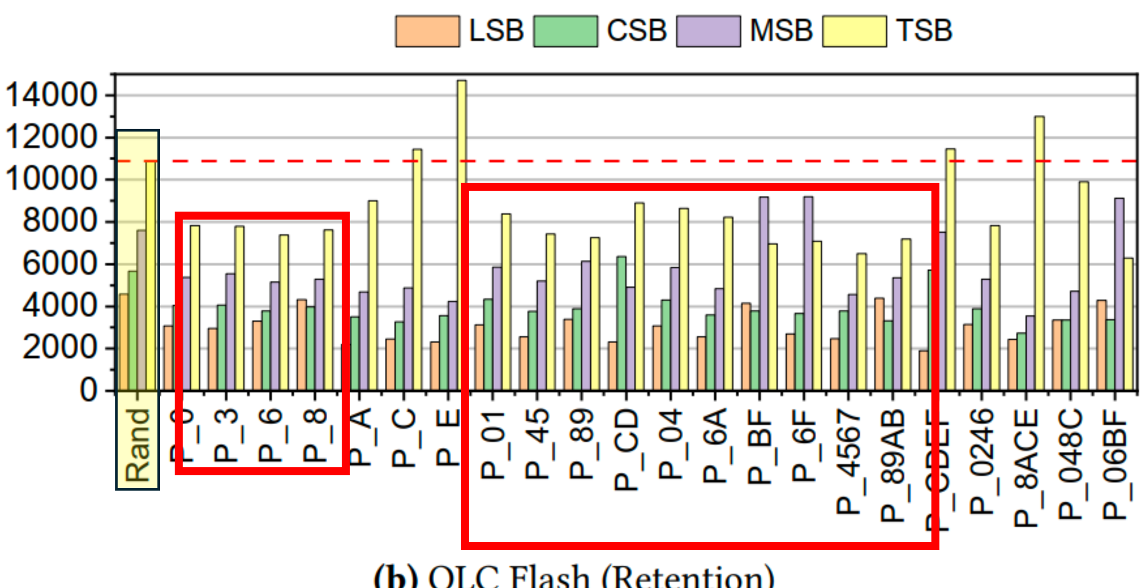
Randomizer is suboptimal

Some data patterns outperform randomized data in RBER





(a) TLC Flash A (Retention)



(b) QLC Flash (Retention)

✓ What randomization does well

Prevents worst-case data patterns that cause extremely high raw bit error rate (RBER). Industry standard for decades.

✗ What randomization misses

Also eliminates best-case patterns that have very low RBER — a significant missed optimization opportunity for cold data.

💡 Part I — Key Insight

Deep physical characterization of 3D NAND flash reveals that many industry-standard assumptions are incomplete or incorrect. Every optimization must start from accurate hardware understanding.

"Accurate understanding of the hardware is the prerequisite for effective optimization."

Collectively these characterizations form the foundation of our optimization work

Part II

Boosting Read Performance

From LDPC optimization to achieving near-zero read retry

LDPC Error Correction in NAND Flash

Modern high-density flash relies on LDPC (Low-Density Parity-Check) codes for error correction. But LDPC introduces significant read latency overhead.

11 10 01 00 V_{th}

LDPC hard decoding

Low RBER ↔ Less Sensing Levels ↔ Low Read Latency

1110 1001 0100

1111 1010 0101 0000 V_{th}

1101 1000 0010

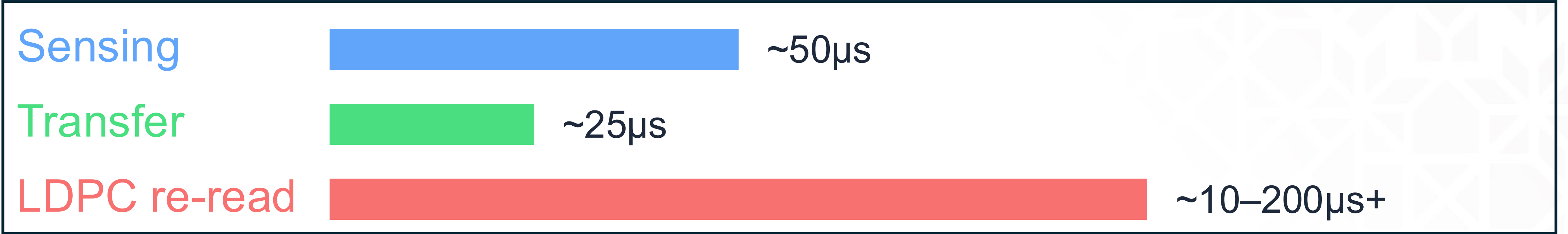
LDPC soft decoding

High RBER ↔ More Sensing Levels ↔ High Read Latency

LDPC Error Correction in NAND Flash

Modern high-density flash relies on LDPC (Low-Density Parity-Check) codes for error correction. But LDPC introduces significant read latency overhead.

