Future Compute Memory
Non Volatile Memory (NVM) in Compute

Al Fazio

Intel Fellow

Director, Memory technology Development

November 12, 2008

Outline

Motivations for NVM in Compute

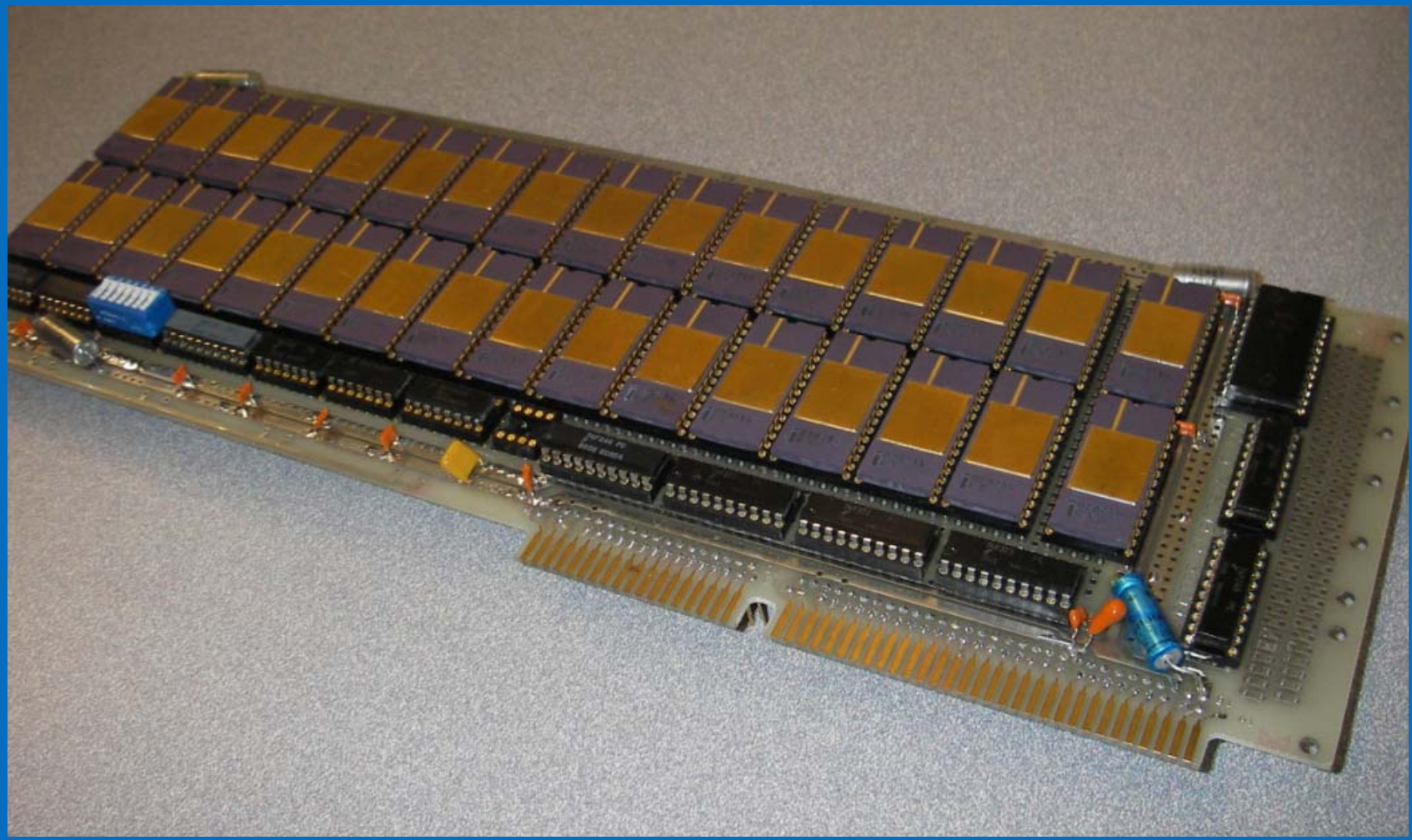
Key Principles of NAND Flash Operation & Device Physics from a compute applications viewpoint

Memory Controller Architecture

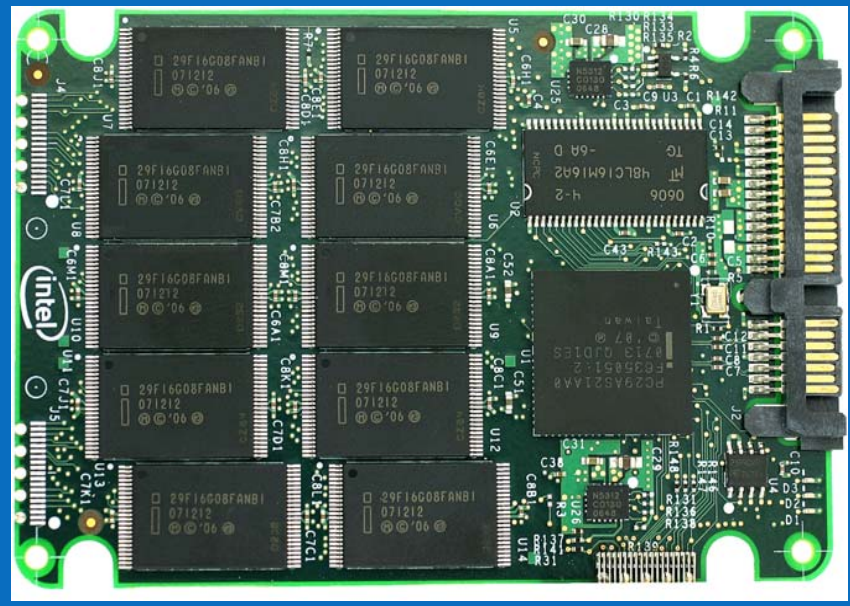
NVM Impact on Compute Applications

Future NVM Technology Trends

1987 View of NVM in Compute



2008 View of NVM in Compute



Form Factor: 2.5”/ 1.8” Standard SATA 3Gb/s

Performance

- World Class SATA I/O Performance
- X2S-E (SLC) Throughput
  - Sustained R/W: 240 / 170MB/s
  - Active (avg): 2.4W; Idle: 0.06W
- X25-M/X18-M (MLC) Throughput
  - Sustained R/W: 240 / 70MB/s
  - Active (avg): 0.25W; Idle: 0.06W

A Full Range of NVM in Compute

Intel® X25-E Extreme
SATA Solid-State Drive

Intel® X25-M and X18-M
Mainstream SATA Solid-State Drive

Intel®
Turbo Memory

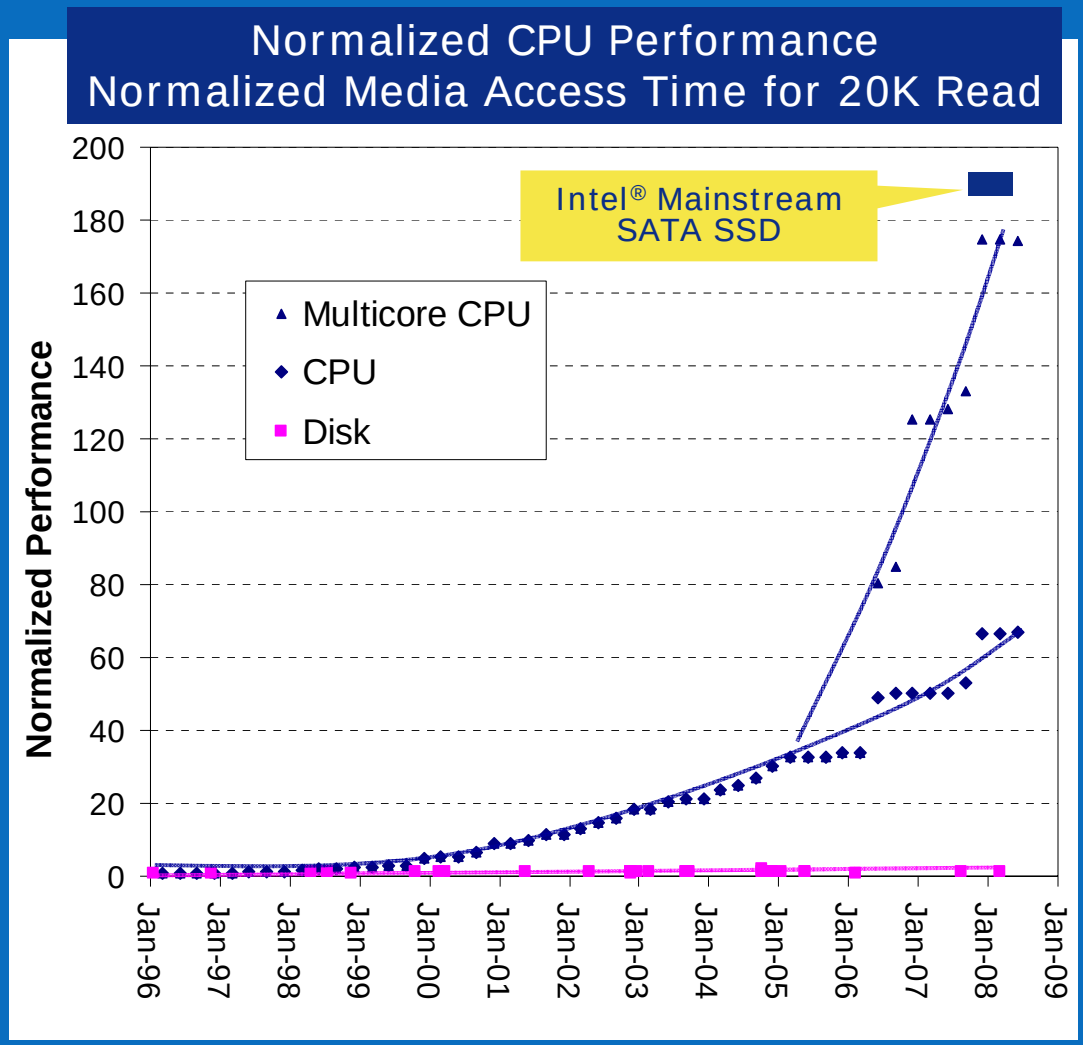
Intel® Z-P230 PATA
Solid-State Drive

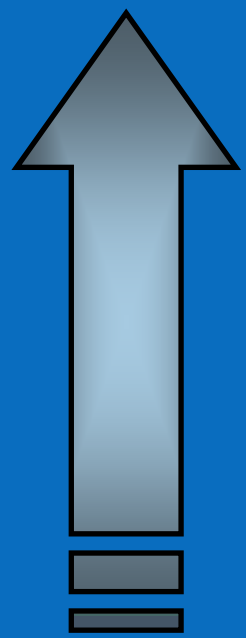
Intel® Z-P140 PATA
Solid-State Drive

Intel® Z-U130 USB
Solid-State Drive

Motivation for NVM in compute:

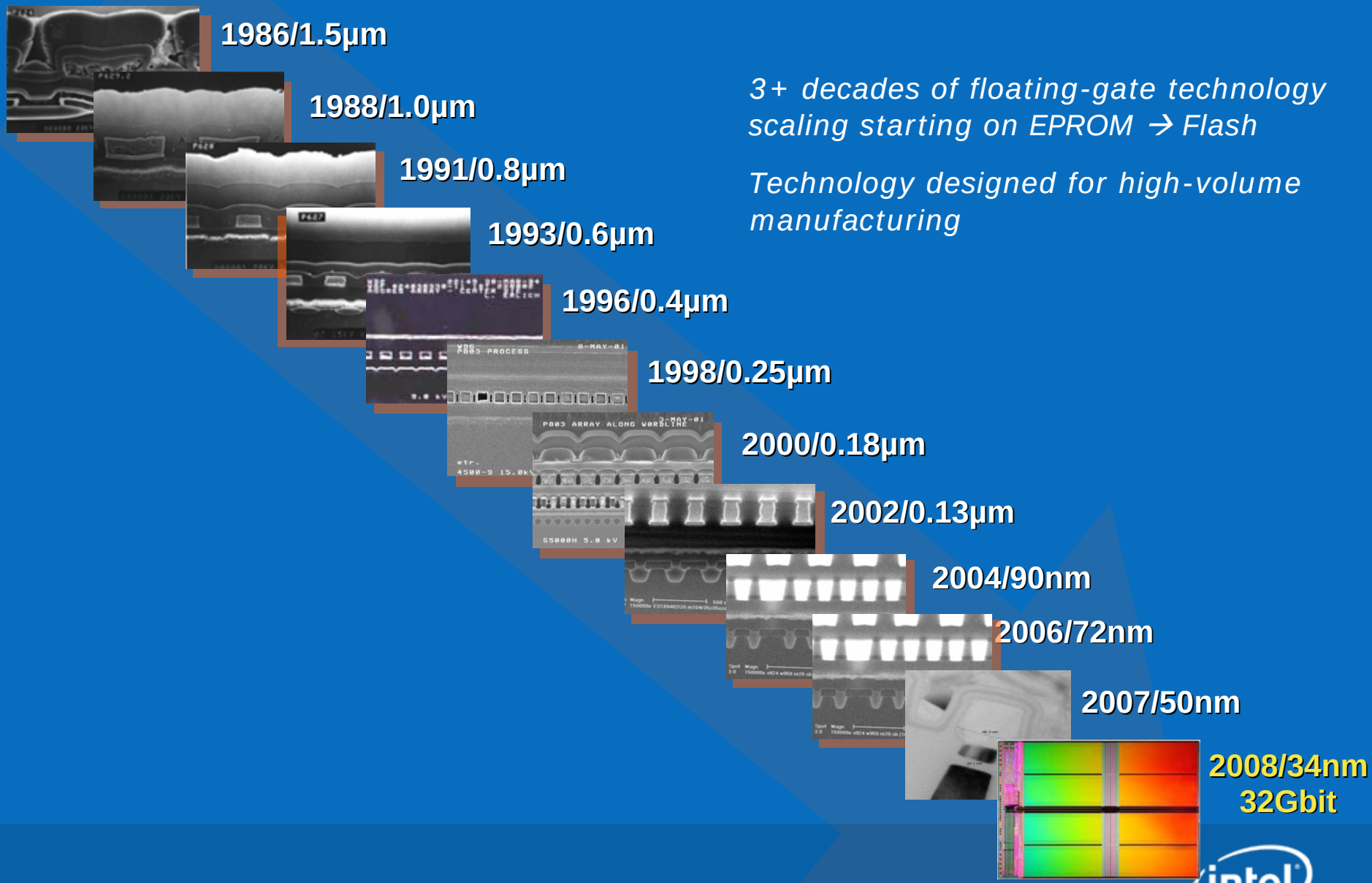
Huge Scaling Discrepancy Between CPU and HDD





1.3X vs 175X in 13 years!

