SNIA Education

NAND Flash Solid State Storage Performance and Capability -- an In-depth Look

lance l. smith, Fusion-io

SNIA Legal Notice

Education SNIA

- The material contained in this tutorial is copyrighted by the SNIA.
- Member companies and individual members may use this material in presentations and literature under the following conditions:
  - Any slide or slides used must be reproduced in their entirety without modification
  - The SNIA must be acknowledged as the source of any material used in the body of any document containing material from these presentations.
- This presentation is a project of the SNIA Education Committee.
- Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.
- The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.

NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.

Abstract

NAND Flash Solid State Storage Performance and Capability

"This presentation provides an in-depth examination of the fundamental theoretical performance, capabilities, and limitations of NAND Flash-based Solid State Storage (SSS). The tutorial will explore the raw performance capabilities of NAND Flash, and limitations to performance imposed by mitigation of reliability issues, interfaces, protocols, and technology types. Best practices for system integration of SSS will be discussed. Performance achievements will be reviewed for various products and applications. "

Moore's continues to beat Newton's Law

- Mechanical Drives have hit their limits
  - Platter stability degrades at higher speeds
  - Short-stroking reduces capacity for seek time
  - Capacity is limited by smaller form factors
- Solid State Storage continues to evolve
  - Greatest bit density (bits per cubic volume)
  - Random IOPS are 250 times greater
  - MLC increases capacity and lowers costs
  - Advanced error correction improves reliability
  - Performance and Capacity are intertwined

Moore's continues to beat Newton's Law

- Mechanical Drives have hit their limits
  - Platter stability degrades at higher speeds
  - Short-stroking reduces capacity for seek time
  - Capacity is limited by smaller form factors
- Solid State Storage continues to evolve
  - Greatest bit density (bits per cubic volume)
  - Random IOPS are 250 times greater
  - MLC increases capacity and lowers costs
  - Advanced error correction improves reliability
  - Performance and Capacity are intertwined

Data Integrity + Performance

There can be no data integrity trade-off for performance

Media Reliability / Availability

Education SNIA

- The GOOD
  - No moving parts
  - Post infant mortality (catastrophic) device failures are rare
  - Predictable wear out
- The BAD
  - Relatively high bit error rate, which increases with wear
  - Higher density and MLC increases bit error rate
  - Program and Read Disturbs
- The UGLY
  - Partial Page Programming
  - Data retention is poor at high temperature and wear
  - Infant mortality is high (large number of parts…)

Controller Reliability Management

- Wear leveling & Spare Capacity
- Read & Program Disturb control
- Data & Index Protection
  - ECC Correction
  - Internal RAID
  - Data Integrity Field (DIF)
- Management

Poor Media + Great Controller = Great SSS Solution

Data Integrity versus Performance

Performance is about ROI

Education SNIA

Lower OpEx

- Less HW Maintenance
- Less SW Maintenance
- Greater Uptime
- Less Power/Cooling
- Fewer Diverse Skills

Lower CapEx

- Fewer CPUs
- Less RAM
- Less Network Gear
- Fewer SW Licenses
- Less Space

HIGHER Productivity

Media Performance

Education SNIA

- The GOOD
  - Performance is excellent (wrt HDDs)
  - High performance per power (IOPS/Watt)
  - Low pin count: shared command / data bus → good balance
- The BAD
  - Not really a random access device
    - Block oriented
    - R/W access speed imbalance
    - Slow effective write (erase/transfer/program) latency
  - Performance changes with wear
- The UGLY
  - Some controllers do read/erase/modify/write
  - Others use inefficient garbage collection

Performance Drivers – SSS Design

Education SNIA

- Number of NAND Flash Chips (Die)
- Number of Channels (Real / Pipelined)
- Interconnect
- Data Protection (internal/external RAID; DIF; ECC…)
- SLC / MLC Flash Type
- Effective Block Size (LBA; Sector)
- Write Amplification Efficiency
- Garbage Collection (GC) Efficiency
- Bandwidth Throttling
- Buffer Capacity & Mgmt

