Flash Trends: Challenges and Future

John D. Davis

work done at Microsoft Researcher- Silicon Valley

in collaboration with

Laura Caulfield*, Steve Swanson*, UCSD*

My Research Areas of Interest

- Flash characteristics
- Flash trends (FAST ‘12)
- Flash specialization (HW/SW co-design) (USENIX ’13, …)
- Flash-based systems (NSDI ‘12, TOCS ‘13, SYSTOR ’13, …)
- PCM endurance (ISCA ‘13)
- Hardware accelerators (HW/SW co-design) (ISCA ‘13)
- Data center efficiency and modeling (ISWC ’12, …)
- Computer Architecture

tape is dead

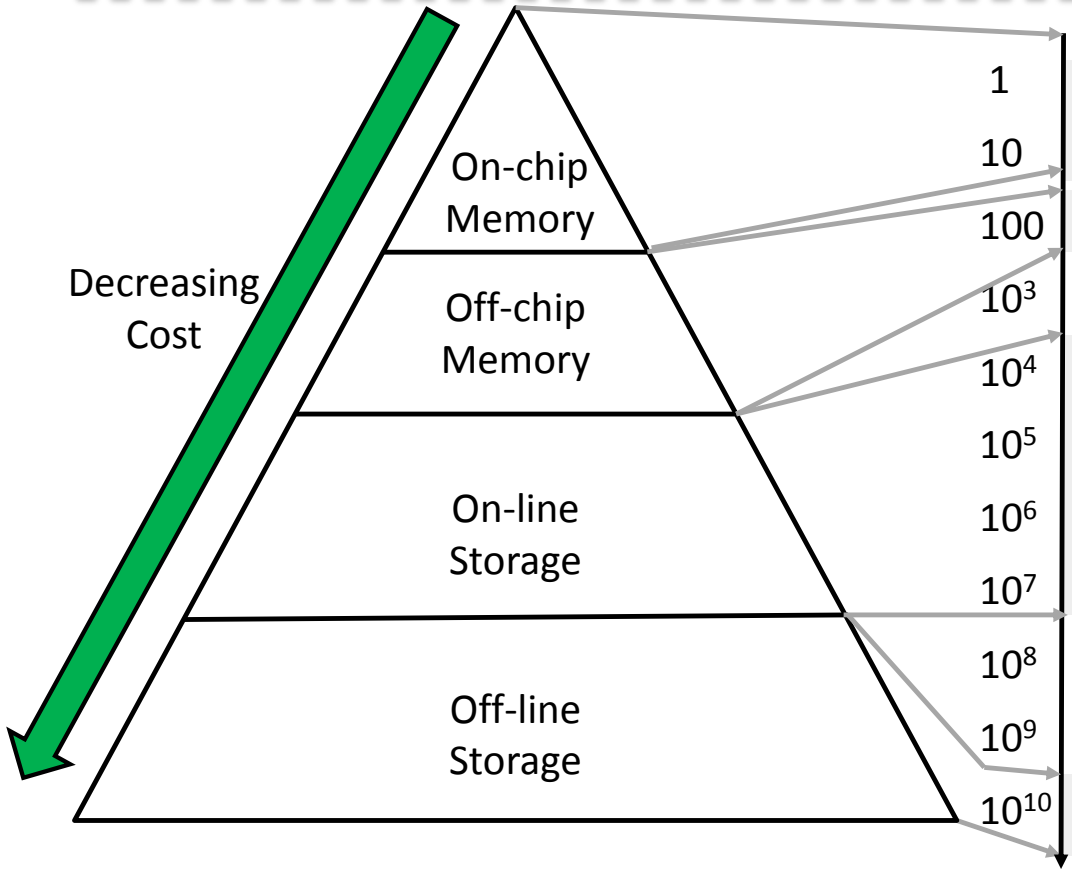
disk is tape

flash is disk

RAM locality is king

- Jim Gray, Dec 2006

The Access-Time Pyramid



Access Time (ns)		Human scale ($t \times 10^9$)
1	CPU operations (1 ns)	second
10	L2 Cache access (10 ns)	
100	DRAM access (60 ns)	minute
10^3		hour
10^4	NAN	
10^5		day
10^6	NAND Fla	week
10^7	HDD read/write (5 ms)	month
10^8		year
10^9		decade
10^{10}	TAPE read (40 s)	millenium

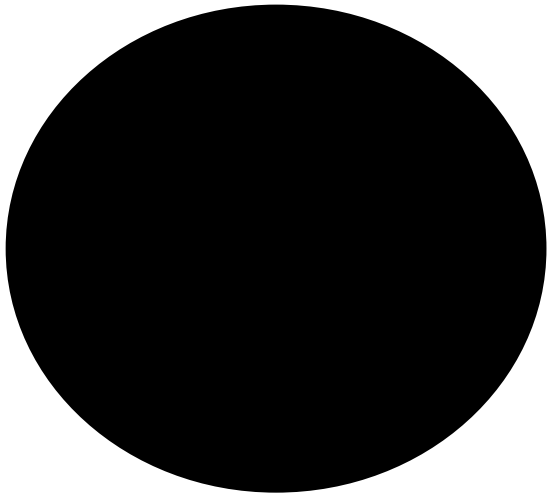
The Gap before NAND Flash!

NAND Flash by the Data Sheet

- Block interface(like an HDD): page = 2KB – 16KB
- Asymmetric read/write time (fast/slow)
  - Sequential writes ☹
  - Random write and sequential write almost the same ☺
  - Random read = fast, no seek penalty! ☺
- Block erase (64 -128 pages) ☹
  - No write-in place & slow
- Low write endurance ($<10^6$) ☹
  - Can wear out fast
- Scalability beyond 1X nm (at 19 nm now) ☹
  - Can't increase density of devices
  - **Toshiba introducing 3D lithography in a couple of years! ???**

The NAND Flash Black Hole in 2008

- How do bits fail?
- What does endurance mean?
- What about data retention and bit rot?
- Average/typical/maximum latency?
- Characteristics change with scaling?
- Write out-of-order?
- Program\read disturb?
- Page layout?