Read Latency Breakdown:



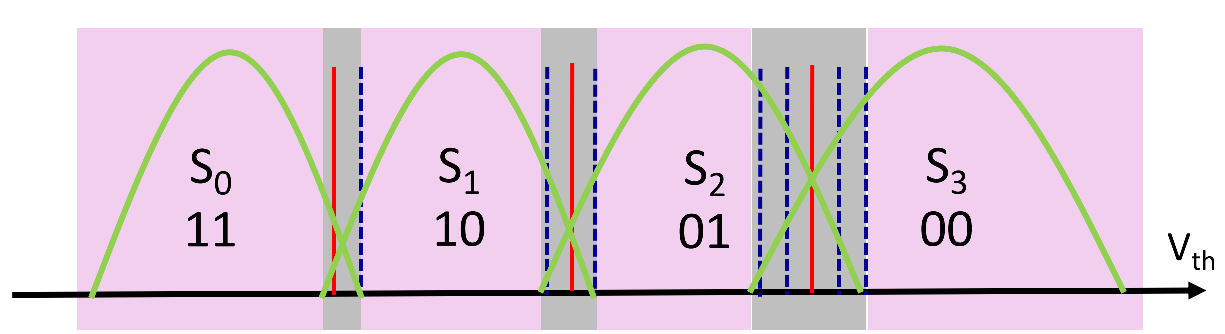
Problem: When LDPC fails at the default level, the SSD must perform read retries with adjusted voltage references — each retry adds 100μs+ of latency. This creates severe tail latency for storage systems.

Our research tackles this from multiple angles: better sensing, better decoding, and ultimately eliminating retries.

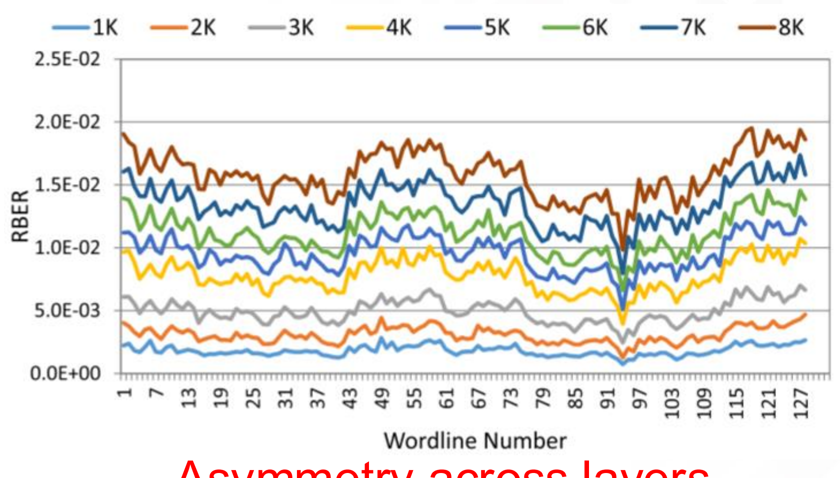
Exploiting Error Asymmetry

Flash errors are not symmetric across voltage distributions.

By placing sensing levels asymmetrically, we minimize raw bit errors and dramatically improve LDPC success rate.



Asymmetry across states



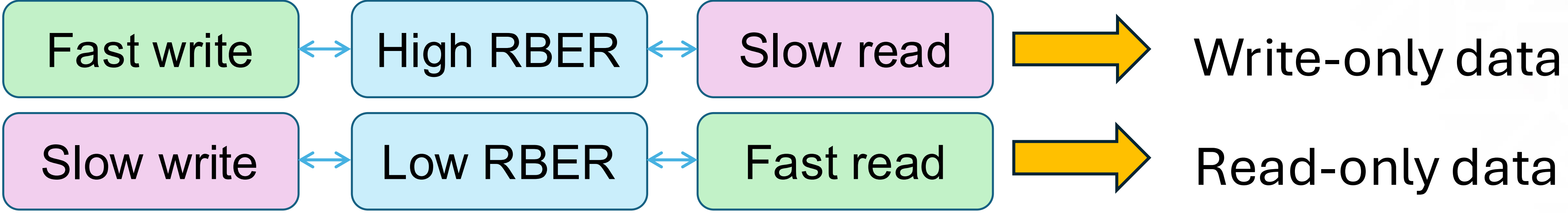
Asymmetry across layers

Key principles: exploiting physical error characteristics at the device level

Li et al., ASP-DAC 2017 · Li et al., TC 2019

Exploiting Access Patterns

Most read requests occur on read-only data; Most write requests occur on write-only data