20+ Years Flash Floating Gate Technology



3+ decades of floating-gate technology scaling starting on EPROM → Flash

Technology designed for high-volume manufacturing

Source: Intel

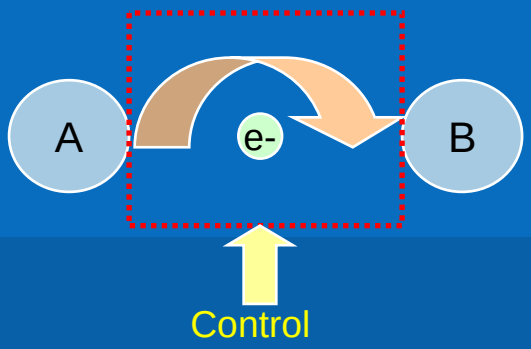
Flash: A License to Disrupt

35mm film, Floppy drives, audio tape…

- Flash use in consumer electronics characterized by:
  - Large block files (.jpg, mp3…)
  - # Writes determined by human interaction (i.e. photos taken)

To disrupt HDD, flash must accommodate compute characteristics:

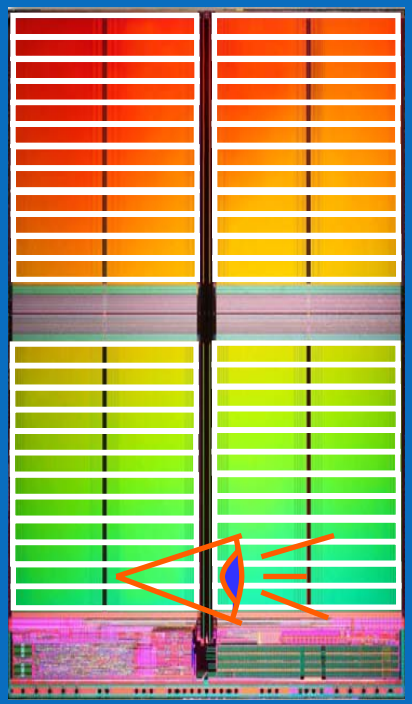
- Small random writes, # writes determine by OS
- Add to this:

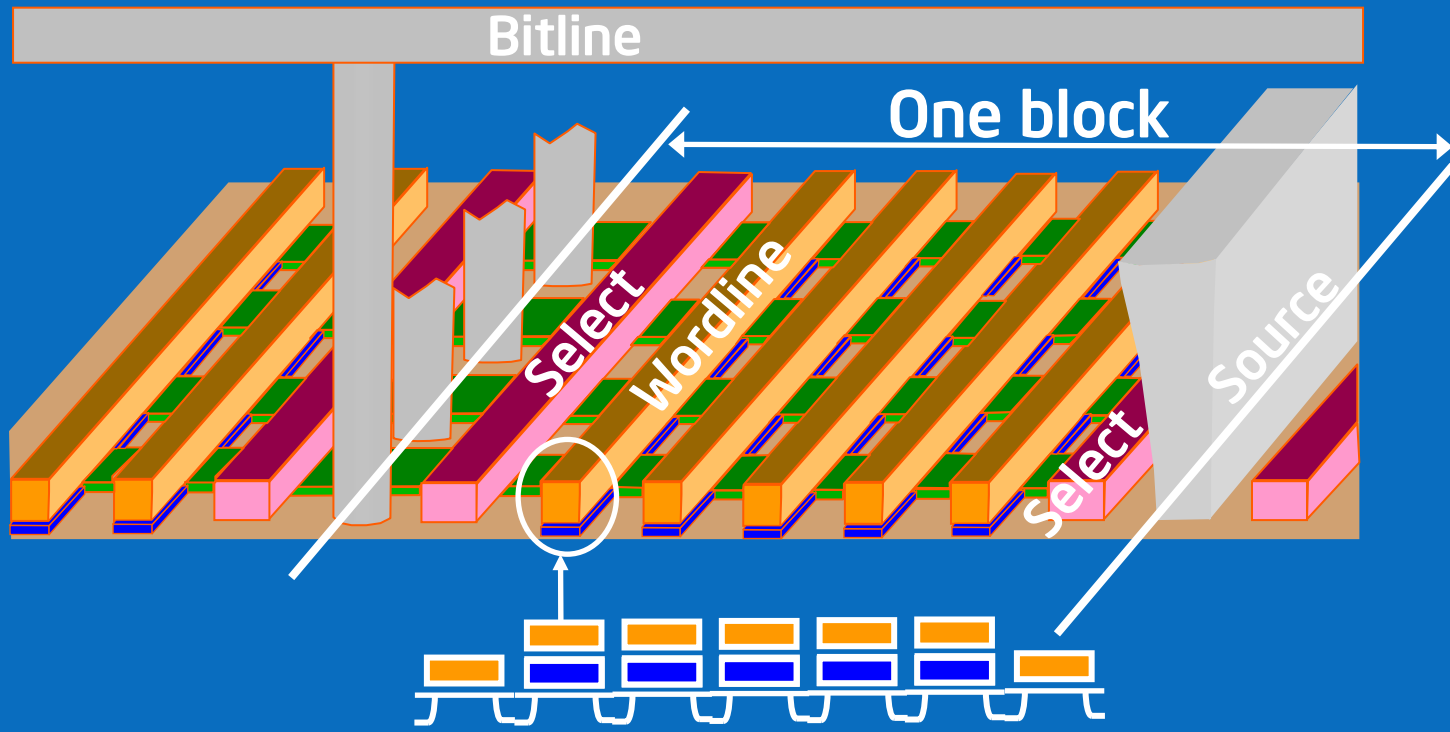


Flash requires high fields to overcome energy barriers for non-volatility

→ Flash reliability dominated by oxide-degradation; result of program/erase

NAND Physical Organization





A block is a sea of cells arranged in a grid

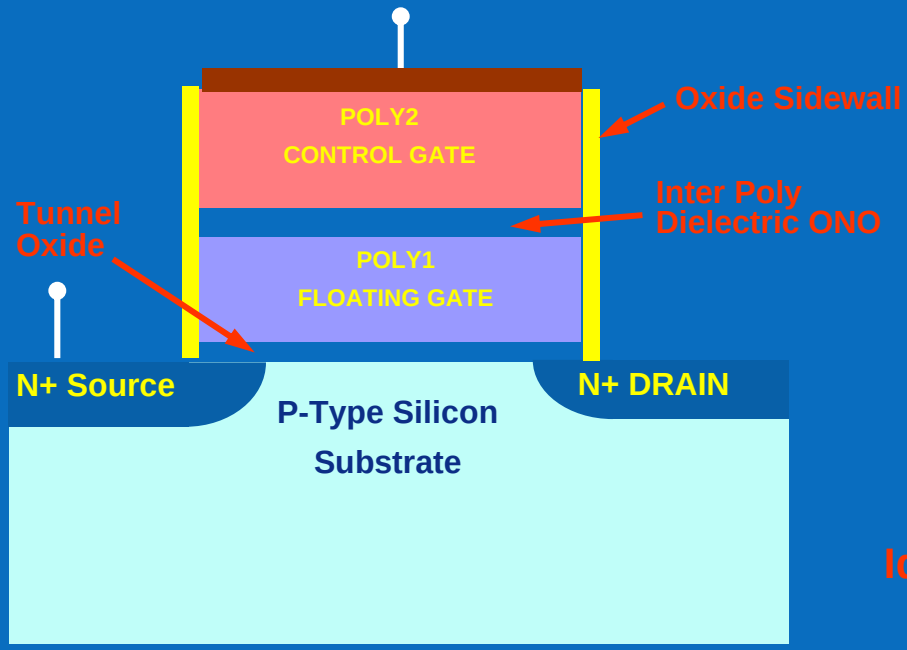
TG's are connected in wordlines (typ. 32, only 5 shown)

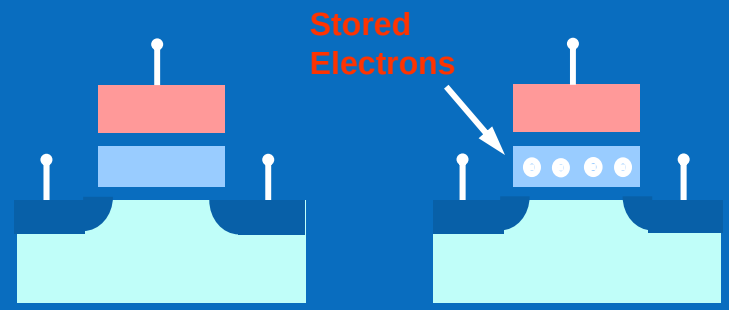
Cells in different wordlines are strung together in series

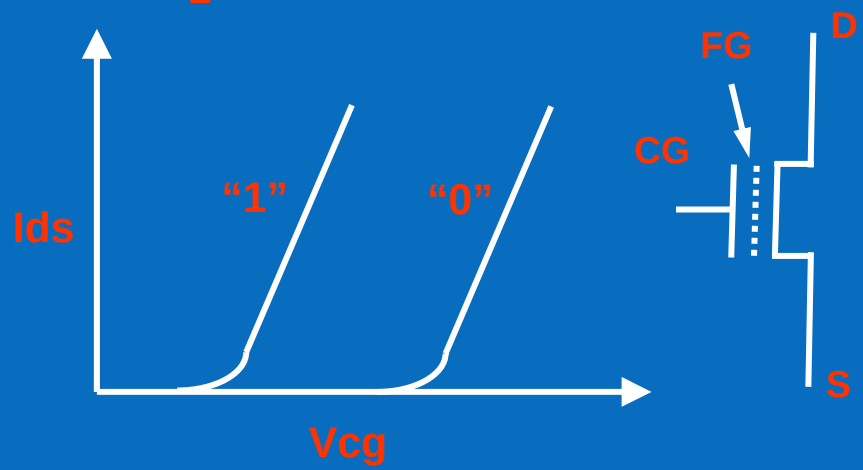
Each string of cells is connected to a bitline at one, source at the other

Select devices control whether the block is connected to bitlines and source

Flash Cell Layout and Cross-Section



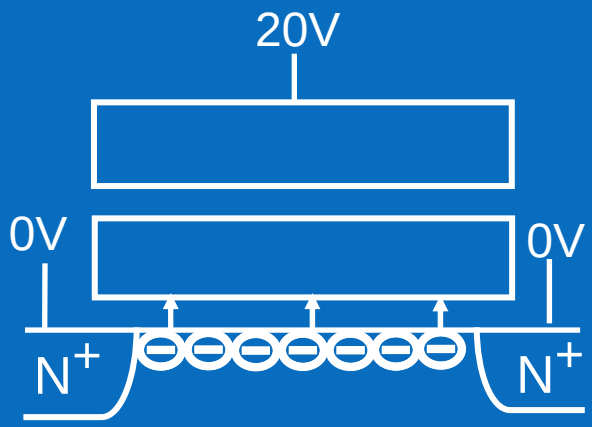




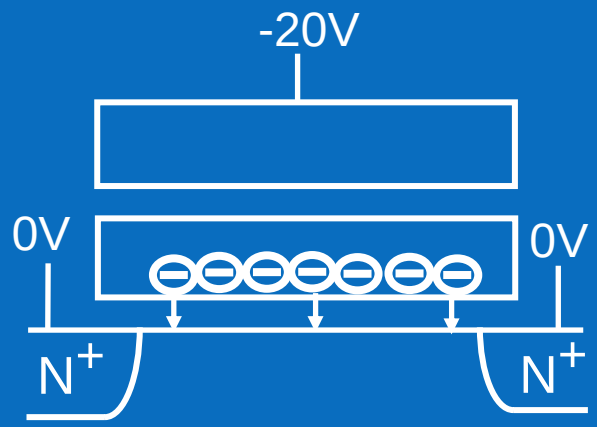
N-Channel MOSFET with a few distinguishing features:

- Isolated floating gate
- Charge storage on Floating gate modulates threshold voltage of underlying MOSFET

Charge Storage: Program and Erase



Programming: NAND



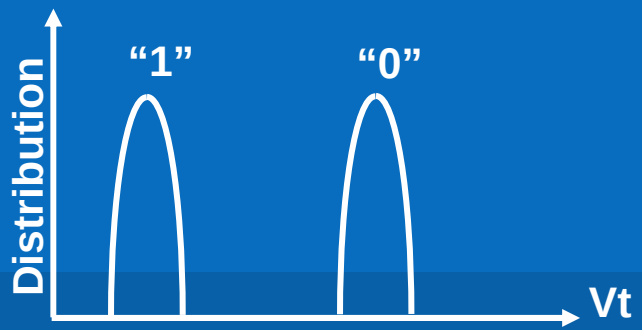
Erase

Programming means injecting electrons to the FG

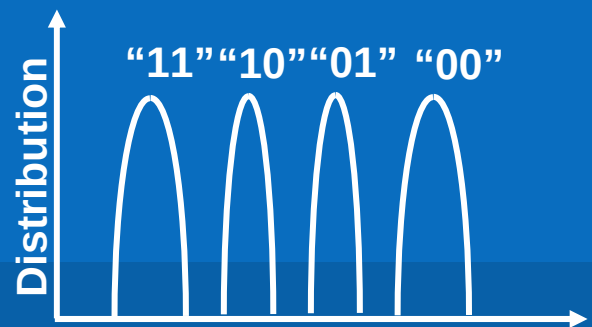
- Fowler-Nordheim Tunneling

Erase: Fowler-Nordheim Tunneling in reverse direction

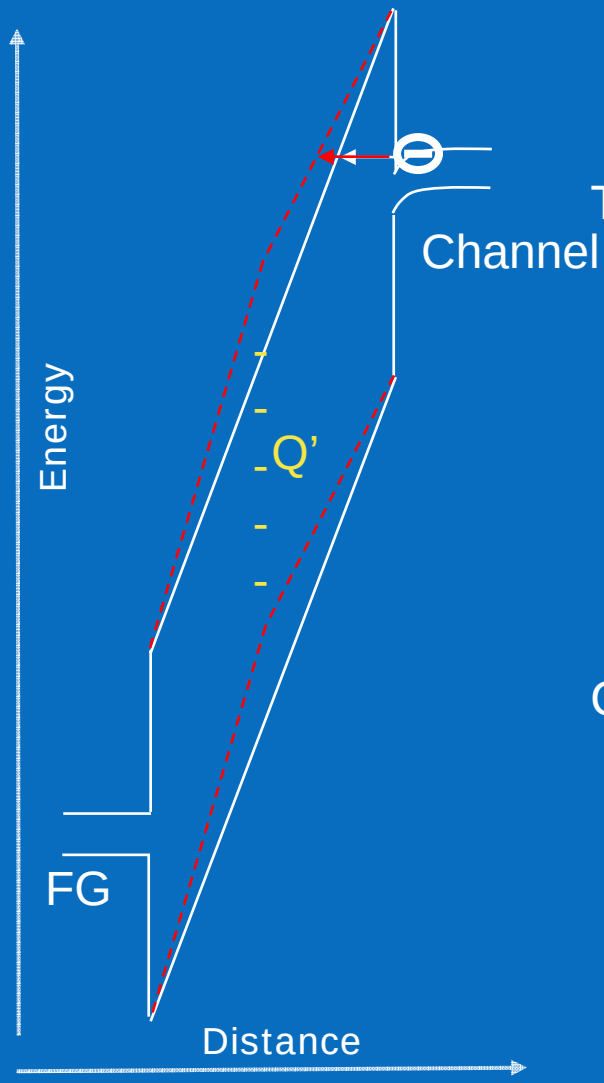
2 Levels => 1 bit/cell



4 Levels => 2 bit/cell



Reliability and Oxide Traps



Normally, F-N tunneling occur only during accelerated stresses done by engineers trying to study oxide degradation…