Simplified Theoretical Analysis

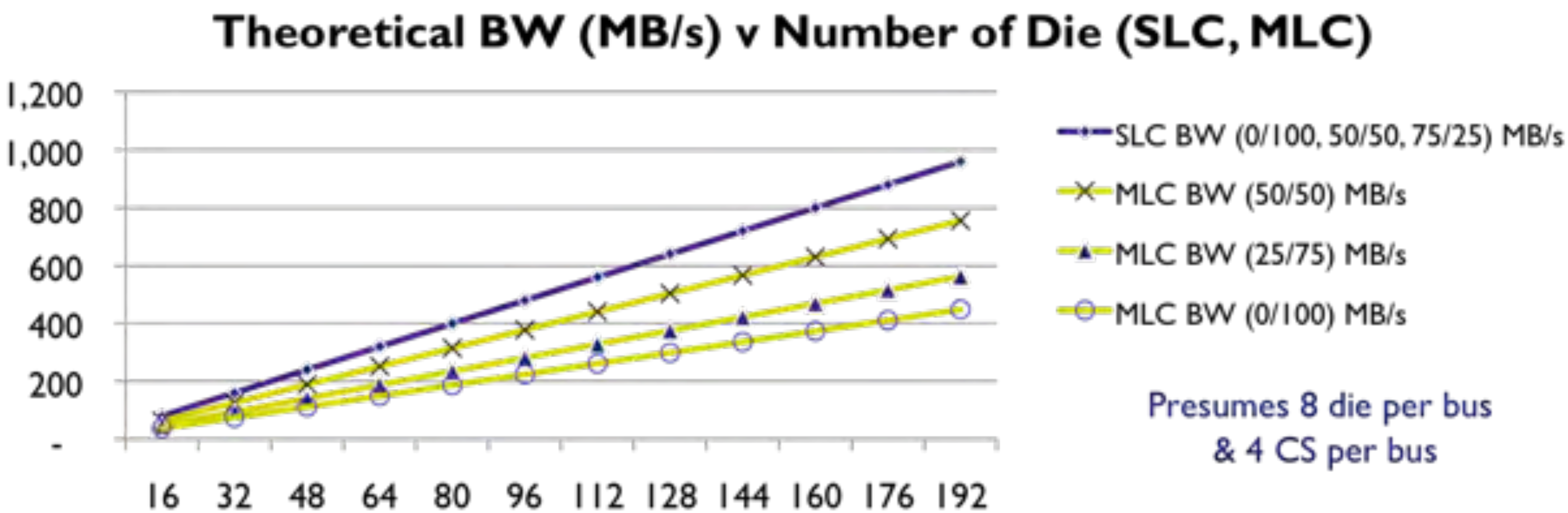
Education SNIA

- Bandwidth Only (Not IOPS)
  - Large Transfers (Data length = Integer times die count)
  - Infinite Buffer
  - Reads/Writes queued for maximum bandwidth
  - No system latency
- Read/Write Ratio %'s fixed
  - 100/0, 75/25, 50/50, 25/75, 0/100
  - Steady State, 100% Efficient GC (EB erase/EB written = 1)
- Maximum Total BW for SATA-II and PCI-e X4
  - No overhead considered