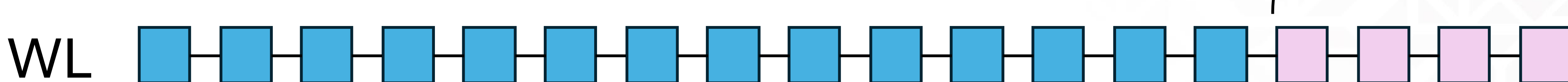


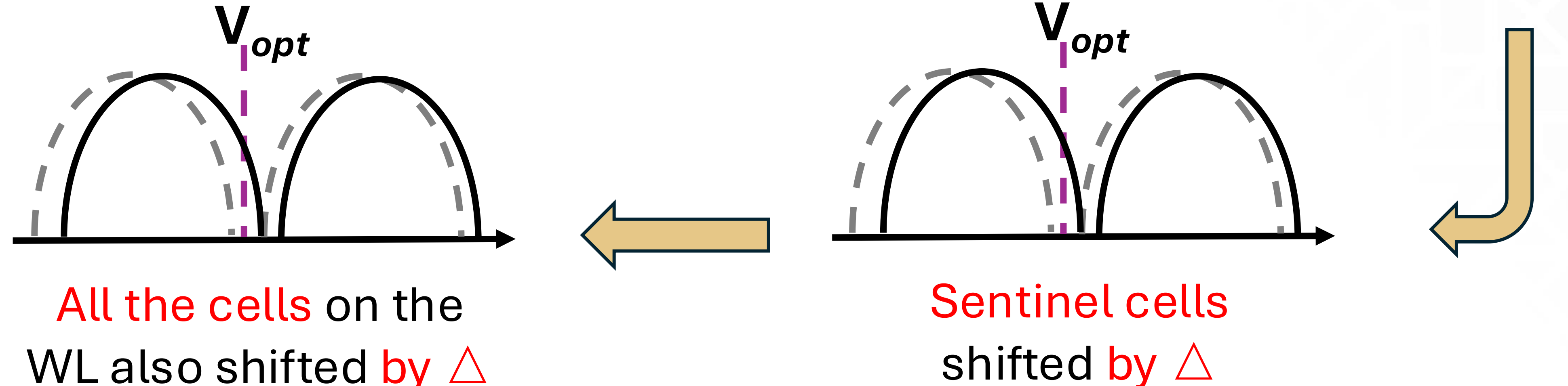
Key principles: exploiting data characteristics to guide flash optimization

Li et al., FAST 2016 · Li et al., TC 2019

Sentinel Cells: Shaving Retries for Fast Read

The cells on the same WL endure the same error conditions and tend to have the same/similar voltage shift





Li et al., "Shaving Retries with Sentinels for Fast Read over High-Density 3D Flash," MICRO 2020

Sentinel Cells: Shaving Retries for Fast Read

The optimal values of different read voltage have a linear relationship

Obtaining the optimal value of one read voltage → Inferring the optimal value of other read voltages

Li et al., "Shaving Retries with Sentinels for Fast Read over High-Density 3D Flash," MICRO 2020

Sentinel Cells: Shaving Retries for Fast Read

Core Idea: Designate a small set of flash cells as "sentinels" that serve as reliability indicators for the entire page.

Traditional Read Retry Process

Read with Default Voltage
ECC Decode Fails
Retry #1 with Voltage Set A
Fails
Retry #2 with Voltage Set B
Fails
...
Retry #N Success
Long Latency
Blind Trial-and-Error

Word line
Word line
Word line
Word line
...
Word line

Same conditions:
P/E cycles
Retention time
Read disturb

Sentinel Cells

Sentinel-Guided Fast Read

Read Sentinel Cells
Analyze Voltage Shift
Calculate Optimal Read Voltage
Direct Read with Optimal Voltage
Success
Low Latency

Threshold Voltage distribution
Shifted
Shifted
Threshold Voltage (V)

Sentinel readings locate optimal read reference voltage
Informed Single-Shot

81% reduction in read retries

Li et al., "Shaving Retries with Sentinels for Fast Read over High-Density 3D Flash," MICRO 2020

Achieving Near-Zero Read Retry

Building on sentinel work, this paper pushes toward the ultimate goal: virtually eliminating read retries in 3D NAND flash.

Initial Offset Point — *Overcrowded Offset Points*

(a) Flash A

Overcrowded Offset Points — *Overlapped Offset Points*

(b) Flash B

Zigzag Pattern

(c) Flash C

(Axes: R3 offset vs. R7 offset)

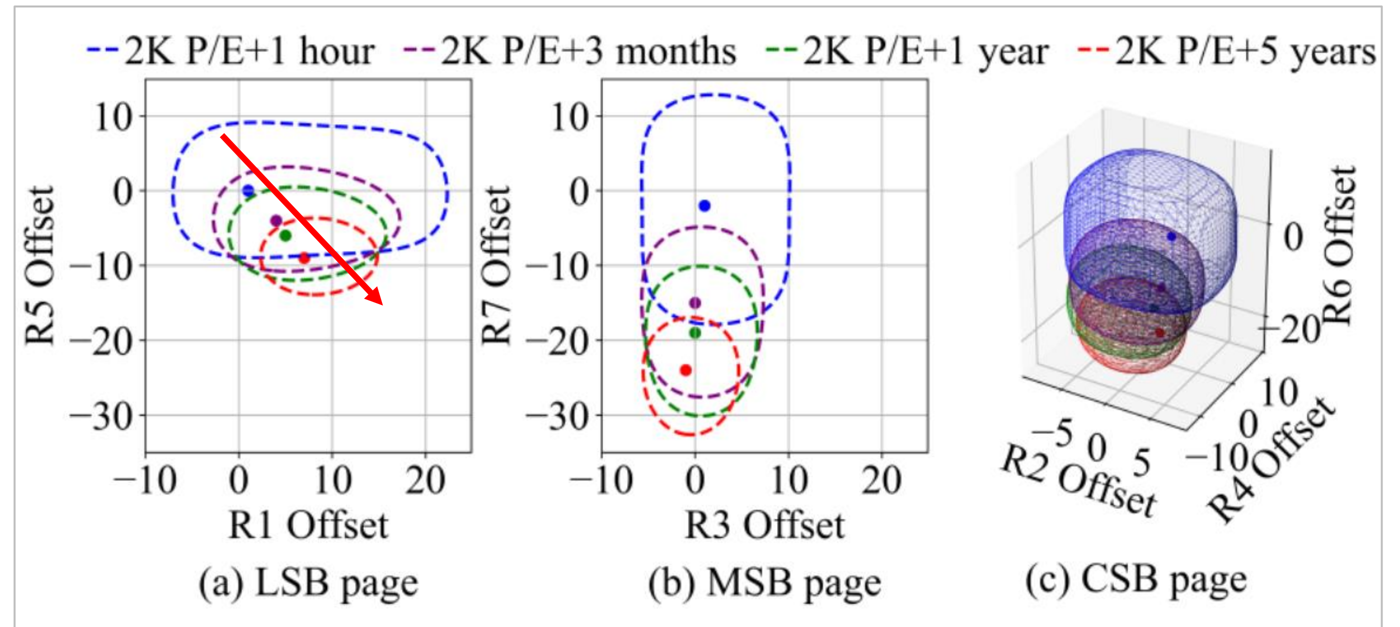
Current RRT in flash controller: overcrowded, overlapped, zigzag

