嶄新記憶體帶來之計算機系統的硬軟體整合設計

郭大維

中央研究院 資訊科技創新研究中心

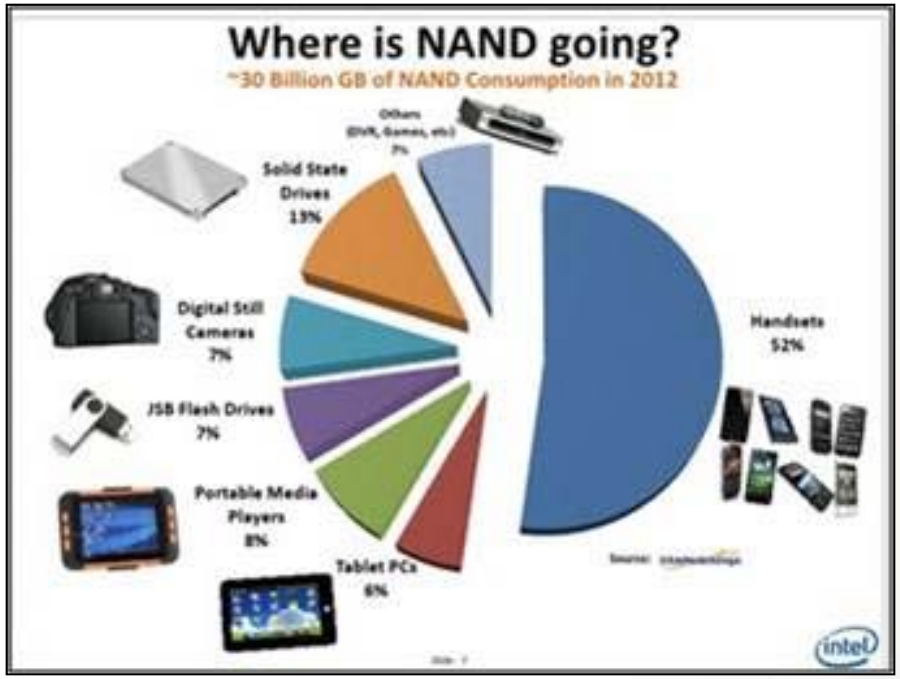
台灣大學 資訊工程學系暨網媒所

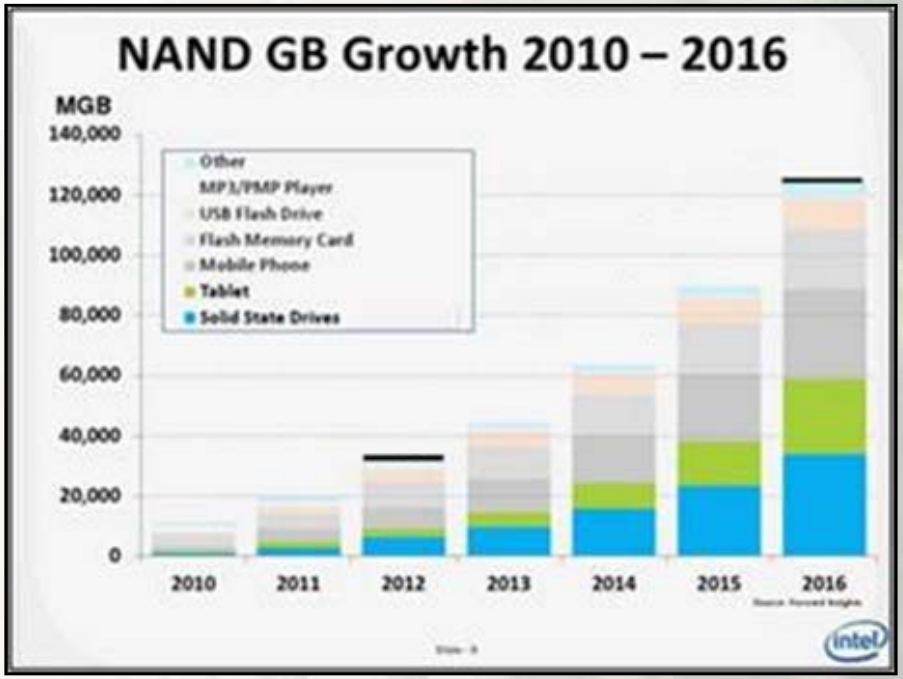
Agenda

- A Global Picture
- More Efficient Chip Programming
- More Reliable Blocks
- Less Data Transfers
- Better System Optimization
- Conclusion

Emerging of Non-Volatile Memory

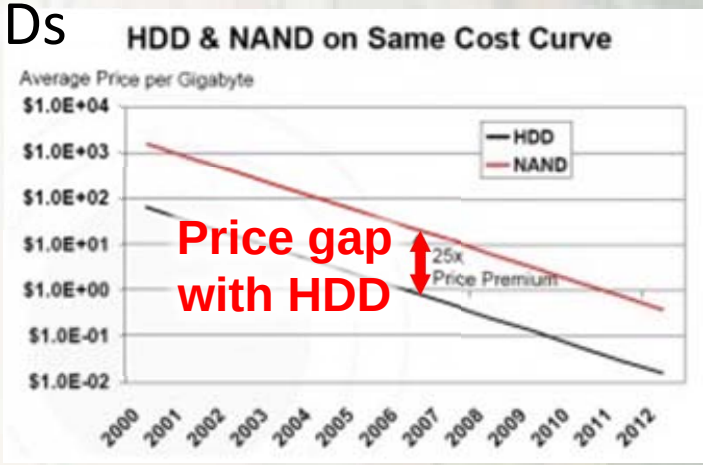
- **Mobile Devices**
  - Smartphones, Tablet PCs, Solid-state Drives, Media Players, etc.
- **Flash Memory**
  - Growth: **4X growth** in the past four years



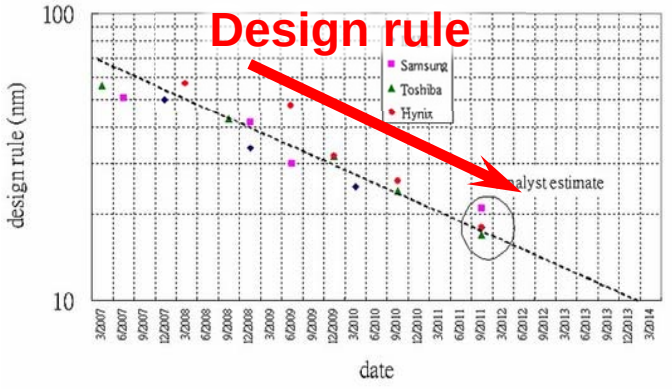


Facts in Technology Trend

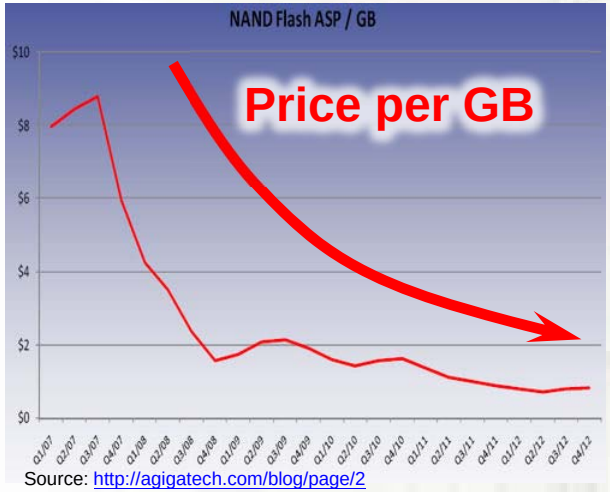
- **Competitiveness in the Flash Price**
  - Dropping Rate and the Price Gap with HDDs
- **Technology Trend over the Market**
  - Improved density
  - Degraded performance
  - Degraded reliability
  - Worsened access constraints



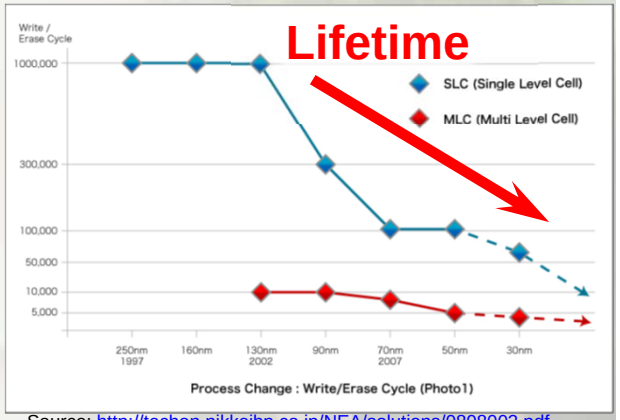
Source: http://agigatech.com/blog/system-uses-for-nand-flash/



Source: http://www.storagenewsletter.com/images/public/sites/StorageNewsletter.com/articles/icono7/intel_and_micron_20nm_f3_540.jpg



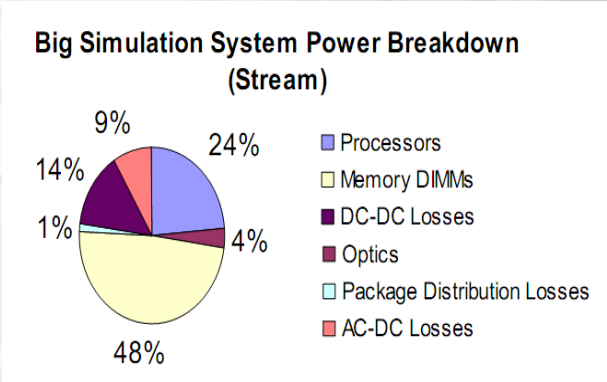
Source: http://agigatech.com/blog/page/2



Source: http://techon.nikkeibp.co.jp/NEA/solutions/0808002.pdf

Why Non-Volatile Memory

- Shock Resistance and Cost
- Performance
- Scalability
  - 53% Execution Reduction by Replacing 8GB DRAM with 32GB PCM [ISCA09]
- Energy Efficiency
  - 65% Energy Saving for Intensive Write Access over Desktops and Web Servers When 4GB DRAM was Replaced with 4GB PCM [ISCA09]



Source: Rajamani *et al.*, ISLPED'08
Estimated power breakdown for two Petaflop supercomputer HPC node designs, each configuration tailored to application class

Why Non-Volatile Memory Again

- Convenience in Data Manipulation
  - Doubling in the Capacity Every 18 Months and Efficiency in Random Access over the Huge Addressing Space.
- Reliability Enhancement
  - Compared to the MTTF of HDD, Such as a Typical Figure 1.5M Hours, the MTTF of a Solid-State Drive Can Be Over 2M Hours.
  - Much Better Flexibility in Realizing Information Redundancy.

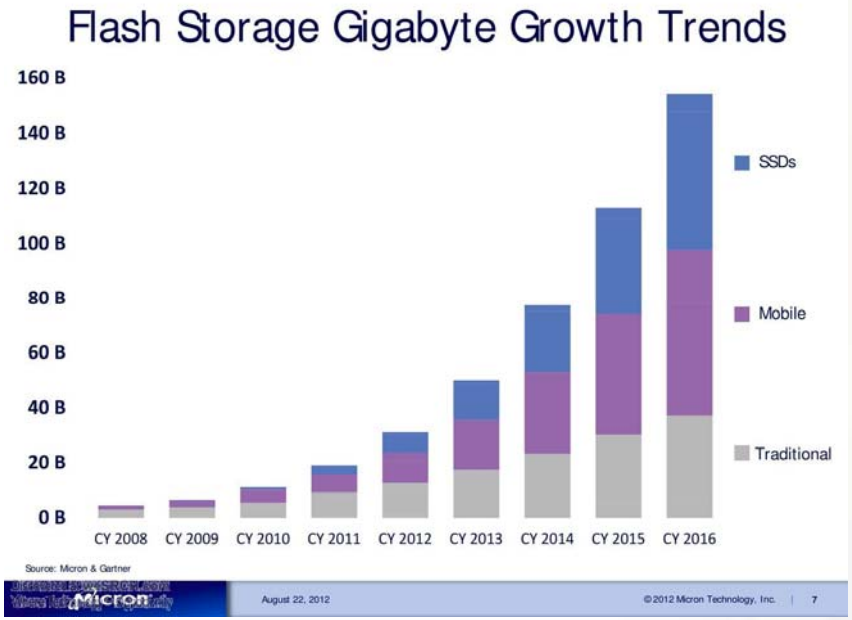
Source: http://www.storagereview.com/ssd_vs_hdd

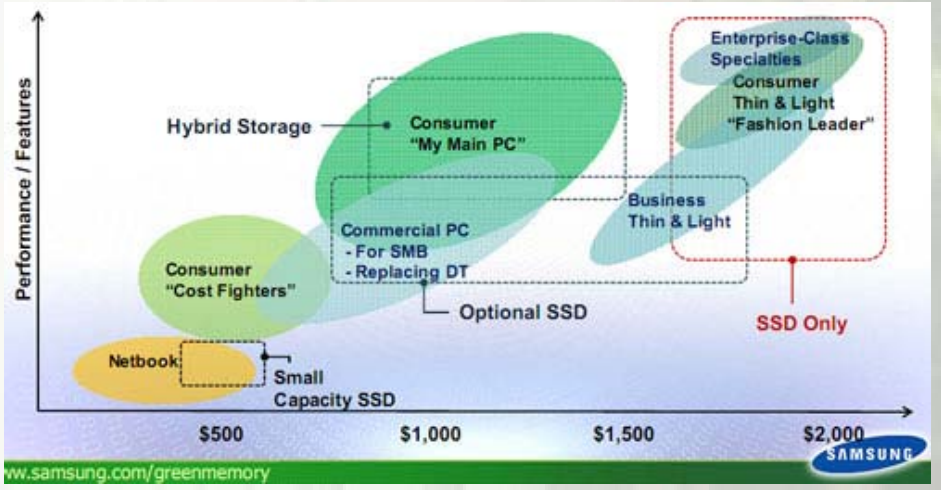
Huge Variety and Potentials

- A Variety of Products
  - Consumer Electronics
  - Servers and Storage Systems
  - Industrial Applications
- Revolutions vs New Challenges
  - New Dimensions in Optimization in Cost, Performance, and Reliability

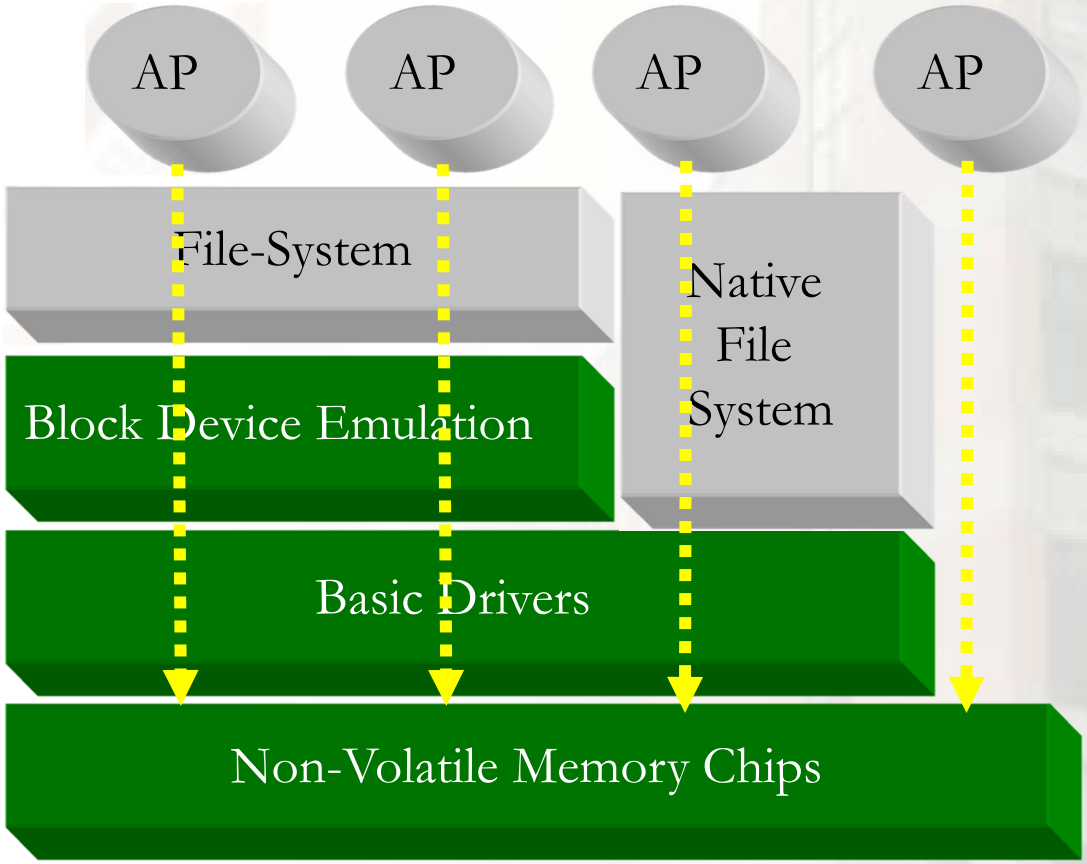
New Business Opportunities and Challenges

- Flourishing in Solid-State-Disk Developments
  - **Top 25 Vendors in 2015Q2**: Diablo Technologies, SanDisk, OCZ, Hitachi Global Storage Technologies (HGST: a WD Company), Seagate, Virtium, Microsemi, Tegile Systems, Violin Memory, Pure Storage, Cactus Technologies, Kaminario, Micron, Samsung, InnoDisk, Intel, IBM, Nimbus Data, Foremay, EMC, BiTMICRO, Marvell, RunCore, Maxta, Primary Data