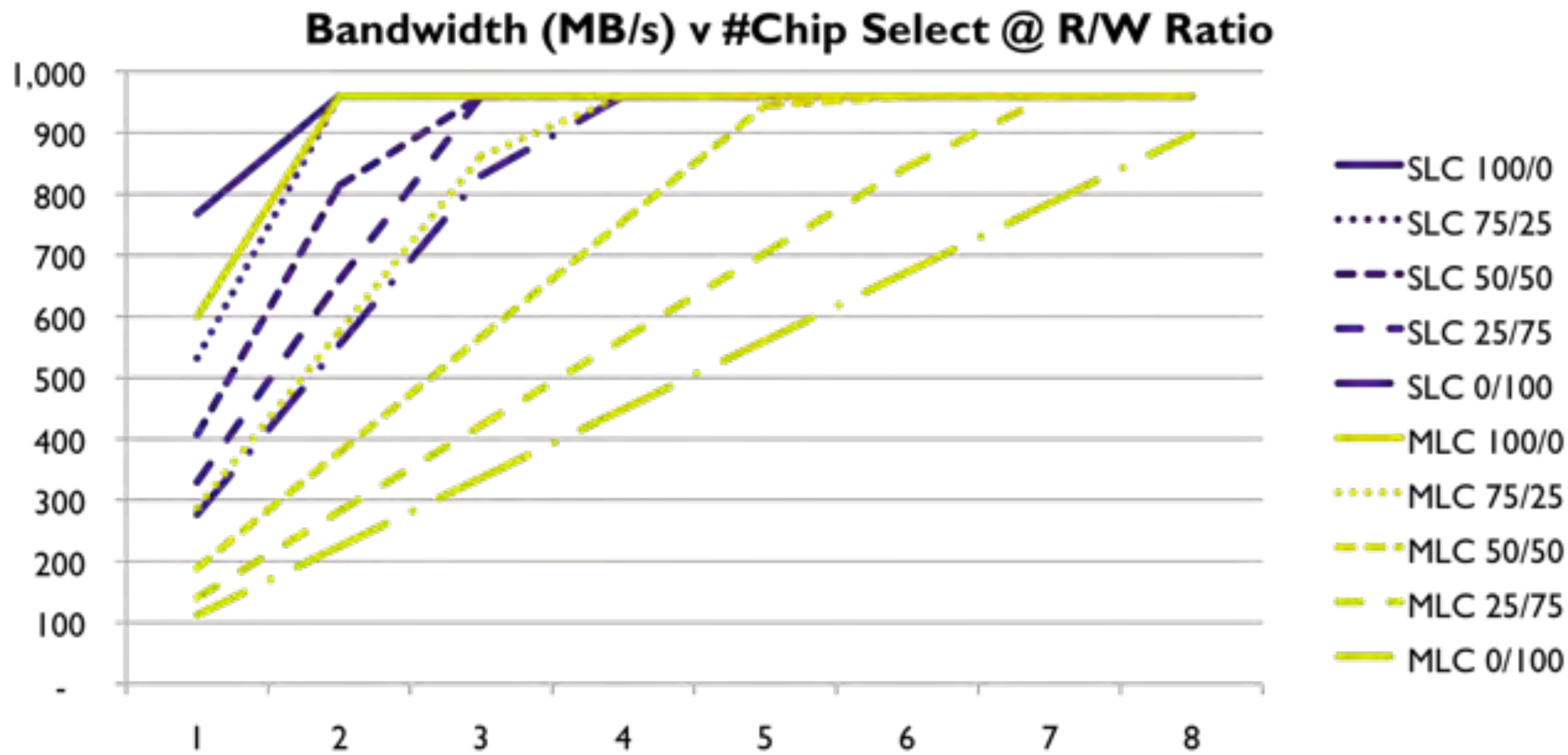
Bandwidth Depends on Die Count

Education SNIA

		SLC	MLC
Transfer Rate (MB/s)	tRC & tWC	400	400
Page Program (us)	tProgram	200	600
EB Erase (us)	tErase	3000	10,000
Load Page (us)	tR (tRead)	25	60
Capacity per die		0.5	1.0



Single-Level versus Multi-Level Cell



Read / write performance imbalance closed with additional banks
Greater R/W imbalance in MLC requires more banks