RRT – Read Retry Table

Ye, Li et al., "Achieving Near-Zero Read Retry for 3D NAND Flash Memory," ASPLOS 2024

Achieving Near-Zero Read Retry

Our work: customize RRTs based on the moving of RRCS
First introduce Read Retry Cover Space (RRCS)



Our Approach
≈ 0%
near-zero read retry

Significance: Near-zero retry means consistent, predictable read latency — critical for latency-sensitive applications in data centers and cloud storage.

Ye, Li et al., "Achieving Near-Zero Read Retry for 3D NAND Flash Memory," ASPLOS 2024

💡 Part II — Key Insight

Read performance evolution: from reducing LDPC iterations → sentinel-guided reads → near-zero retry.

Each layer of understanding enables the next breakthrough.

" Each layer of understanding enables the next breakthrough in read performance."

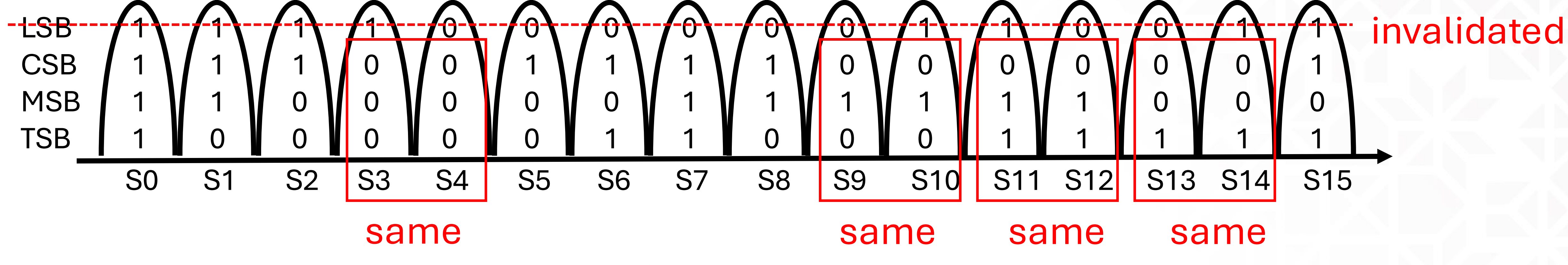
Part III

Smart Data Encoding

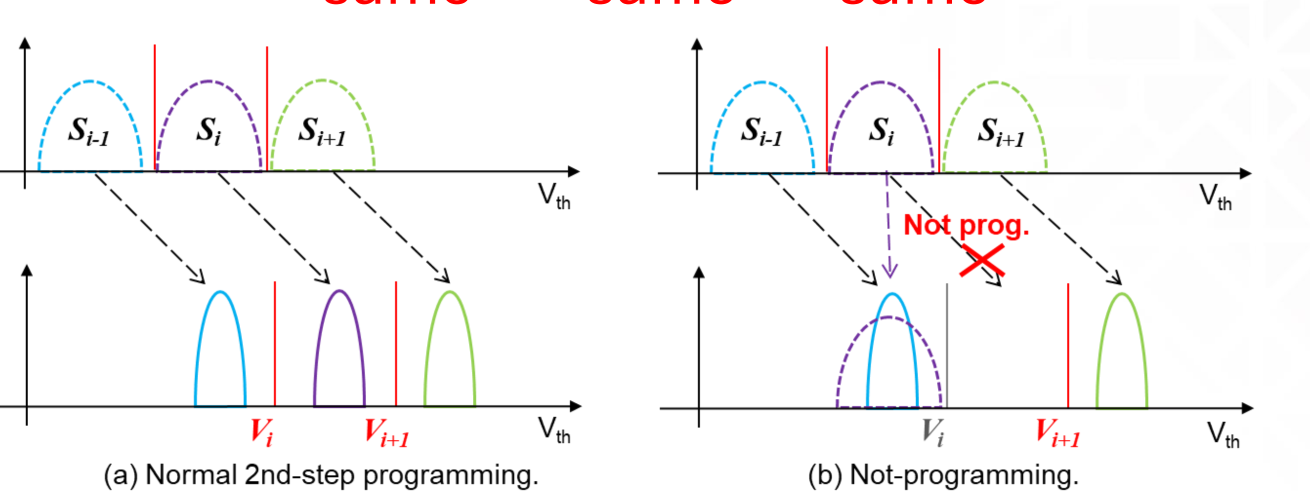
How data is encoded determines flash reliability and lifetime

Midas Touch: Invalid Data Assists Flash

When data is logically invalidated (overwritten) before the second programming step, can we turn this "waste" into an advantage?



Key idea: Enlarge the noise margin between voltage states to improve data reliability of valid pages.



(a) Normal 2nd-step programming.