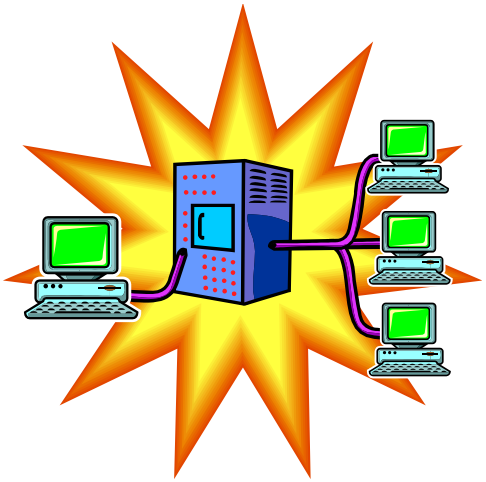


FLASH BASICS

Flash's Future: Bright

Reliability

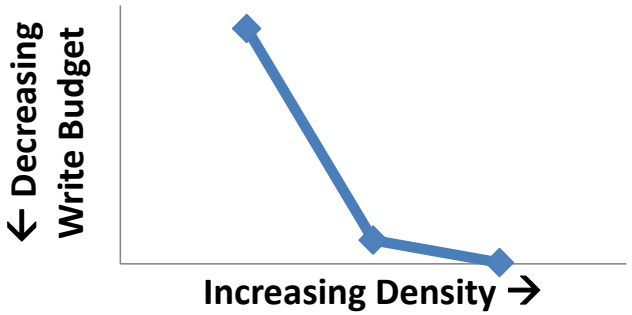
Performance



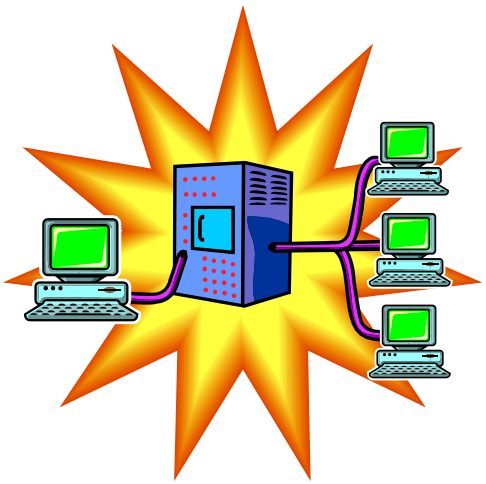
Cost Per Capacity

Flash's Future: ~~Bright~~ Bleak

Reliability



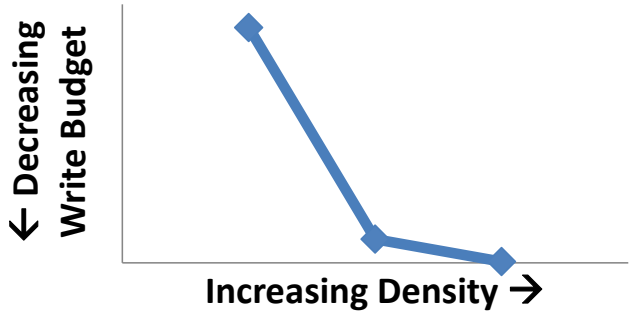
Performance



Cost Per Capacity

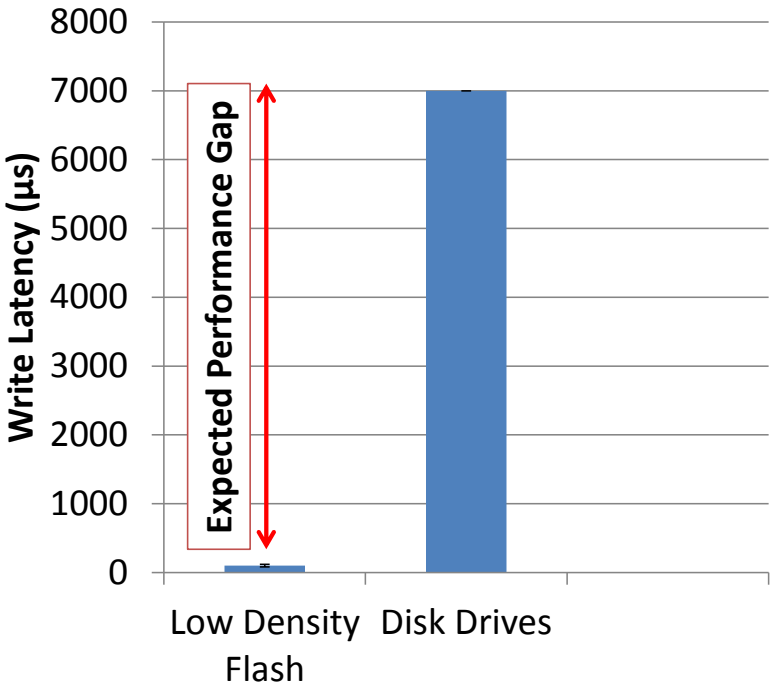
Flash's Future: ~~Bright~~ Bleak

Reliability



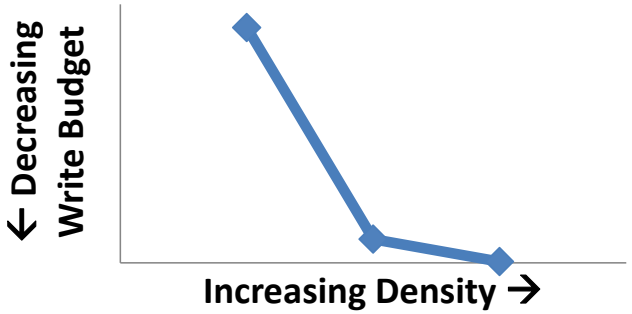
Cost Per Capacity

Performance



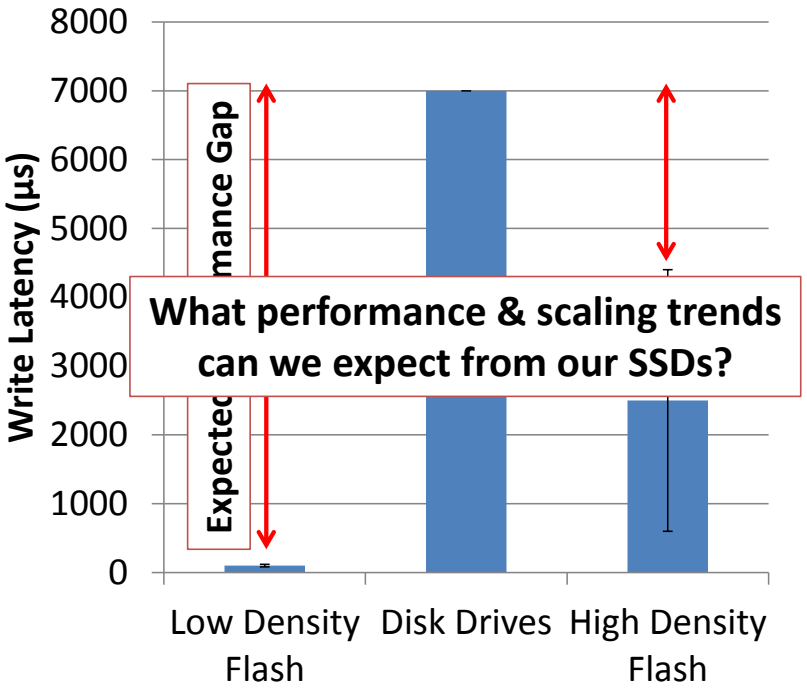
Flash's Future: ~~Bright~~ Bleak

Reliability

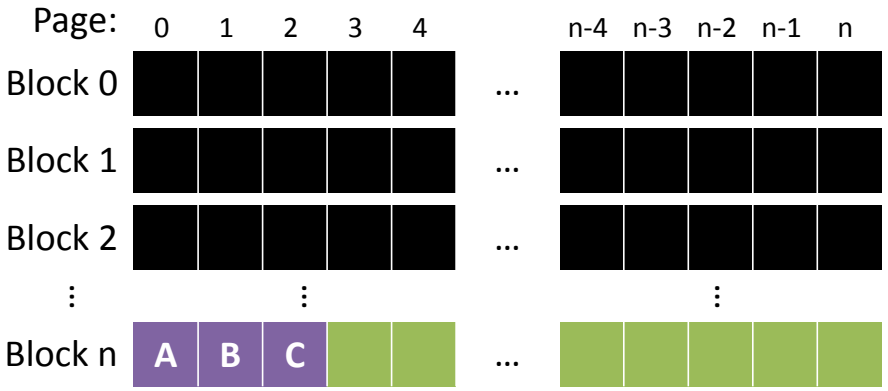


Cost Per Capacity

Performance



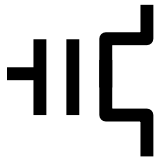
Flash Chip Operation



Erase Blocks Before Programming

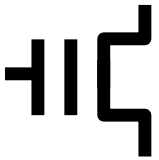
Program Pages In Order

SLC: Single Level Cell



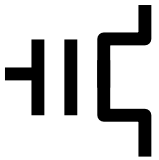
== 1 bit

MLC: Multi Level Cell



== 2 bits

TLC: Triple Level Cell

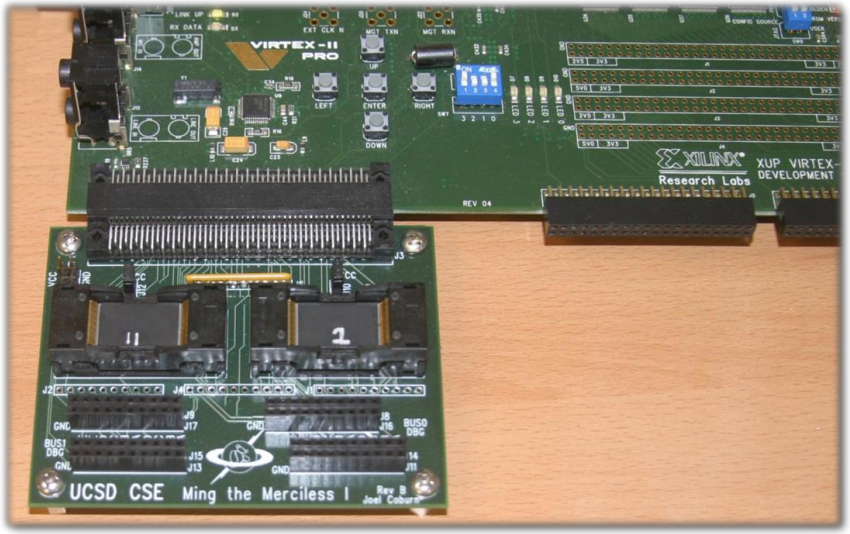


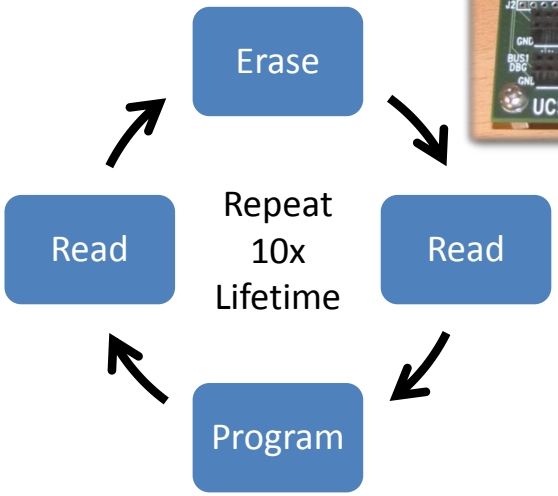
== 3 bits

Collecting Flash Latency Trends

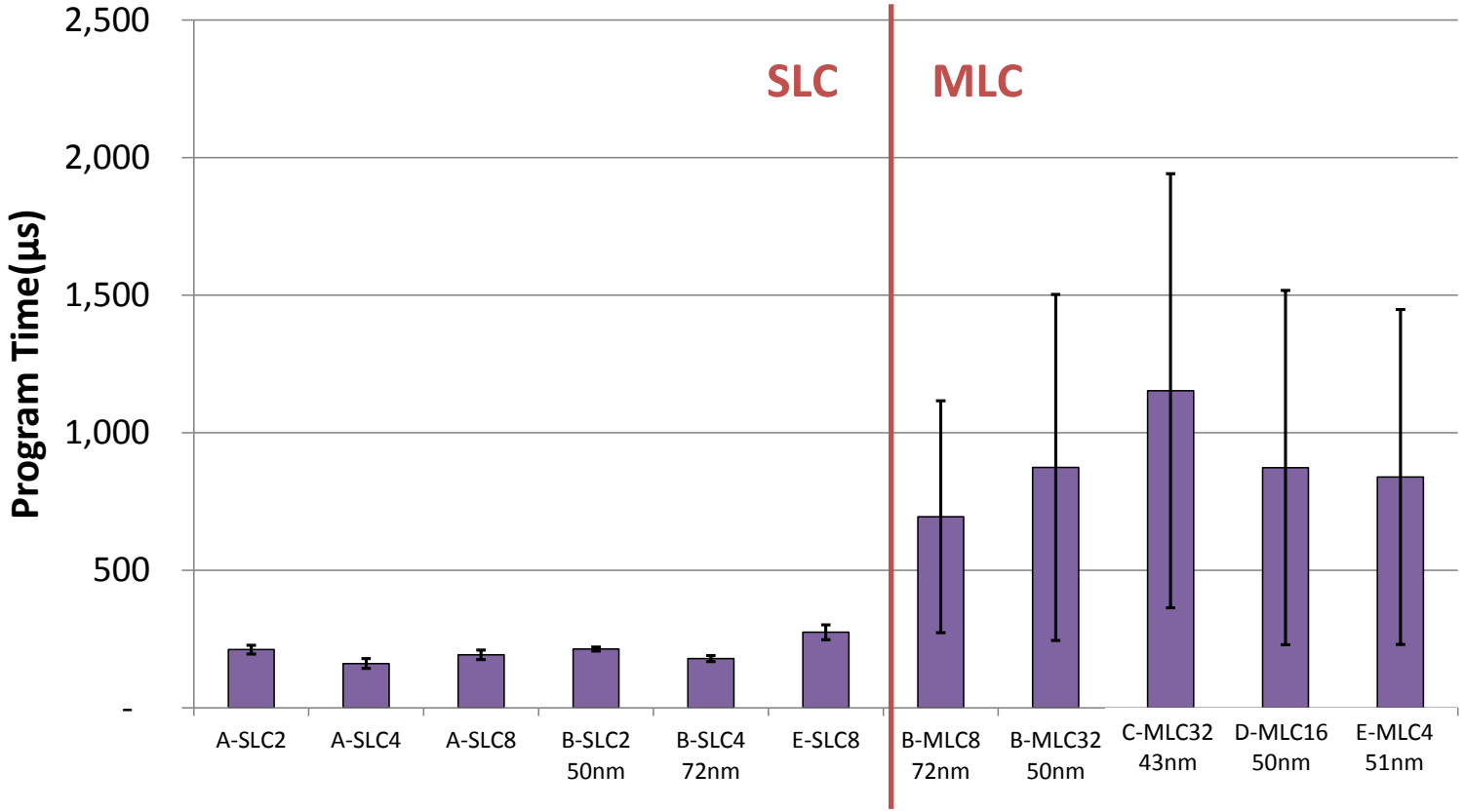
- Xilinx XUP Board
- Custom Built Daughter Board
- Full-fledge Linux
- Kernel Module
  - 10ns resolution