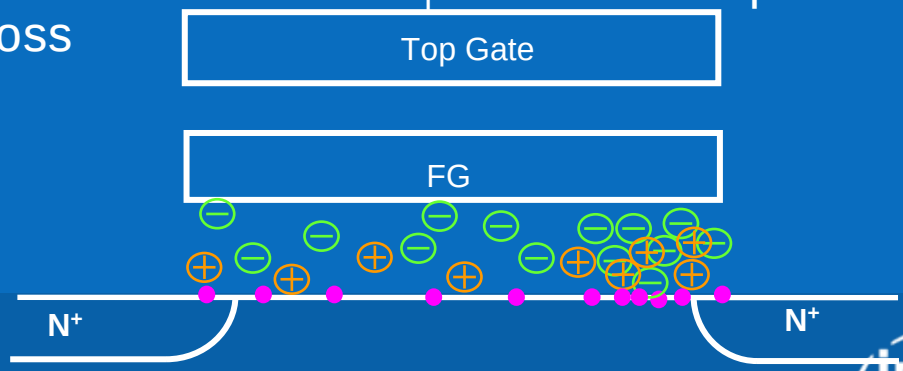
- Flash memories: basis device operation itself

This fact has two fundamental implications:

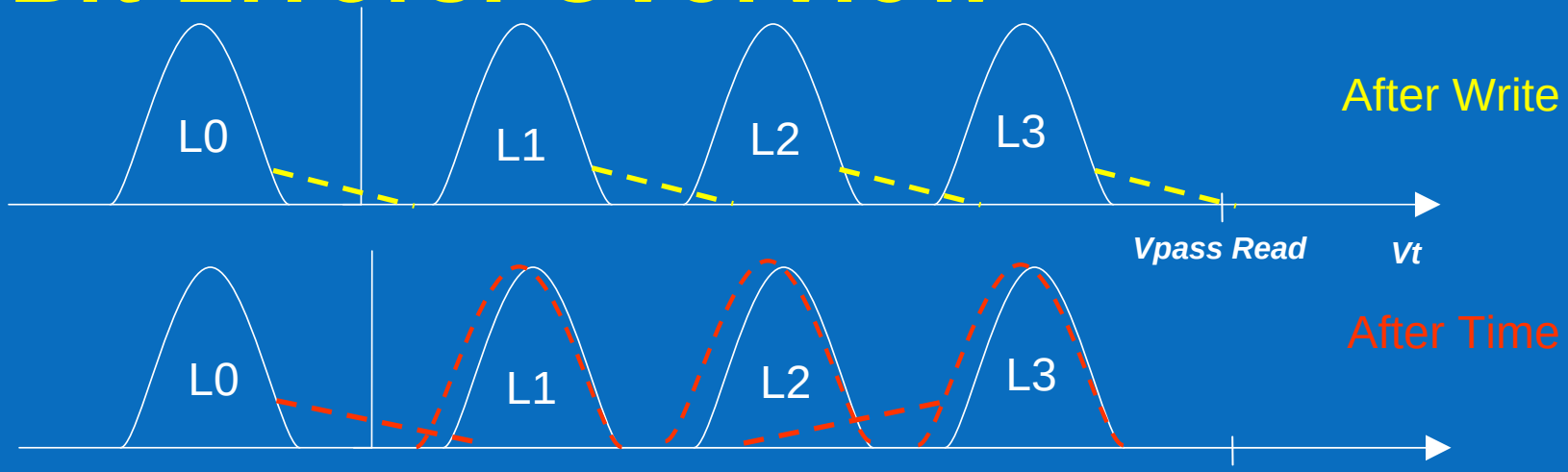
- Flash reliability is dominated by oxide-degradation effects, notably trap buildup in the tunnel oxide, which occur as a result of program/erase cycling
- More than any other IC technology, developing a Flash technology centers around obtaining acceptable reliability

Over time, charges can detrap

- Effect will cause V_T to shift and possible data loss



Bit Errors: Overview



At any instant, some fraction of bits are in the wrong data state, typically 1E-9 to 1E-6, called the “raw bit error rate” or RBER

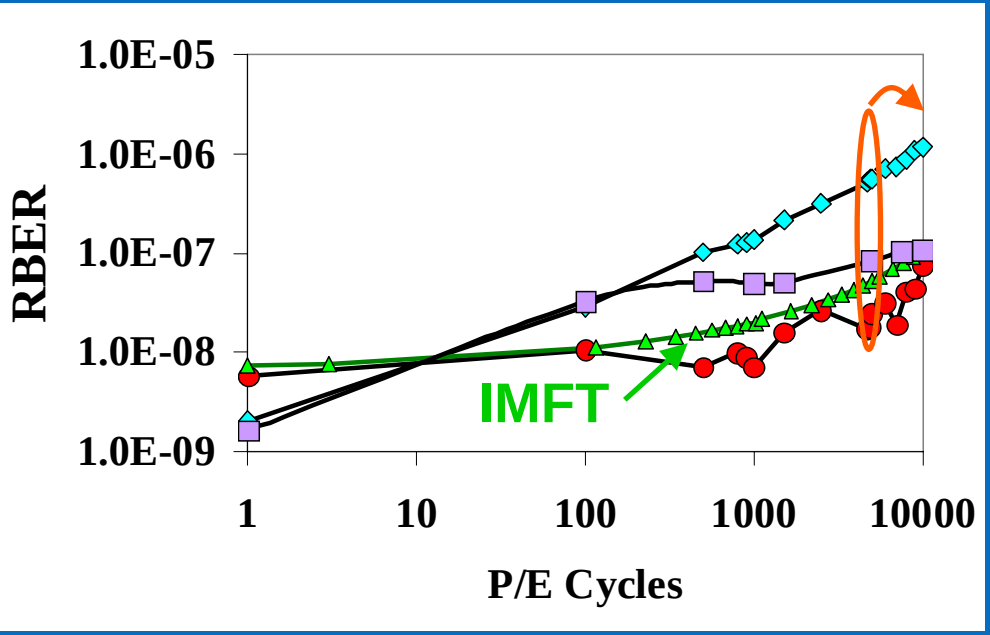
These failing bits *develop with use*

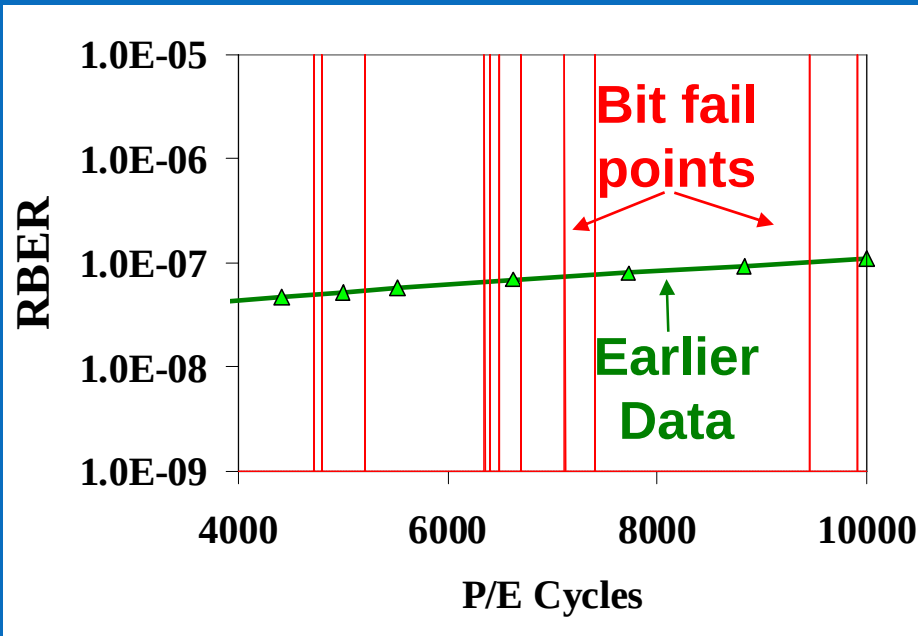
- During write, some bits program when they shouldn’t, or program higher than they should
- Cells shift in V_T over time, because of simply time (“data retention”) or of repetitive read operations (“read disturb”)
- Both kinds increase with more program/erase cycles
- Several mechanisms cause bit errors, each with its own dependence on cycles, time, temperature, etc.

This complexity means that RBER is a number, but not like pi:

- like temperature: a # for specific set of conditions, location, instant

Erratic Nature of Write Errors





Errors are erratic: Most bits failing at 5K didn't fail at 10K

Explanation: oxide traps are transient

Data verified only at symbols: did we miss errors in between?

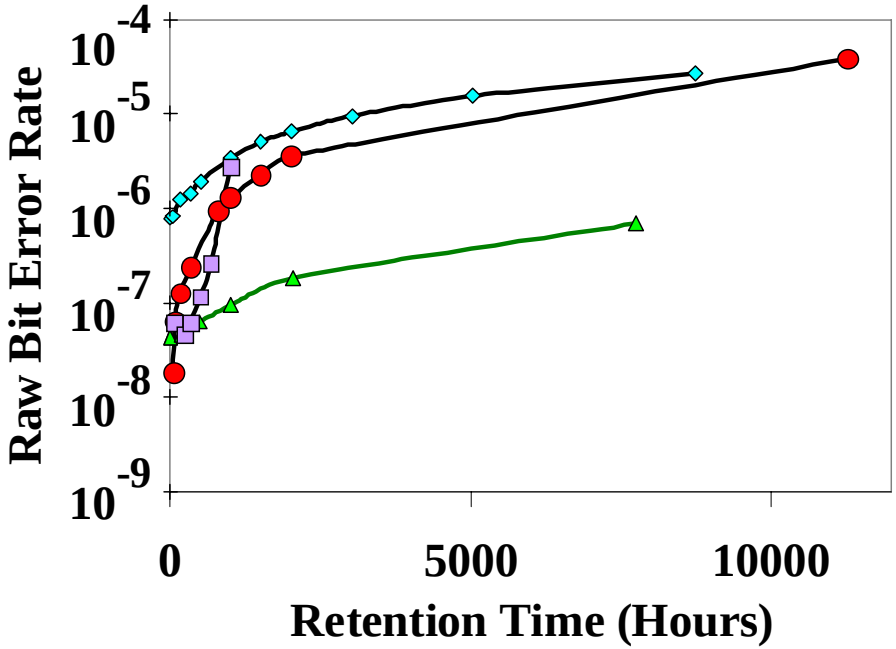
Ran experiment to verify data after every cycle

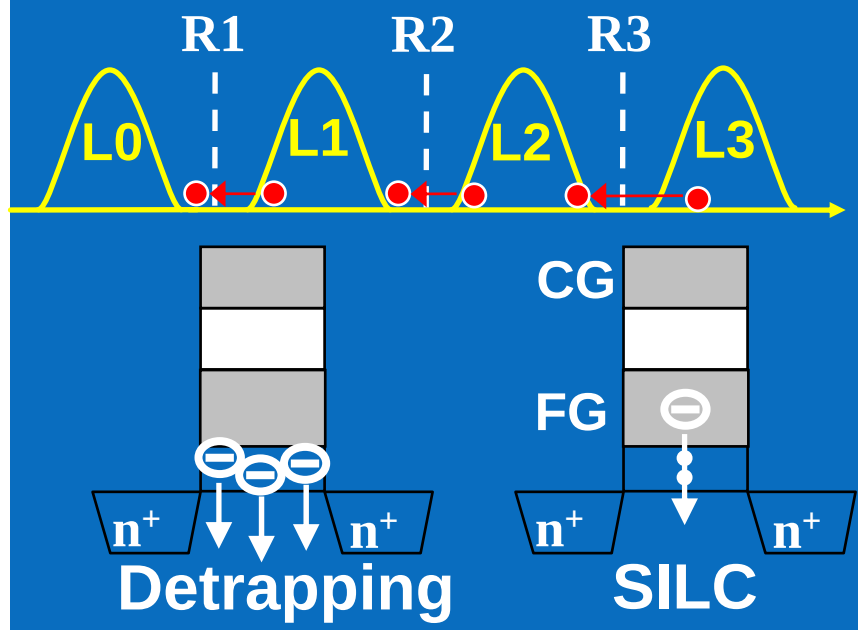
- Example bit failed 11 times, never at previous verify points
- Previous verifies detected only 0.6% of failing bits

Standard "test after stress" qualifications miss most errors!

Next Several Slides are based on 70nm results from: Mielke, N., et. al., "Bit error rate in NAND Flash memories", IEEE International Reliability Physics Symposium, 2008

Data-Retention Errors





After cycling, RBER increases over time without bias

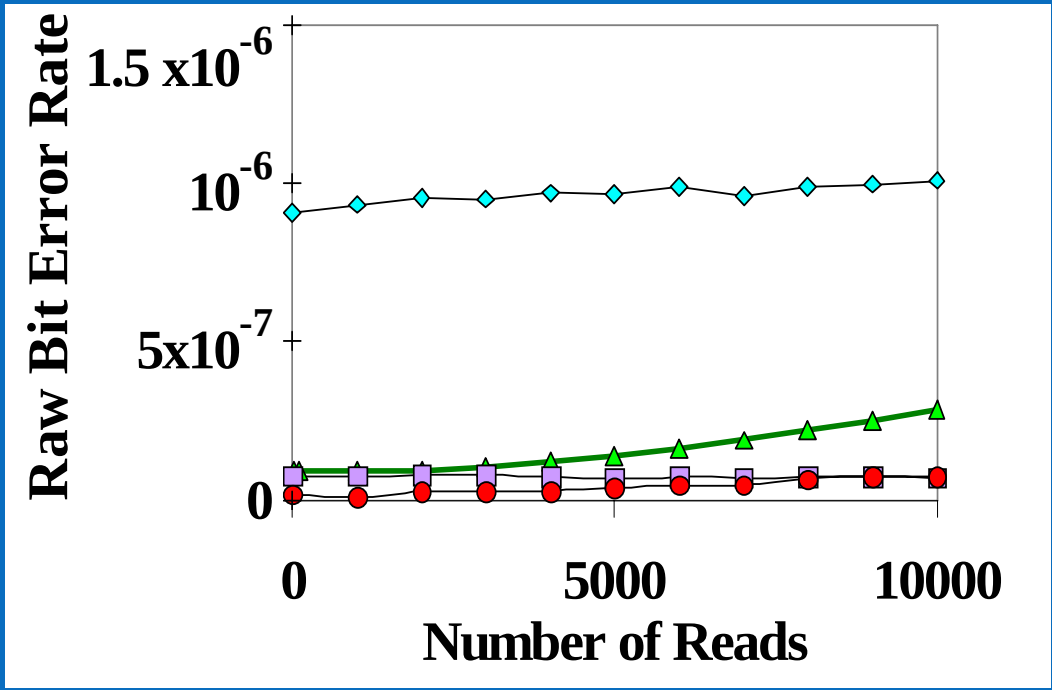
Error transitions show cells are losing V_T ("charge loss")

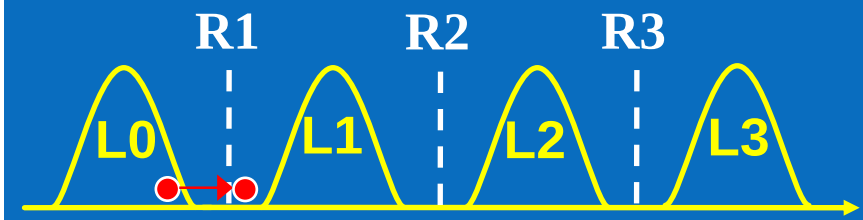
Two products dominated by upper state (L3), others by L1 & L2