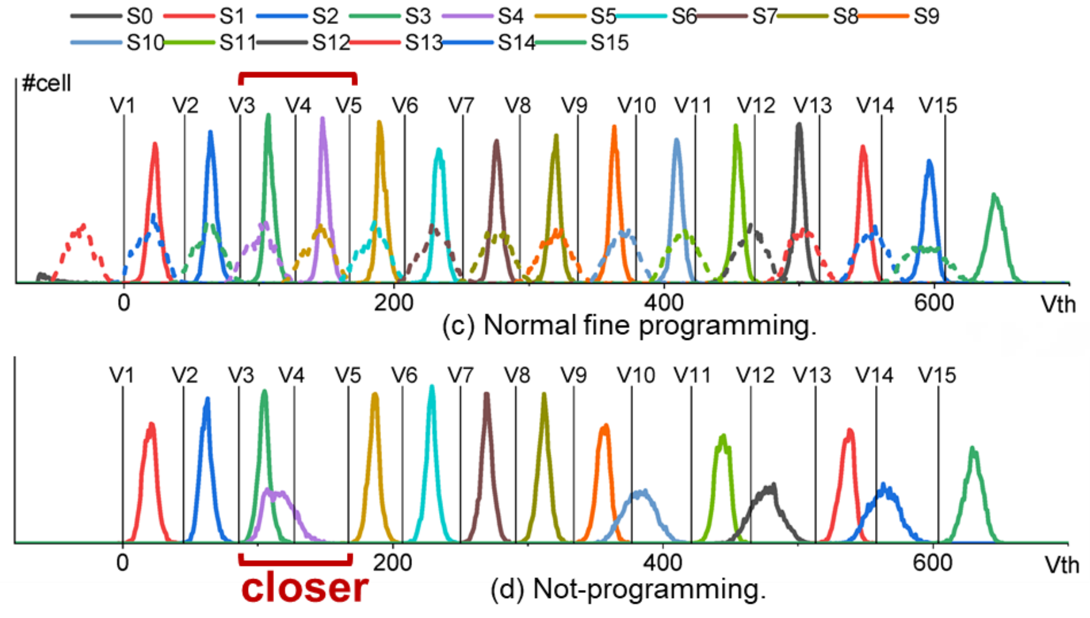
(b) Not-programming.

Li et al., "Midas Touch: Invalid-Data Assisted Reliability and Performance Boost for 3D High-Density Flash," HPCA 2024

Midas Touch: Invalid Data Assists Flash

When data is logically invalidated (overwritten) before the second programming step, can we turn this "waste" into an advantage?

Real-device voltage distribution



Turns invalid data from a nuisance into an asset — "everything it touches turns to gold." Zero cost, pure gain.

Li et al., "Midas Touch: Invalid-Data Assisted Reliability and Performance Boost for 3D High-Density Flash," HPCA 2024

ColdCode: Data Encoding for Cold Storage

While randomization in NAND flash prevents worst-case RBER, it misses low-RBER patterns.

Original data

Randomizer

Evenly distributed

Number of bit errors

800
700
600
500
400
300
200
100
0

1205

Much lower error rate than randomization

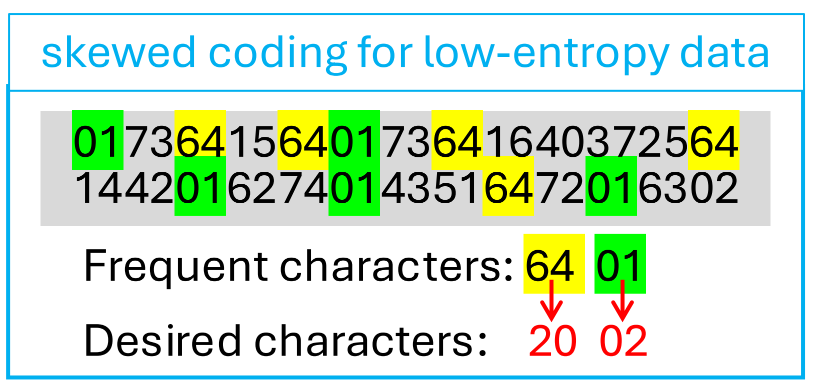
Rand PS0 PS1 PS2 PS3 PS4 PS5 PS6 PS7 PS01 PS12 PS23 PS34 PS45 PS56 PS67 PS07 PS16 PS25 PS03 PS14 PS35 PS46 PS24

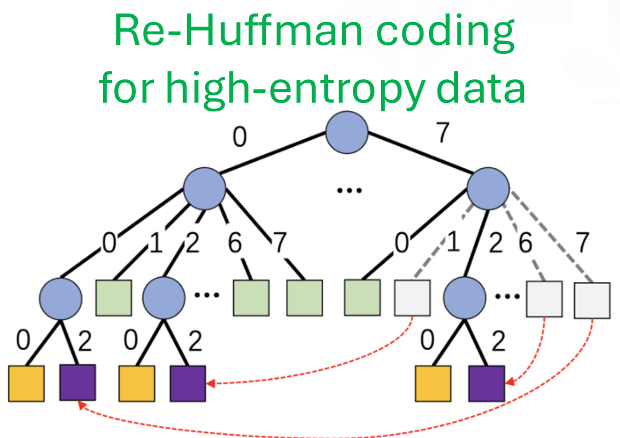
For cold data that won't be frequently rewritten, replace randomization with carefully chosen encoding to push RBER toward the best-case patterns.