- Chip Collection
  - 45 chips
  - 6 companies
  - 25nm-72nm
  - SLC, MLC, TLC

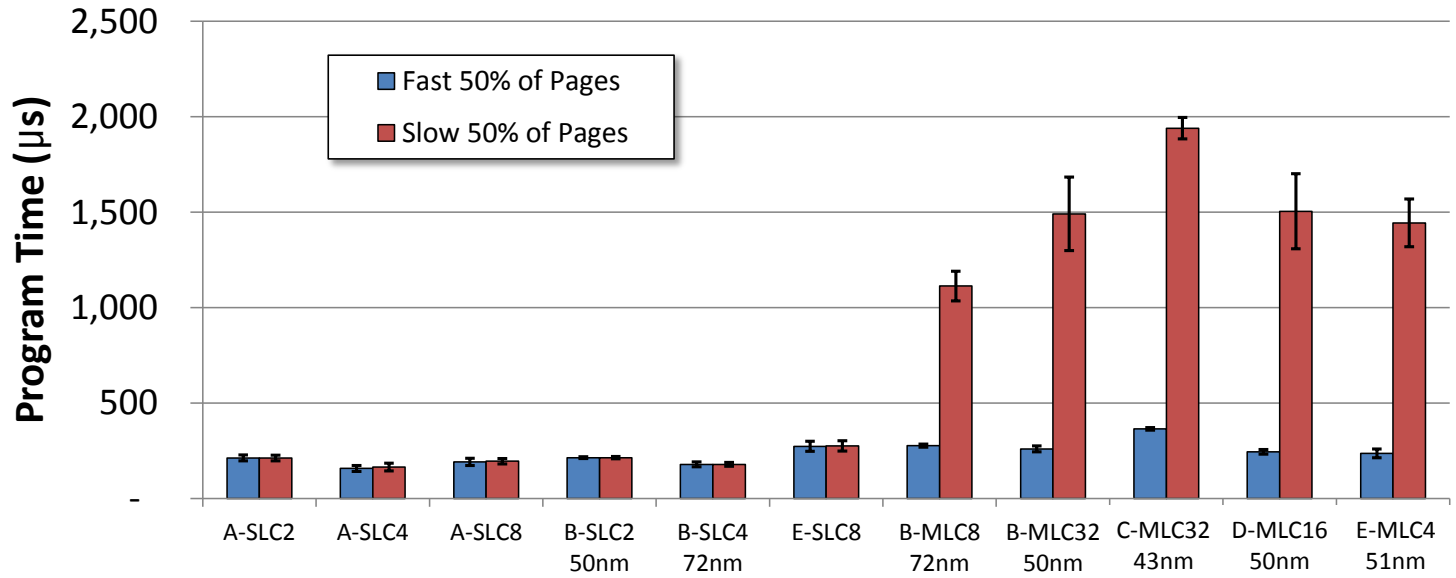


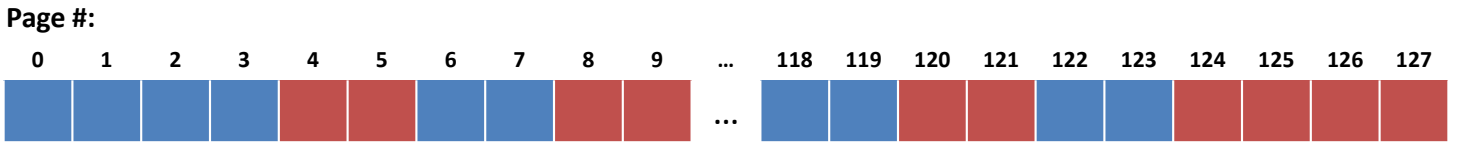


Program Operation Latency

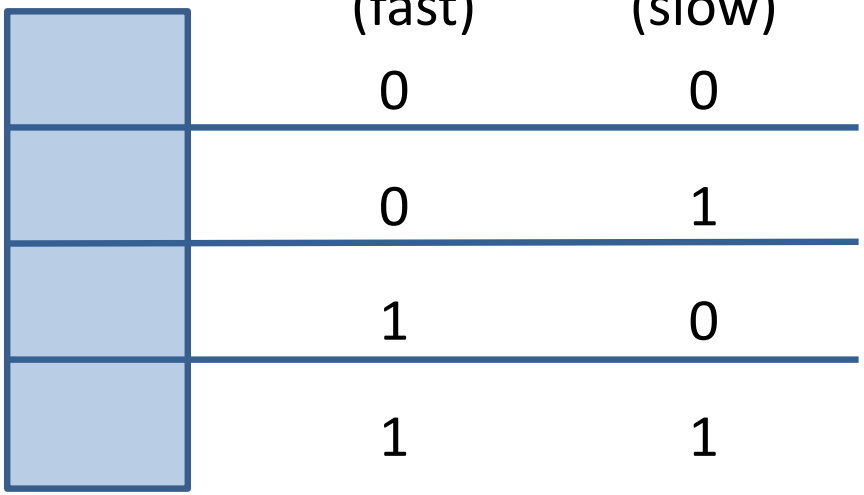


Program Latency Anomaly





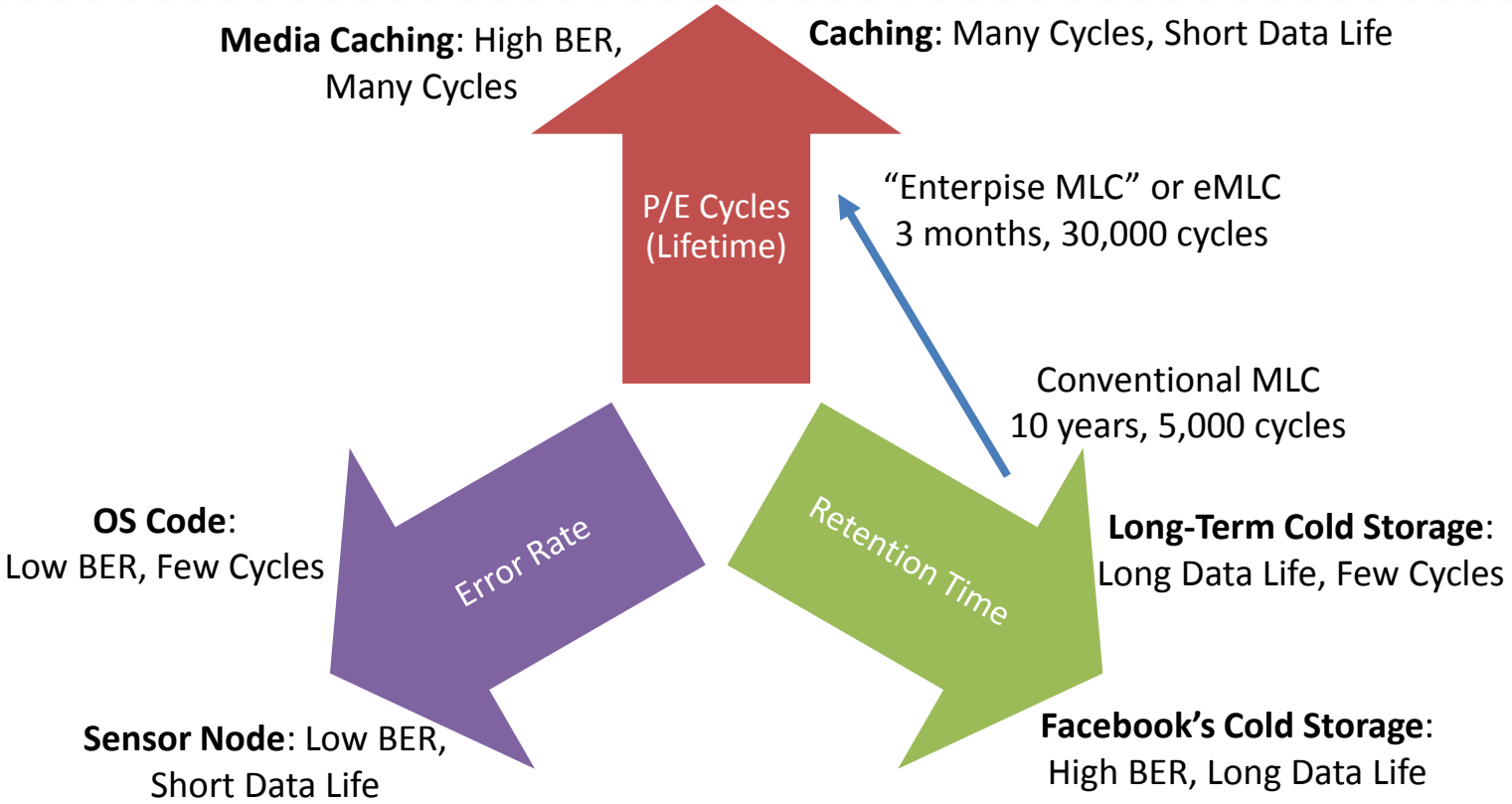
Paired Pages: High & Low Order Bits



Different Bits, Different Behaviors
Harey Tortoise [USENIX13]

Industry Impact?
Datasheets Before and After Fault Tolerance Study

Flexible Dimensions of Reliability

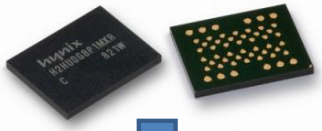


FLASH TRENDS

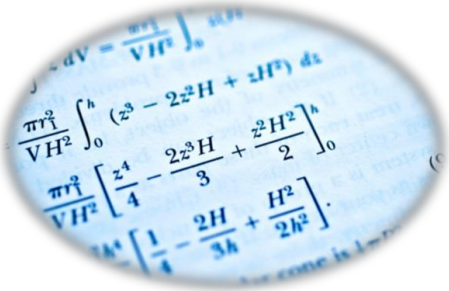
Predicting Future Flash-Based SSDs

Fixed SSD Architecture

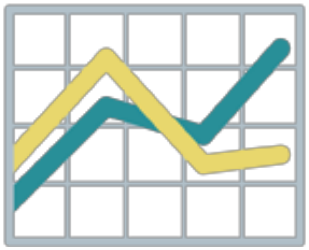
Flash Chip Trends



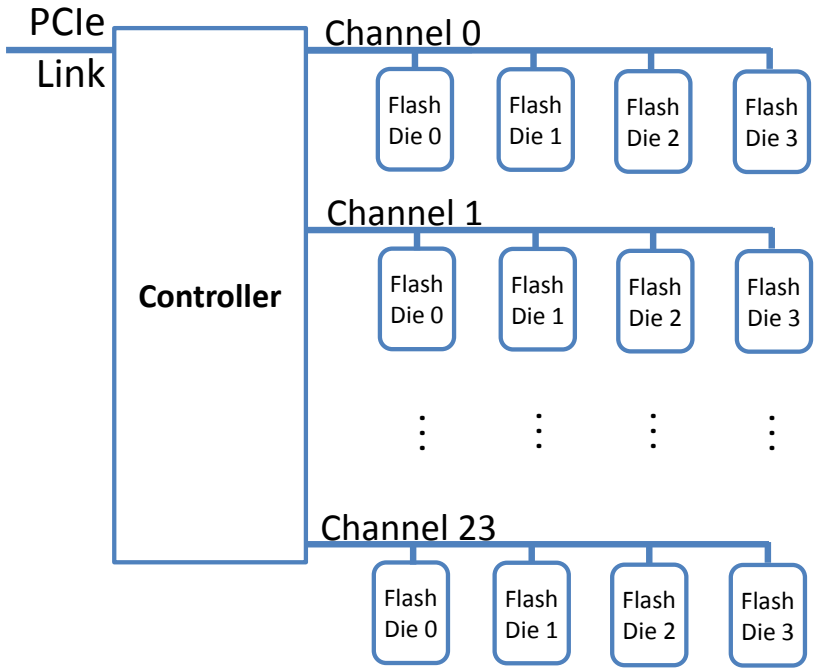
Model's Equations



SSD Trends

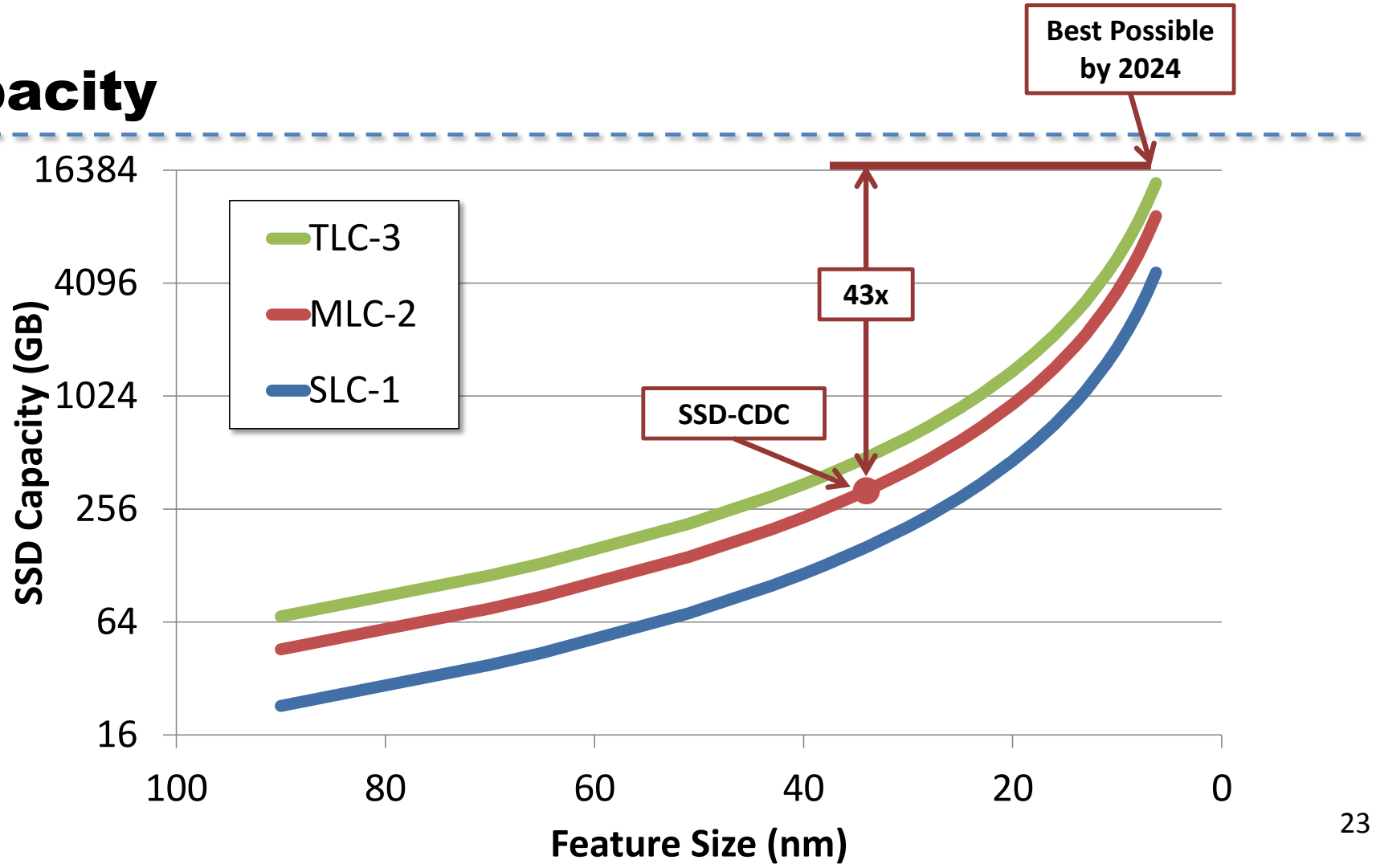


The Constant-Die-Count SSD (SSD-CDC)