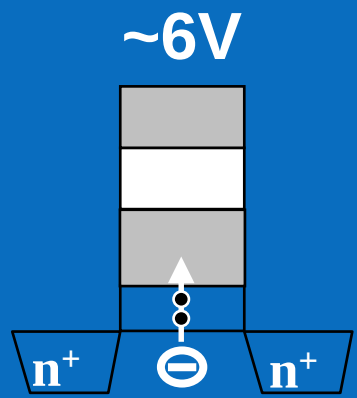
Characteristics:

- L1 & L2: Detrapping from the tunnel oxide
- L3: SILC (trap-assisted tunneling) leakage off FG

Read Disturb Errors







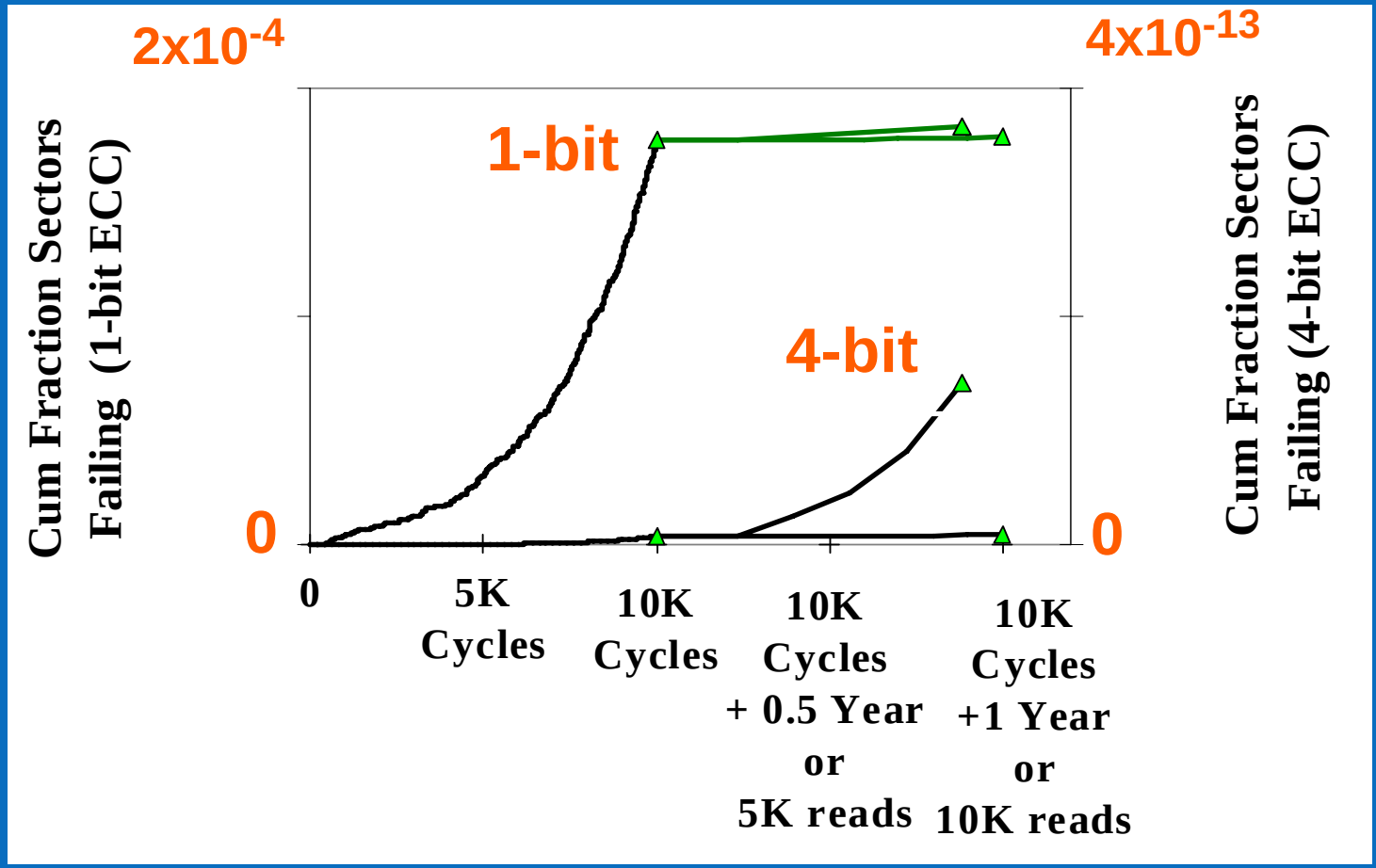
SILC

After cycling, RBER increases with repetitive reading

Error transitions show erased cells gaining V_T

Mechanism is well known: SILC under read bias

Effect of ECC



Failures drop several orders of magnitude, ~10^{12}x over no ECC

Curves get steeper (because of Ecc power law)

Dominant mechanism switches to retention (because of underlying error distribution)

Workable UBER Definition for NAND

UBER = Uncorrectable Bit Error Rate

$$\text{UBER} = \frac{\text{Cum Fraction Sectors Failing}}{(\text{bits per sector}) \cdot (\#\,\text{reads per sector})}$$

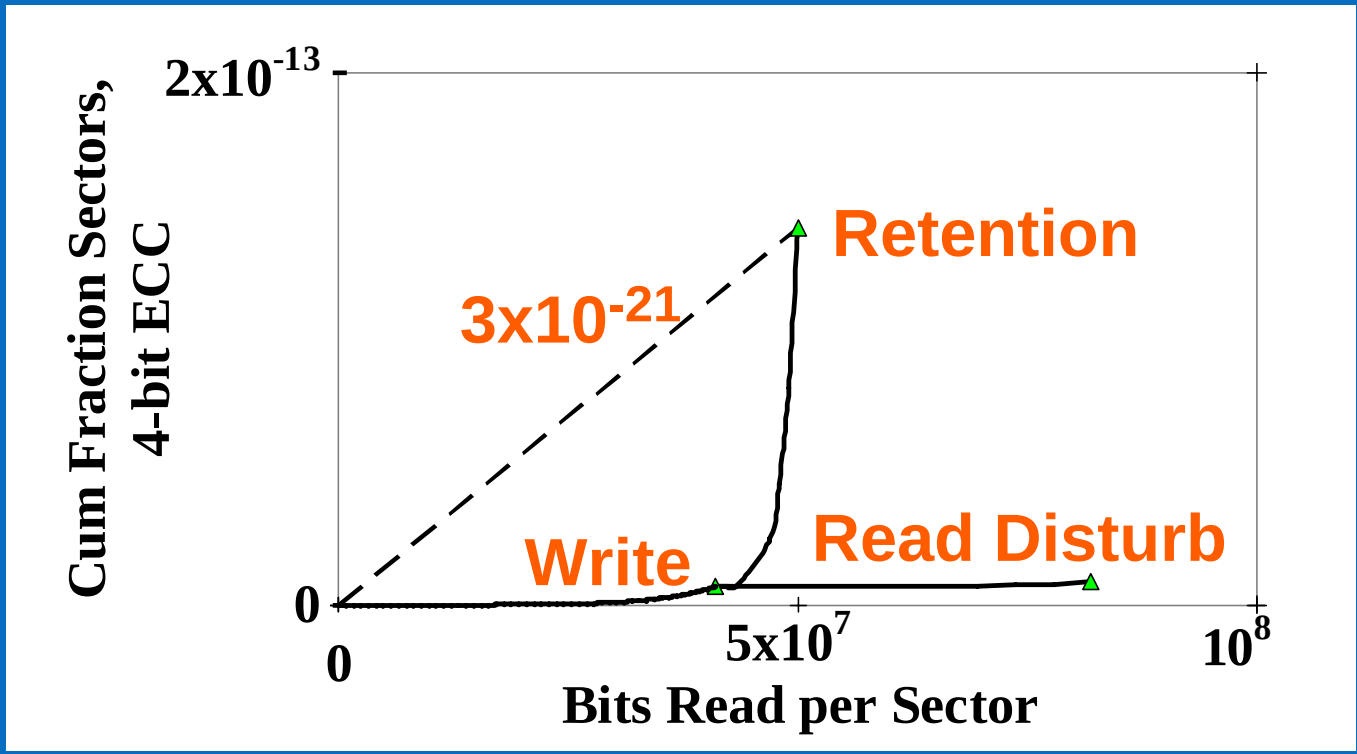
$$= \frac{\text{Cum Fraction Sectors Failing}}{(\text{bits per sector}) \cdot (N_{CYC} \bullet \#\,\text{Reads/Cycle} + N_{\text{Post - Cyc}})}$$

Worst case:1

Read Disturb: #reads in stress

Unbiased: Impute same rate as in cycling

UBER Estimate



Data re-plotted vs. # bits read

UBER at any point is the slope of line to the origin

UBER is very low $3x10^{-21}$ at worst-case point (retention)

UBER increases with greater use, so use range must be stated when UBER is specified

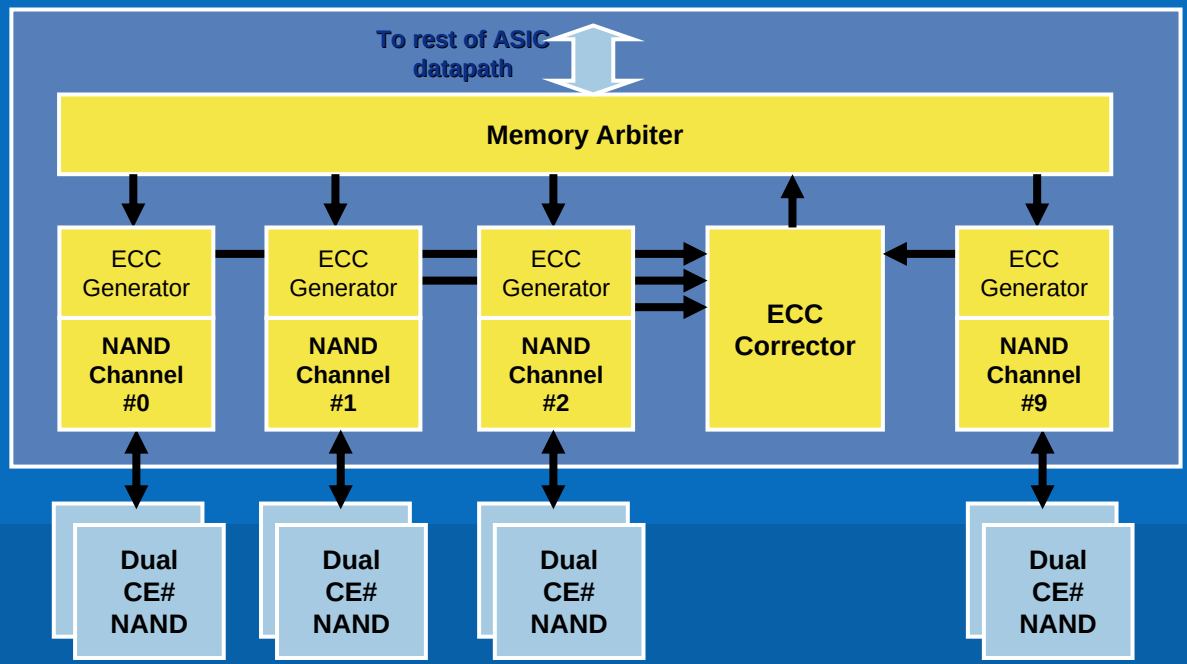
Concurrency in Intel® SSD ASIC

10 external physical NAND channels

- Up to 2 NAND components per channel
- Component = Dual Die or Quad Die Packages

Each channel supports multiple outstanding tasks

- Each NAND channel fully hardware automated/accelerated
- Hardware fully overlaps & pipelines commands
- Automated ECC generators & correctors



Algorithmic Efficiency

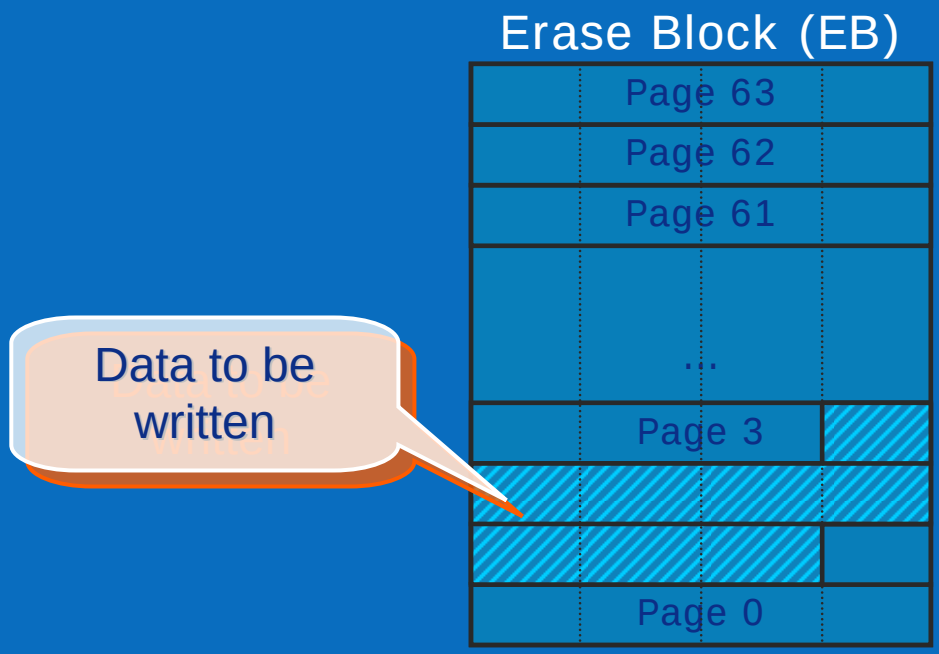
A high-performance NAND controller is necessary *but not sufficient*