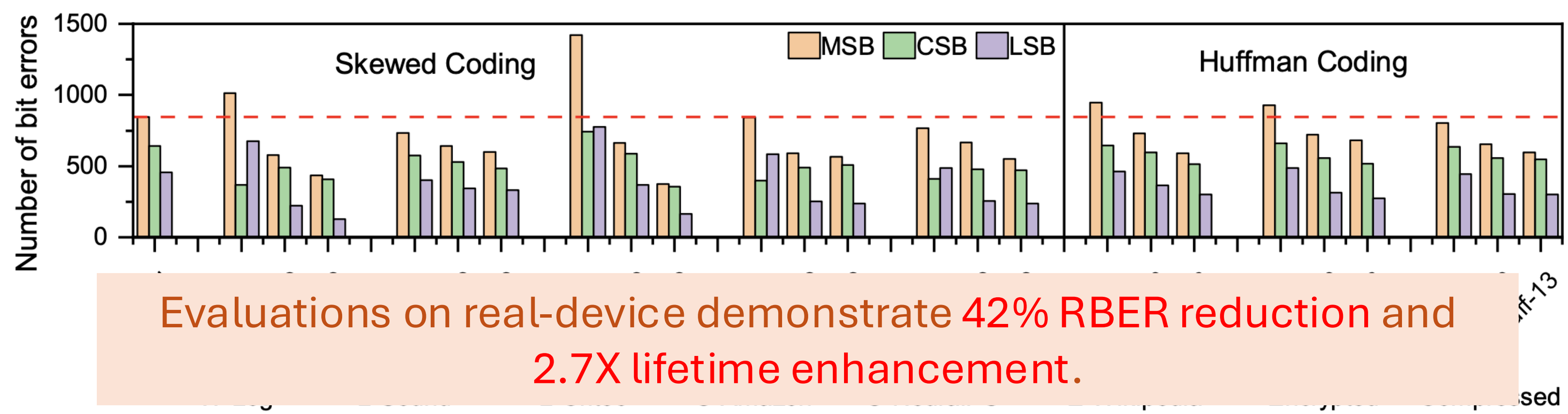
ColdCode: Cold Data Encoding for Enhanced Reliability and Lifetime in 3D NAND Flash, EuroSys 2026

ColdCode: Data Encoding for Cold Storage

We propose **entropy-aware** encoding schemes for cold data: skewed coding and reversed Huffman coding for low- and high-entropy data, respectively.







Evaluations on real-device demonstrate 42% RBER reduction and 2.7X lifetime enhancement.

ColdCode: Cold Data Encoding for Enhanced Reliability and Lifetime in 3D NAND Flash, EuroSys 2026

💡 Part III — Key Insight

Data encoding is a powerful and under-explored lever. From invalid data patterns (Midas Touch) → cold data encoding (ColdCode) — each approach unlocks reliability and lifetime gains.

> *"The bits you write determine how long your flash will last."*

Part IV

The Reprogramming Revolution

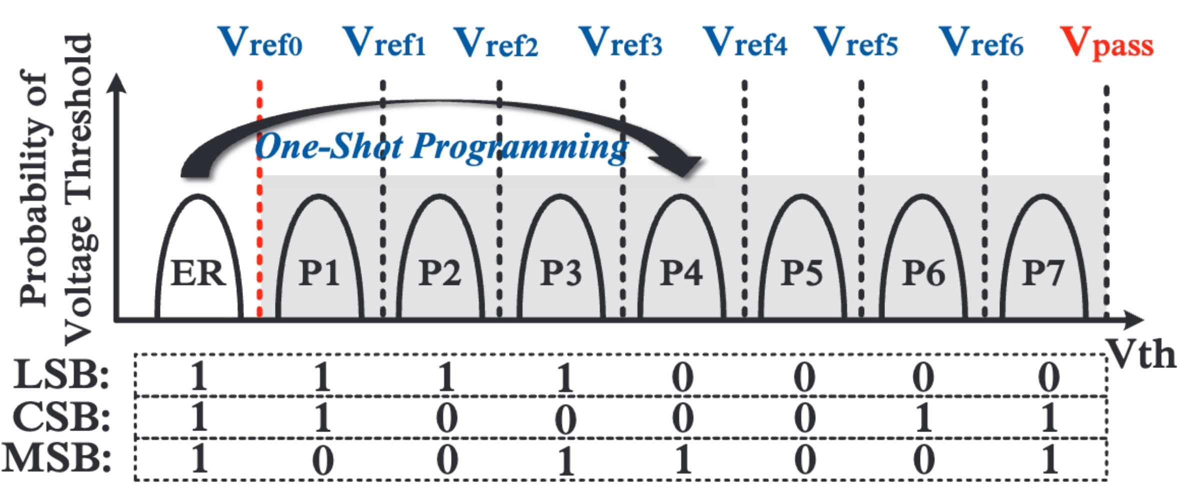
Breaking the sequential programming constraint of NAND flash