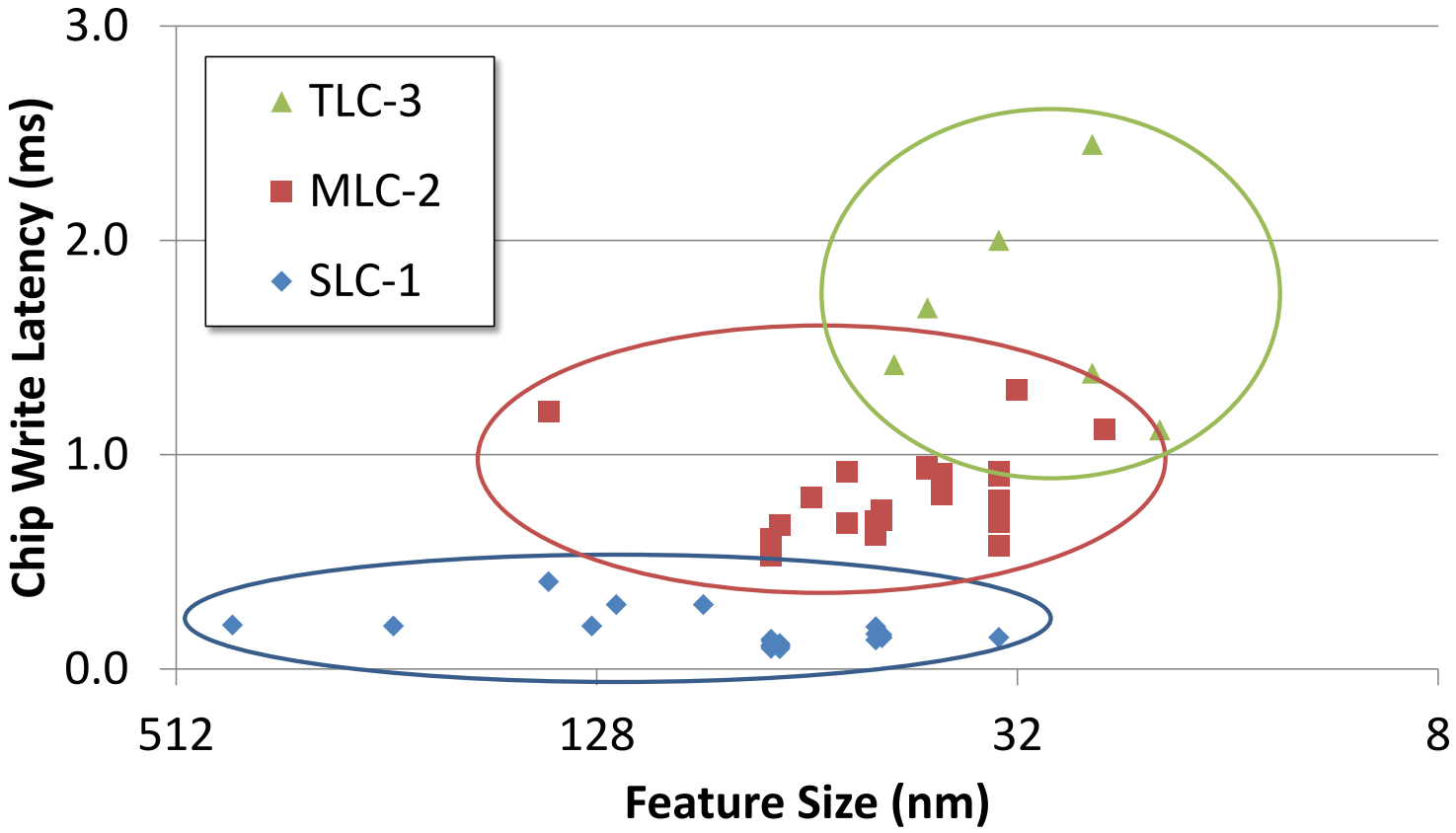


- Represents High-End (FusionIO, Virident, OCZ)
- Baseline
  - 96 dies
  - 320 GB
  - 34nm, MLC
- Scaling from Chips
  - SLC, MLC, TLC
  - To 6nm by 2024 (ITRS)
- Assumptions
  - Constant die count
  - Unlimited PCIe Link
  - Channel Speed: 400MB/s

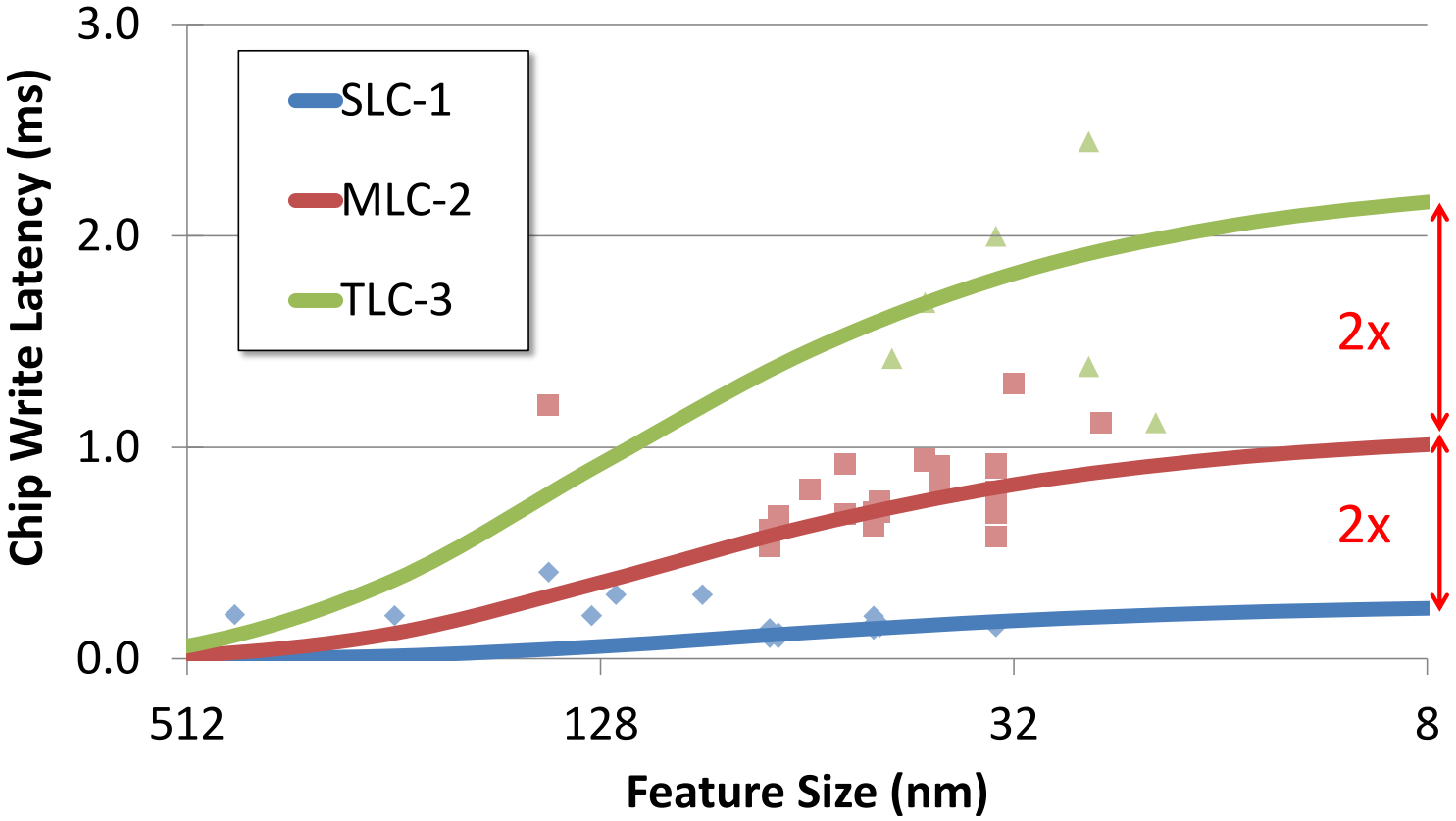
Capacity


Best Possible by 2024
16384
4096
1024
256
64
16
SSD Capacity (GB)
TLC-3
MLC-2
SLC-1
43x
SSD-CDC
100
80
60
40
20
0
Feature Size (nm)