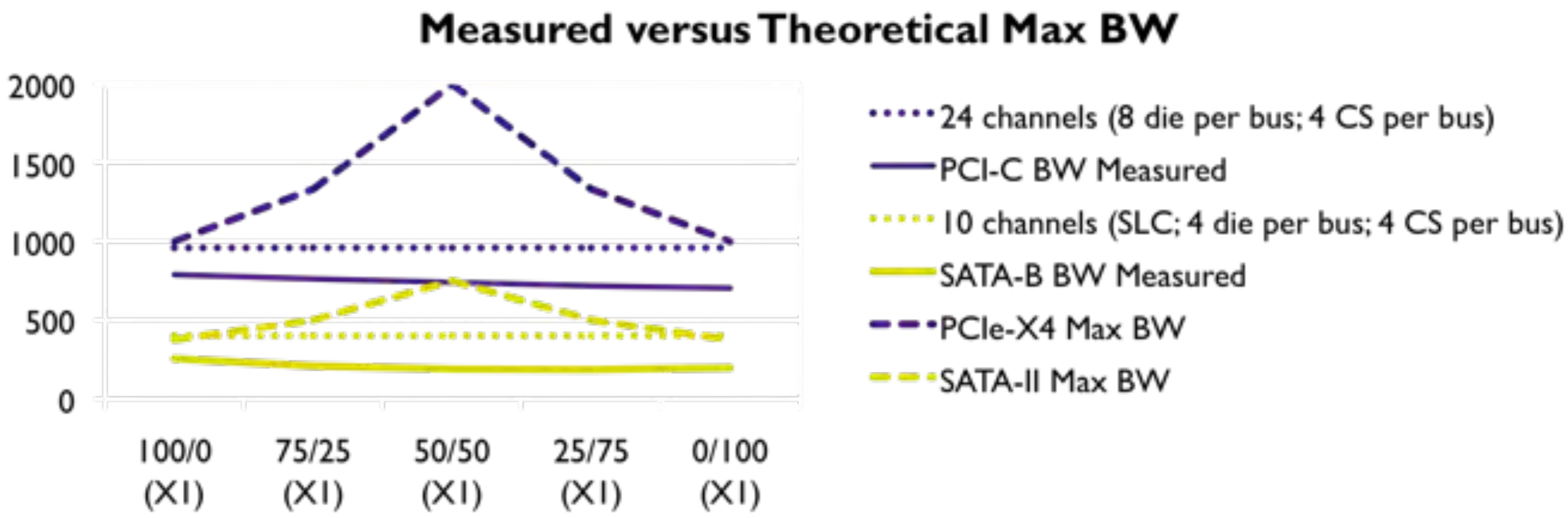
Product Comparison under Test

Education SNIA

Features directly affecting performance measurements

	SATA (A)	SATA (B)	PCI (C)
Capacity (GB)	32	32	160
Bus/Link	SATA-II (3 Gb/s)	SATA-II (3 Gb/s)	PCI-E X4 1.1
Memory Type	SLC	SLC	SLC
Adjustable Reserve Capacity	No	No	Yes
SSS Internal RAID -- Running during test	No N/A	No N/A	Yes Yes
K-IOPS (RMS)	8	27	88
K-IOPS (RMS) / WATT	3	?	7
Bandwidth (RMS, MB/s)	56	208	743
ECC correction	7 bits in 512B	4 bits in ?	11 bits in 240B