Primary impact on overall performance is algorithmic efficiency

- *Especially* the case for small random writes

Write Amplification

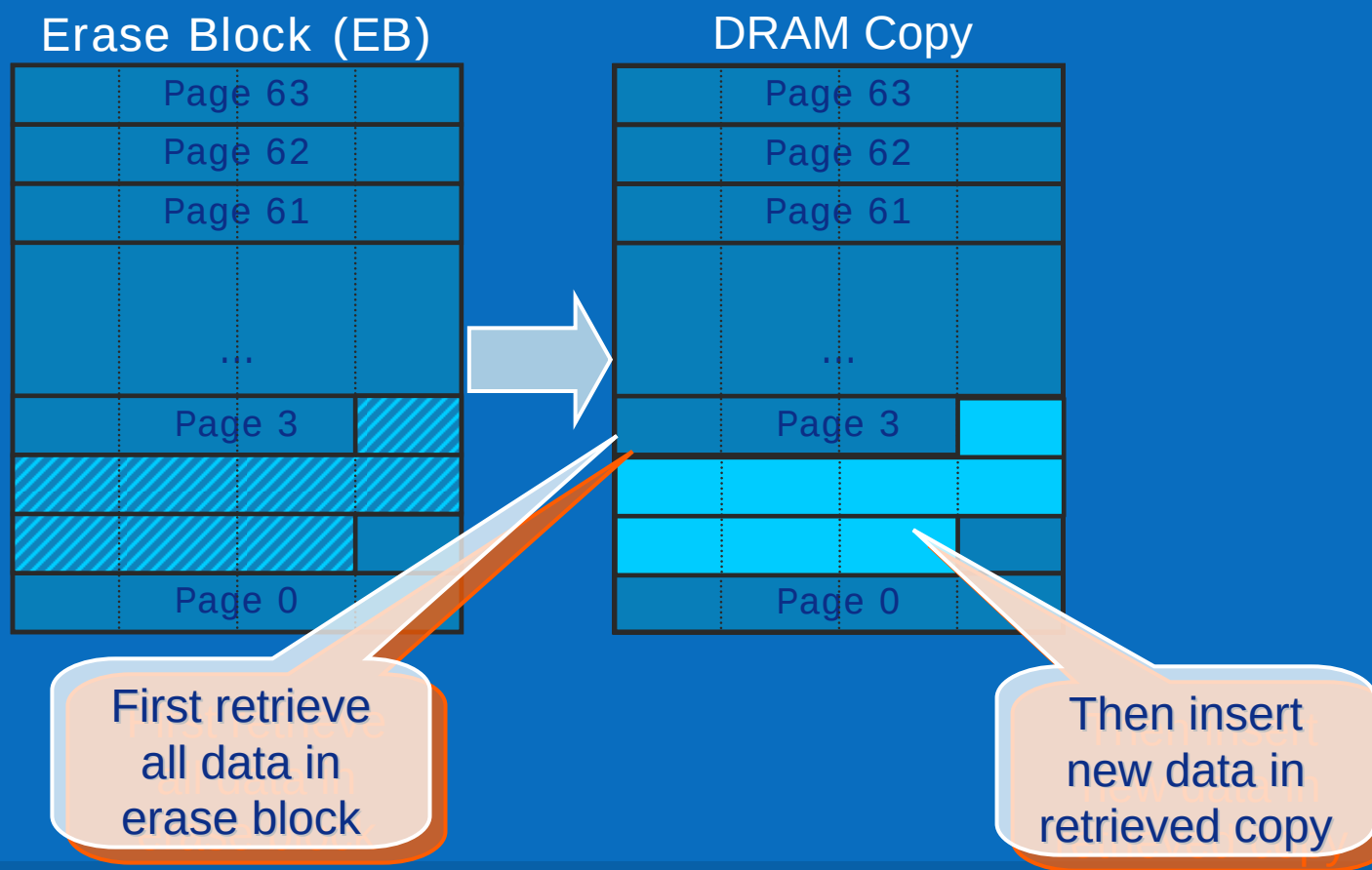
Write Amplification is the amount of NAND written for a requested amount of write from host



*Simplified example to illustrate the write amplification effect. Specific algorithms vary greatly.

Write Amplification

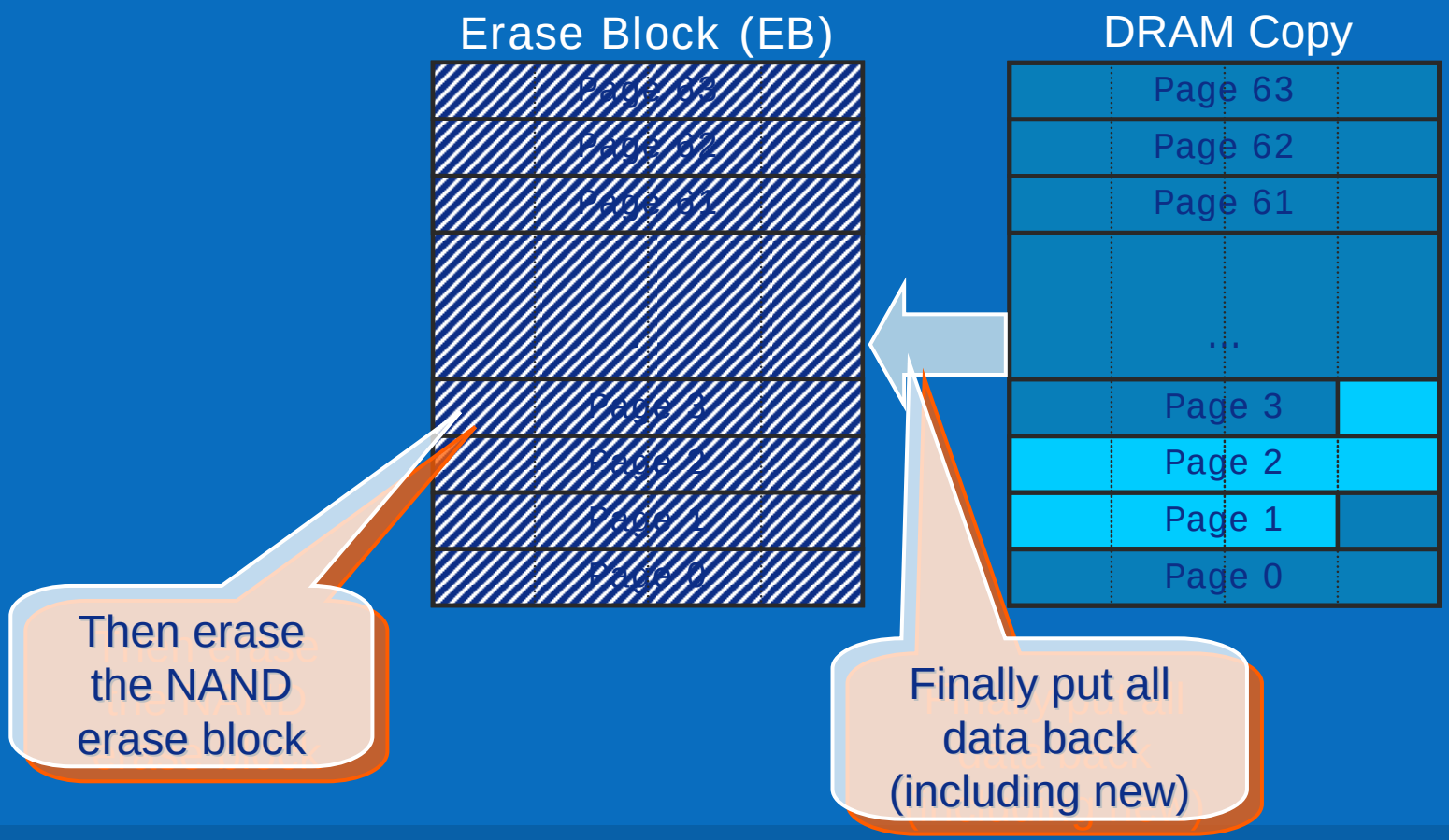
Write Amplification is the amount of NAND written for a requested amount of write from host



*Simplified example to illustrate the write amplification effect. Specific algorithms vary greatly.

Write Amplification

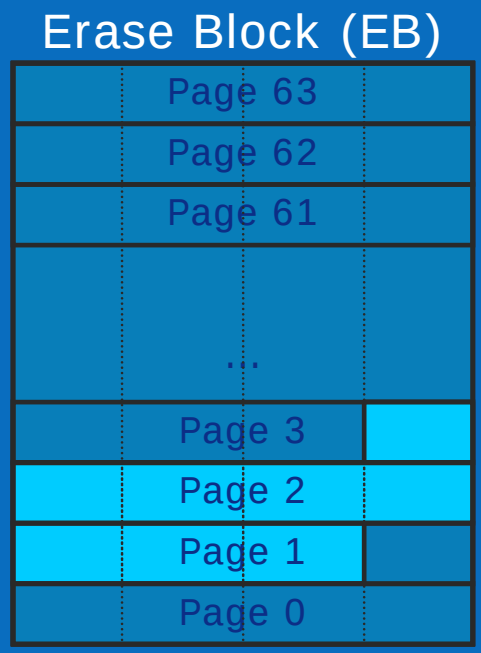
Write Amplification is the amount of NAND written for a requested amount of write from host



*Simplified example to illustrate the write amplification effect. Specific algorithms vary greatly.

Write Amplification

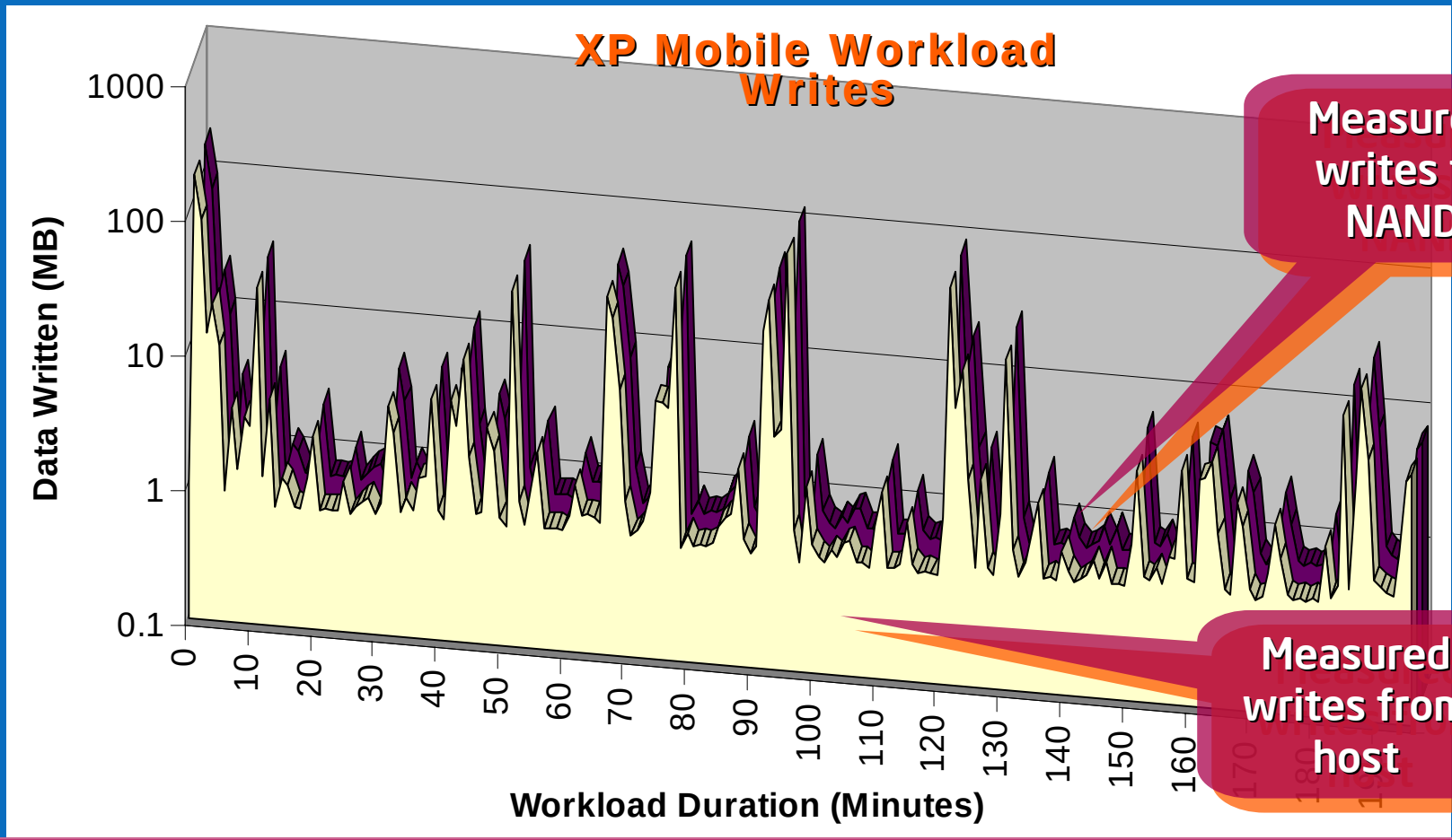
Write Amplification is the amount of NAND written for a requested amount of write from host



Example amplification is 32 (32X NAND written for host request). Traditional schemes have amplification of approx 20-40X.

*Simplified example to illustrate the write amplification effect. Specific algorithms vary greatly.

Client Workload Write Amplification

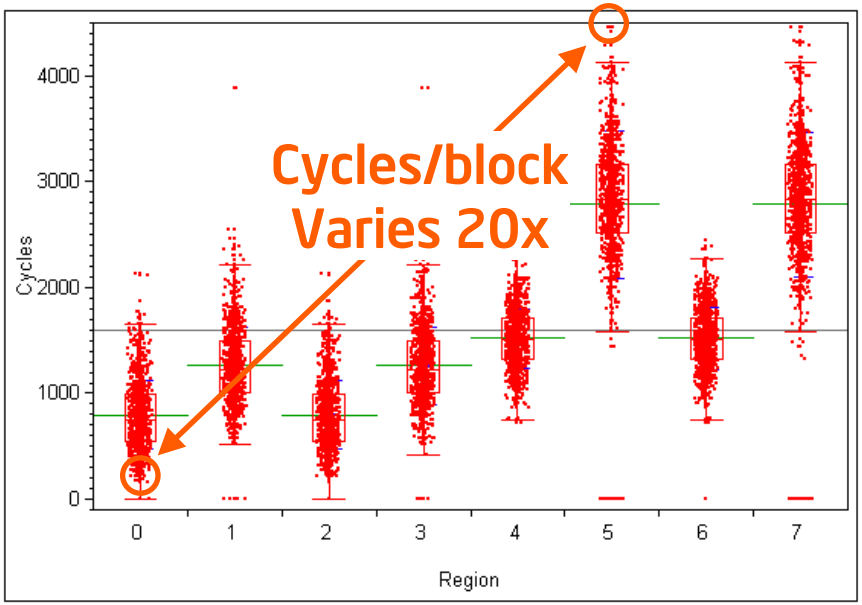


Intel® High-Performance SATA SSDs typical write amplification <1.1 for client workloads (this example <1.05)

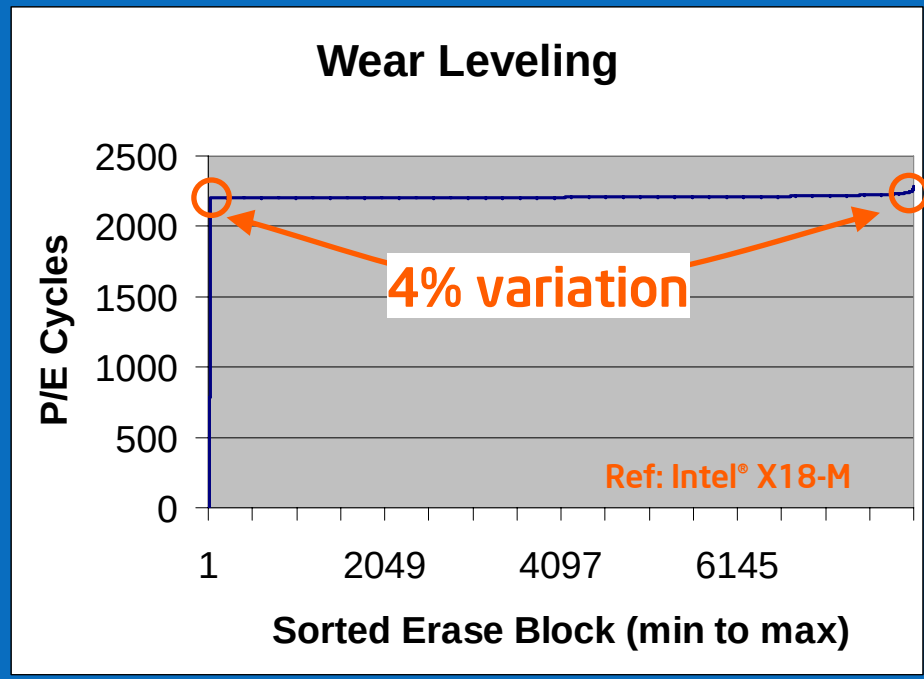
Performance measurements are made using specific computer systems and/or components and reflect the approximate performance of the technology as measured by those tests. Any difference in system hardware or software design or configuration may affect actual results.