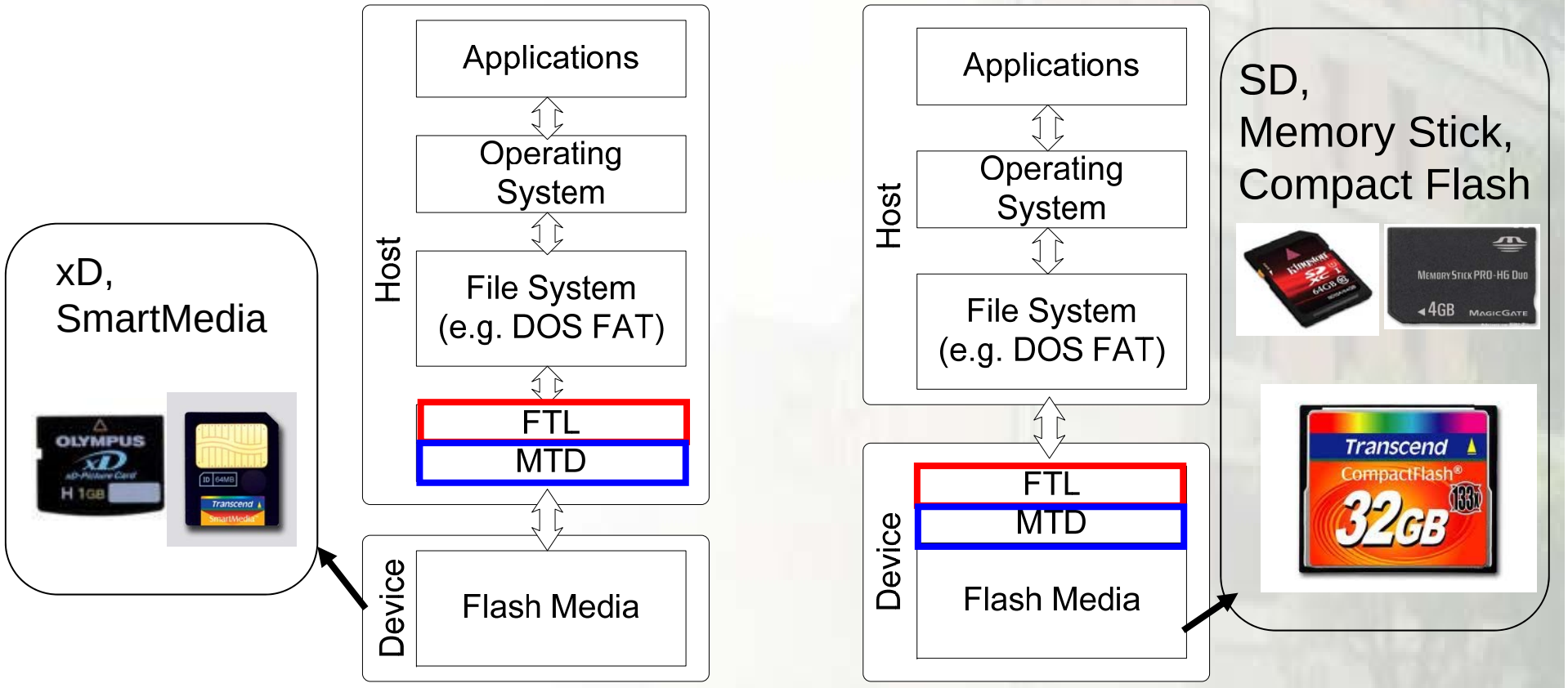




Storage-System View Point

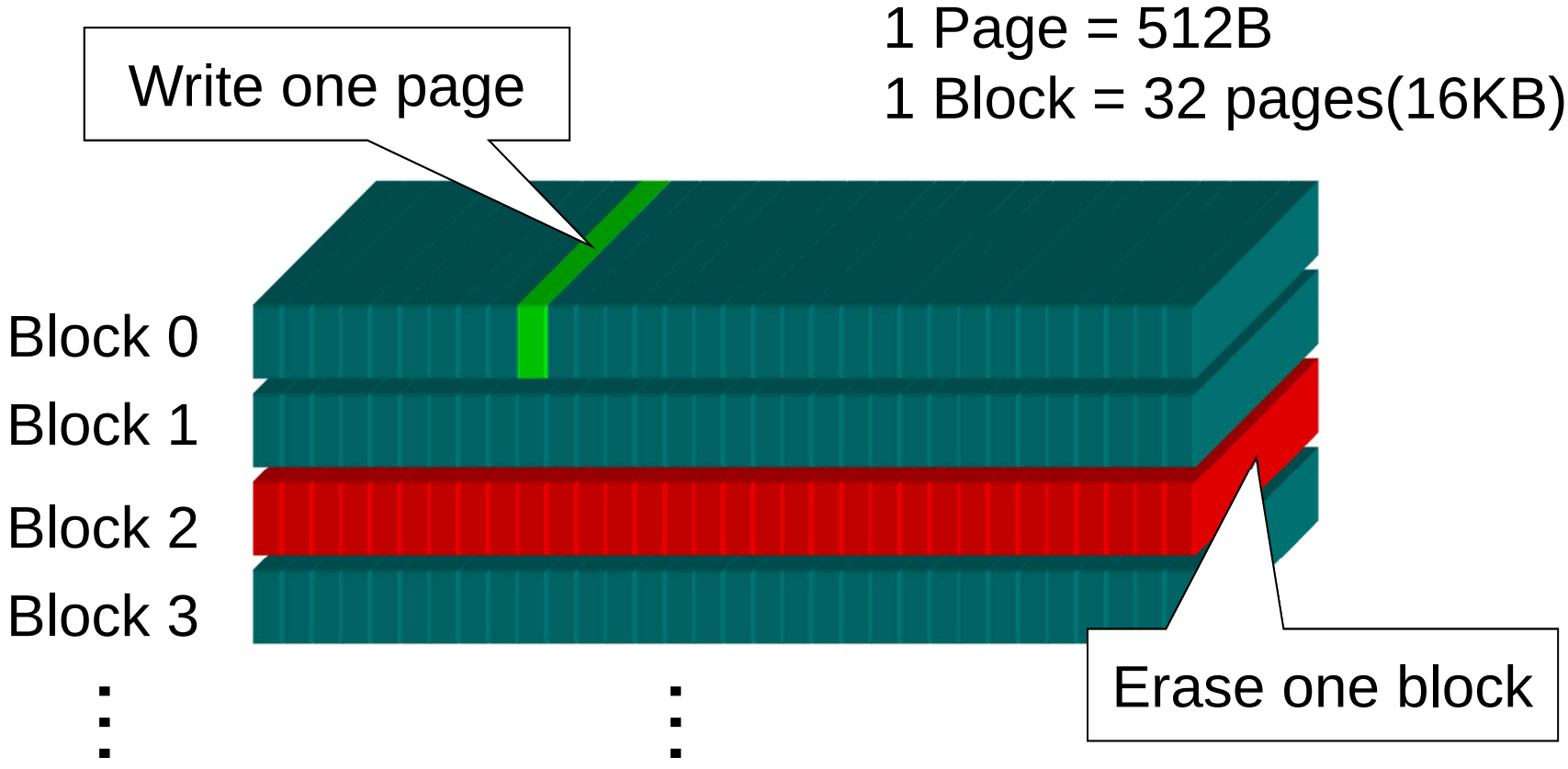


Layered System Architecture



*FTL: Flash Translation Layer, MTD: Memory Technology Device

Flash-Memory Characteristics



Flash-Memory Characteristics

- **SLC Flash Access Constraints**
  - **Write-Once**
    - No writing on the same page unless its residing block is erased!
    - Pages are classified as valid, invalid, and free pages.
  - **Bulk-Erasing**
    - Pages are erased in a block unit to recycle used but invalid pages.
  - **Wear-Leveling**
    - Each block has a limited lifetime in erasing counts.
- **Additional MLC Flash Access Constraints**
  - Prohibition of partial page programming
  - Serial page programming in a block
  - 3D access constraints

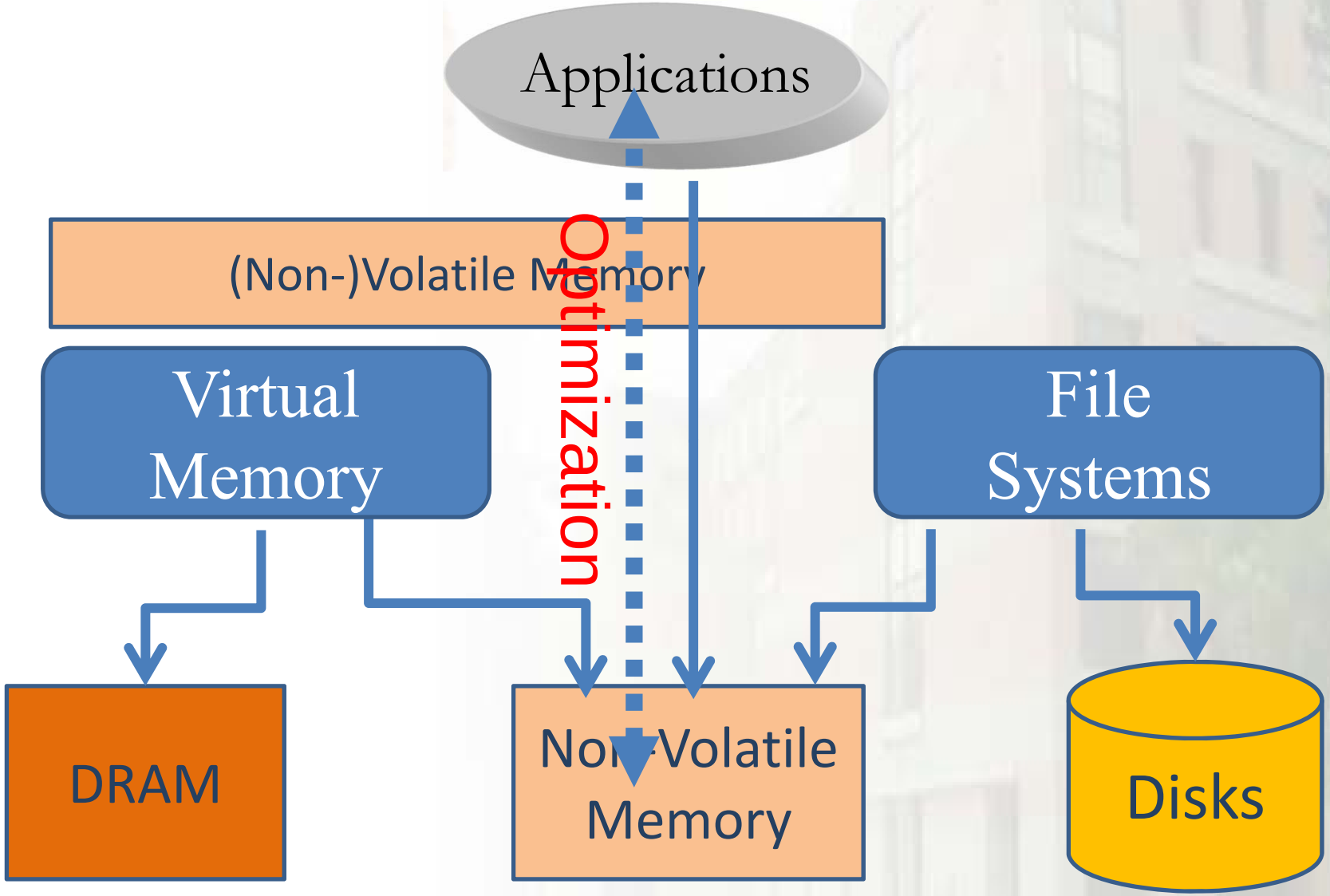
A Bird View of Different Memory Media

Metrics \ Media	DRAM	PCM	NAND Flash	HDD
Read Energy	0.8J/GB	1J/GB	1.5J/GB	65J/GB
Write Energy	1.2J/GB	6J/GB	17.5J/GB	65J/GB
Idle Power	~ 100mW/GB	~ 1mW/GB	1-10mW/GB	~ 10W/TB
Endurance	∞	$10^6 – 10^8$	$10^4 – 10^5$	∞
Page Size	64 B	64 B	4 KB	512B
Read Latency	20-50 ns ~ 1-3us (4KB)	~ 3us (4KB)	~ 25us (4KB)	~ 40ms (4KB)
Write Latency	20-50ns (64B) ~ 1-3us (4KB)	~ 1us (64B) ~ 64us (4KB)	~500us (4KB)	~ 5ms (512B) ~ 40ms (4KB)
Erase Latency	N/A	N/A	~ 2ms	N/A

X ns : Y ns : Z us : W ms

Reference: "Rethinking Database Algorithms for Phase Change Memory", CIDR 2011

A Global Picture Again

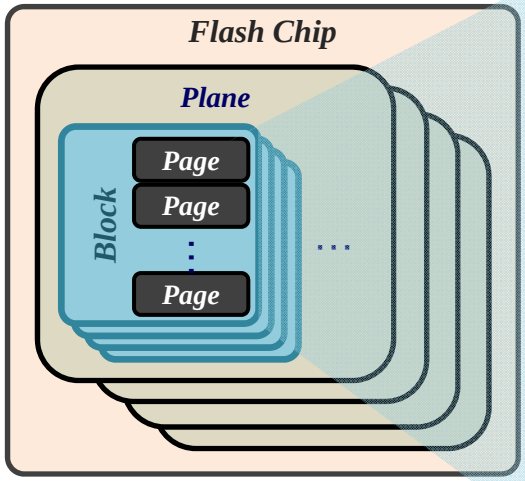


Agenda

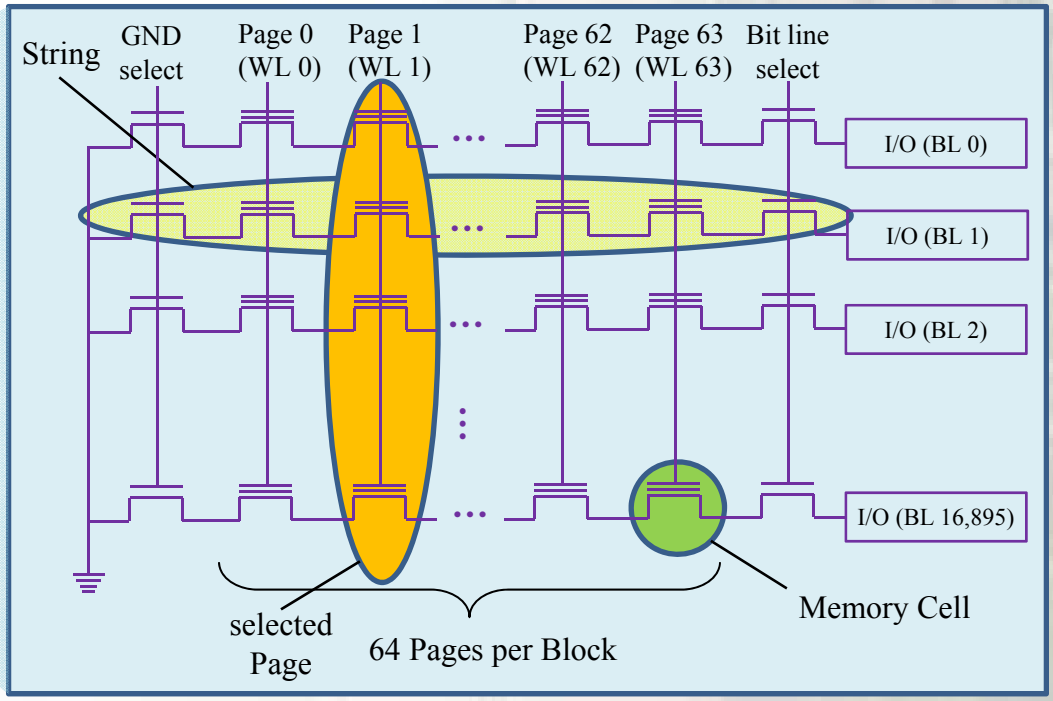
- A Global Picture
- More Efficient Chip Programming
- More Reliable Blocks
- Less Data Transfers
- Better System Optimization
- Conclusion

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.

NAND Flash Memory



- ***Page***: Read/Write Unit
- ***Block***: Erase Unit

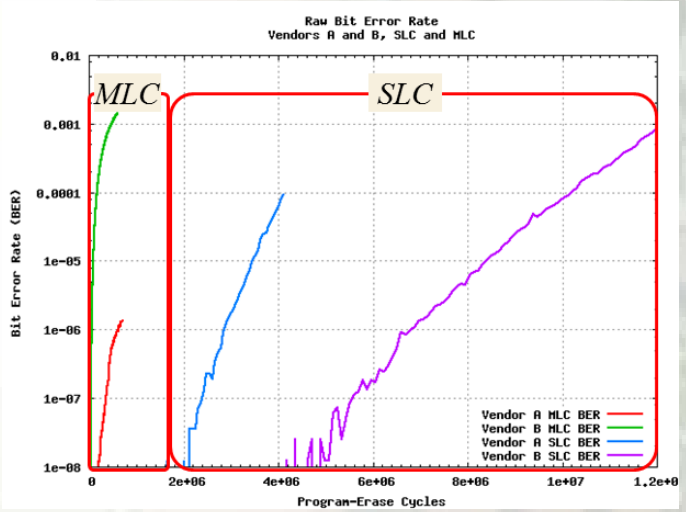


- ***WL***: Word Line
- ***BL***: Bit Line

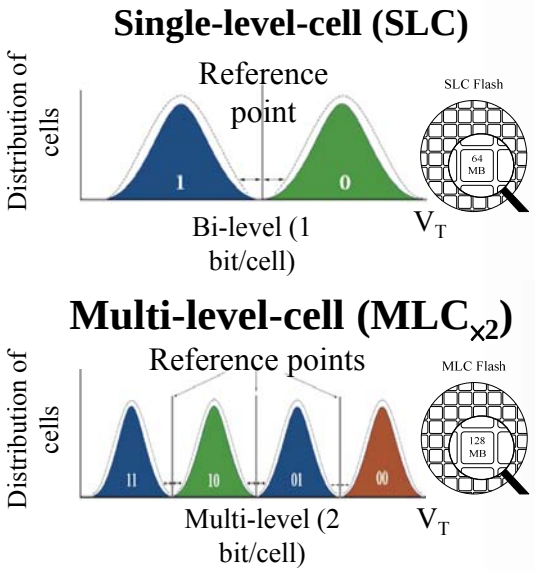
- Flash memory is
  - ***Non-volatility***
  - ***Shock-resistance***
  - ***Power-efficiency***
- Constraints
  - ***Write-once property***
  - ***Bulk-erasure***

Motivation

- **Multi-Level Cell (MLC)** technology
  - MLC_{xn}: N bits stored in one memory cell
  - ***higher capacity*** and ***lower price***
  - ***lower performance*** and ***lower endurance***