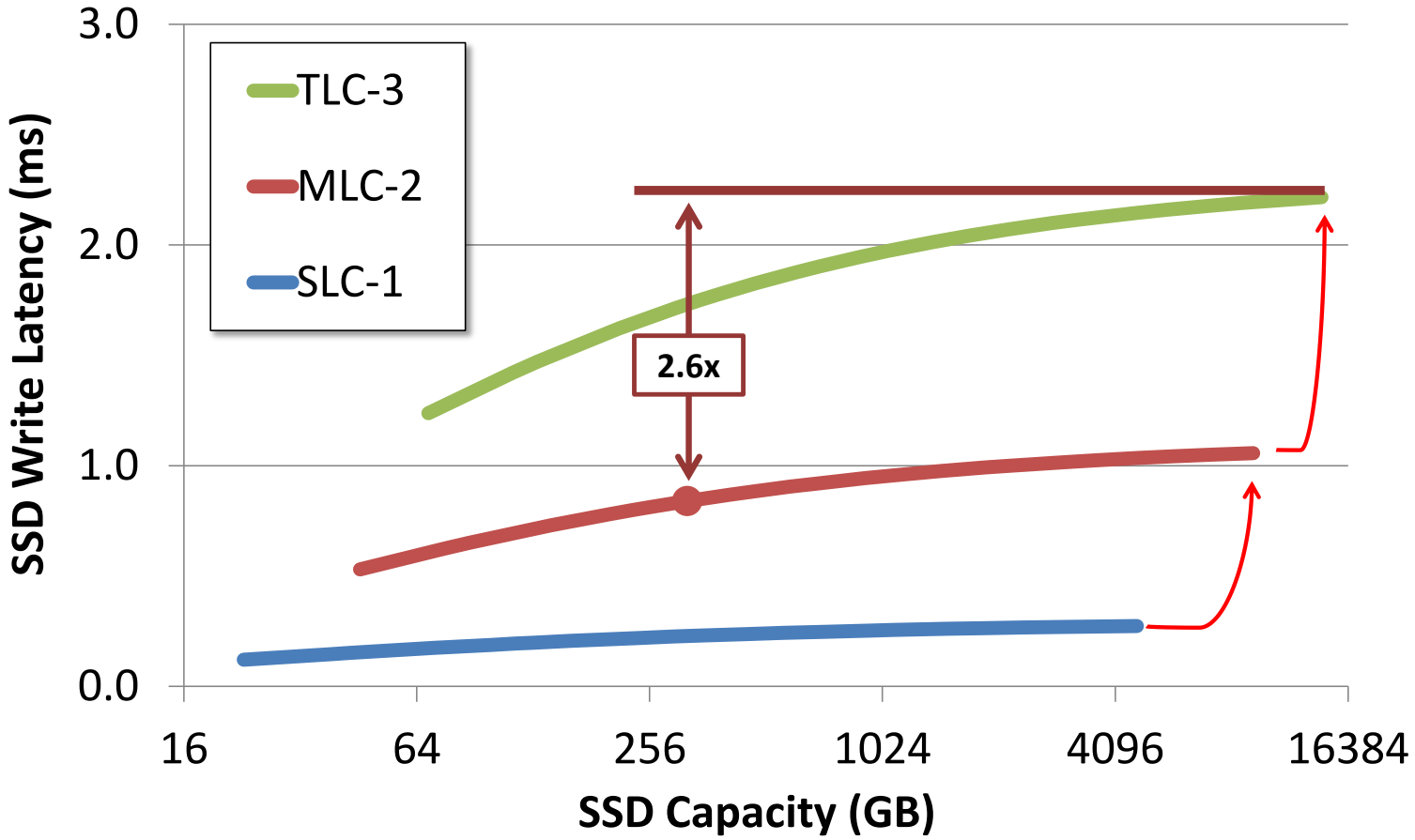
Empirical Data



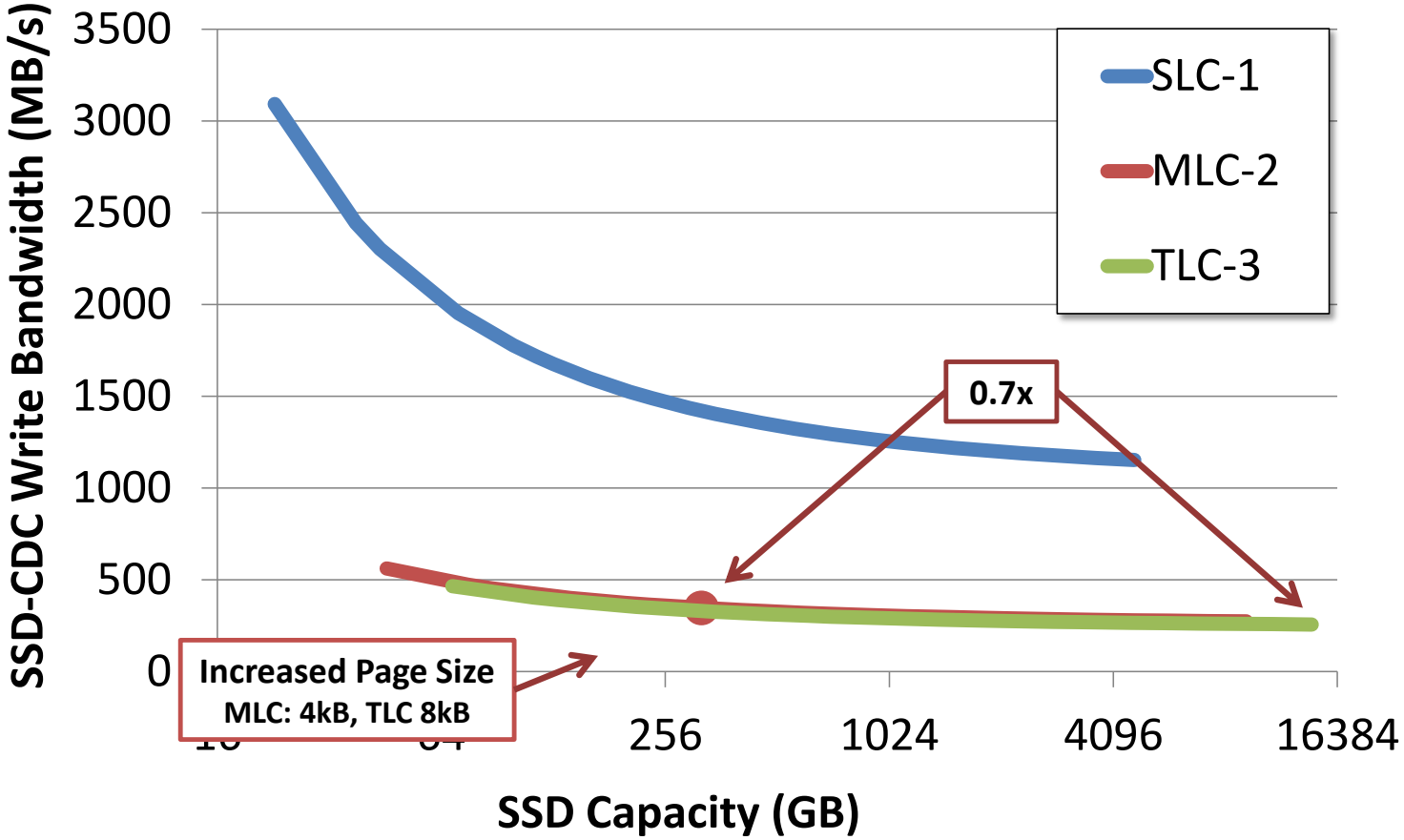
Scaling Trends in Empirical Data



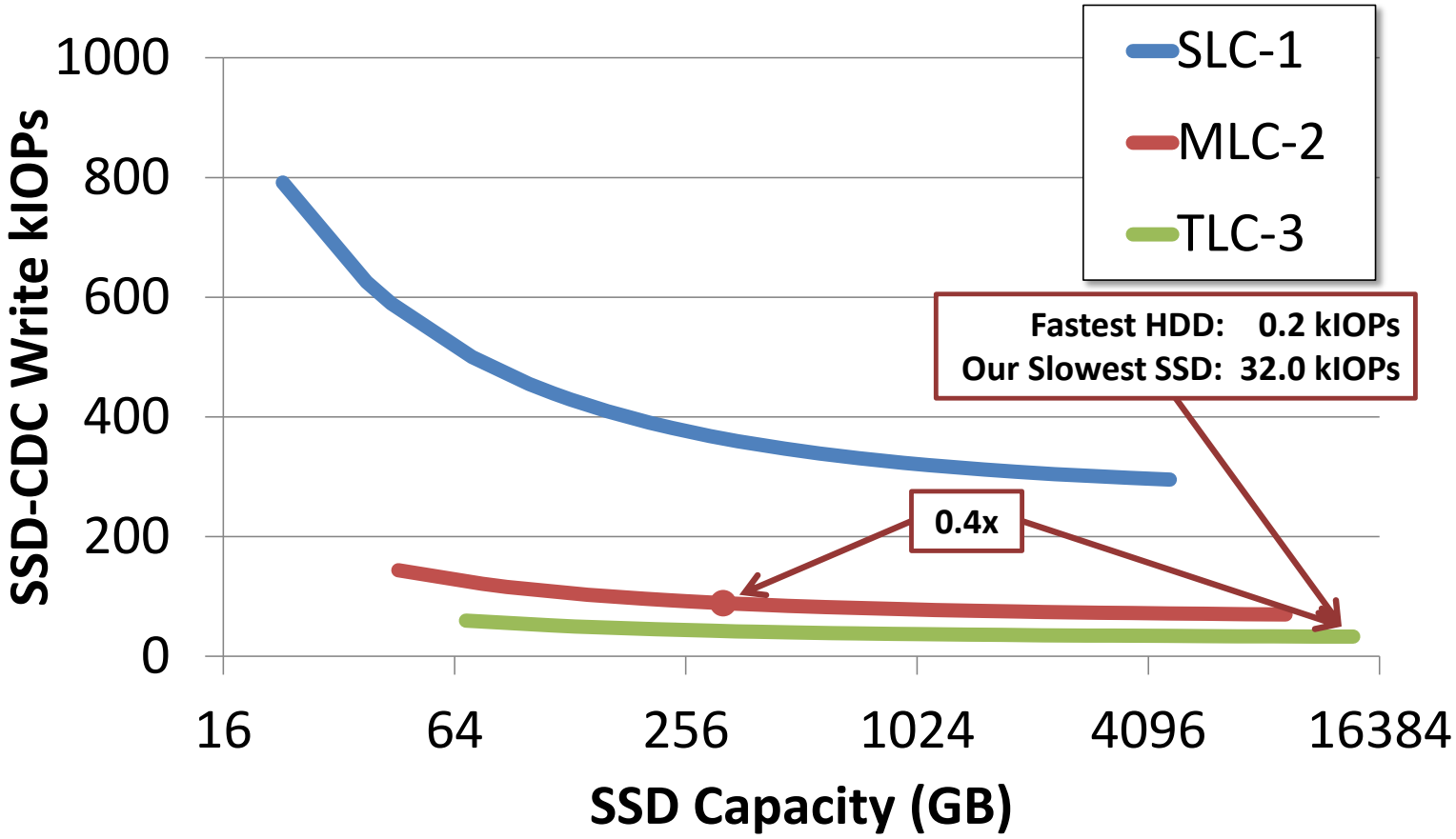
Write Latency of Fixed-Sized SSD



Reduced Bandwidth

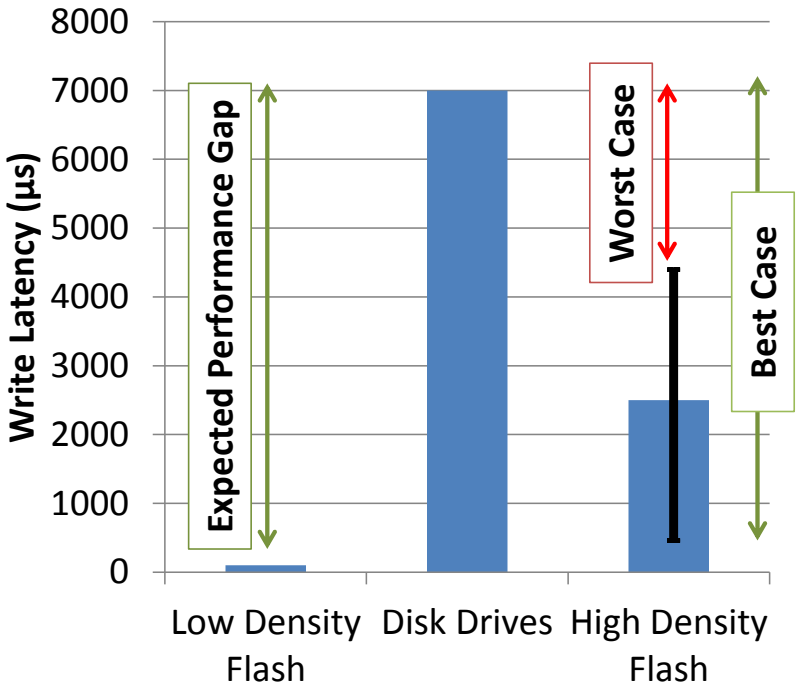


IOPs – 512B Random Accesses



SPECIALIZING FOR FLASH

High Density, Large Variation



Advantages:

- Two Distinct Performance Bins
- Equal Division of Pages Between Bins
- Regular pattern
- Consistent Between Vendors

Challenges:

- Pattern of fast and slow areas
- SSD management algorithms

Variation-Aware Interface

Extend Interface of Page-Mapped FTL

"High priority" == write to fast page

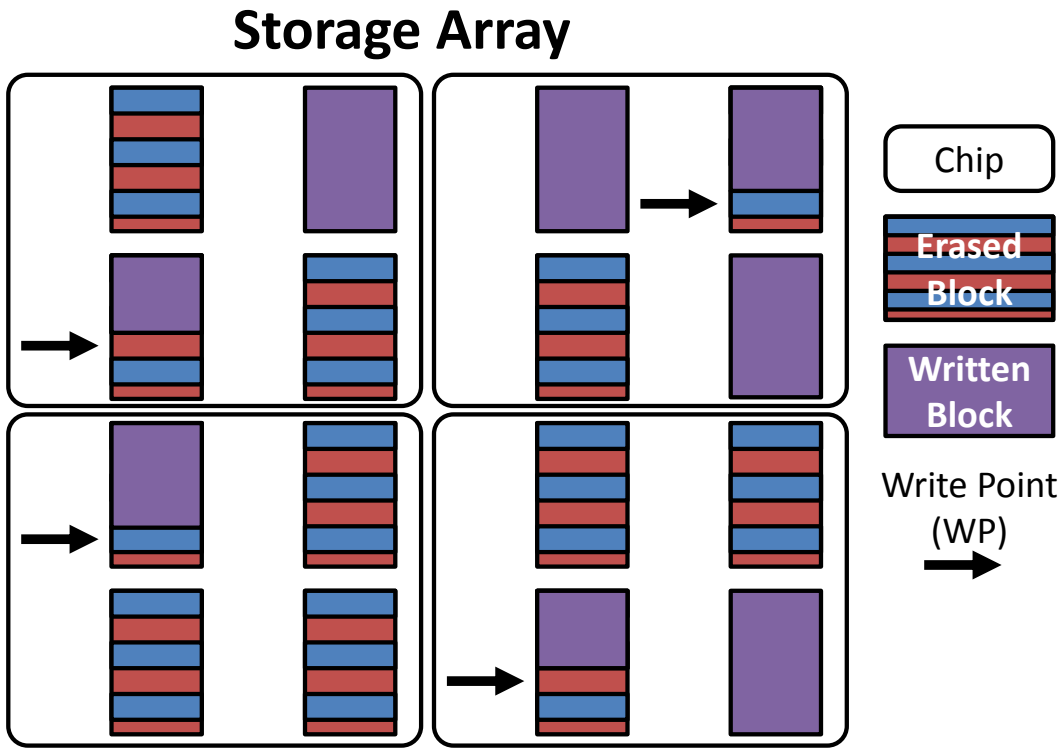
"Low priority" == write to any page

+ Lower latency when it matters
- Increase wear

Multi-Chip Variability Aware FTL:
Many-Write Points for Increased Flexibility

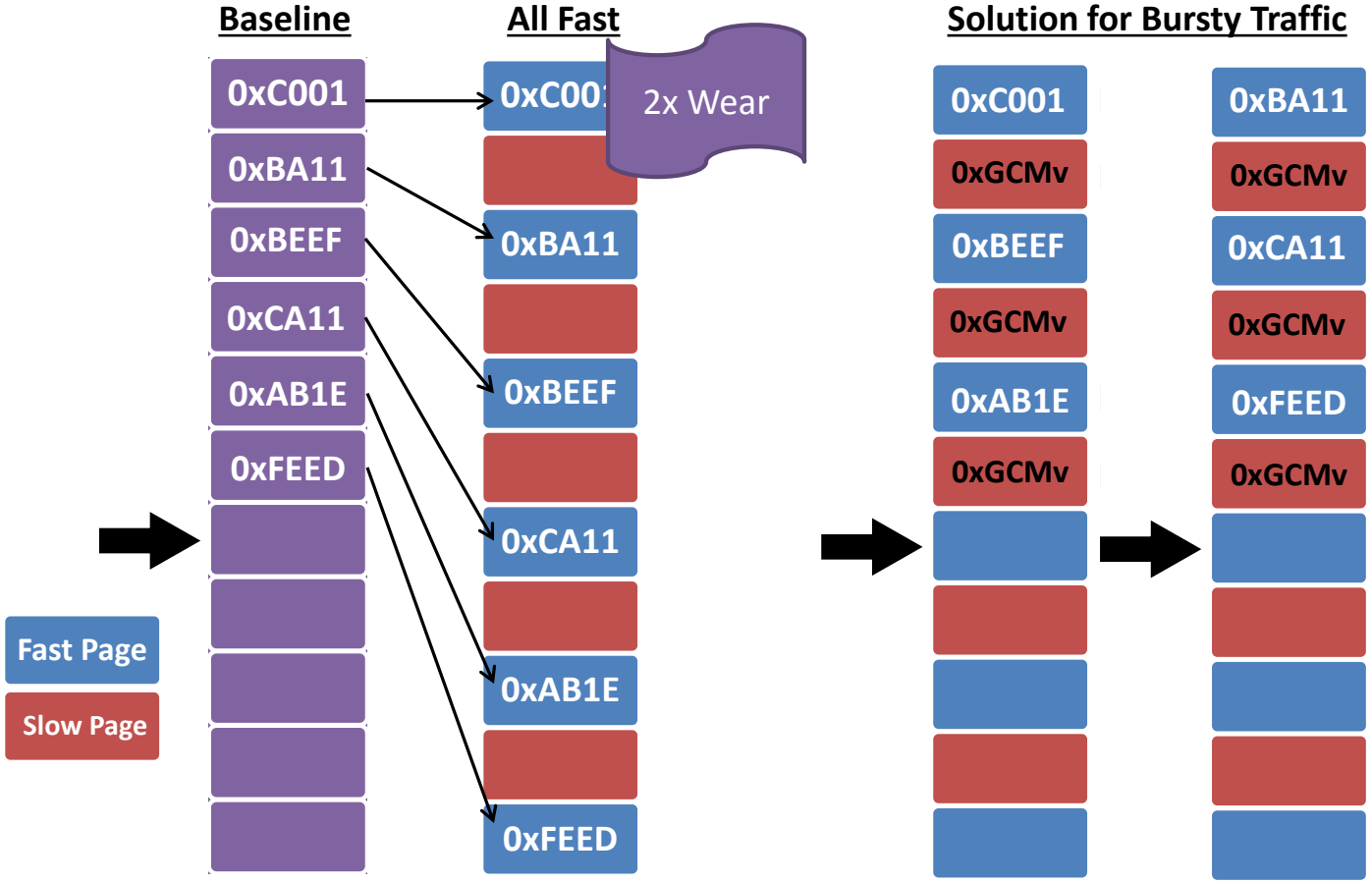
This Work:

- Leverage write variability in how WP is chosen

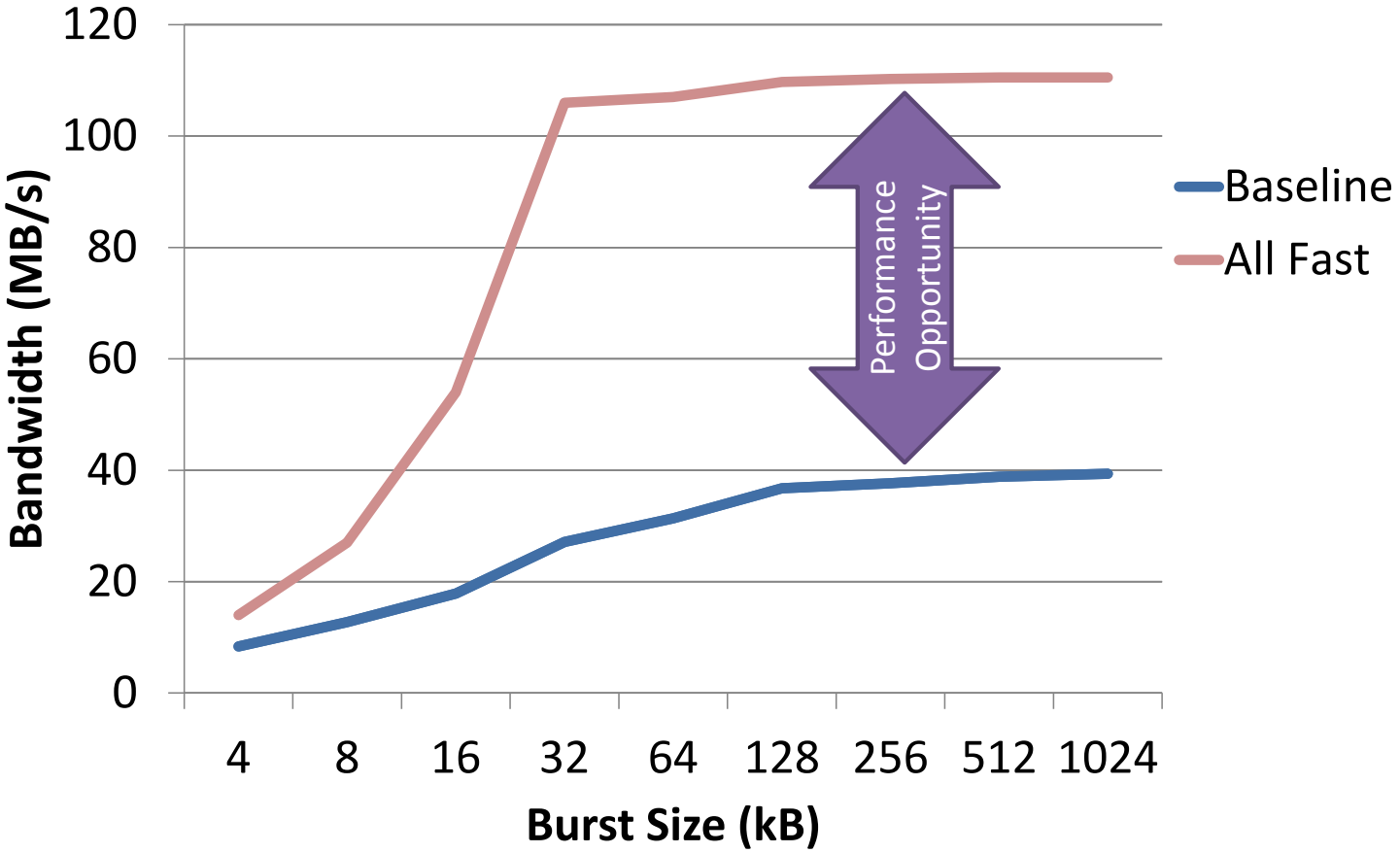


Slow Page Latency == 4.6x Fast Page Latency

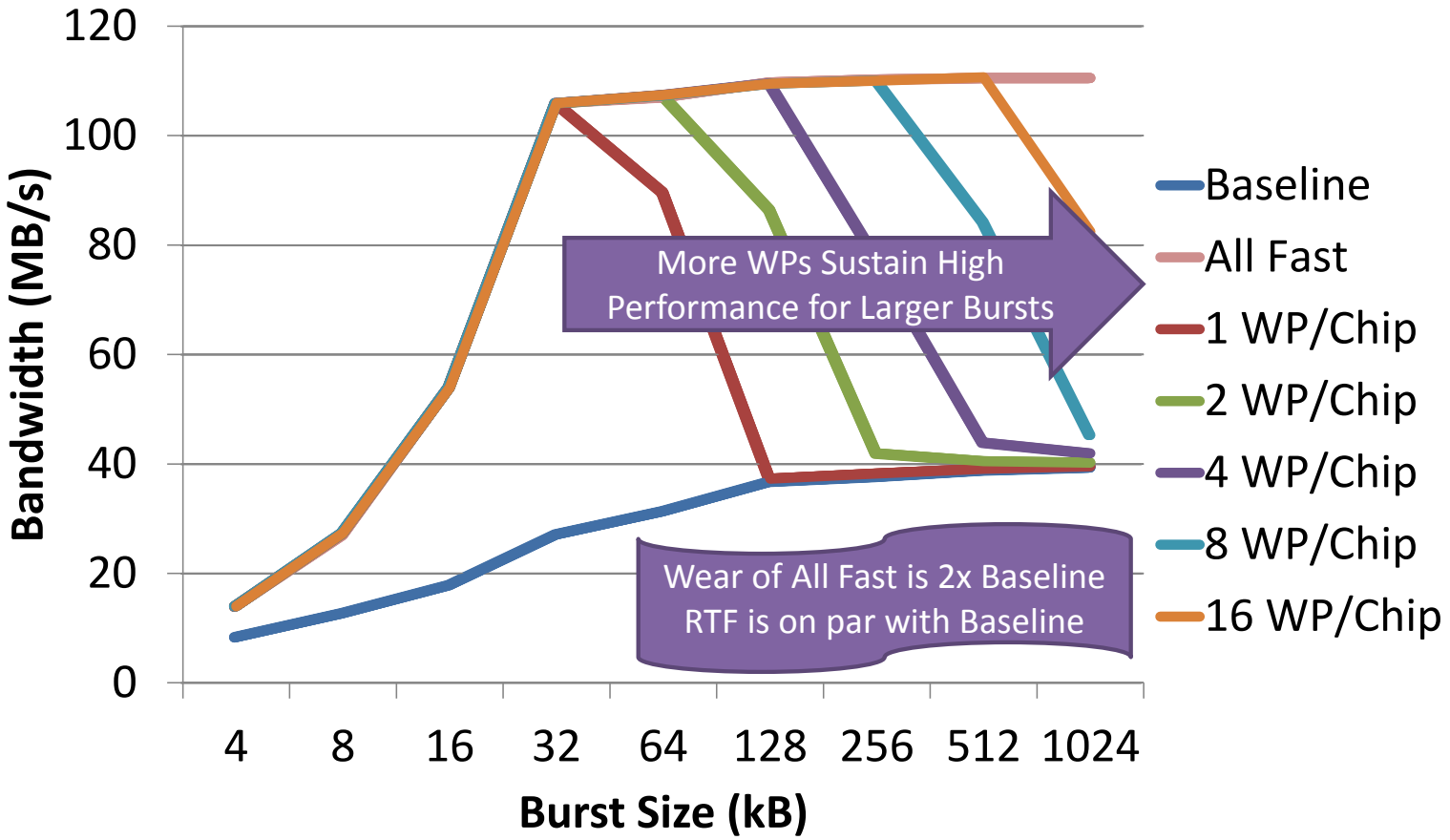
For Bursts: Return To Fast (RTF)