*Third party marks and brands are the property of their respective owners

Variability in Wear Leveling



Unsophisticated regioned scheme



More sophisticated scheme

Controllers vary in in wear-leveling effectiveness

Poor wear leveling can have high impact

20x in cycles can be 10x or more in RBER

10x in RBER is 10^{ECC+1} in ECC failure rate: 100,000x for 4-bit ECC

Putting it together: SSD Reliability Metrics

SSD UBER values can be << 10-15

UBER ∝ usage: program/erase/read & subsequent retention

Intel® X18-M and X25-M Mainstream SATA SSD (80GB)

- 10 Channels Architecture with 50nm MLC ONFI 1.0 NAND
- 5 years usage, 1000G, 1.2million hrs MTBF
- GB/day client workload @ 1e-15 UBER → >>100GB/day, 5 years

Intel® X25-M and X18-M Mainstream SATA SSDs deliver >5X accepted requirement for clients (20GB/day)

Intel® X25-E Extreme SATA SSD (32GB)

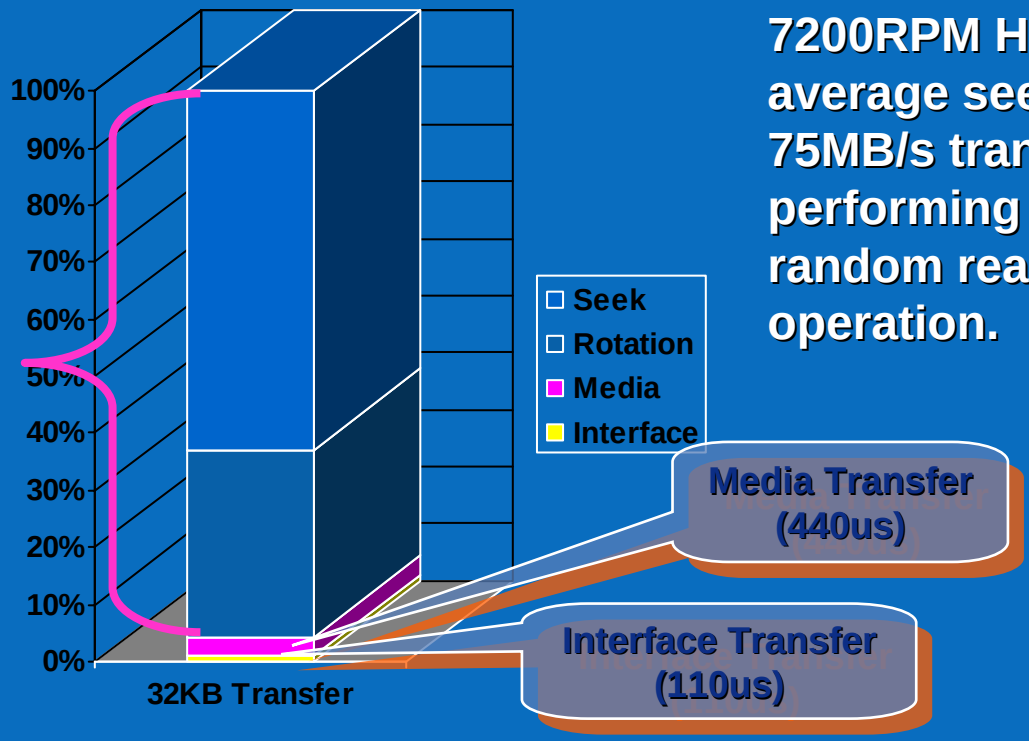
- 10 Channels Architecture with 50nm SLC ONFI 1.0 NAND
- 1000G, 2Million hrs MTBF
- Intel SLC SSD support > 7000 8K 2:1 R/W Random IOPs 24/7, 5 years

Intel X25-E SLC SSDs support the endurance required to replace many 15K RPM HDDs for IOPS applications

Why Random Performance Matters (more than sequential transfer rate)

Most requests are *non-sequential*

For non-sequential accesses, >95% of total HDD service time is mechanical latency

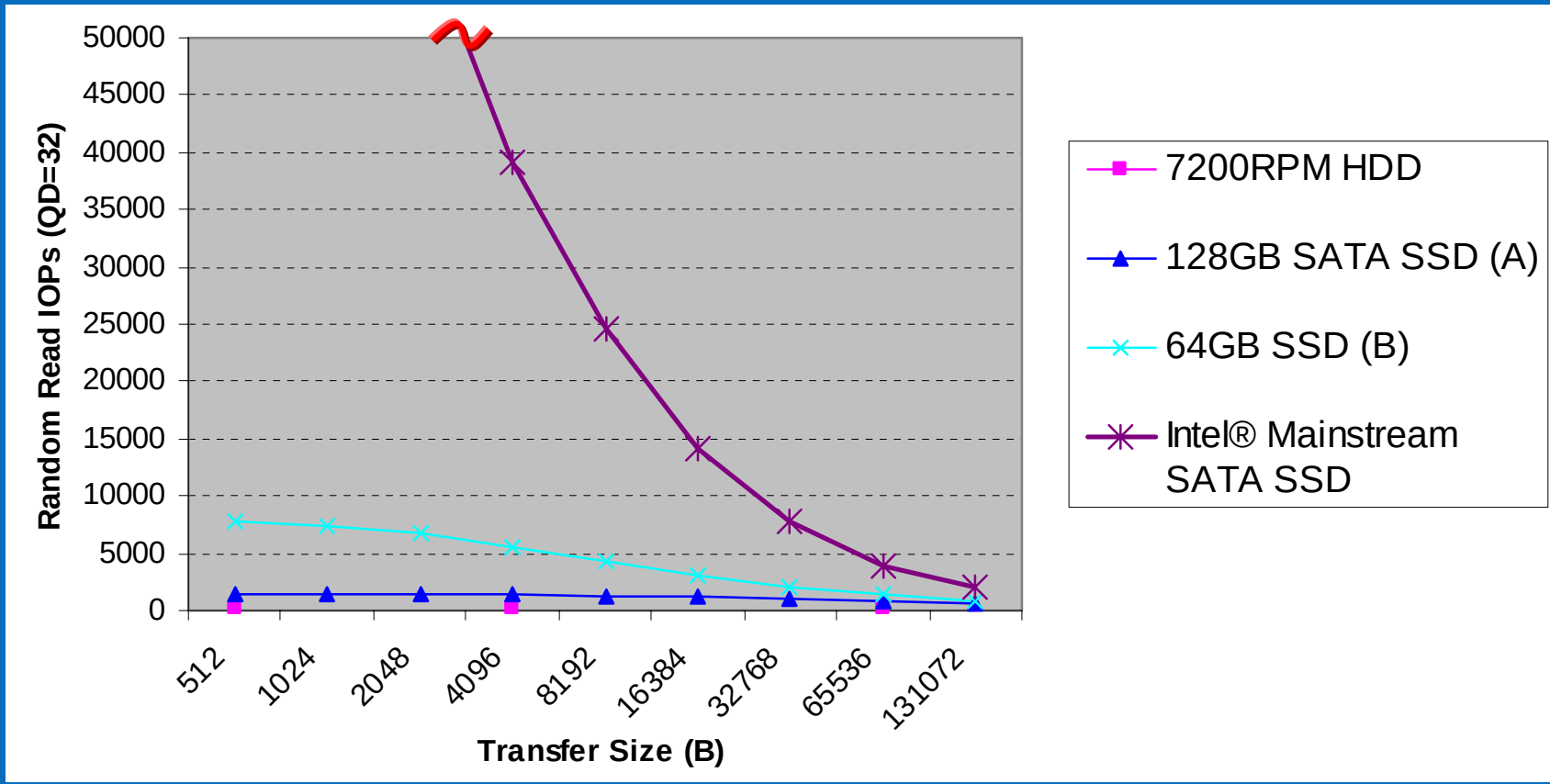


Approximate service time breakdown for 7200RPM HDD w/ 8ms average seek time and 75MB/s transfer rate performing 32KB random read operation.

Most requests are non-sequential where the non-transfer time component is dominant

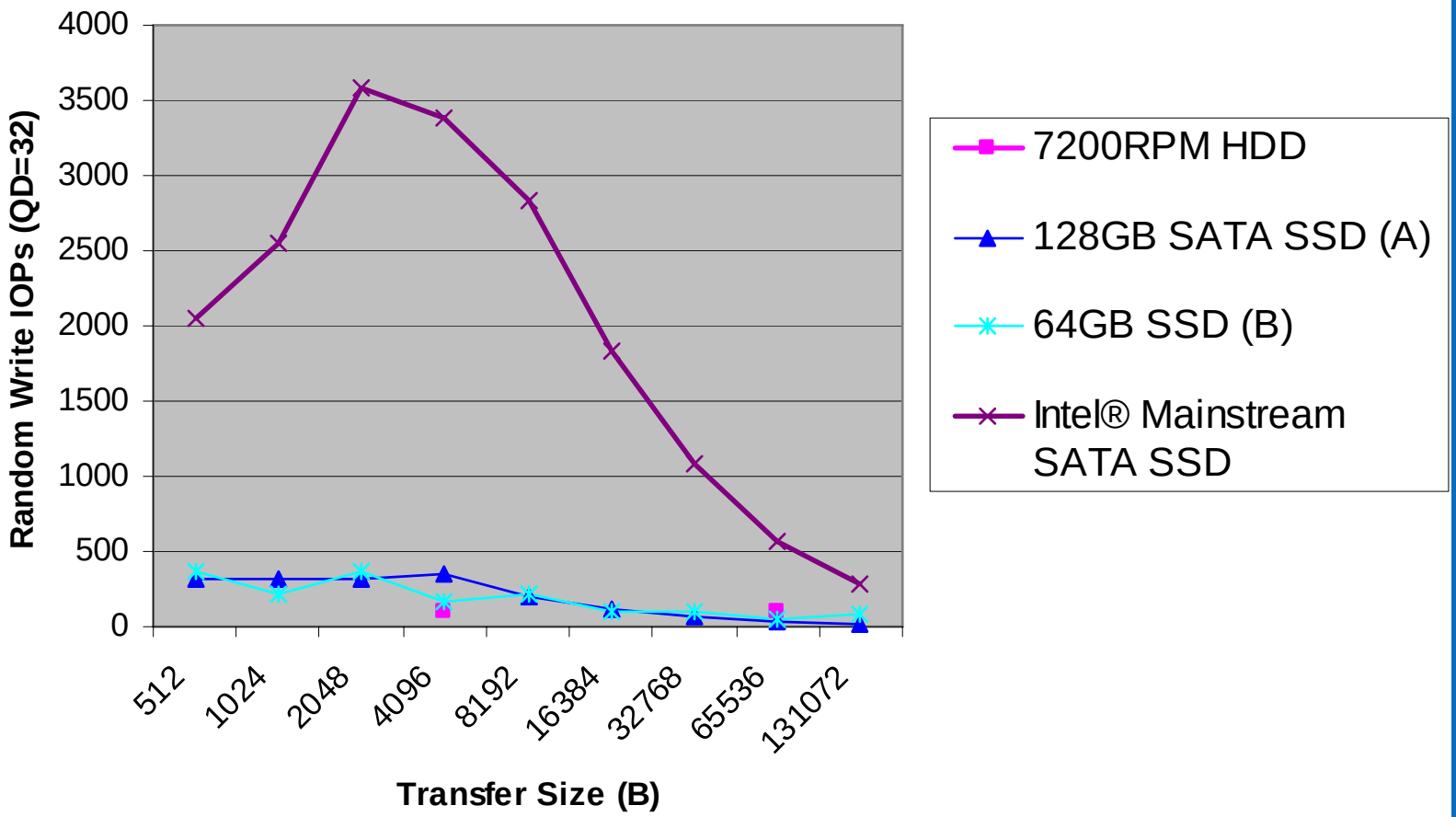
Intel® Mainstream SATA SSD Bridges the HDD Performance Gap

Random Read Performance

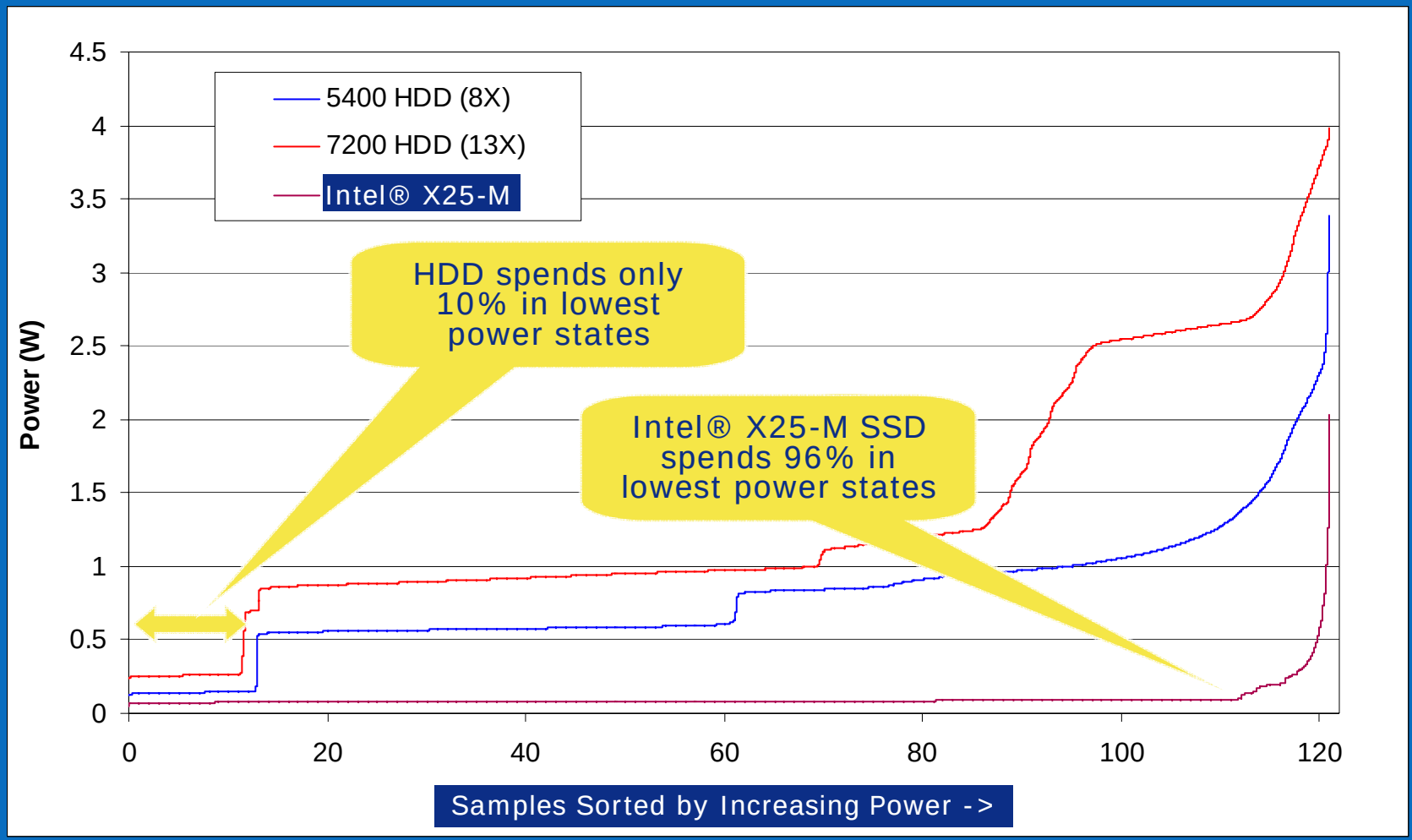


Intel® Mainstream SATA SSD Bridges the HDD Performance Gap (cont'd)