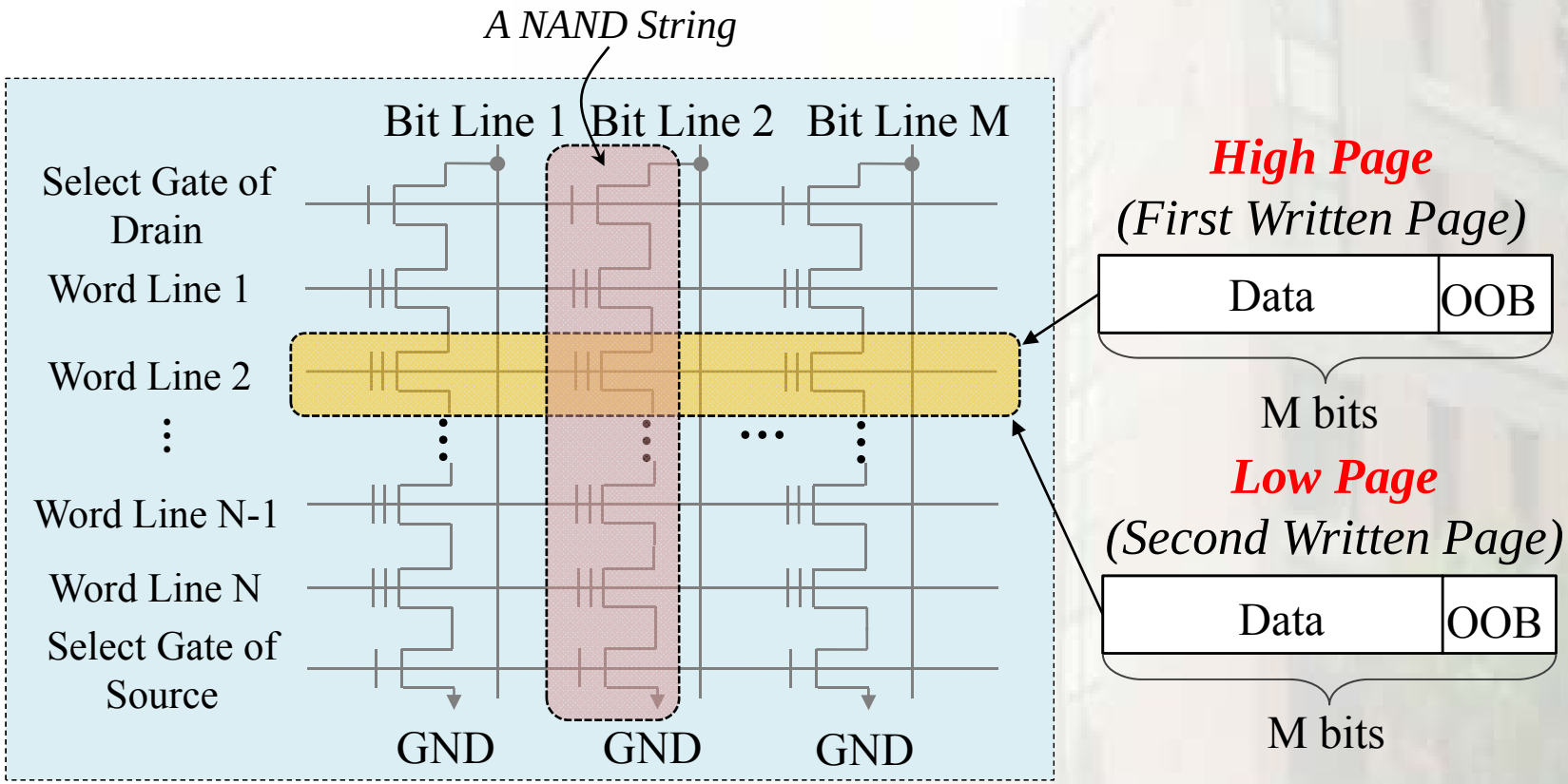


Source: http://liobaashlyritchie.blogspot.tw/



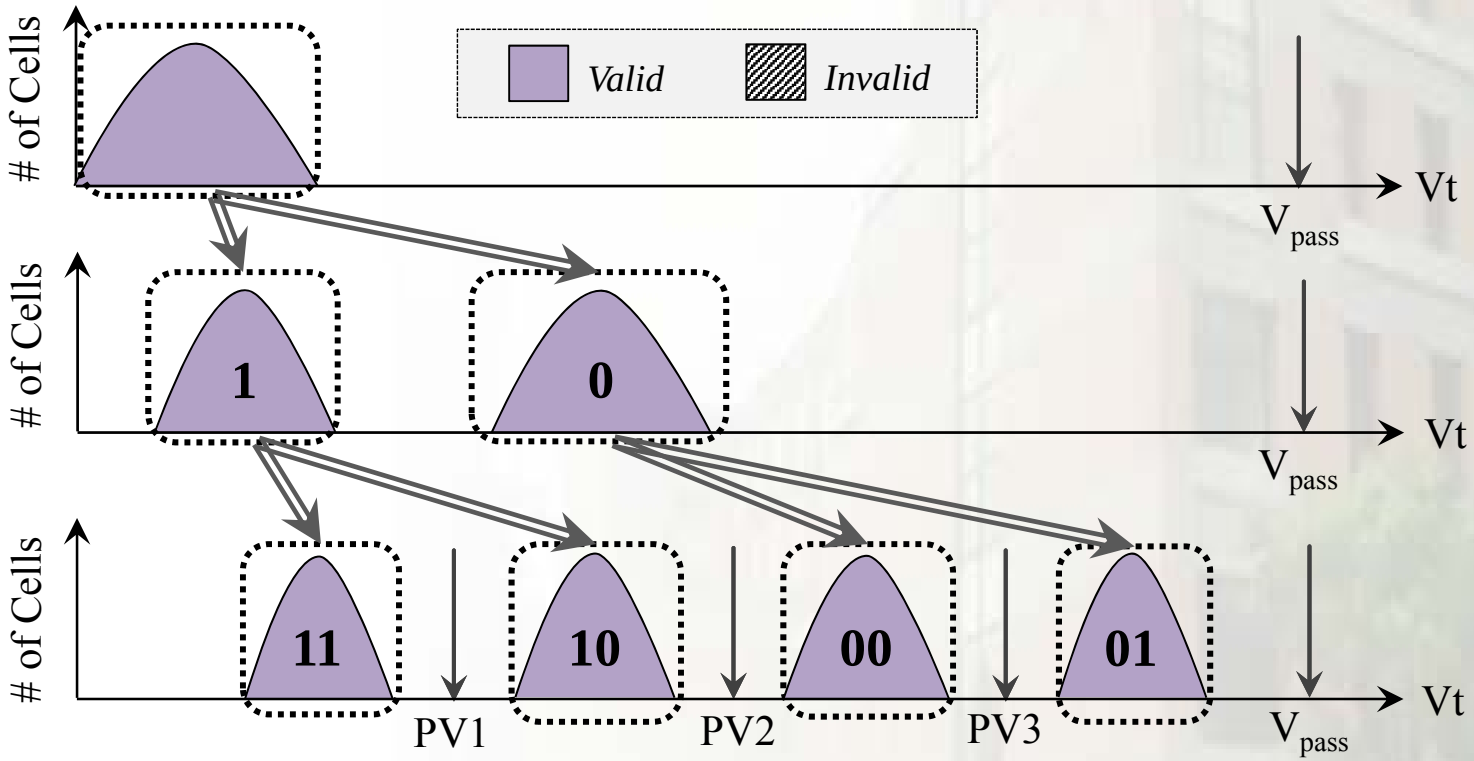
	SLC	MLC_{x2}
Page size	4KB	4KB
Pages per block	64	256
Random read	25µs	50µs
Write time (per page)	200µs	H: 400us L: ***1500us***
Erase time (per block)	1.5ms	3ms
ECC required	1bits/512B	***12bits***/512B
Endurance (W/E cycles)	100,000	3,000~5,000

MLC Flash Block Organization



- Each page consists of M-bit data
- A block is composed of 2×N pages
- The size of a block is **2×M×N** *bits*

MLC Programming

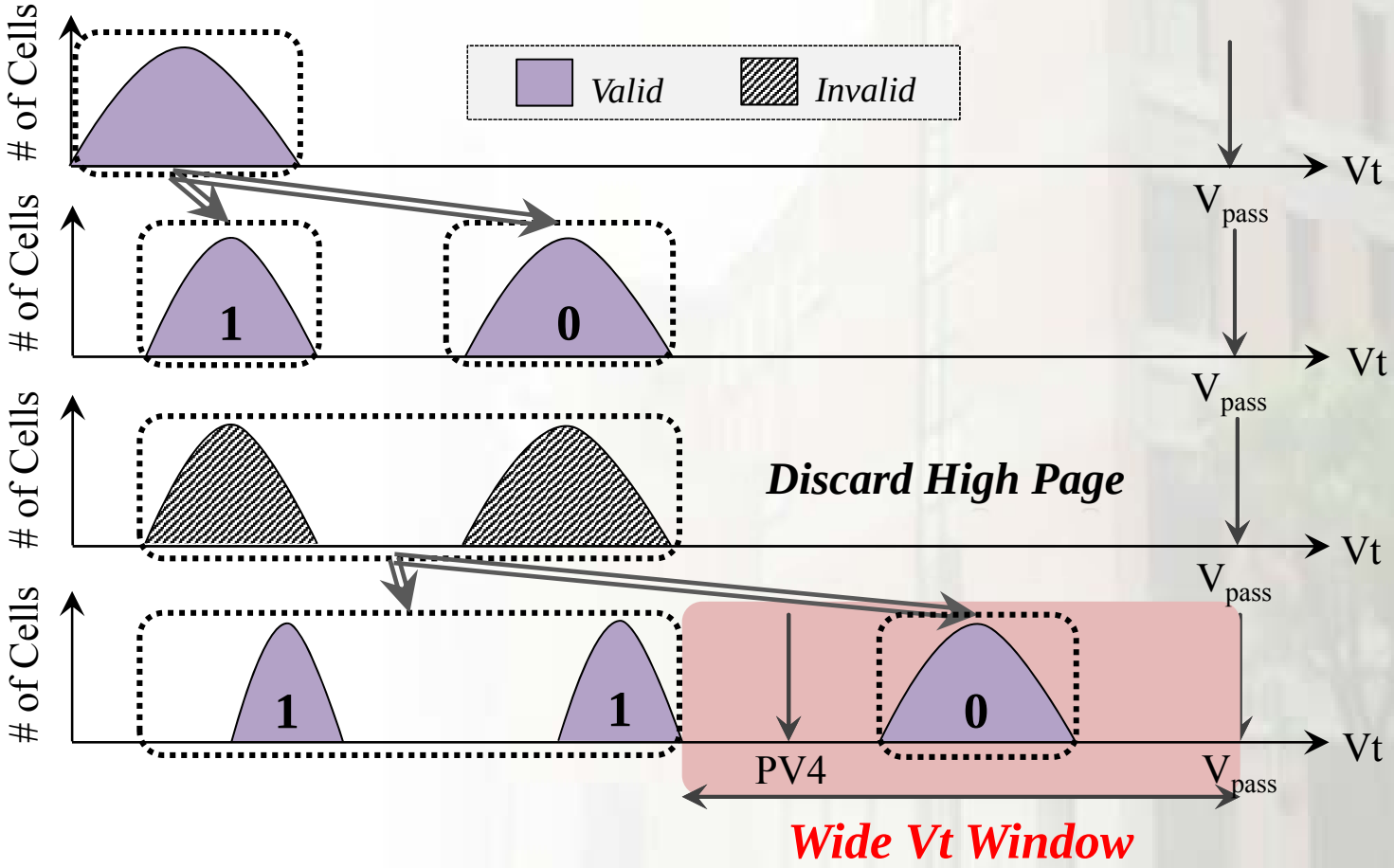


PV: Program Verify Voltage
V_{pass}: A Pass-gate Voltage
Vt: Threshold Voltage

1. ***Low efficiency in programming low pages***
2. ***High bit error rate in low pages***

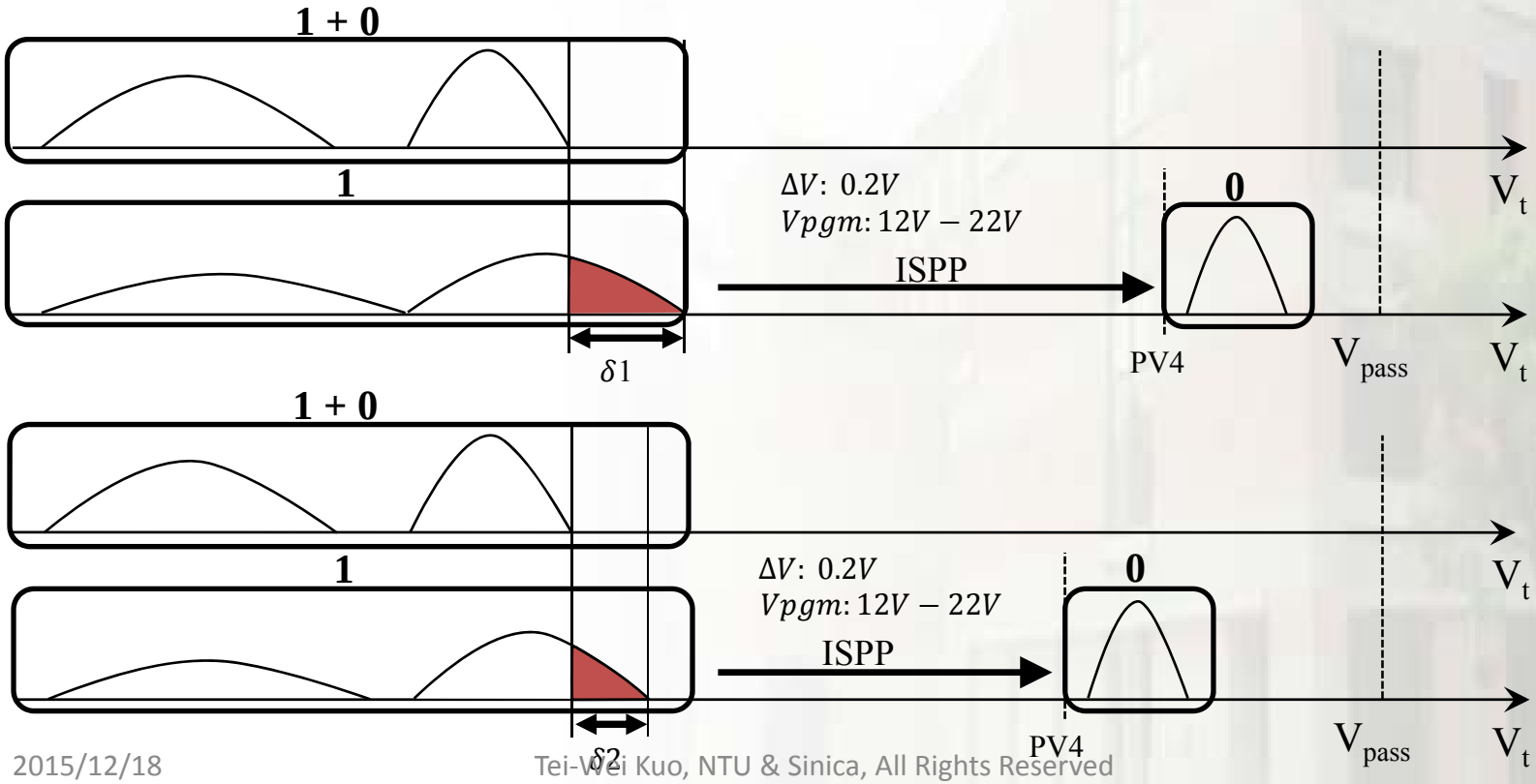
An SLC-like Programming Method

*Strategically exploits the system knowledge, **data validity**, to provide an considerably large threshold voltage window similar to SLC memory via discarding some pages with out-of-date data*



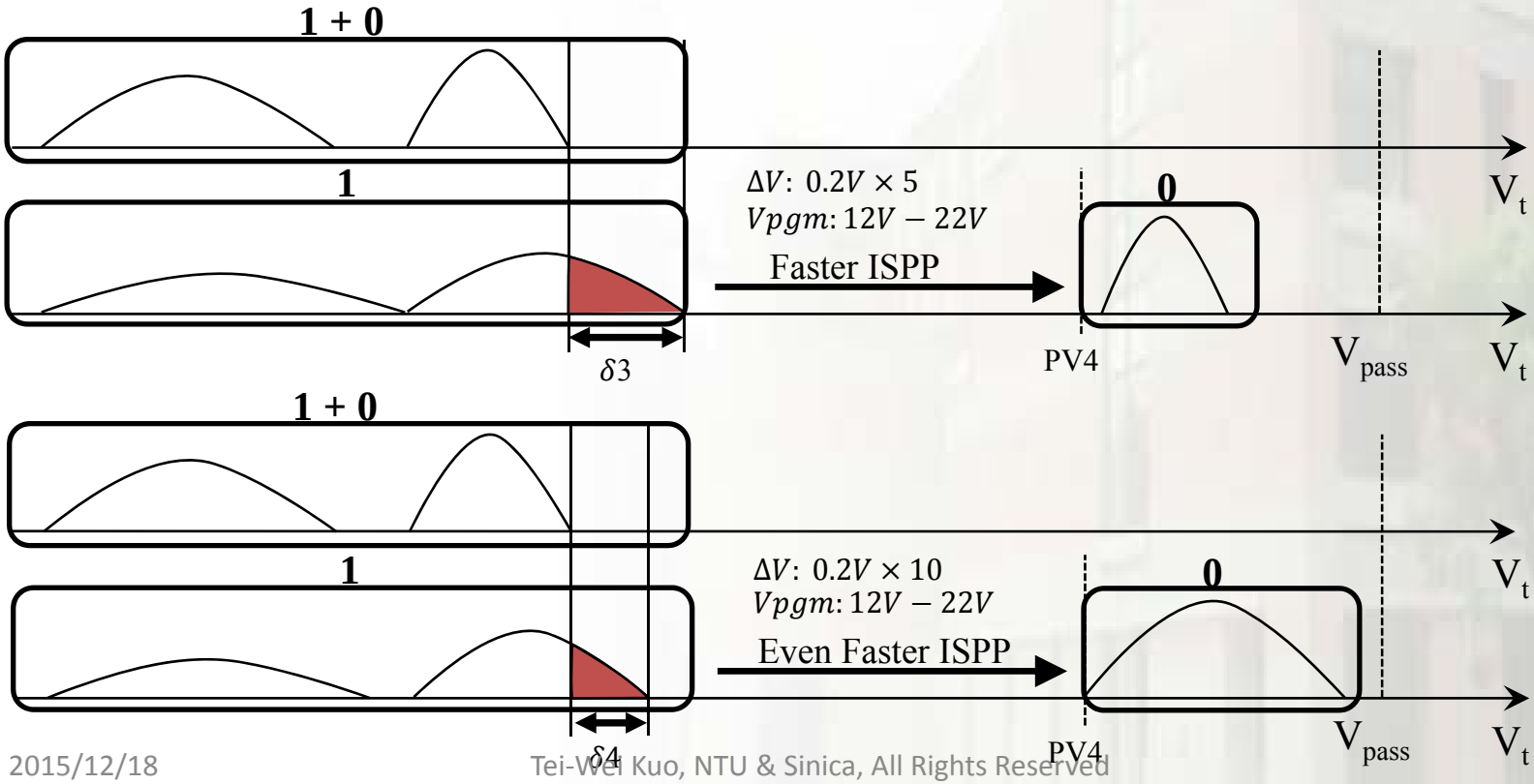
Design Detail with Large Vt Window - The Selection of PV4

- Large PV4 -> Increased Program Shots -> Large tail bit disturbance
- Small PV4 -> Reduced Program Shots -> Small tail bit disturbance