Measured vs Theoretical Bandwidth

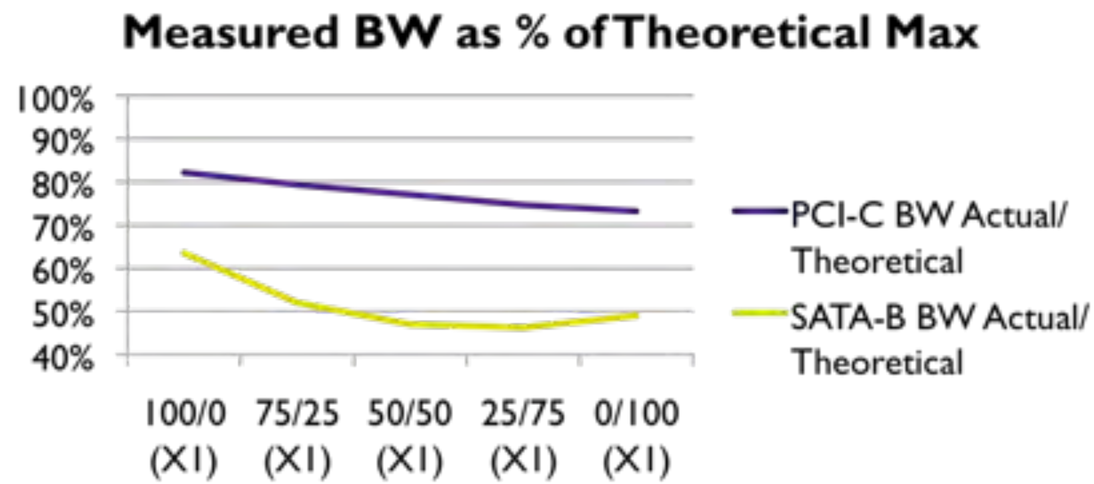


Note: Theoretical Max BW with 24 channels (4 die per bus, 4 CS per bus) is identical to the PCI-C, 24 channel shown in these charts.

Capacity Multiplier:

SATA-B: 1

PCI-C: 2



Access Process (Physics Ignored)

Education SNIA

- Read Access
  - Address Chip / EB / Page
  - Load Page into Register
  - Transfer Data From Register 1-byte per cycle

Typical NAND Flash Die:
- 2000 Erase Blocks (EB)
- 64 Pages per EB
- 4000 Bytes per Page
- 500 MByte Total Capacity

- Write Access
  - Address Chip / EB
  - Erase EB
- …some time later…
  - Address Chip / EB / Page
  - Transfer Data To Register 1-byte per cycle
  - Program Register to Page

Example 1: Read/Erase/Modify/Write

Education SNIA

Time = t1

Starting State

Page	Erase Block 1			
0	b	c	--	--
1	j	--	k	l
2	m	--	--	--
3	--	--	q	r

Time = t2

Write Buffer & W,X,Y

Page	Erase Block 1			
0	b	c	W	X
1	j	Y	k	l
2	m			
3			q	r

Time = t3

Write Buffer & Z,A,B',C',R'

Page	Erase Block 1			
0	B'	C'	w	x
1	j	y	k	l
2	m	Z	A	
3			q	R'

Buffer holds data while EB-1 Erased

Page	Erase Block 1			
0				
1				
2				
3				

Buffer holds data while EB-1 Erased

Page	Erase Block 1			
0				
1				
2				
3				

Example 2: Read/Modify/Write

Education SNIA

Time = t1	Time = t2	Time = t3
Starting State	Data to Buffer (not shown) Erase EB-1 (not shown) Write Buffer & W,X,Y to EB-1	Data to Buffer (not shown) Erase EB-1 (not shown) Write Z,A & Replace b,c,r with B',C',R' & Write EB-1

Page	Erase Block 1			
0	b	c	--	--
1	j	--	k	l
2	m	--	--	--
3	--	--	q	r

Page	Erase Block 2			
0	b	c	W	X
1	j	Y	k	l
2	m			
3			q	r

Page	Erase Block 3			
0	B'	C'	w	x
1	j	y	k	l
2	m	Z	A	
3			q	R'

Implicit wear leveling; EB-1 → EB-2 → EB-3

Presumes that destination EB-2 & EB-3 erased prior to transfer of data → higher performance (than previous "Read/Erase/Modify/Write" example)