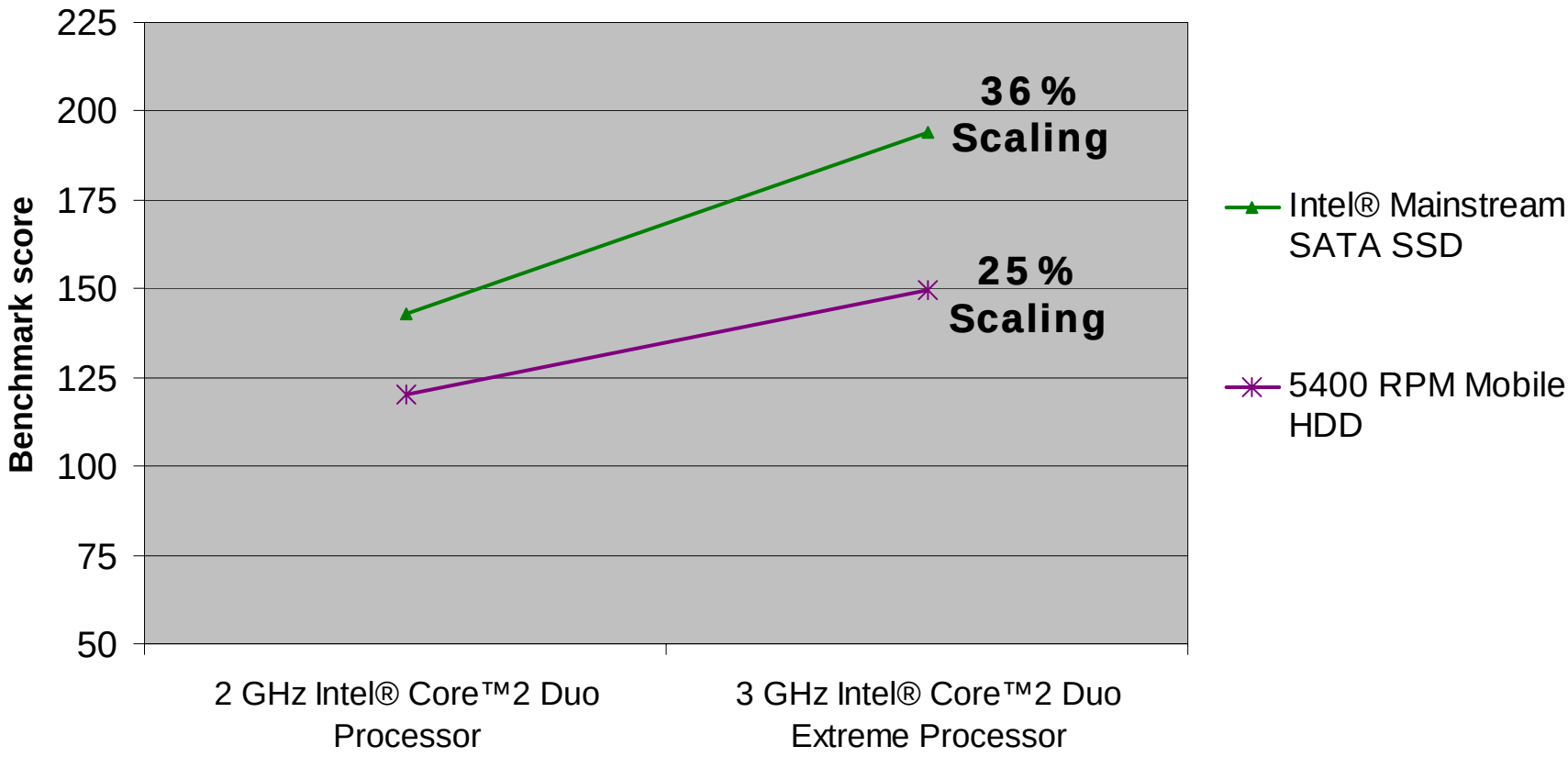
Random Write Performance



Intel® Mainstream SATA SSDs Save Power: SATA Power Rails With 2 Hour Mobile Workload



Intel® Mainstream SATA SSDs Mean Better Mobile CPU Scaling



Intel® Mainstream SATA SSDs maximize end user's processor performance

Performance measurements are made using specific computer systems and/or components and reflect the approximate performance of the technology as measured by those tests. Any difference in system hardware or software design or configuration may affect actual results.

SSDs in Data Center

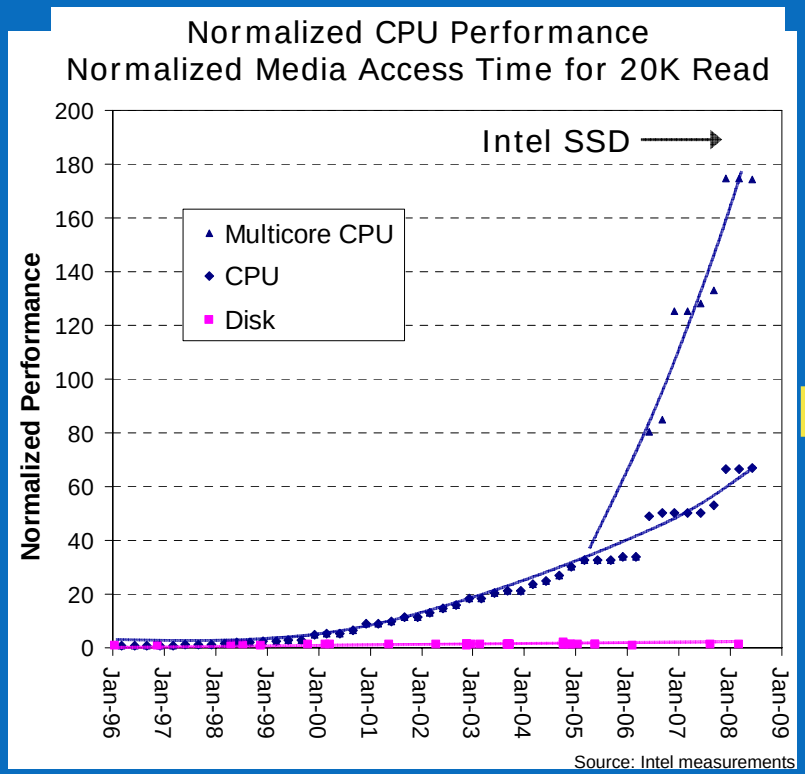
Data center value proposition:

- Performance, especially IOPS performance
  - IOPS = Input/Output Operation Per Second
- Fewer devices needed to meet IOP need, saving money
- Lower power consumption
- Higher system reliability

SSD Value:

A lower cost, greener, more reliable data center

Enterprise HDD Performance Gap Results in Multiplication of HDDs



Source: Intel measurements

TPC-C Report
http://www.tpc.org/

HP Integrity Superdome - Itanium2/1.6GHz/24MB iL3 - 64p/128c

TPC-C Revision 5.8
Report Date: August 21, 2007

Total System Cost	TPC Throughput	Price Performance		Availability Date
USD $11,978,134	4,092,799 tpmC	USD $2.93/tpmC		August 6, 2007
Server Processors/Cores/Threads	Database Manager	Operating System	Other Software	Number of Users
64/128/256 Intel Itanium2 1.6GHz	Oracle Database 10g Release 2 Enterprise Edition with Partitioning	HP-UX 11i v3	TUXEDO 8.1	3,226,200

Server

Client 1

Client 95

1000 Base T

10GbE

6 Procurve Switches

Fiber Channel

126 HP StorageWorks MSA 1500
- 504 HP StorageWorks Enclosures MSA30
- 4,200 36GB 15K RPM Disk Drives
- 2,856 73GB 15K RPM Disk Drives

HP Integrity Superdome
Itanium2/1.6GHz/24MB iL3 – 64p/128c
2,048GB Memory

System Components	Server (HP Integrity Superdome - Itanium2/1.6GHz/24MB iL3 - 64p/128c)		each Client (PA-8620)	
	Qty	Type	Qty	Type
Processors/Cores/Threads	64/128/256	1.6GHz Intel Itanium2	2/2/2	1.3GHz Intel Itanium2
Cache Memory	2	12MB iL3 Cache	each	3M iL3 cache
Memory	2048	GB	8	GB
Disk Controllers	40	PCI-X 4Gb Fibre Channel Adapter (dual port)	1	Ultra2 SCSI LVD
Disk Drives	336	HP StorageWorks Modular Smart Array 1500 with 4200 36GB 15K RPM and 2856 73GB 15K RPM	1	36 GB
	1	HP Storestore Disk System 2120 with 1 36GB 15K HDD		
Total Storage		320974 GB		
Tape Drives	1	DVD ROM		
Terminals	1	HP Server Thin Client		

TPC Benchmark C Full Disclosure Report for HP Integrity Superdome

7056 15K RPM HDDs

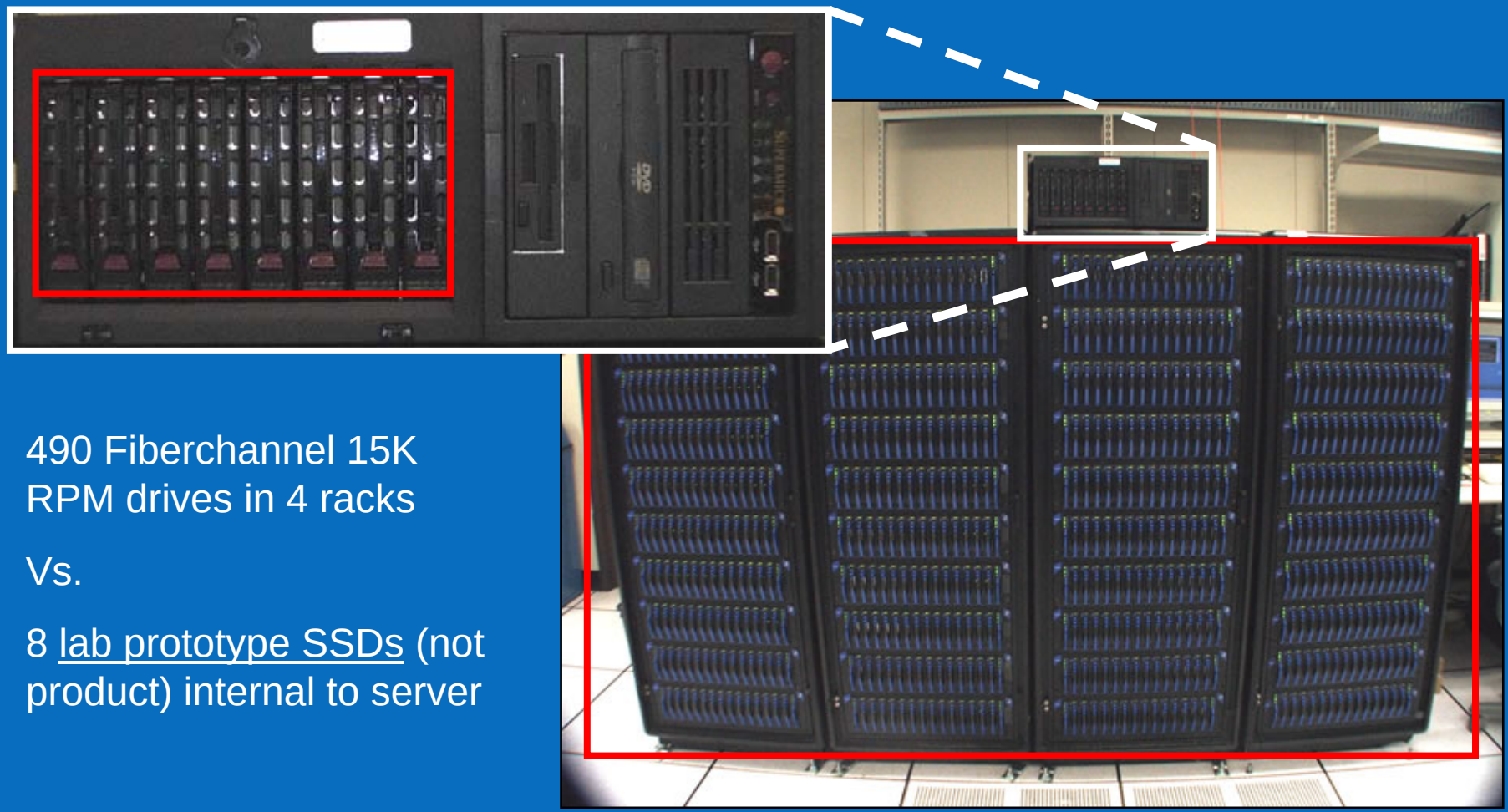
7056 HDDs are expensive

7056 HDDs are hard to manage

7056 HDDs fail often

7056 HDDs burn a lot of power

Similar I/O Performance For IOPS Intensive Workload



490 Fiberchannel 15K RPM drives in 4 racks

Vs.