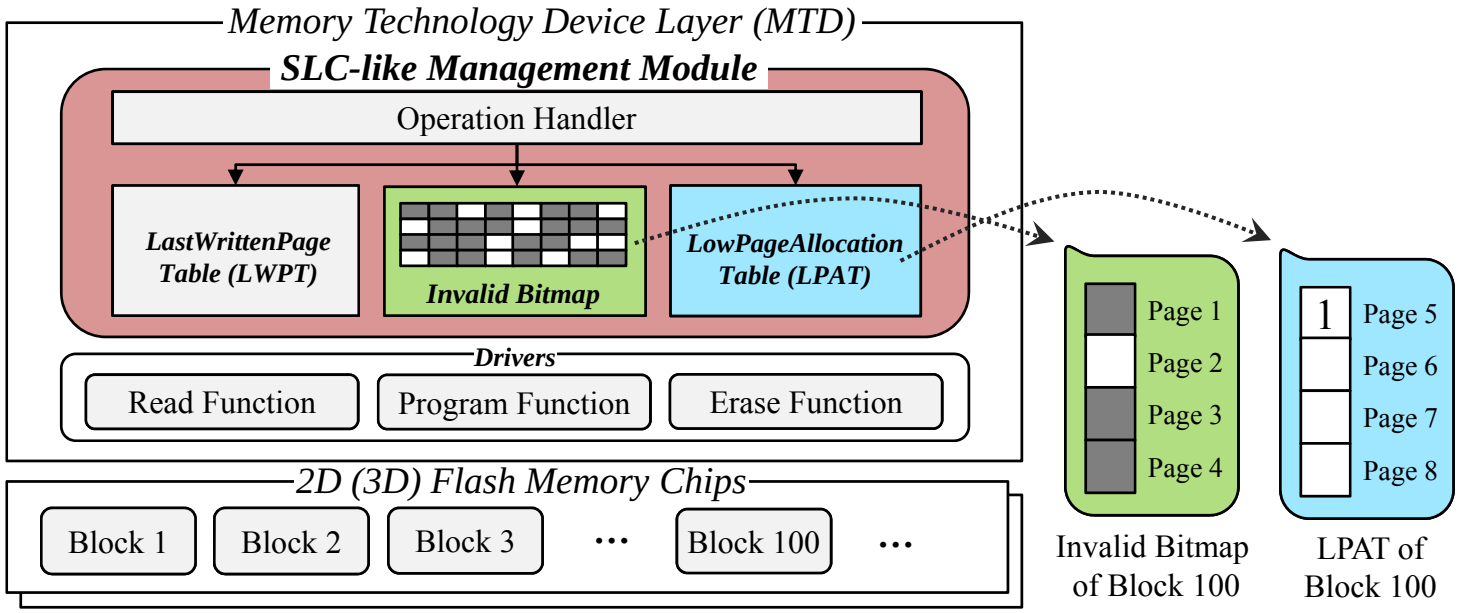


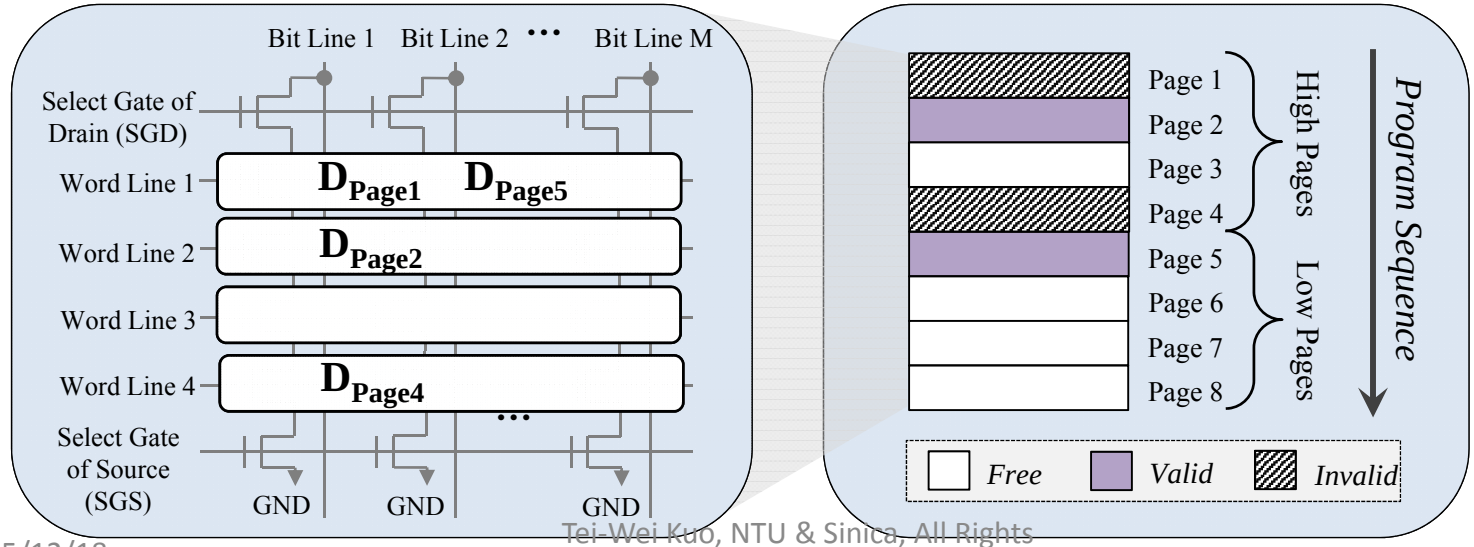
Design Detail with Large Vt Window - The Selection of ΔV

- Small ΔV -> Increased Program Shots -> Large tail bit disturbance
- Large ΔV -> Reduced Program Shots -> Small tail bit disturbance



System Co-design With SLC-like Programming Method

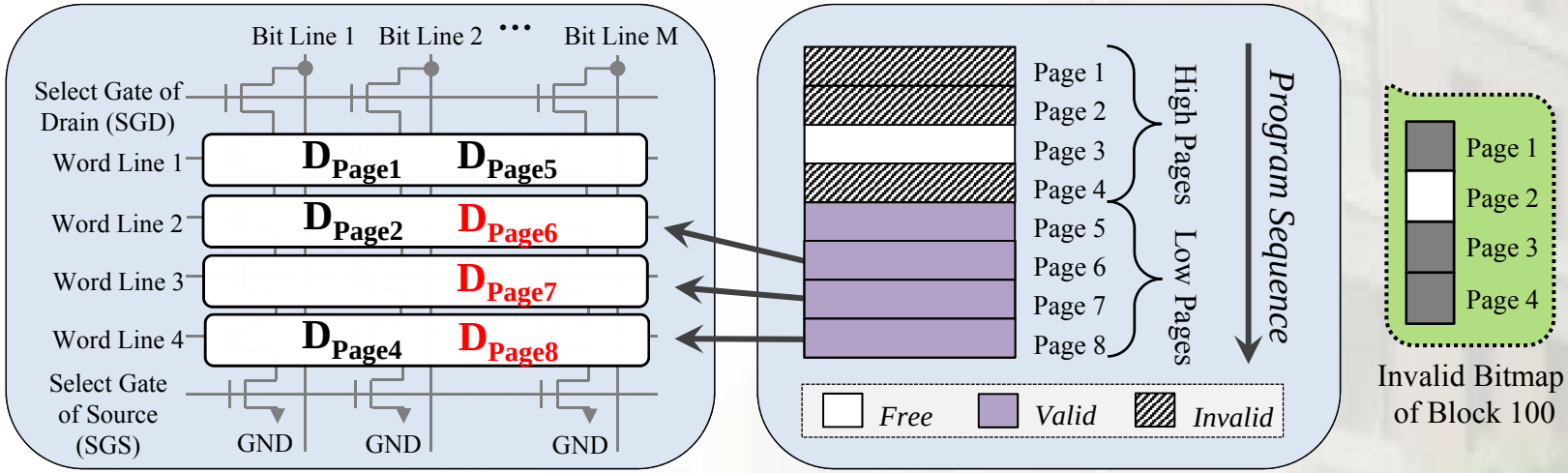




Physical View of Block 100

Logical View of Block 100

An Example of Static SLC-like Management Design

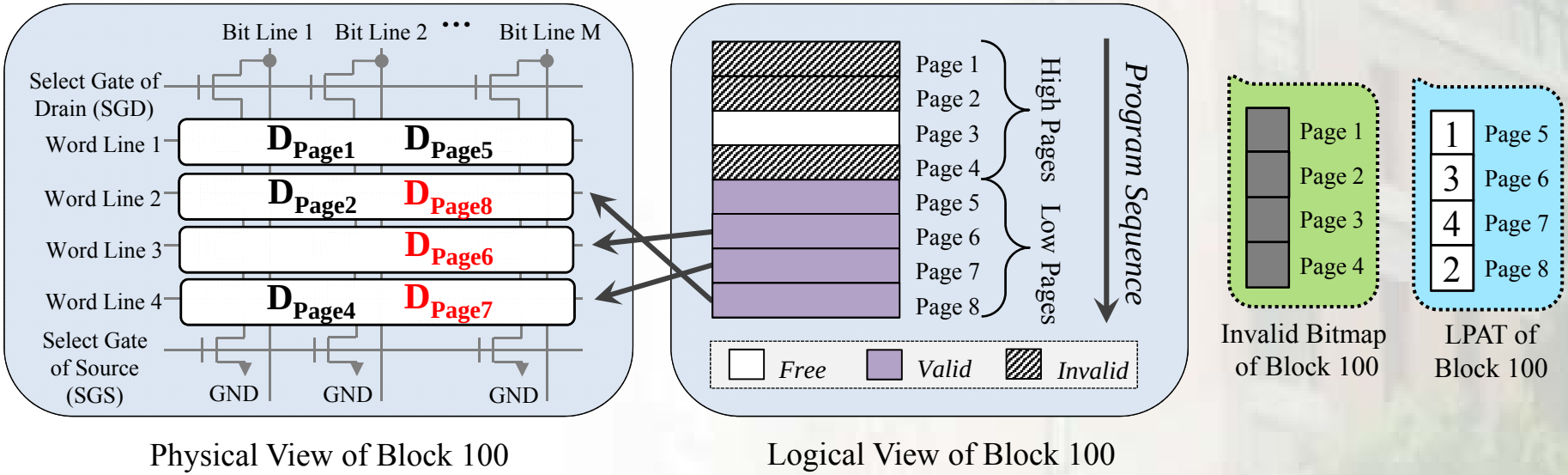


Physical View of Block 100 Logical View of Block 100

- ***Write page 6***
  - *The high page of word line 2 currently contains **valid** data*
  - *Program page 6 data with typical MLC programming method*
- ***Write page 7***
  - *The high page of word line 3 contains **no** data*
  - *Program page 7 data with SLC-like programming method*
- ***Write page 8***
  - *The high page of word line 4 contains invalid data*
  - *Program page 8 data with SLC-like programming method*

1 times of MLC operations
2 times of SLC-like operations

An Example of Dynamic SLC-like Management Design



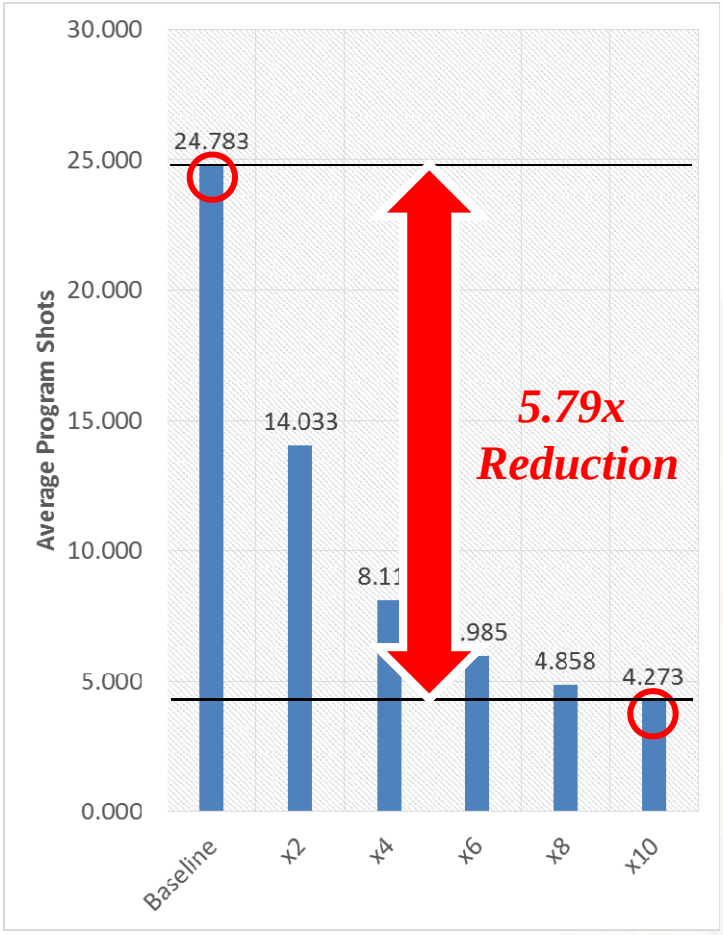
Physical View of Block 100 Logical View of Block 100

- ***Write page 6***
  - *The high page of word line 2 currently contains valid data*
  - *The high page of word line 3 is a free page (should be considered as an invalid page)*
- ***Write page 7***
  - *The high page of word line 2 still contains valid data*
  - *The high page of word line 4 is invalid (best candidate)*
- ***Write page 8***
  - *The high page of word line 2 becomes invalid*
  - *Program low page of word line 2 could be achieved with SLC-like programming method*

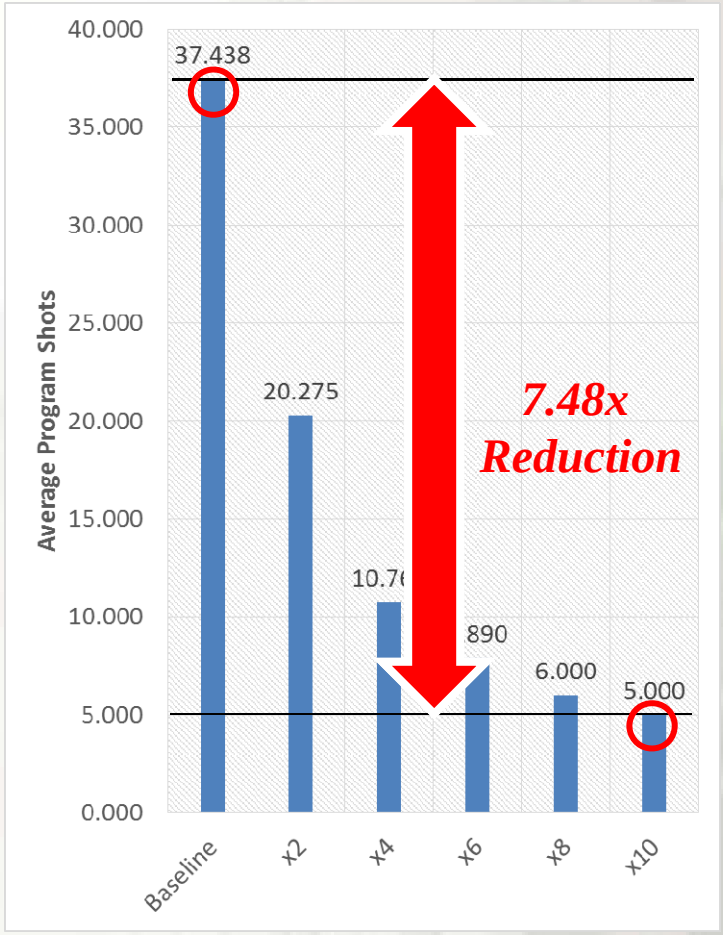
0 times of MLC operations
3 times of SLC-like operations

Experimental Results
- Program Efficiency Improvement

(with Static Programming Order)



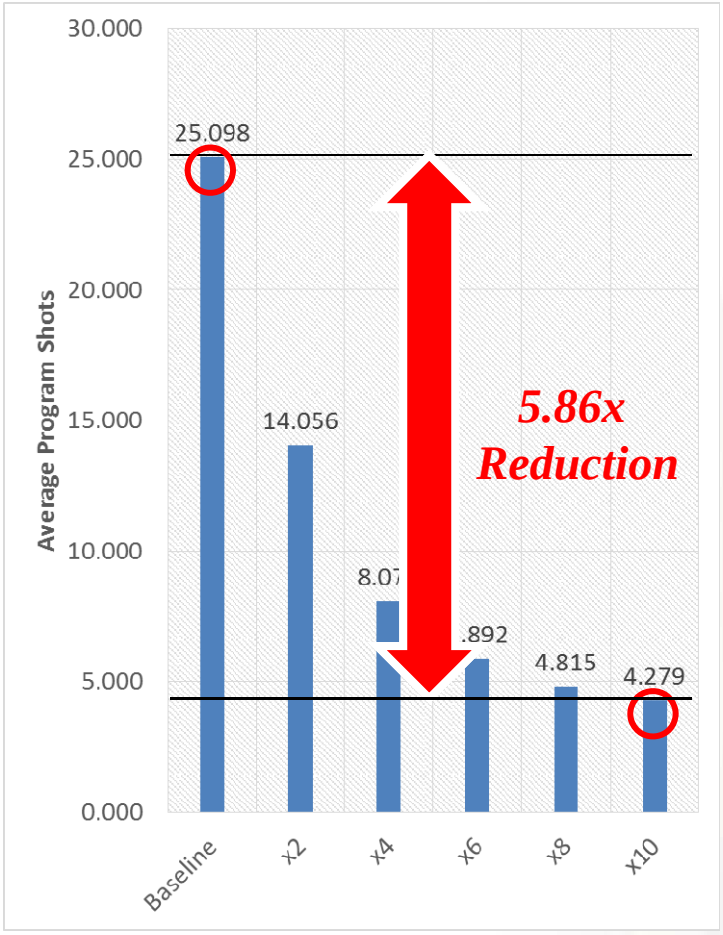
PV4: 1V



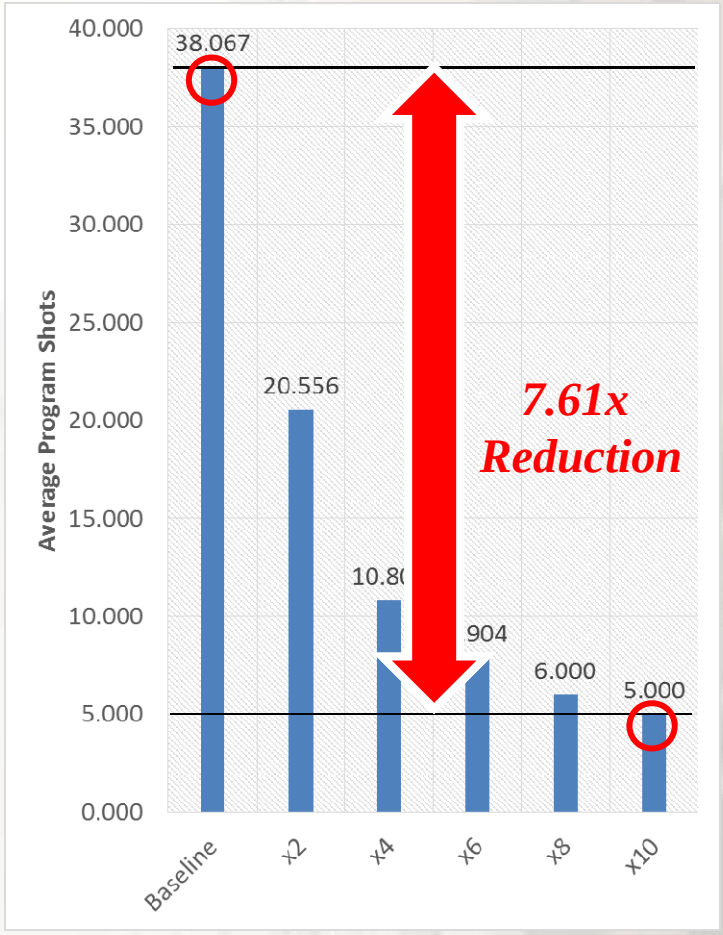
PV4: 3V

Experimental Results
- Program Efficiency Improvement

(with Dynamic Programming Order)