Results: Improved Performance



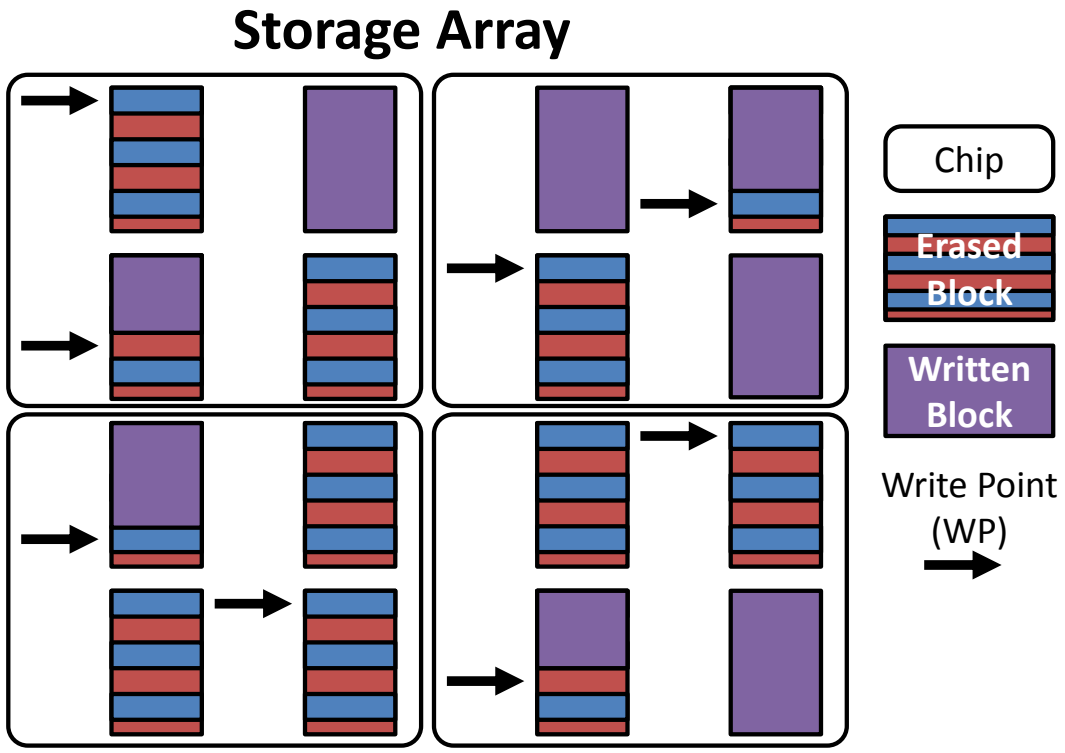
Results: Improved Performance



Multi-Chip Variability Aware FTL:
Many-Write Points for Increased Flexibility

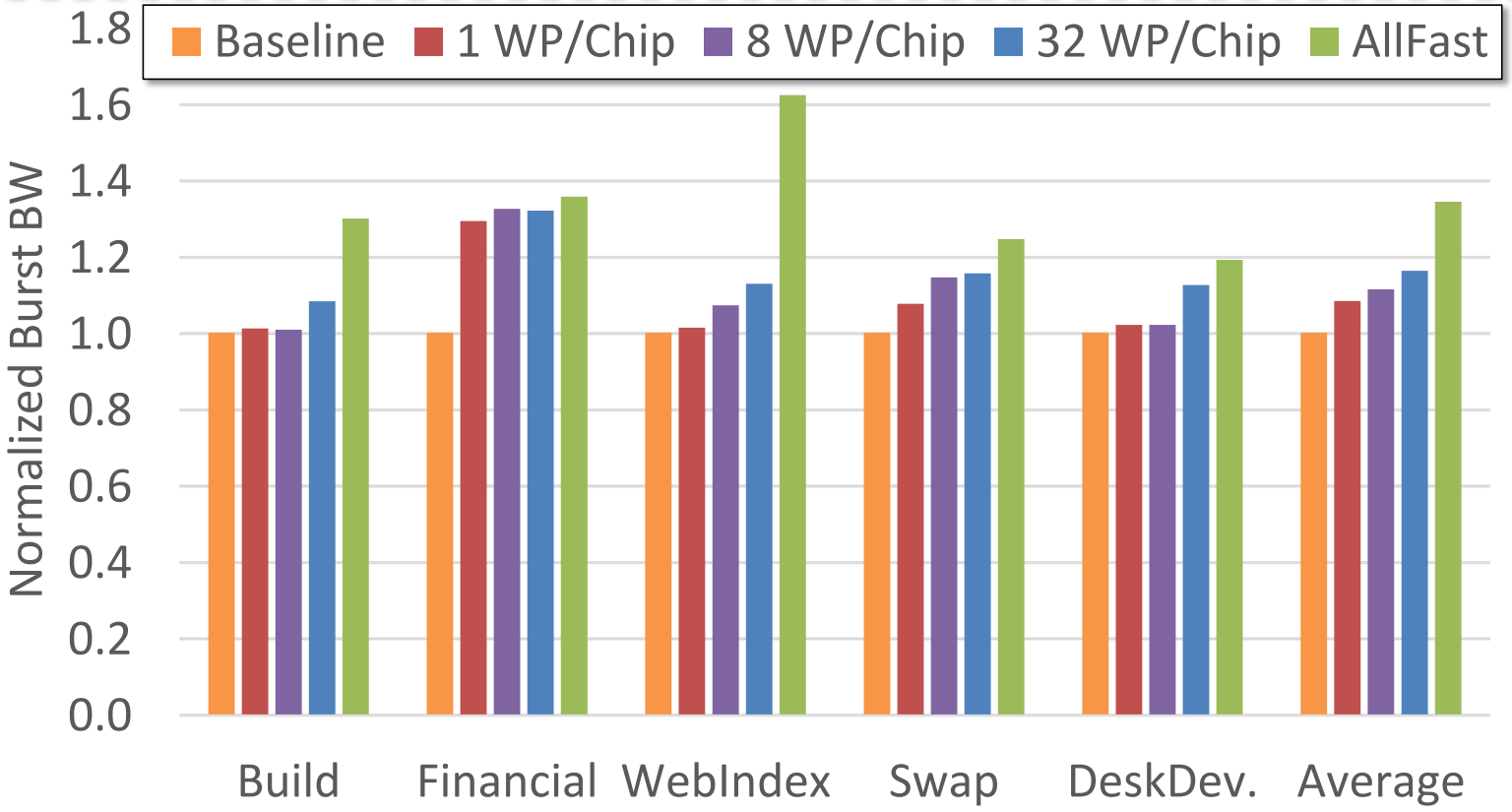
This Work:

- Leverage write variability in how WP is chosen
- Technique for coping with alternating page latencies: More Write Points (WPs)

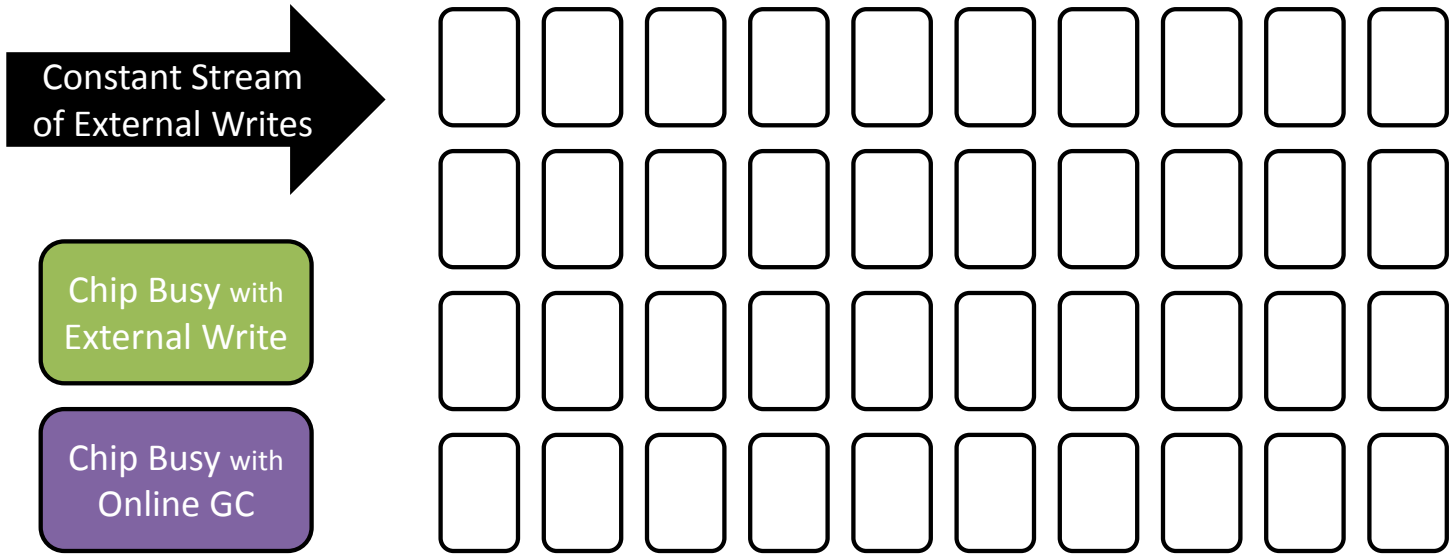


Slow Page Latency == 4.6x Fast Page Latency

Workloads


Baseline
1 WP/Chip
8 WP/Chip
32 WP/Chip
AllFast
Normalized Burst BW
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Build
Financial
WebIndex
Swap
DeskDev.
Average