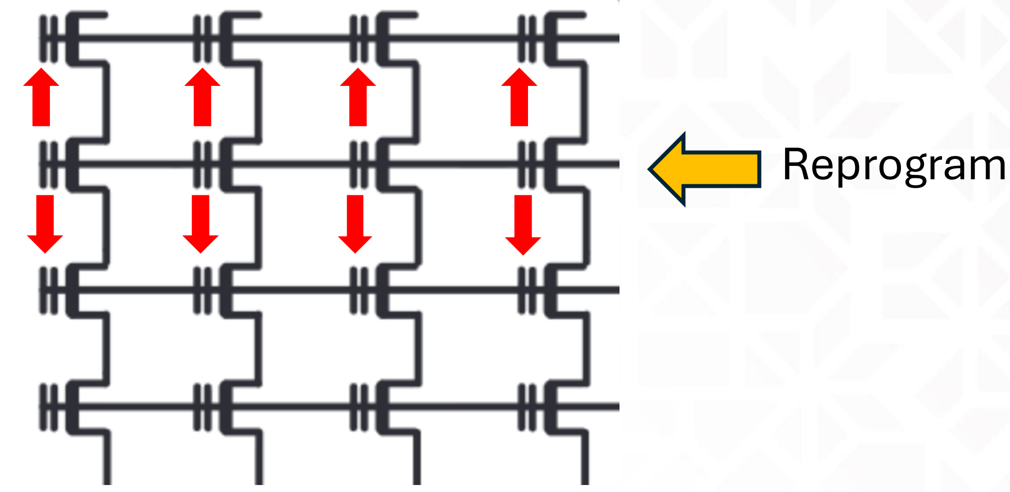
ReSSD: Reprogramming 3D TLC Flash

NAND flash enforces strict sequential page programming.

Once a page is written, it typically cannot be modified.

But what if we could reprogram cells?





Severe interference to neighbors

For update-intensive scenarios, previous data is quickly invalidated in reprogrammable block

Gao, Ye, Li et al., "Constructing Large, Durable and Fast SSD System via Reprogramming 3D TLC Flash Memory," MICRO 2019

ReSSD: Reprogramming 3D TLC Flash

First program: use the first four states

Reprogram: one/two bits become invalid, use 1/2 more voltage states

(a) Reprogram(LSB/LSB) (b) Reprogram(LSB/MSB) (c) Reprogram(MSB/MSB) (d) Reprogram(MSB/LSB)

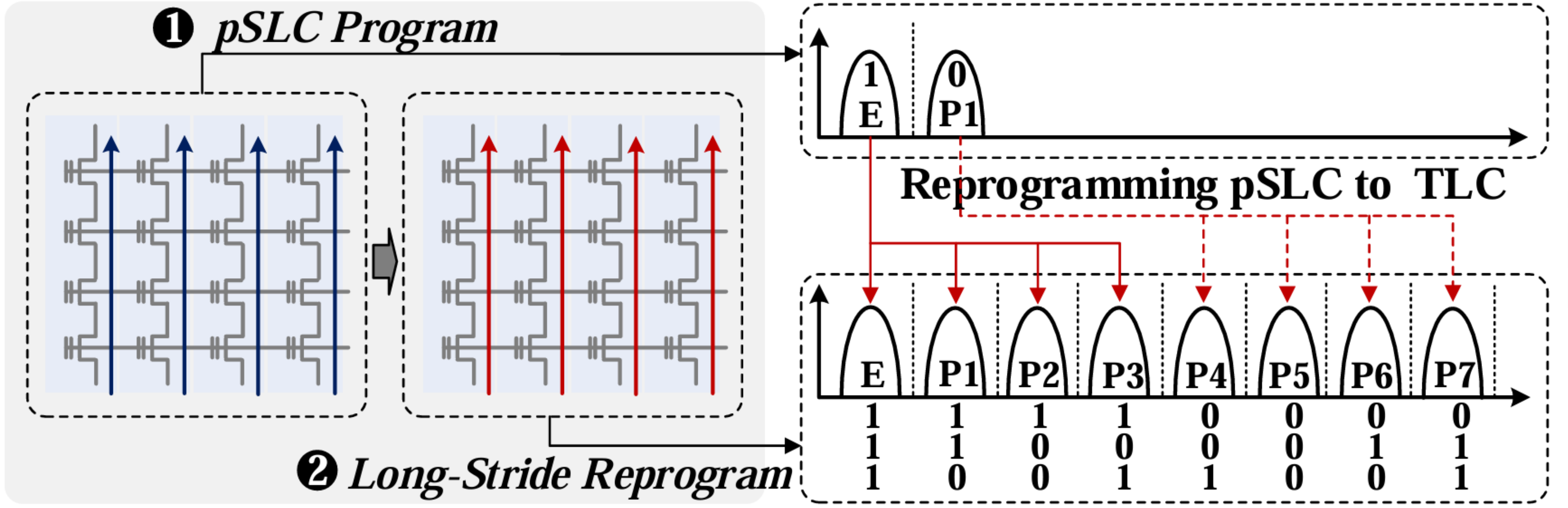
Case study on RAID 5: improve the endurance by 30.3%, boost write performance by 16.7%

Gao, Ye, Li et al., "Constructing Large, Durable and Fast SSD System via Reprogramming 3D TLC Flash Memory," MICRO 2019

LOONG: Breaking the Stride Limitation

Prior reprogram work is limited: only a small subset of wordlines per block can be reprogrammed (small stride). LOONG expands this to entire blocks.

Key Innovation: Spatially decouple the two program steps across the entire block.