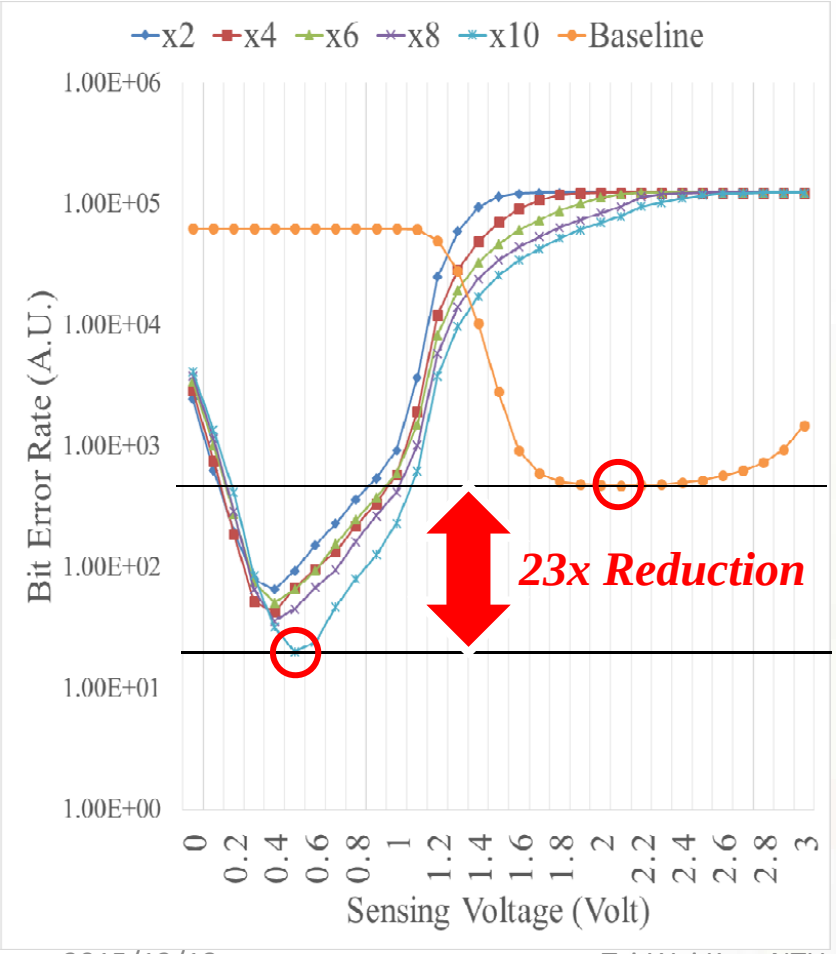
PV4: 1V



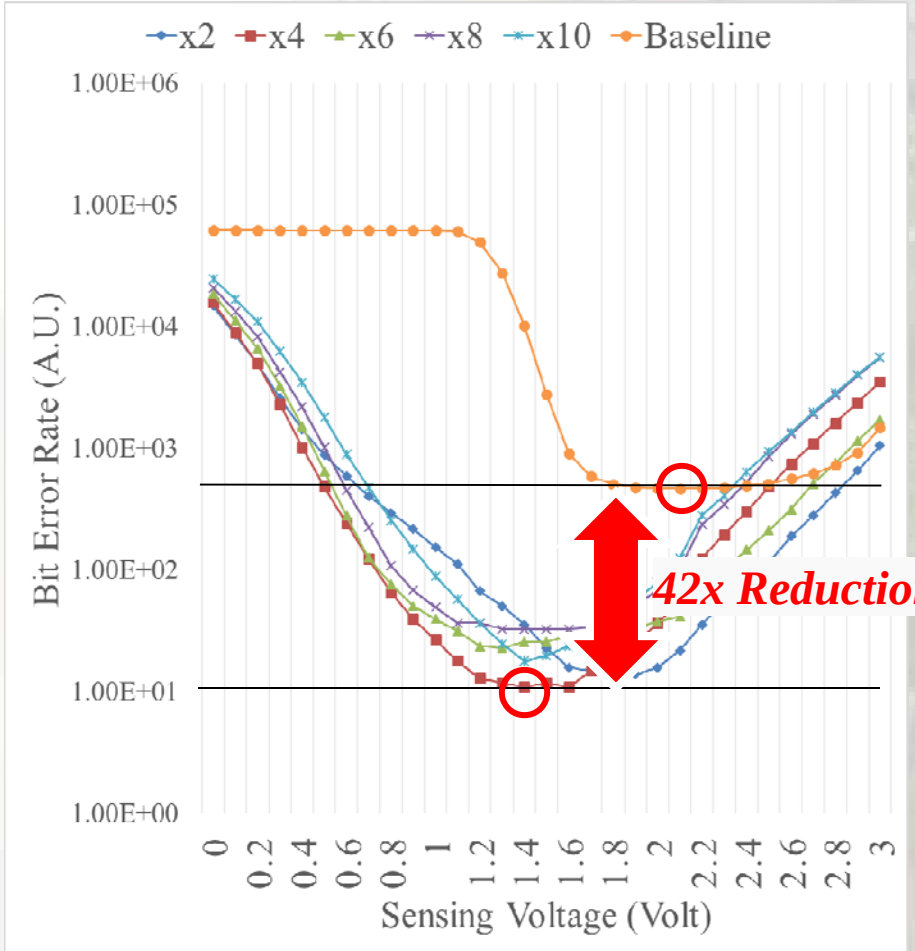
PV4: 3V

Experimental Results
- Bit Error Rate Improvement

(with Static Programming Order)



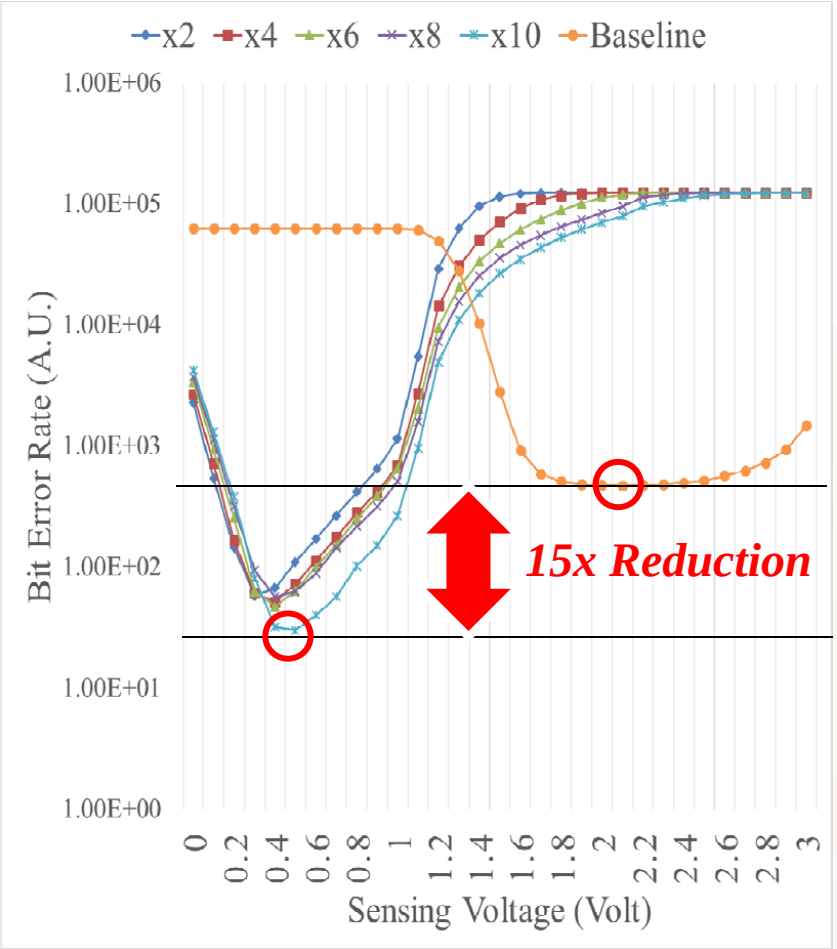
PV4: 1V



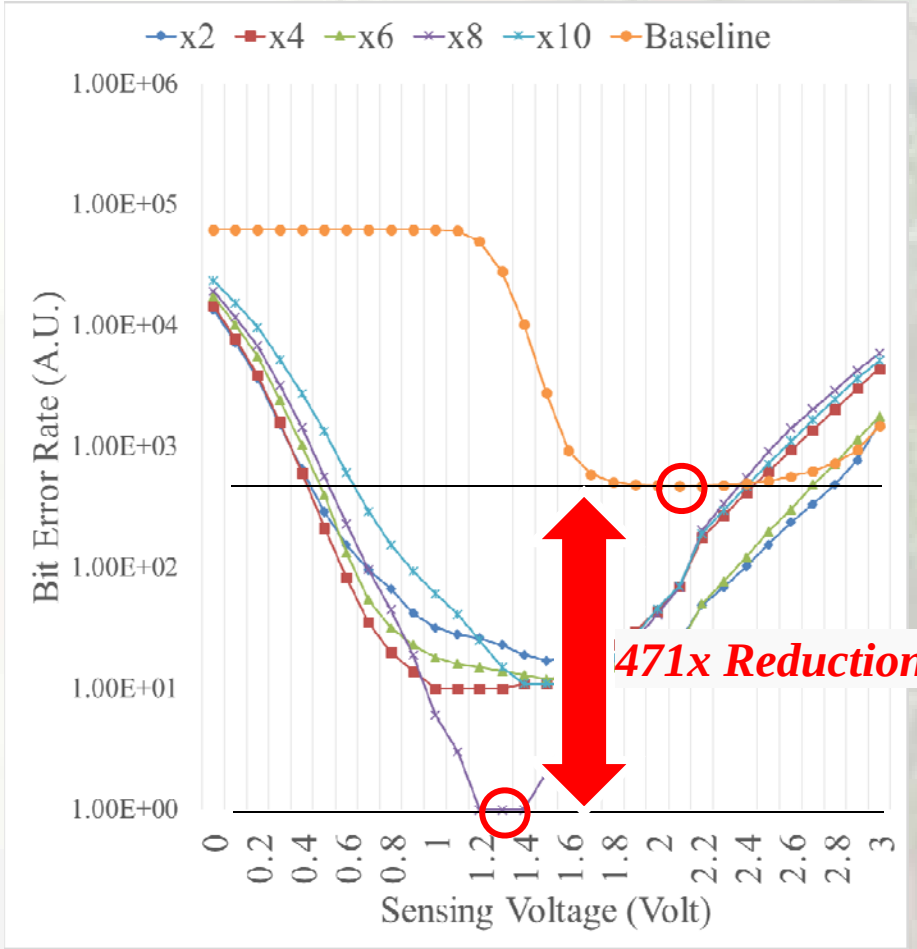
PV4: 3V

Experimental Results
- Bit Error Rate Improvement

(with Dynamic Programming Order)



PV4: 1V



PV4: 3V

Summary

- An ***SLC-like programming method*** is presented to greatly enhance the reliability, as well as improve the program efficiency of MLC flash memories via exploiting the system knowledge, data validity
  - Decrease BER up to ***471%***
  - Decrease average program shots up to ***761%***
- ***A modular design***, implemented at MTD layer without (or with limited) modification to existing FTL designs, could ***bring the proposed programming method into full play***
  - Increase average write bandwidth up to ***263%***

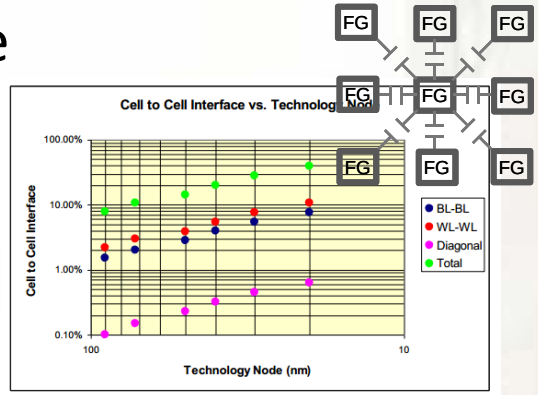
Agenda

- A Global Picture
- More Efficient Chip Programming
- More Reliable Blocks
- Less Data Transfers
- Better System Optimization
- Conclusion

Yu-Ming Chang, Yuan-Hao Chang, Tei-Wei Kuo, Hsiang-Pang Li and Yung Chun Li, 2013, "A Disturb-Alleviation Scheme for 3D Flash Memory," ACM/IEEE 2013 International Conference on Computer-Aided Design (ICCAD), San Jose, USA, November 18-21, 2013.

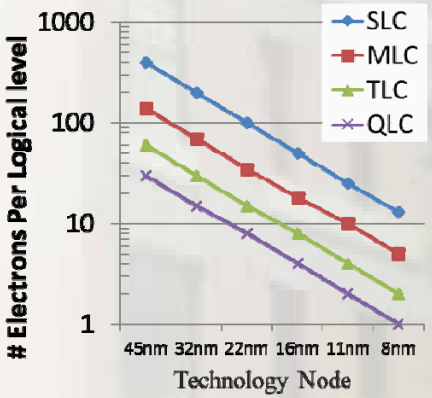
Tremendous Challenges to 3D High-Capacity Flash Memory

- Increasing Disturbance
  - Coupling Effects
  - Less Electrons (for each logical state)



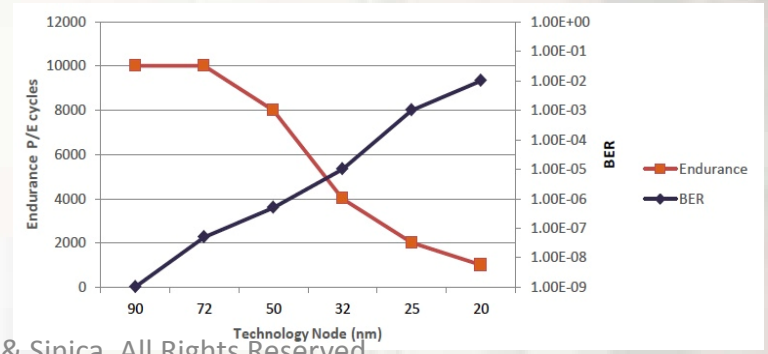
Source: http://www.reddoggamers.net/

*Increasing floating gate **interference** with shorter distance among memory cells*



*More susceptible to any read/program **disturbance** (Physics limitation: hard to breakthrough)*

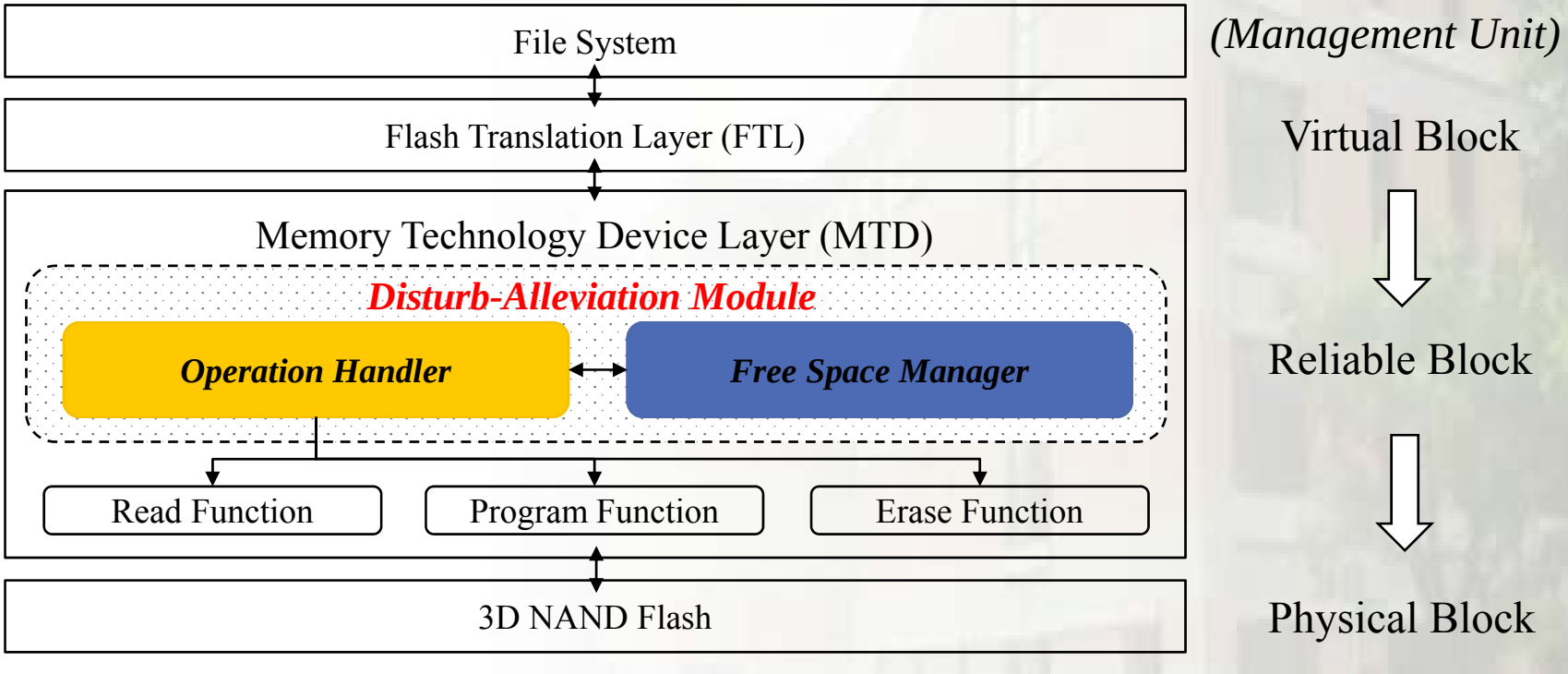
- Decreasing Endurance (Program/Erase Cycles)
  - ***Wear Leveling*** is commonly adopted to even erase cycles among all blocks in a flash chip