8 lab prototype SSDs (not product) internal to server

IOPS Application Optimization

HDD

- 64,000 IOPS
- 490 HDDs
- 35 drive shelves
- 24 sq ft
- 14 kW
- 4.6 IOPS/W

Picture not to scale

SDD

- 120,000 IOPS
- 8 SSDs
- 1 drive shelf
- 1 sq ft
- 0.6 kW
- 200 IOPS/W

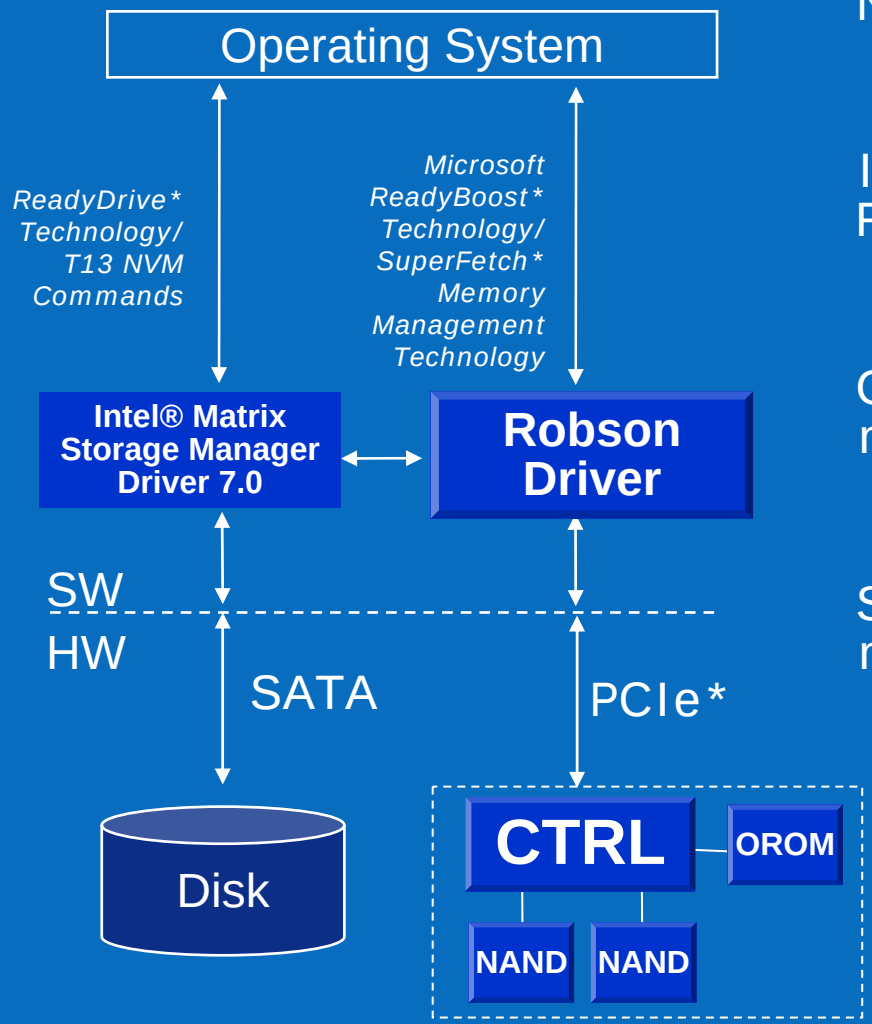
Storage Cost
8X
REDUCTION

Floor Space
24X
REDUCTION

Energy Costs
23X
REDUCTION

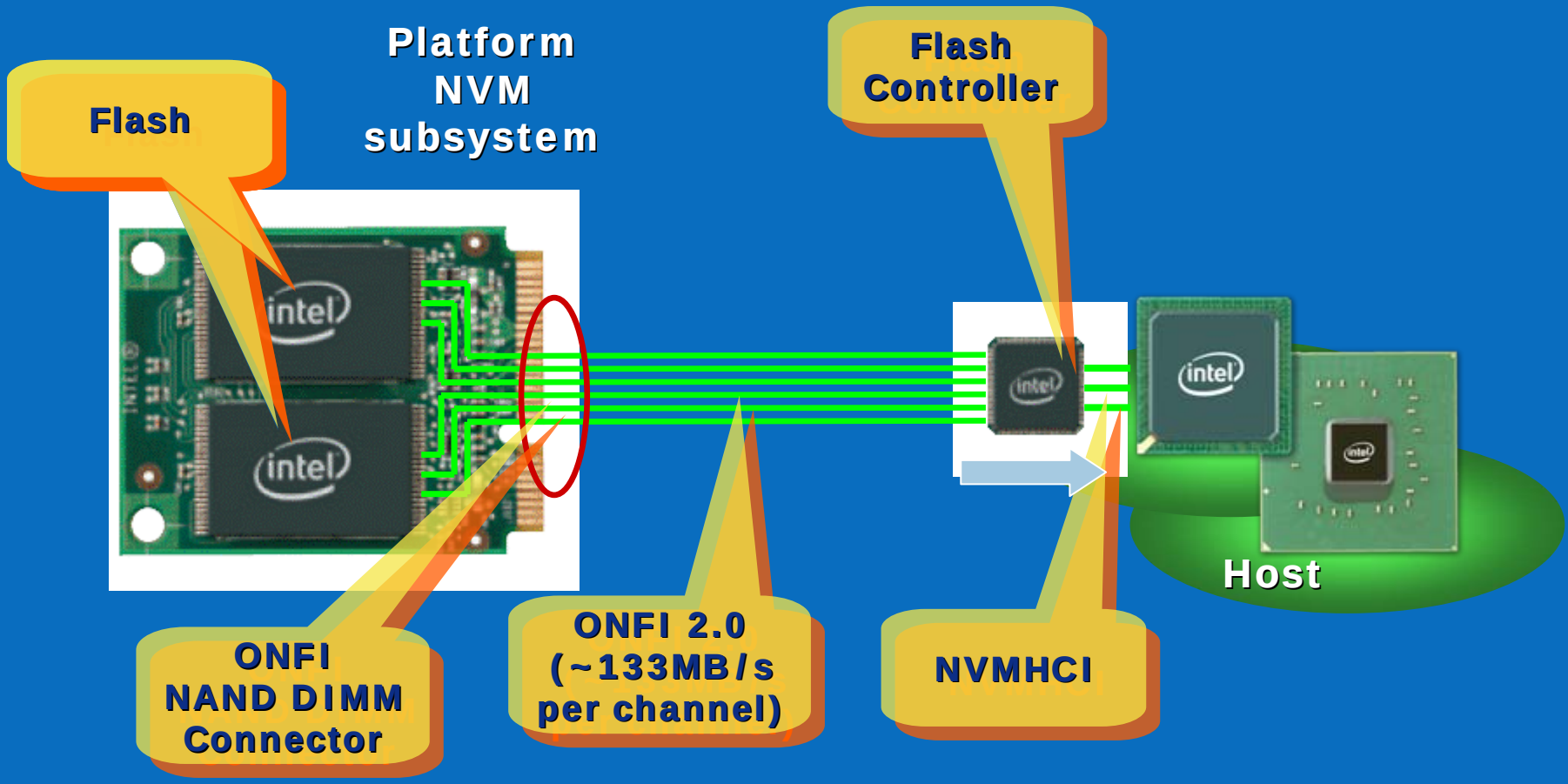
Energy Savings
$9K
SAVINGS/year

Intel® Turbo Memory (NAND Cache)



- NAND flash solution on PCI-e bus
- Intel driver interfaces to Microsoft ReadyBoost* and ReadyDrive*
- O-ROM handles pre-driver load cache management
- Supports on-motherboard and minicard solutions

Standardized High Performance NAND Platform



All elements necessary for standardized high-performance platform NAND solution

*Future platform evolution forecasted

NAND in the Platform

NAND in the platform has started with modules plugged in on PCIe

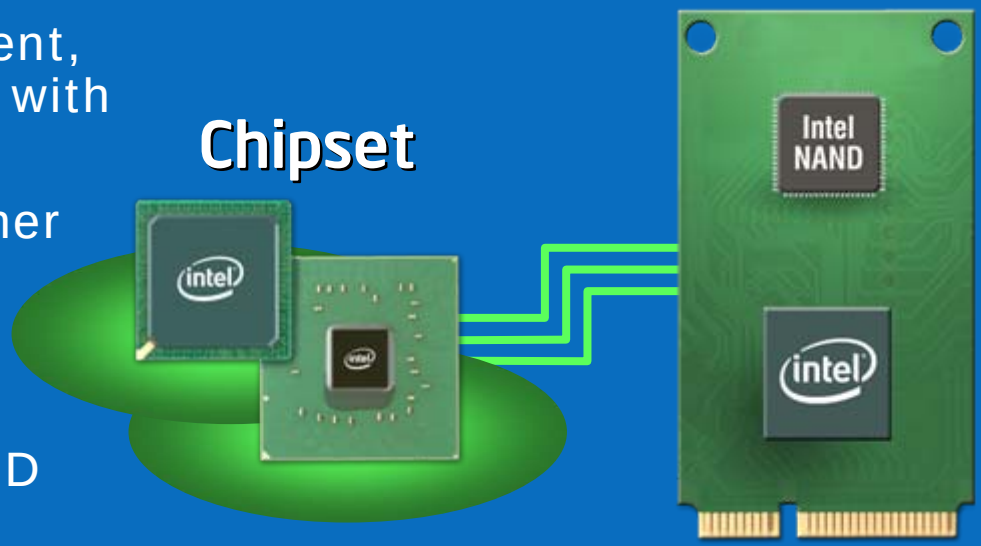
As NAND becomes more prevalent, the controller will be integrated with the platform

- Down on motherboard or higher levels of integration

OEMs want to offer customers capacity/feature choice, so NAND will remain on a module

Issue: How to plug a NAND-only module into a PC platform?

- NAND does not talk PCIe*



Connector for NAND-only Modules

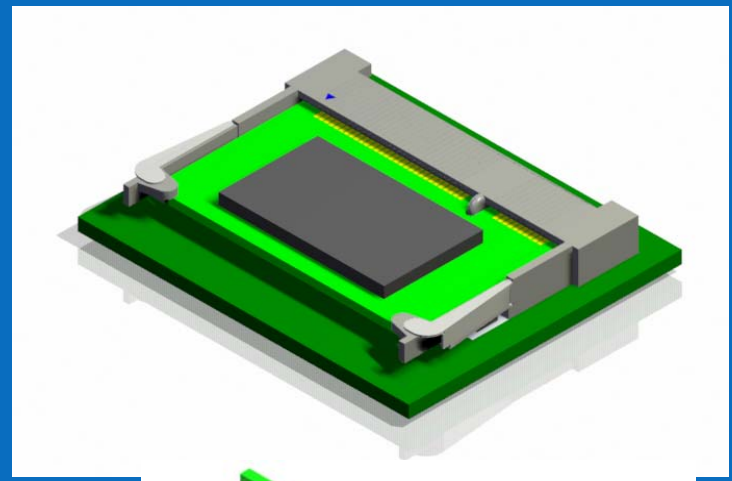
To offer capacity choice, ONFI is defining a standard connector

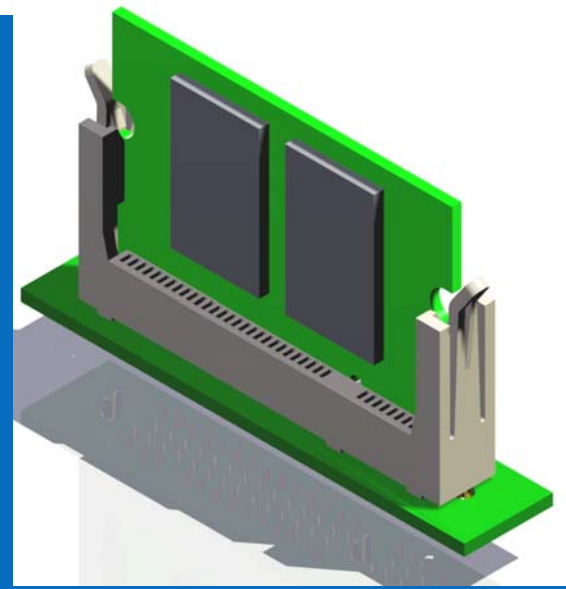
- Enables OEMs to sell NAND on a module
- Like an unbuffered and unregistered DIMM

The ONFI connector effort is leveraging existing DRAM standards

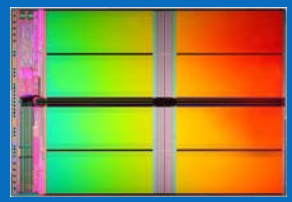
- Avoids major connector tooling costs
- Re-uses electrical verification
- Ensures low cost with quick time to market

Both right-angle and vertical entry form factors are being delivered





NAND Technology Future Scaling Trends: $$/GB = Bit Area & Reliability Scaling



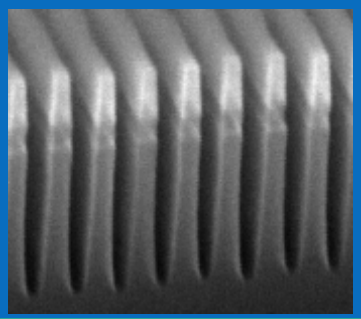
34nm 32Gb MLC NAND

NAND Pathfinding Programs

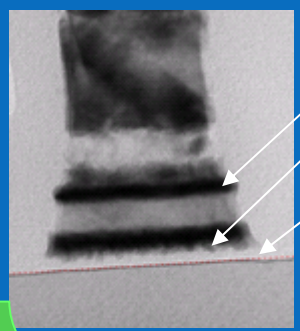
NAND scaling: multiple potential vectors

Pipeline for next several years

Traditional FG NAND



Trap Based Flash



Vertical Integration

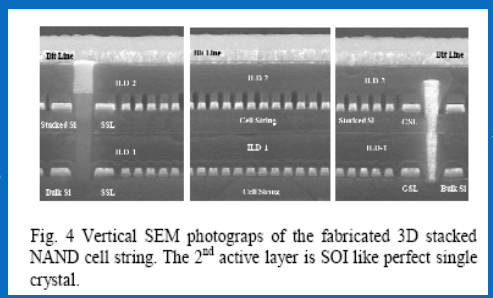
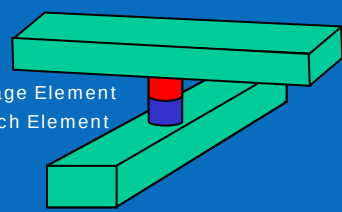


Fig. 4 Vertical SEM photograps of the fabricated 3D stacked NAND cell string. The 2nd active layer is SOI like perfect single crystal.

Ref: S. Jung, IEDM 2006

Non-NAND NVM



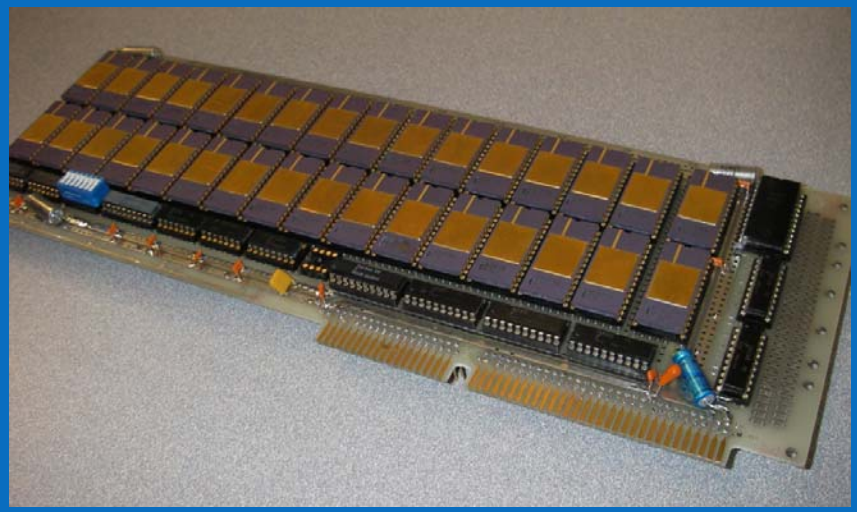
More Evolutionary: Higher production probability, but less scalable

Less Evolutionary: Lower production probability, but potentially more scalable

intel

NVM in Compute…

20+ Year Vision drive by Moore's Law



Now we can start…