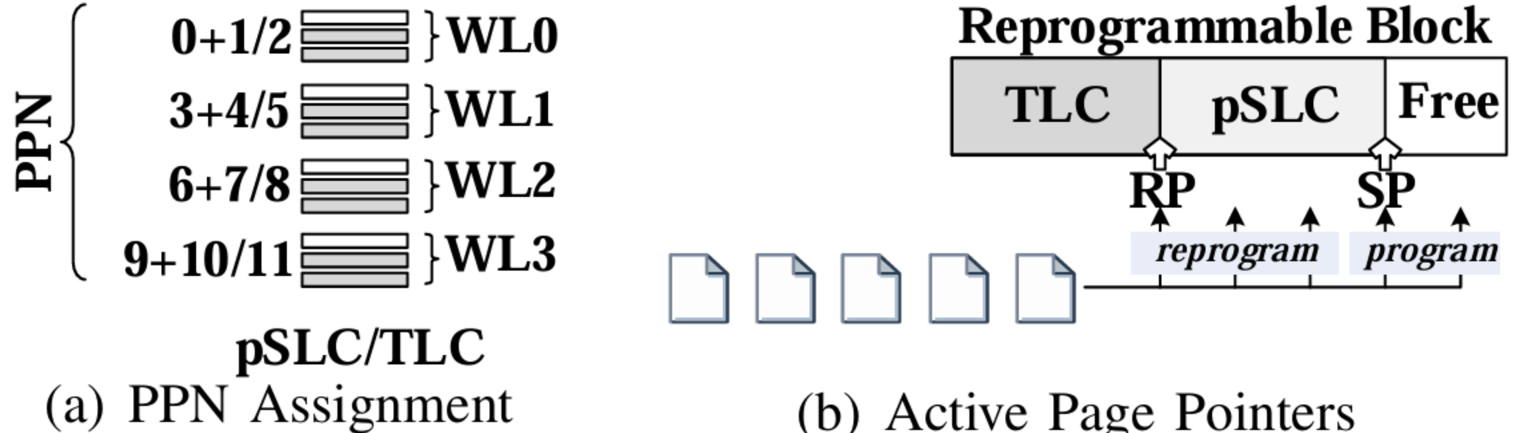
LOONG: Breaking the Stride Limitation

1 **Sequential SLC Program (All Wordlines)**

Program all WLs in a block in SLC mode (1 bit/cell). Fast and maximizing write throughput.

2 **Uniform TLC Reprogram (All Wordlines)**

Reprogram all pages from SLC to TLC mode, preserving full storage capacity (3 bits/cell).



(a) PPN Assignment (b) Active Page Pointers

SLC-speed writes → TLC-density storage. Best of both worlds.

LOONG: Breaking the Stride Limitation

Evaluated on two key SSD use cases

Average Latency Reduction:

GC Acceleration: **37.5%**

Program Opt.: **18.1%**

GC Acceleration

Valid page movement during garbage collection benefits from fast SLC writes + deferred TLC reprogram. Significant reduction in GC-induced latency spikes.

Program Optimization

Normal write operations enjoy SLC-speed initial writes, with TLC conversion batched efficiently during idle periods.

💡 Part IV — Key Insight

Reprogram operations, once considered impractical, can harness Flash for major performance and lifetime gains: from small-stride reprogram → full-block reprogram.

> *"Challenging fundamental constraints often yields the greatest rewards."*

Part V

Future Challenges & Opportunities

What's next for efficient and performant storage systems?

The Cross-Layer Optimization Philosophy

Our decade of work converges on a unified principle: effective flash optimization requires cross-layer co-design.

Application / File System	Coldness tags, access patterns, data characteristics
SSD Controller / FTL	Encoding, scheduling, prediction, LDPC management
NAND Flash Chip	Program modes, error characteristics, reprogram operations
Device Physics	Charge trap behavior, process variation, self-healing effects

↕ Information flows across all layers ↕

Every breakthrough came from understanding one layer deeply and connecting it with another.

Open Challenges & Opportunities

QLC/PLC Scaling

4–5 bits/cell demand exponentially tighter margins. New encoding, error correction, and management strategies are essential for continued scaling.

AI-Driven Storage Management

Machine learning for LDPC prediction, wear leveling, and workload-adaptive management. Moving from heuristics to intelligent, self-optimizing systems.

CXL & Composable Storage

New memory/storage interfaces change the optimization landscape. Co-optimizing flash with CXL-attached memory opens new design spaces.

Sustainability & Green Storage

Extending flash lifetime directly reduces e-waste and energy consumption. Encoding and self-healing contribute to environmentally sustainable storage.

Workload-Specific Flash Optimization

AI training data, genomics, financial data — each has unique access patterns and reliability needs. One-size-fits-all approaches are leaving significant performance and lifetime on the table. The CHEOPS community is uniquely positioned to address this.

Mohamed bin Zayed University of Artificial Intelligence

Key Takeaways

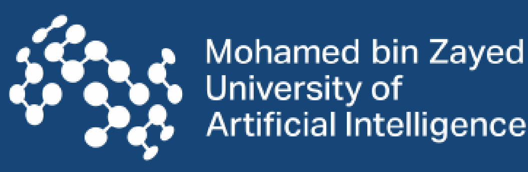
🔬 **Deep physical characterization is the foundation — standard assumptions often mislead.**

⚡ **Read performance can be pushed to near-zero retry through sentinel-guided approaches.**

💾 **Smart data encoding (ColdCode) can extend lifetime by up to 3× with no hardware changes.**

🔄 **Full-block reprogram (LOONG) reduces latency by up to 37.5%, deployable via firmware.**

📐 **Cross-layer co-design is the unifying principle — every optimization bridges layers.**



Thank You!

Questions & Discussion

Chun Jason Xue