Source: EDN Network

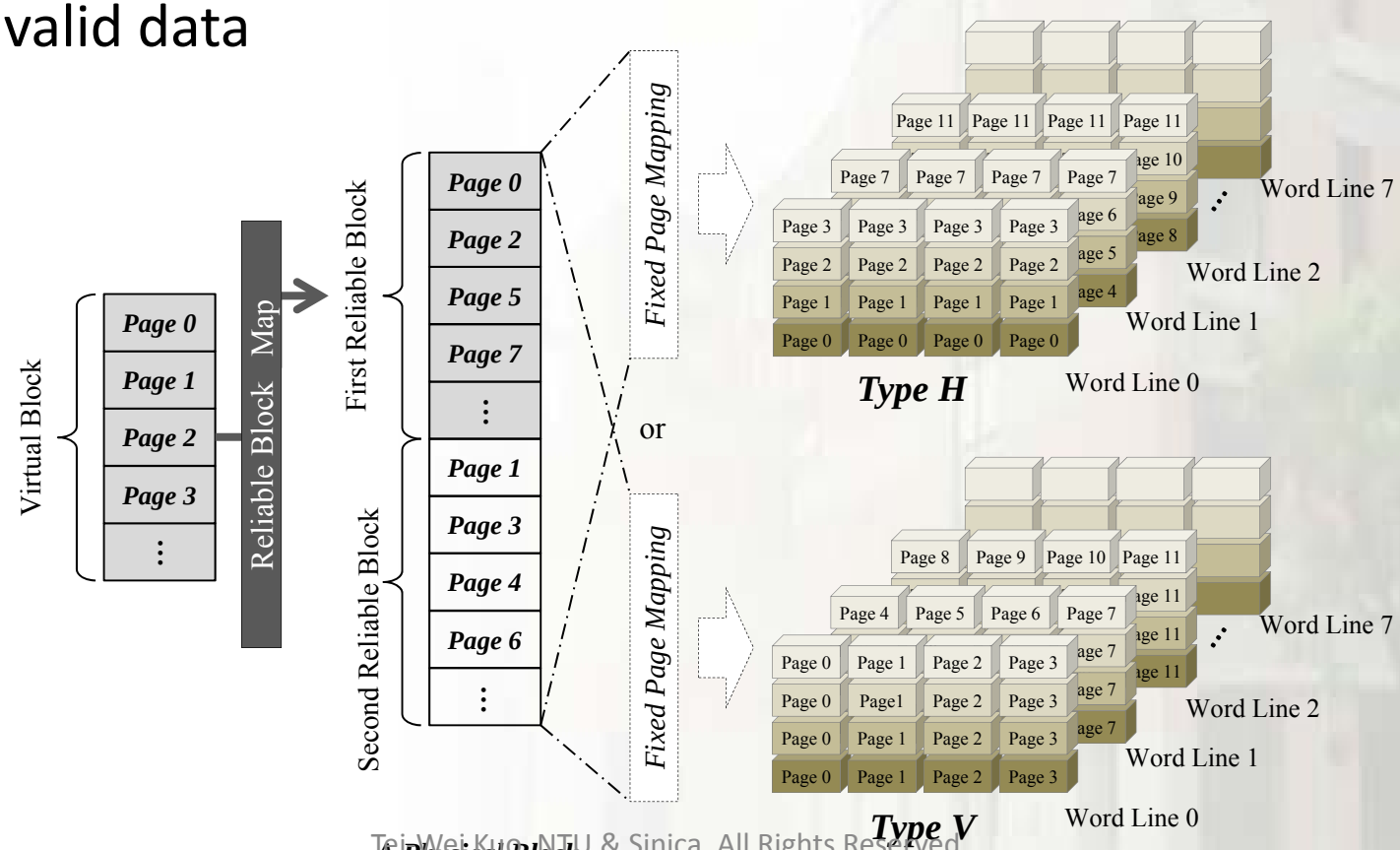
A Disturb-alleviation Scheme

- The proposed scheme can be realized in the MTD as a module
  - Without any modification to existing FTL designs or file system



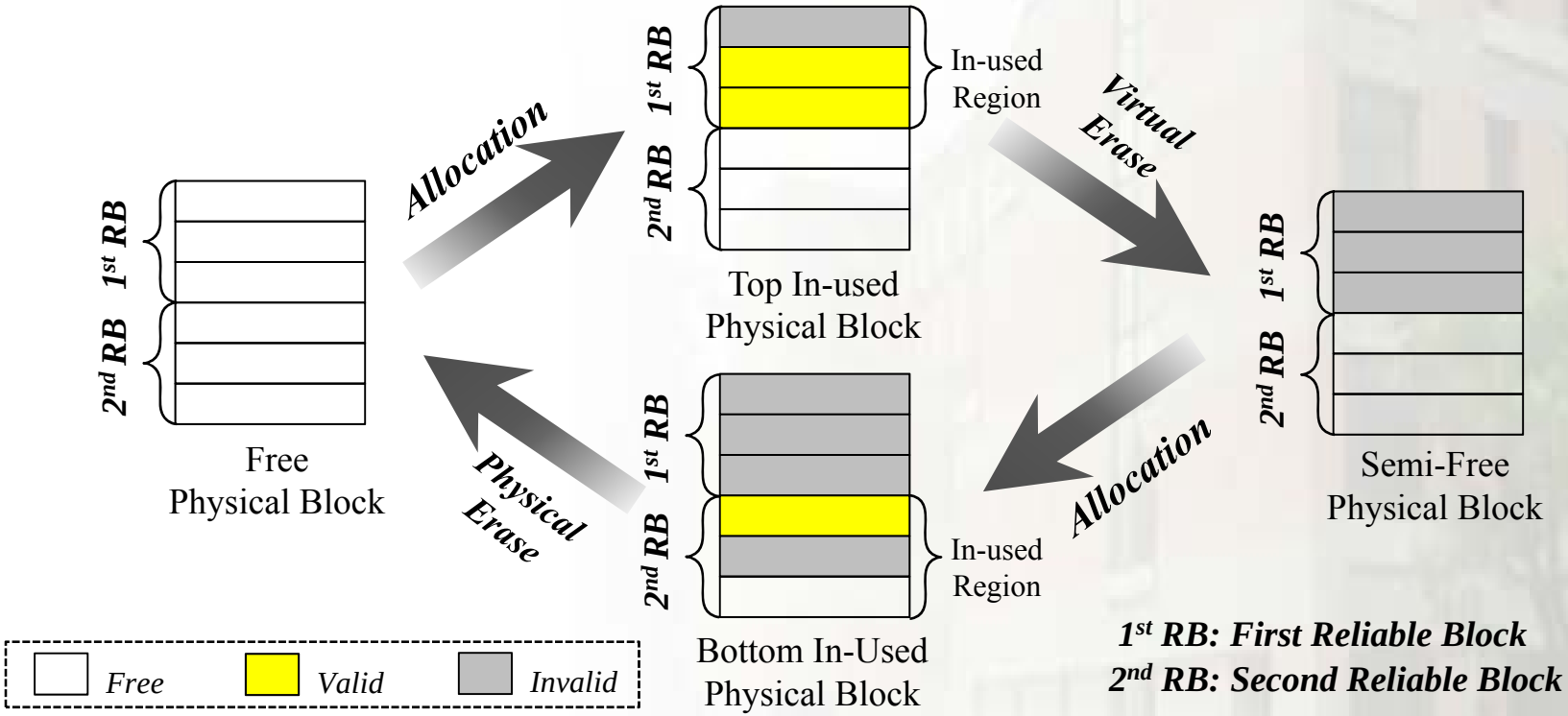
Reliable Blocks

- The creating of two reliable blocks out of physical blocks
  - Each reliable block consists of the pages that are not adjacent to each other so as to avoid writing to a page with adjacent pages of valid data



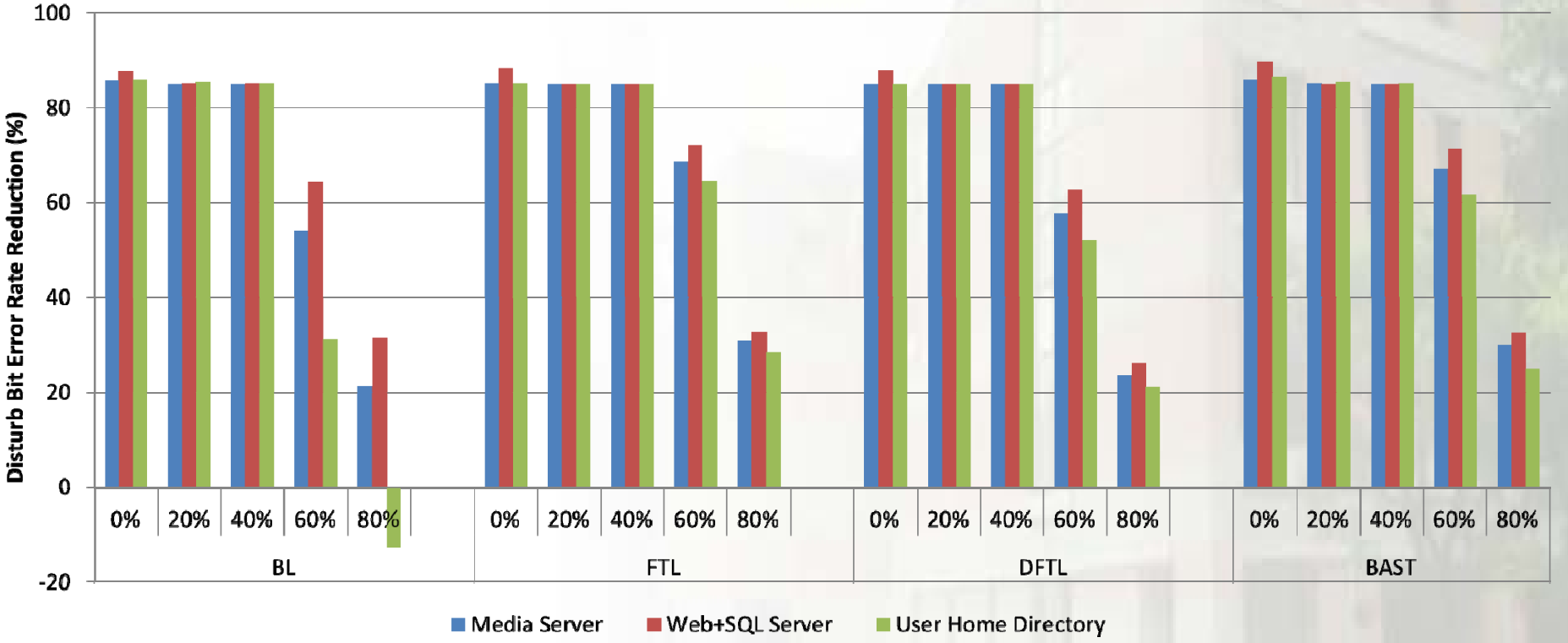
Reliable Blocks

- The life cycle of a physical block – Delaying of the using a free reliable block whose sibling reliable block is in-used



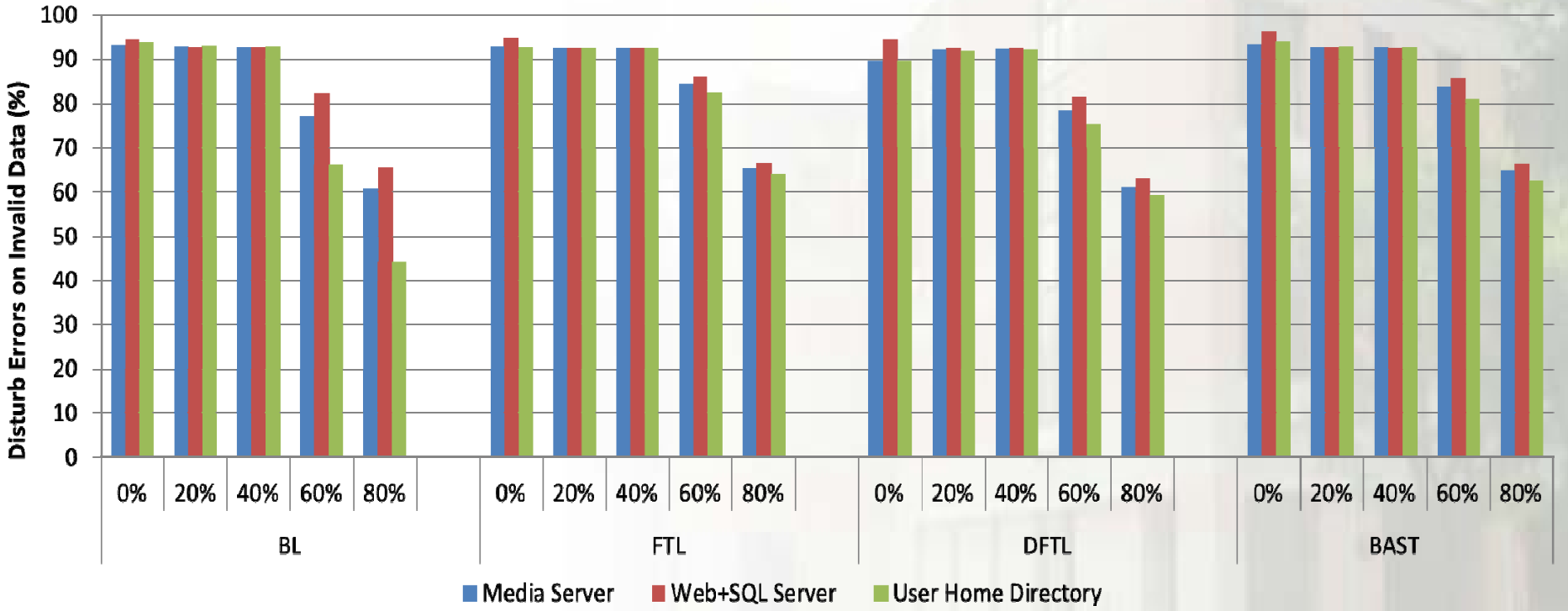
Performance Evaluation

- Significantly reduce the bit error from 64%-71%



Performance Evaluation

- 44%-96% disturb errors could be distributed to invalid pages



Summary

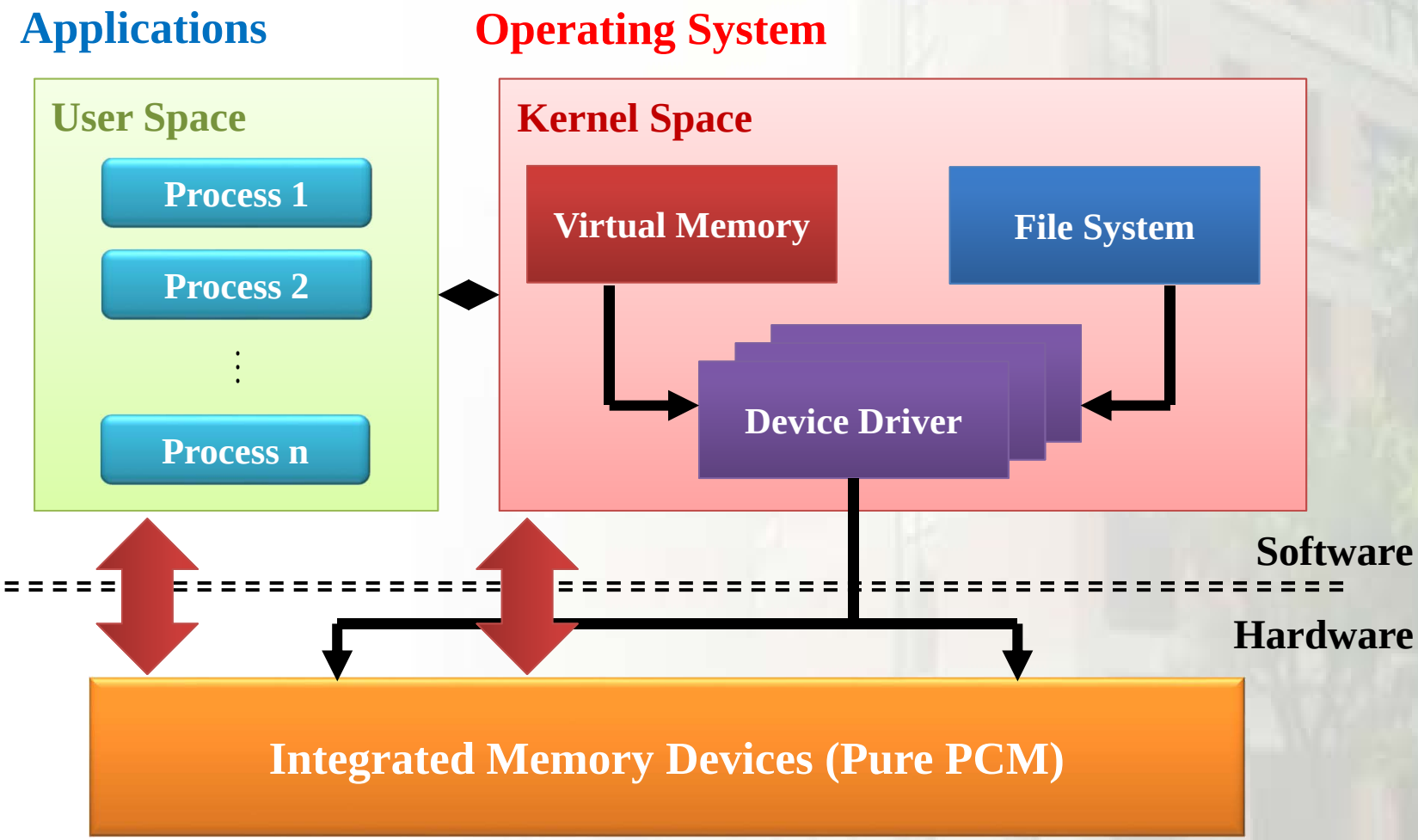
- A Disturb-Alleviation Scheme for NAND Flash Memory
  - It is to enhance the reliability of storage system and reduce the negative effect caused by program disturb.
  - There is no need to any modification to a higher level FTL or a file-system.
- Half-Block Management for Flash Storage Devices
  - The bit error is significantly reduced from 64%-71%.
  - 44%-96% disturb errors could be distributed to invalid pages.

Agenda

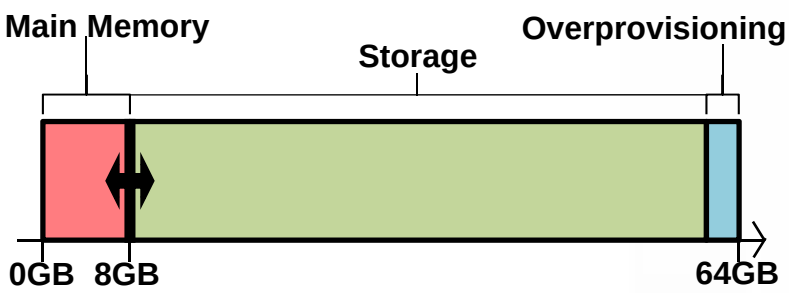
- A Global Picture
- More Efficient Chip Programming
- More Reliable Blocks
- Less Data Transfers
- Better System Optimization
- Conclusion

Bing-Jing Chang, Yuan-Hao Chang, Hung-Sheng Chang, Tei-Wei Kuo, and Hsiang-Pang Li, 2014, "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.