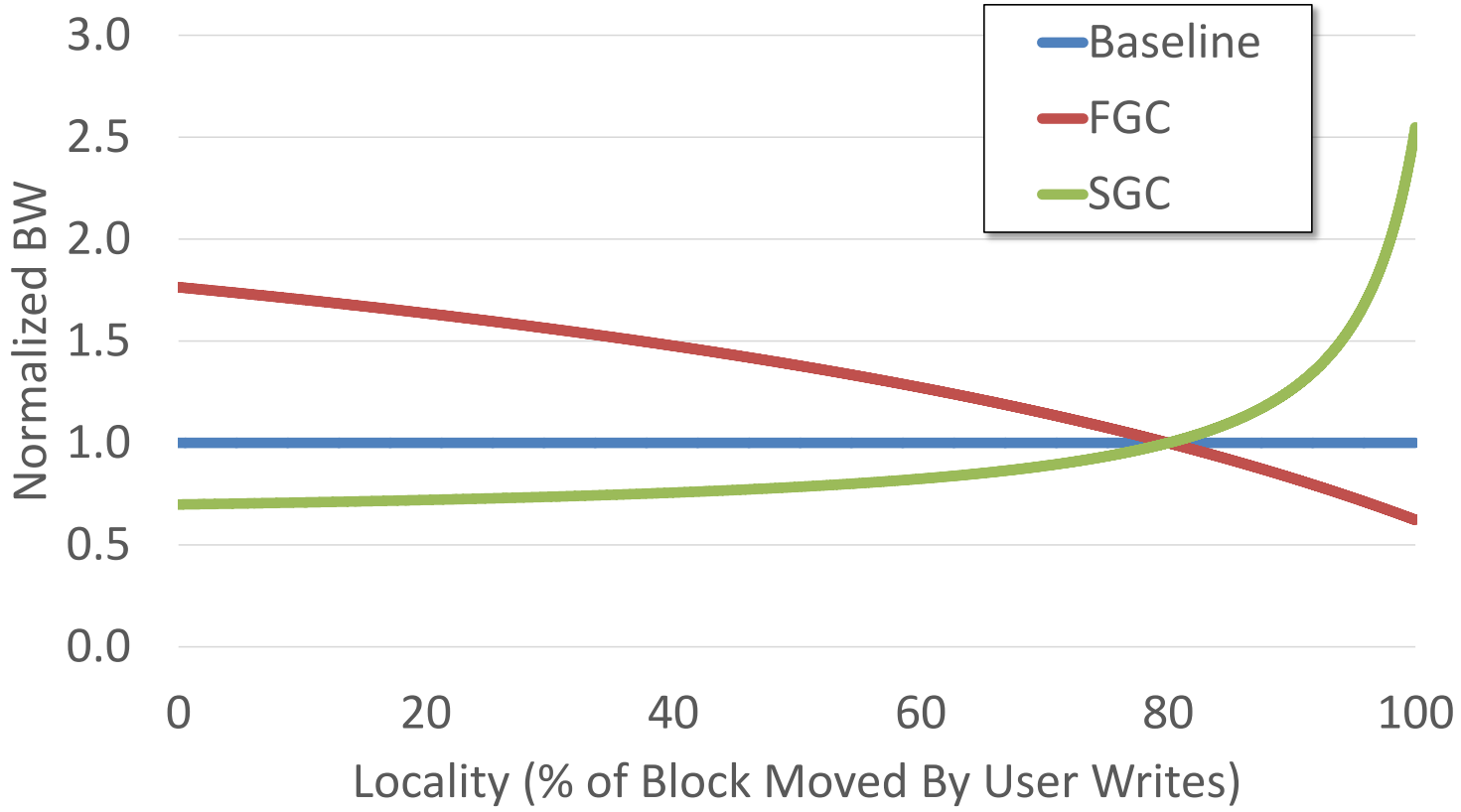
FTL under sustained write load



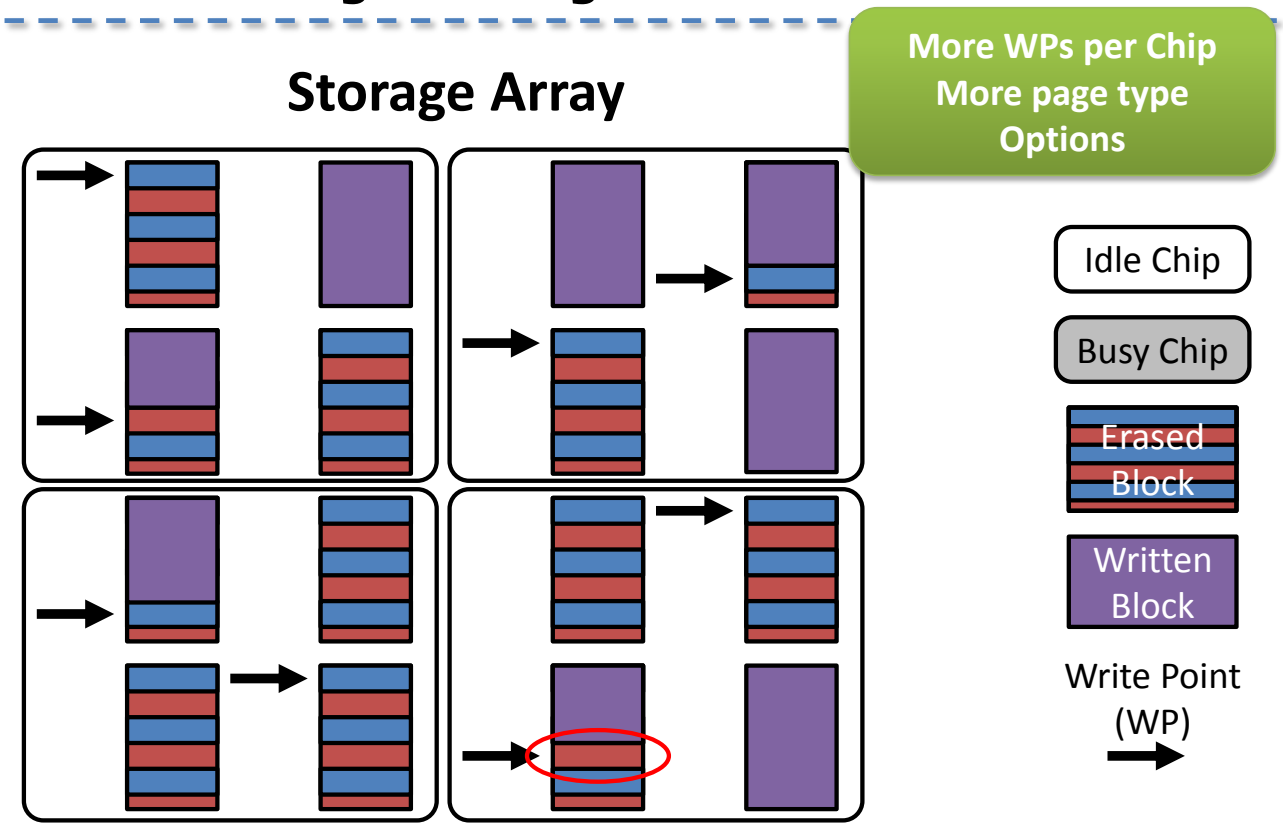
Requirement: Match rates of external writes and GC

Variability: Which page type for external and GC writes?

Locality-Dependent Choice

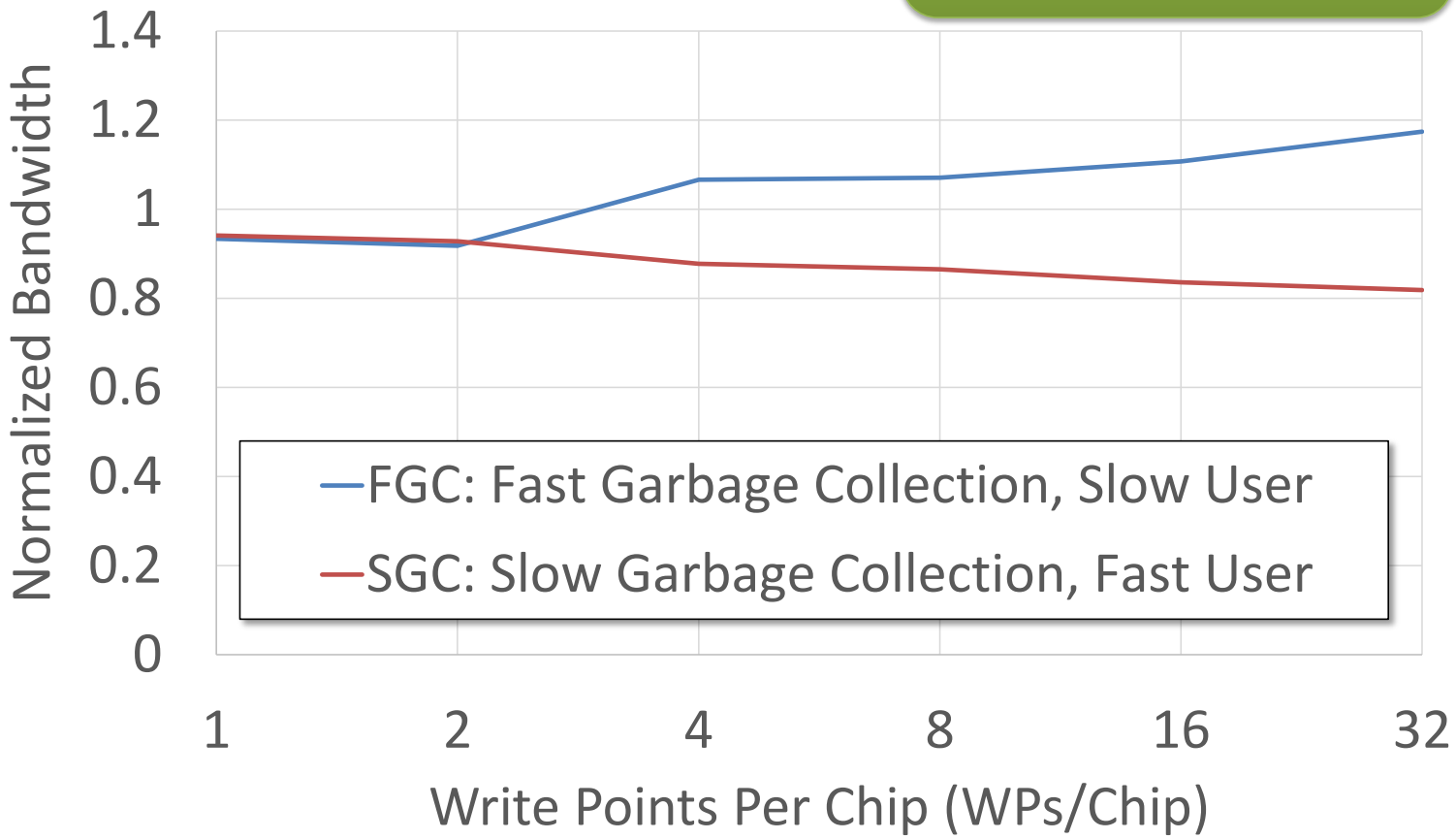


Page Choice in Busy Array

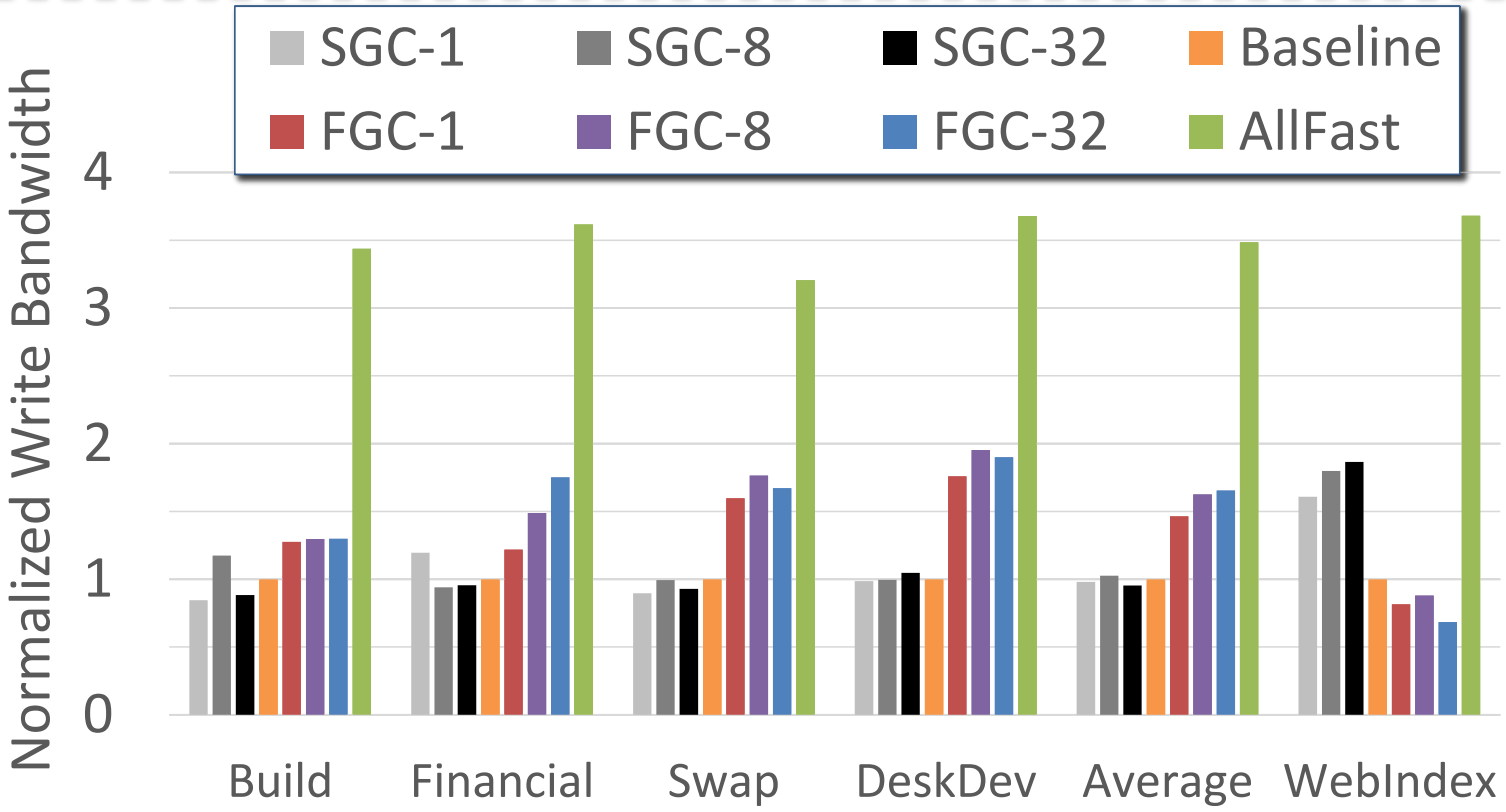


Write Points' Affect on FGC and SGC

FGC and SGC affects amplify with more write points per chip



Workload Performance



Flash Conclusion

- Chip Characterization:
  Revealing a New Landscape to the NAND Flash Community

- Leveraging large-scale flash drives:
  Write point per chip for parallelism

- Flash Trends:
  Capacity: 43x, Latency: 2.6x, Bandwidth: 40%

- Chip & Application Symbiosis: SLC performance in MLC parts
  - > 1 write point per chip
  - Bursts: up to 100% (89% on average) of SLC performance
  - Sustained: up to 95% (65% on average) performance improvement

QUESTIONS?

john.davis@gmail.com