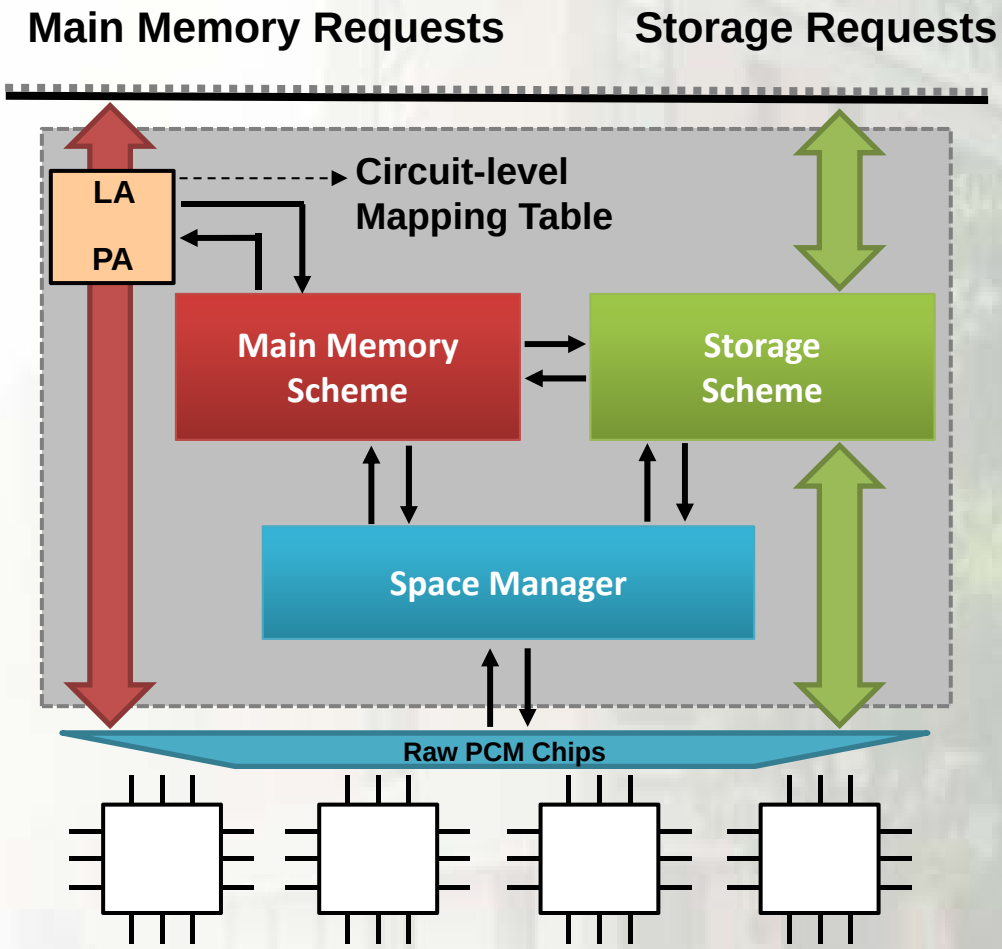
Joint Management of Memory and Storage



Non-volatile Memory Translation Layer (NTL)

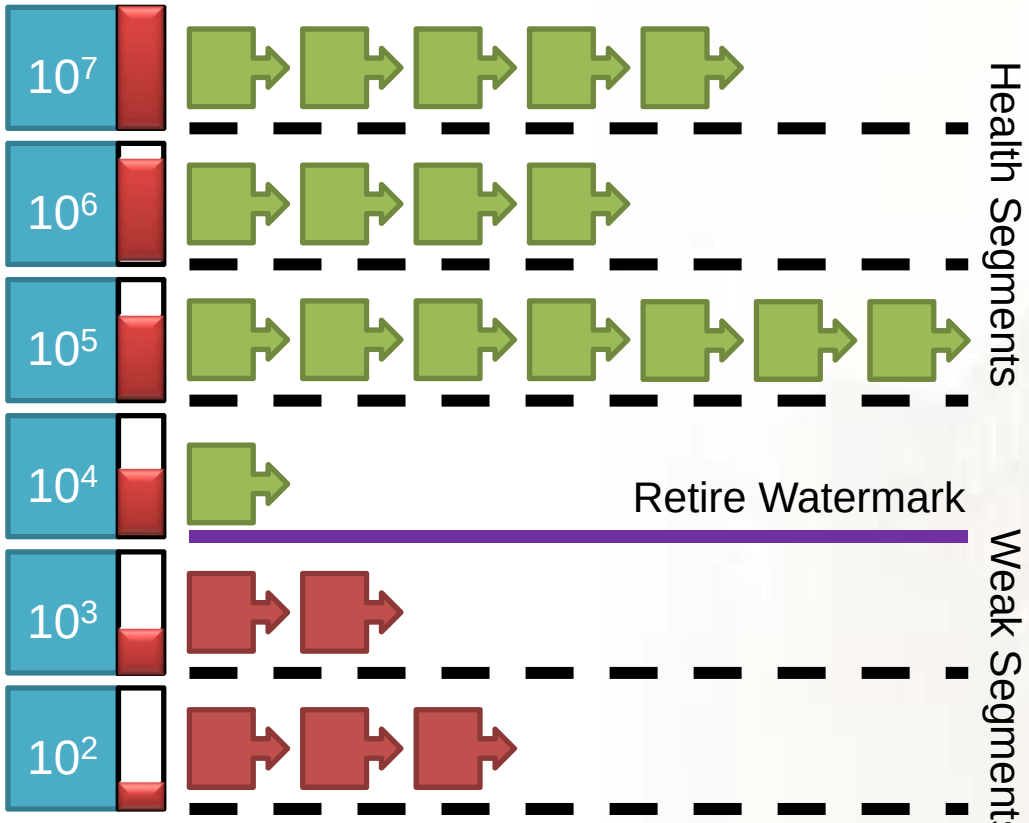


Logical View
(An example of 64GB PCM)

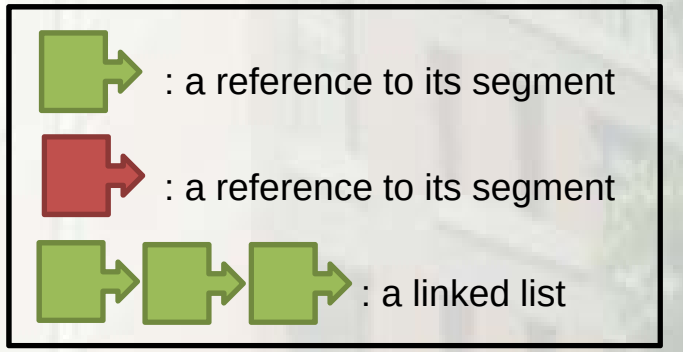


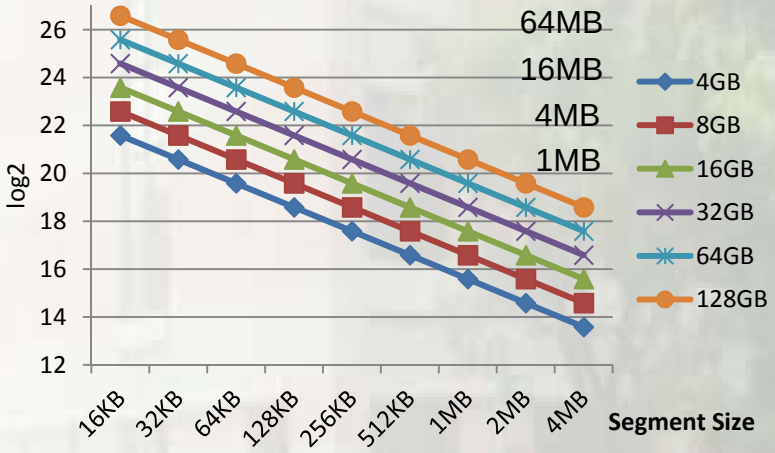
- Benefits:
  1. **Better Performance and Endurance as Memory**
     - Have smart wear leveling
  2. **Excellent Performance for Storage**
     - Extremely fast storage I/O
  3. **Knowledge Sharing**
     - Have cross-layer knowledge: New command

Space Manager with Wear-leveling



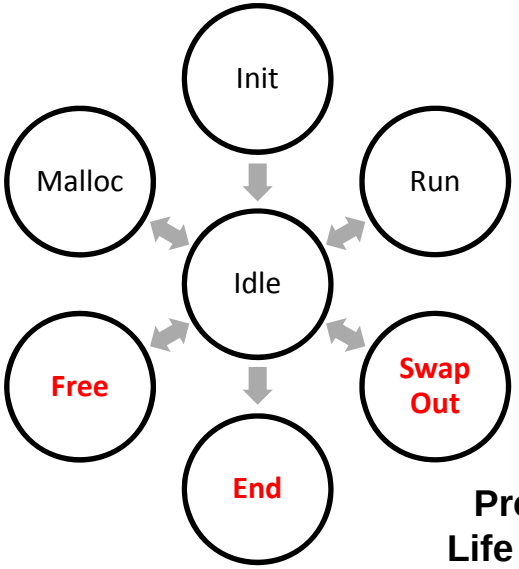
Bucket-based doubly-linked lists





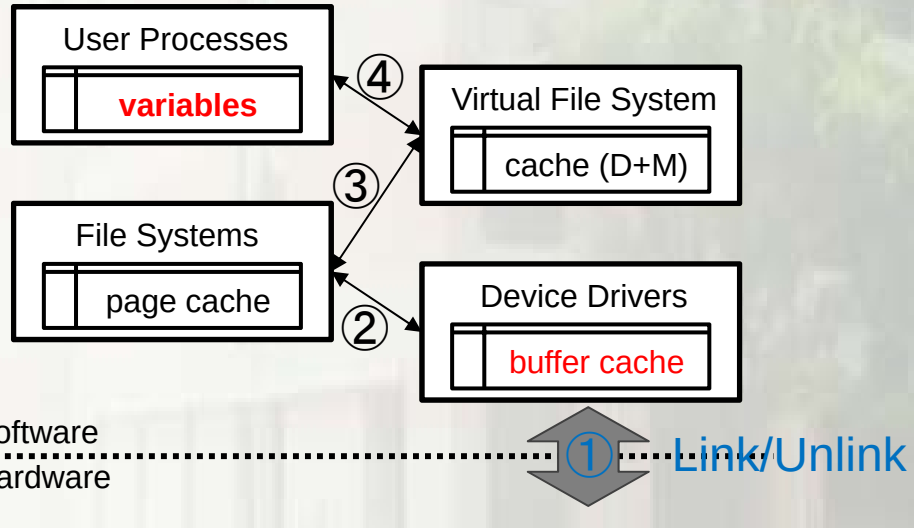
New Commands - Link, Unlink, & Free

- Repartition Command
  - Reset the ratio of main memory to storage
- **FREE Command**
  - **Eliminate swap overheads**



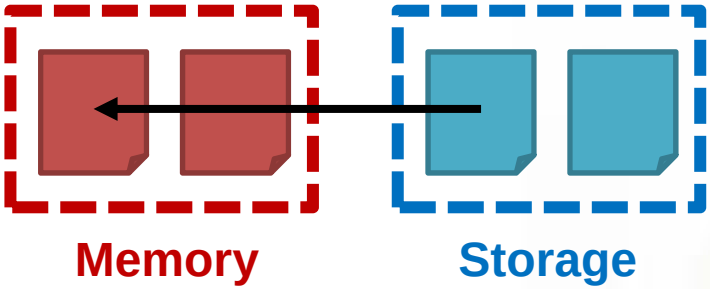
Process Life Stages

- **LINK/UNLINK Command**
  - **Provide extremely fast I/O**

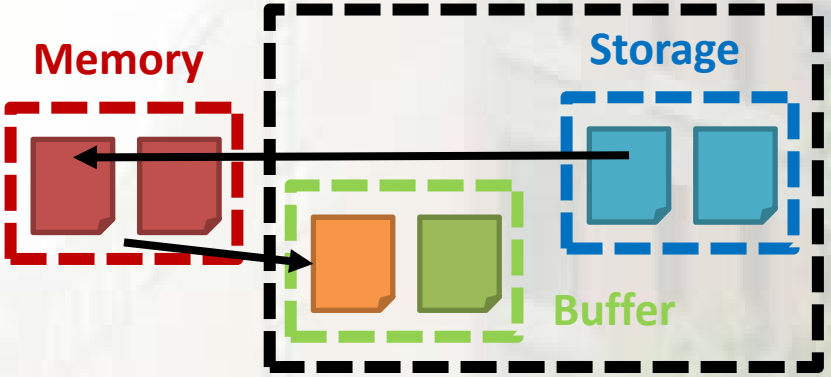


New LINK/UNLINK Commands

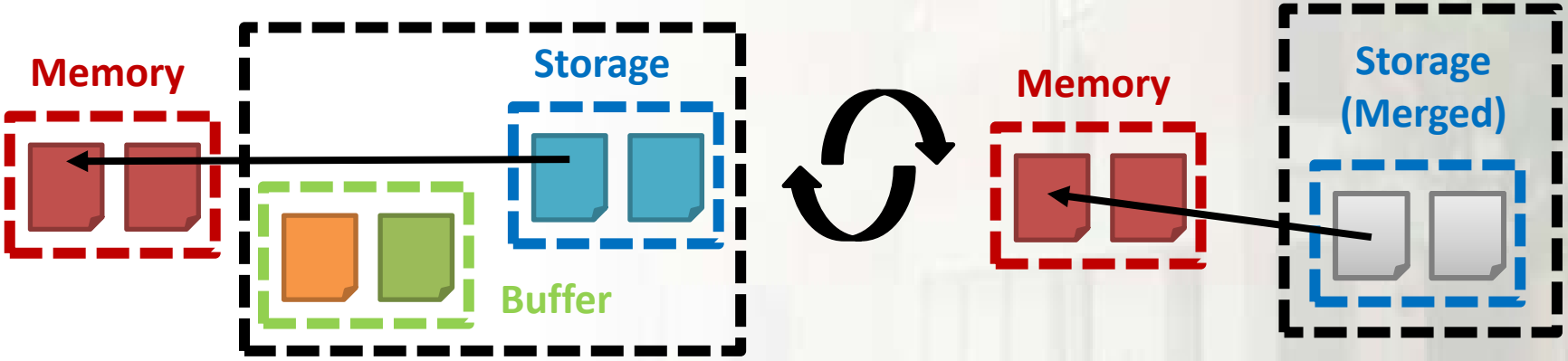
- **Read a file**



- **Create/modify a file**

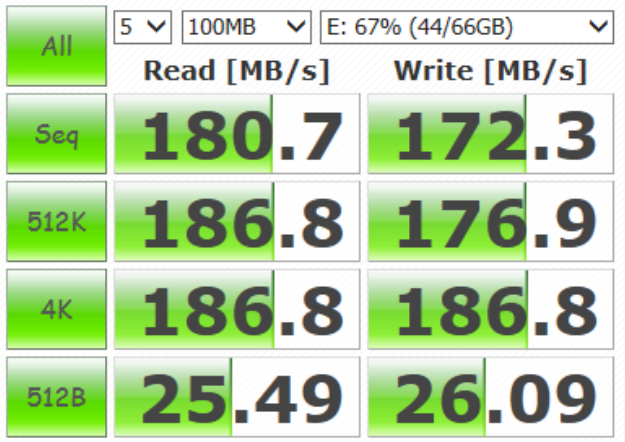


- **Periodically update a file**

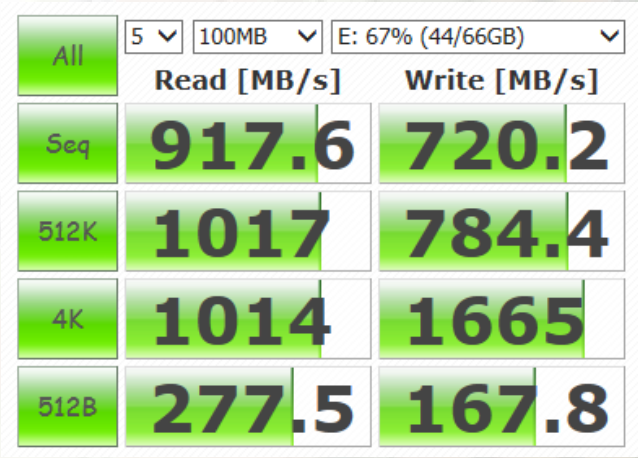


Performance

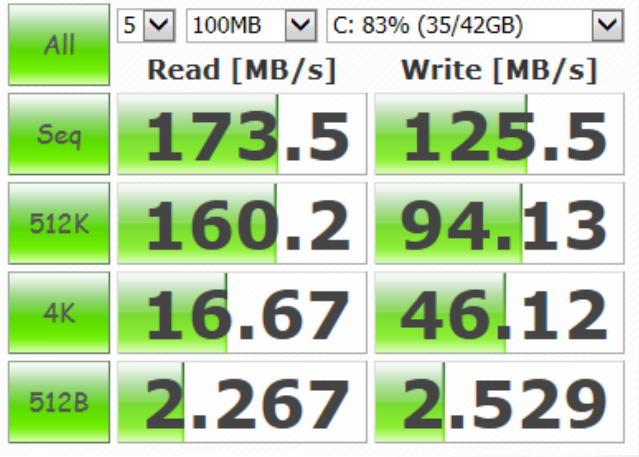
NTL (Native)



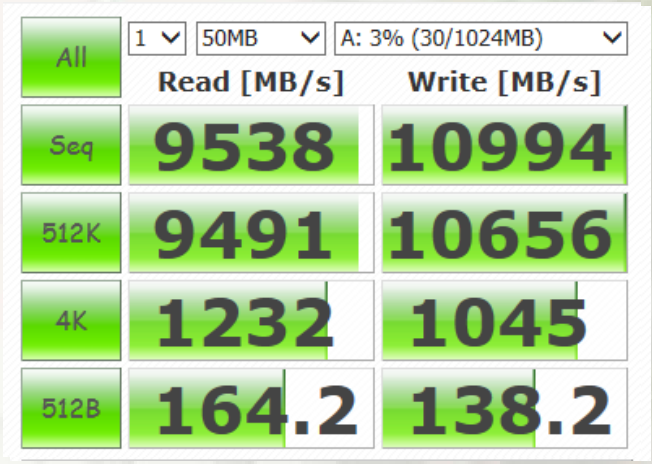
NTL (commands)



DRAM 1600MHZ + SSD (MacBook Air 2010)



RAMDISK (DDR3 1600MHZ)



Chrystal Benchmark

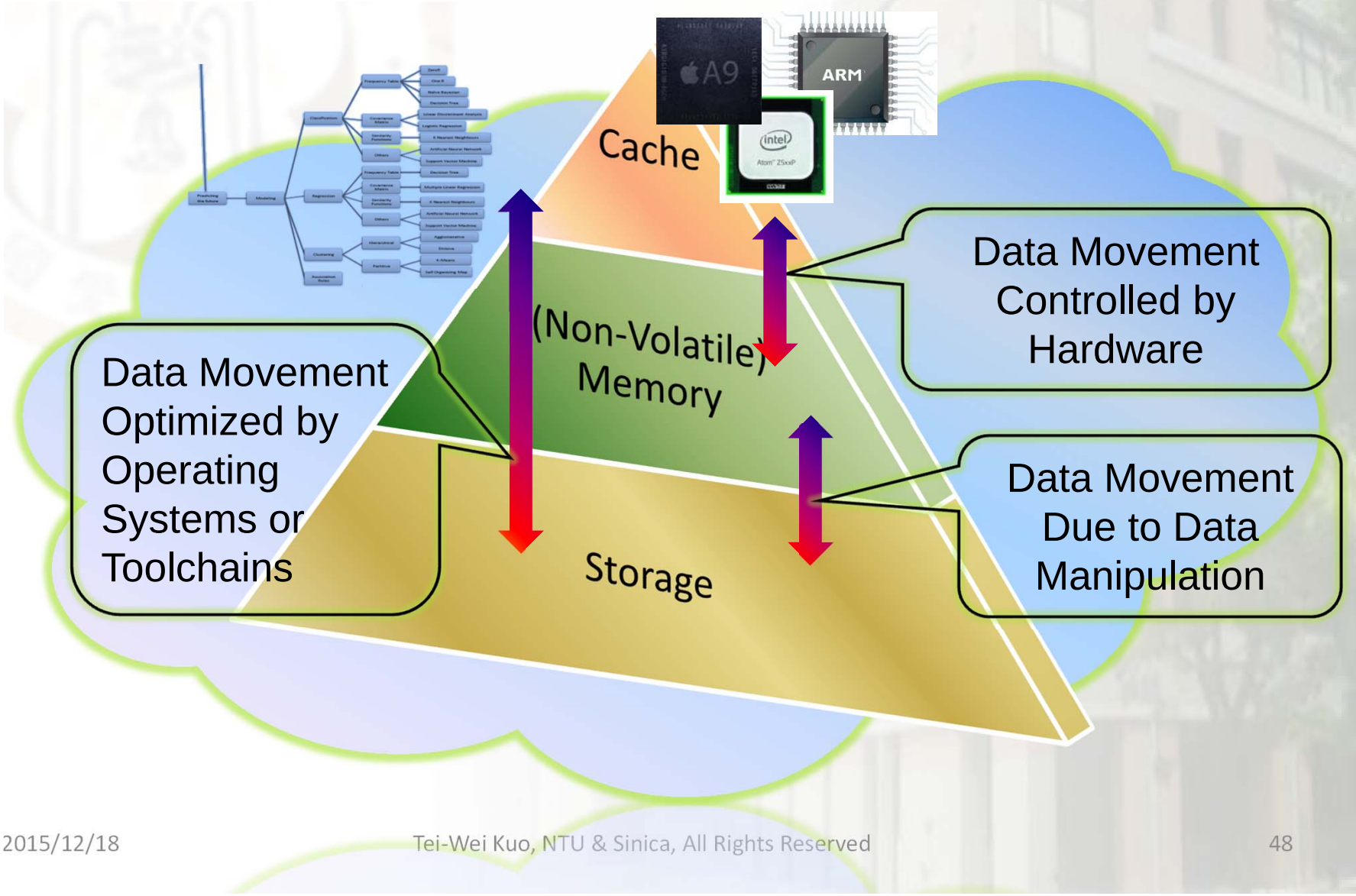
Summary

- Non-volatile memory translation layer with commands is exploited for phase change memory to serve both as the main memory and storage.
- The performance improvement is about **4~10X** of the native performance of PCM.
- The average lifetime guarantee of 6 high load benchmarks is about **3 years** for 1MB segment size .

Agenda

- A Global Picture
- More Efficient Chip Programming
- More Reliable Blocks
- Less Data Transfers
- Better System Optimization
- Conclusion

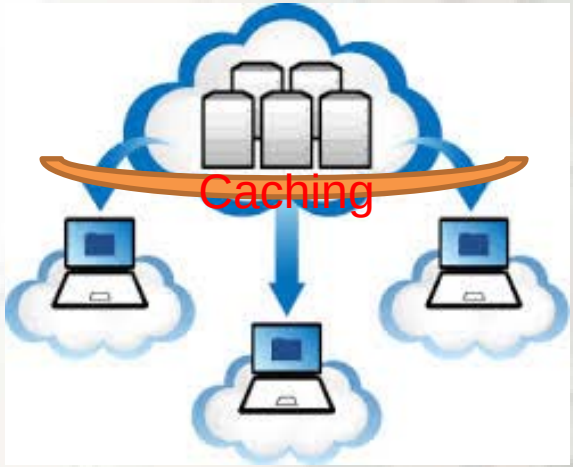
Opportunities and Challenges



Big-Data Challenges in Computing

Example System Designs

- Smarter Caches
  - Huge But Low Power Cache
- Unified Memory and Storage
  - New I/O Concepts/Designs
- Smarter Devices
  - The possibility in having a lower $-per-byte storage device with the De-Duplication or info. compression technology

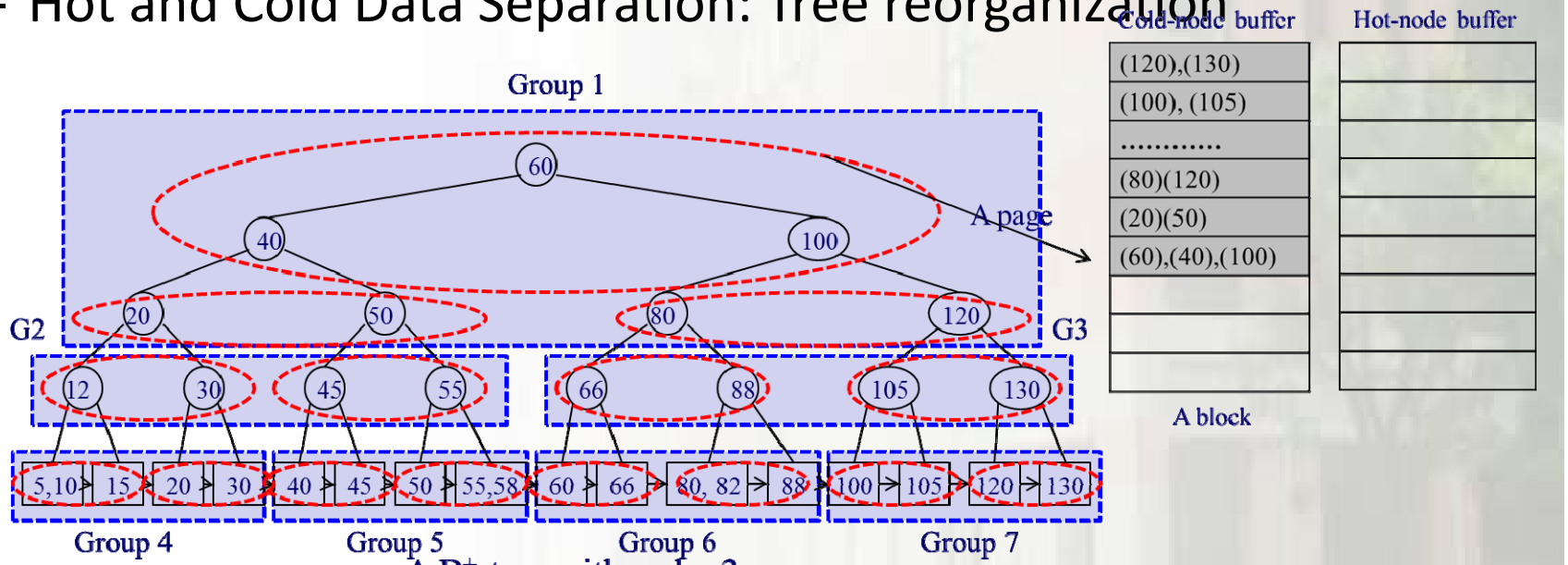


Some Key Issues in Computing

- **Data Manipulation**
  - Application Semantics and Algorithm Designs
  - APIs and Chip/System Interfaces
- **Cross-Layer/Media Optimization**
  - Data Caching, Buffering, Data Placement, and Migration
  - Index Designs, Logging and Recovery
- **System Design Concepts**
  - Rethinking of I/O and System Architectures
  - Programming Models and Support

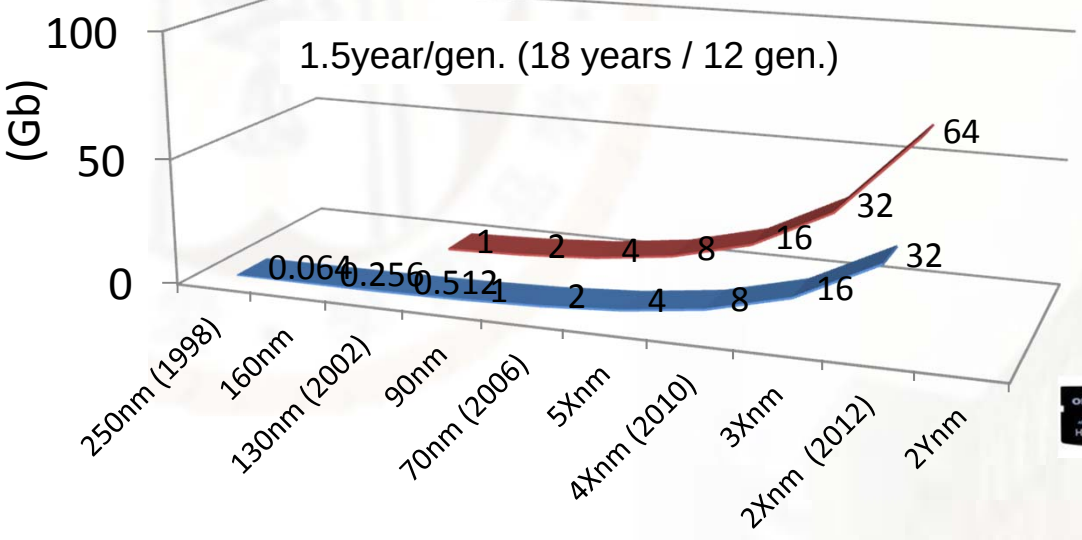
An Adaptive Endurance-Aware B+-Tree for Flash Memory Storage Systems

- Main idea of a Durable B+-tree (DB-tree)
  - Data Compaction: Sibling nodes are saved in the same page
  - Lazy Update: Two kinds of data structures: Cold-node buffer & Hot-node buffer
  - Hot and Cold Data Separation: Tree reorganization

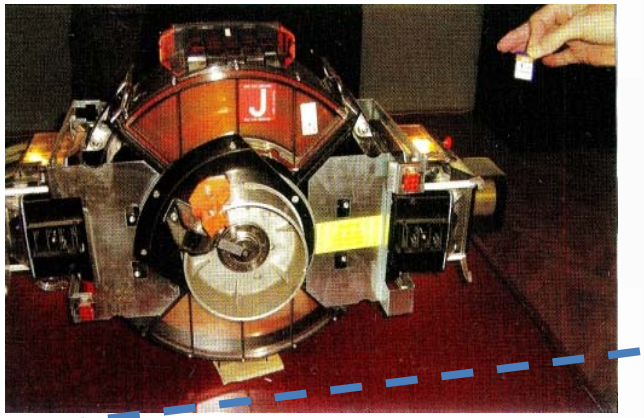


Hua-Wei Fang, Mi-Yen Yeh, Pei-Lun Suei and Tei-Wei Kuo, 2011, "A flash-friendly B+-tree with endurance-awareness," IEEE 9th Symposium on Embedded Systems for Real-Time Multimedia (ESTIMedia), Taipei, Taiwan, Oct 13-14, 2011.

Storage-Class Memory – Storage Innovation



Read Latency in us
Throughput: 500MBps
Overheads: 5%



2005

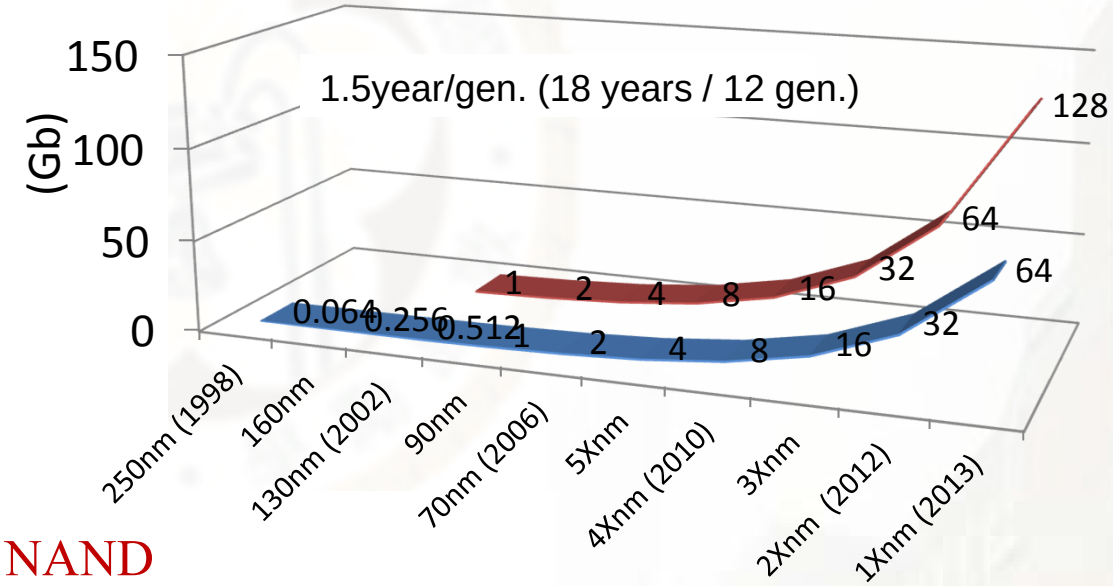
2015

Read Latency in ms
Throughput: 50MBps
Overheads: 95%

A Global Picture

- From One Breakthrough to Another…

Flash Memory – Now and Future



SLC

MLC

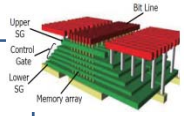
Samsung
64Gb 27nm TLC
(single die)

Micron
512Gb 2Xnm MLC
(8 die)

Toshiba
128Gb 32nm

2D NAND

3D NAND



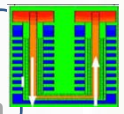
Stacked NAND IEDM 2006

BiCS VLSI Symposium

Multi TFT IEDM 2006

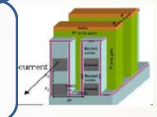
Si Pillar 3D NAND Semi. International

P-BiCS VLSI Symposium



TCAT VLSI Symposium

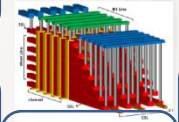
VG-NAND VLSI Symposium



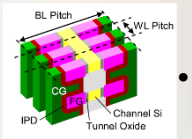
VSAT VLSI Symposium

VG-TFT VLSI Symposium

DC-SF VLSI Symposium

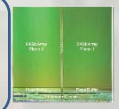


VCSTAR ISDRS



HC-FG IEDM

VNAND 128GB ISSCC



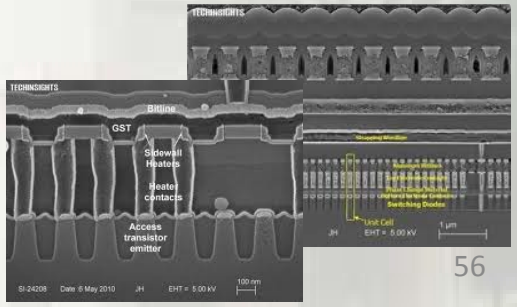
- Stack type
  - Gate Stack Type
    - ✓ P-BiCS, TCAT
  - Channel Stack Type
    - ✓ Vertical Gate (VG)
- Unit Cell Structure
  - Double Gate
  - Gate-All-Around (GAA)

2006 2007 2008 2009 2010 2011 2012 2013 2014

PCM – Now and Future

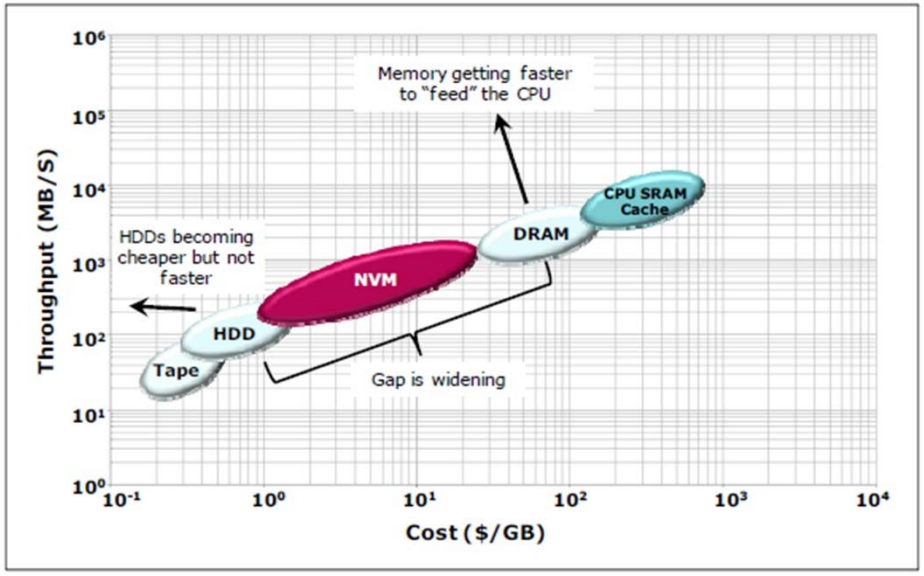
- Samsung
  - 64Mb, 012um (2004), 256Mb, 0.1um (2005), 512Mb 90nm(2006), 512Mb, 65nm (2010), 1Gb, 58nm, LPDDR2 (2011), 8Gb, 20nm, LPDDR2 (2012)
  - Samsung GSM Phone – GT-E2550 with 65nm 512Mb NOR Compatible PCM (2010)
- Micron
  - 128Mb, 90nm (2006), 1Gb, 45nm (2009; Mass Production in 2012)
- Hynix (SK Telecom)
  - 512Mb, 54nm (2010), 1Gb, 42nm (2011)
- More to Come…..

Intel/Micron
3D-Xpoint
1000X Faster
1000X Endurance
10X Denser

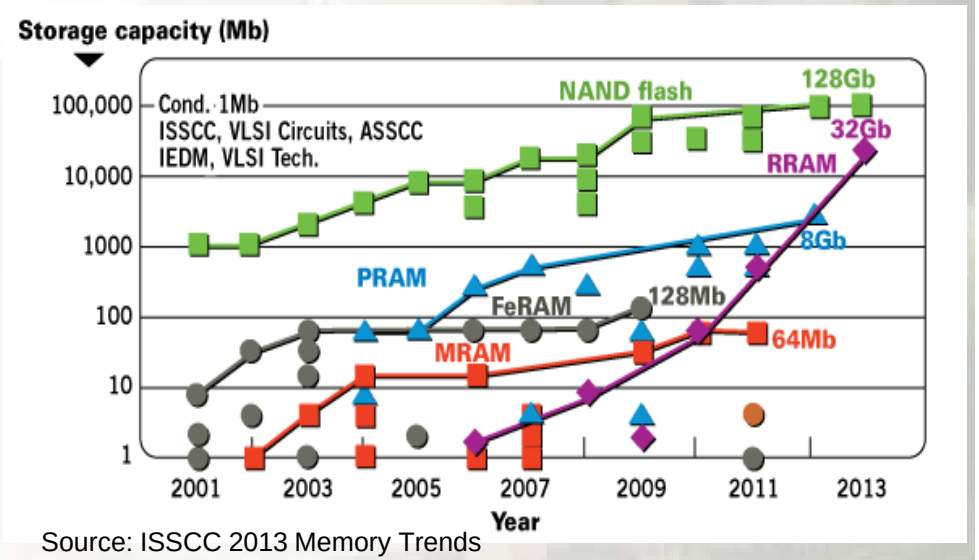


Some More Facts

- Rapid Widening in Usages of Non-Volatile Memory (NVM)
- Rapid Increasing of the Capacity of NVMs



Source: Chipnastics http://www.chipdesignmag.com/bursky/?p=62



Source: ISSCC 2013 Memory Trends
http://www.electroiq.com/articles/sst/2013/02/isscc-2013--memory-trends.html

Conclusion

- The Emerging of Non-Volatile Memory
  - System Architectures and Chip Interfaces
  - Memory Management
  - File System Designs, Interface Designs, Database and Application Designs
  - In-Memory Computing
- The Positioning of Different Memory Technology
  - Applications, Systems, and Devices

Contact Information

- **Professor Tei-Wei Kuo**
  - ktw@csie.ntu.edu.tw
  - URL: http://csie.ntu.edu.tw/~ktw
  - Office: +886-2-23625336-257
  - Fax: +886-2-23628167
  - Address:
    - Research Center for IT Innovation
      Academia Sinica, Taipei, Taiwan 115
    - Dept. of Computer Science & Information Engr.
      National Taiwan University, Taipei, Taiwan 106

Questions or Comments?