Example 3: Garbage Collection

Education SNIA

Time = t1

Start Garbage Collect EB-1

Page	Erase Block 1			
0	b	c	--	--
1	j	--	k	l
2	m	--	--	--
3	--	--	q	r

Page	Erase Block 2			
0				
1				
2				
3				

Time = t2

EB-1 GC'd to EB-2

W,X,Y added

Page	Erase Block 1			
0	b	c	--	--
1	j	--	k	l
2	m	--	--	--
3	--	--	q	r

Page	Erase Block 2			
0	W	b	c	X
1	Y	j	k	l
2	m	q	r	
3				

Time = t3

EB-1 erase

b,c,r replaced by B',C',R'

Page	Erase Block 1			
0				
1				
2				
3				

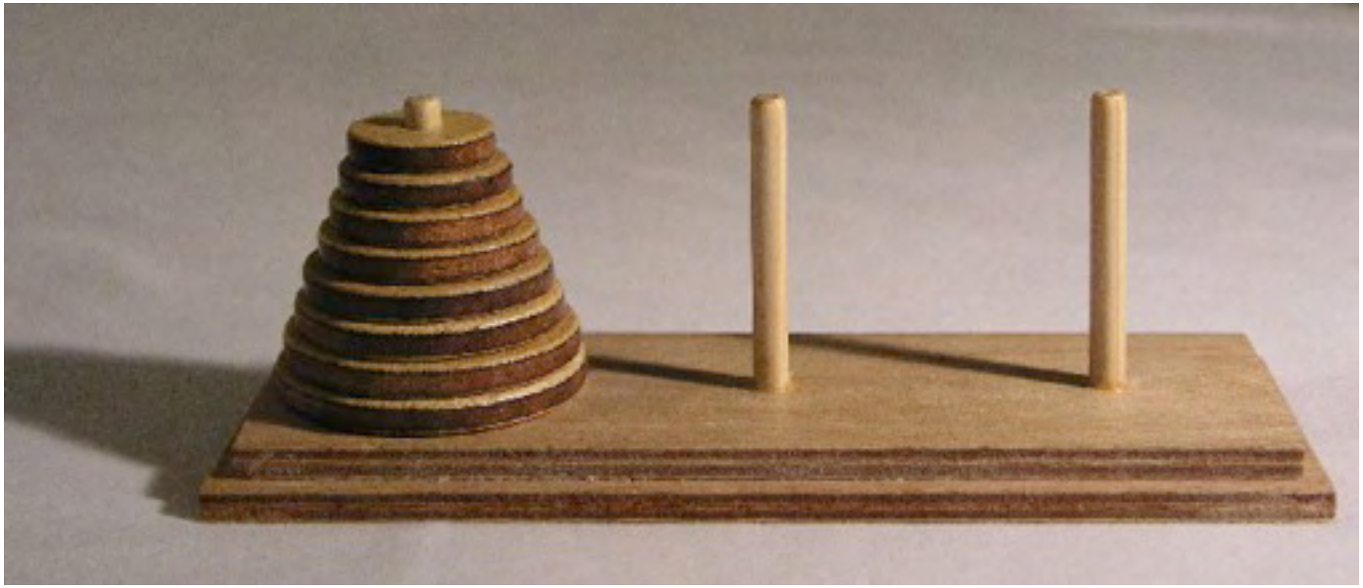
Page	Erase Block 2			
0	w	--	--	x
1	y	j	k	l
2	m	q	--	--
3	B'	C'	R'	--

GC Performance Impact

Education SNIA

- In this example,
  - COPIED DATA: {b, c, j, k, l, m, q, r} 8 blocks
  - NEW DATA {W, X, Y, B', C', Z, A, R'} 8 blocks
  - 50% (8 of 16) writes are user initiated
  - 50% (8 of 16) writes are internal movement (overhead)
- Important:
  - 50% of EB-1 was "invalid data"
  - What if only 10% had been "invalid data?"
  - GC efficiency is dependent upon % of reserve capacity

Tower of Hanoi



Want to do this in fewer moves?
Add more pegs!

GC: Pathological Write Conditions

- If a high percentage of total storage capacity utilized

AND

- A High percentage of data has no correlation-in-time

AND

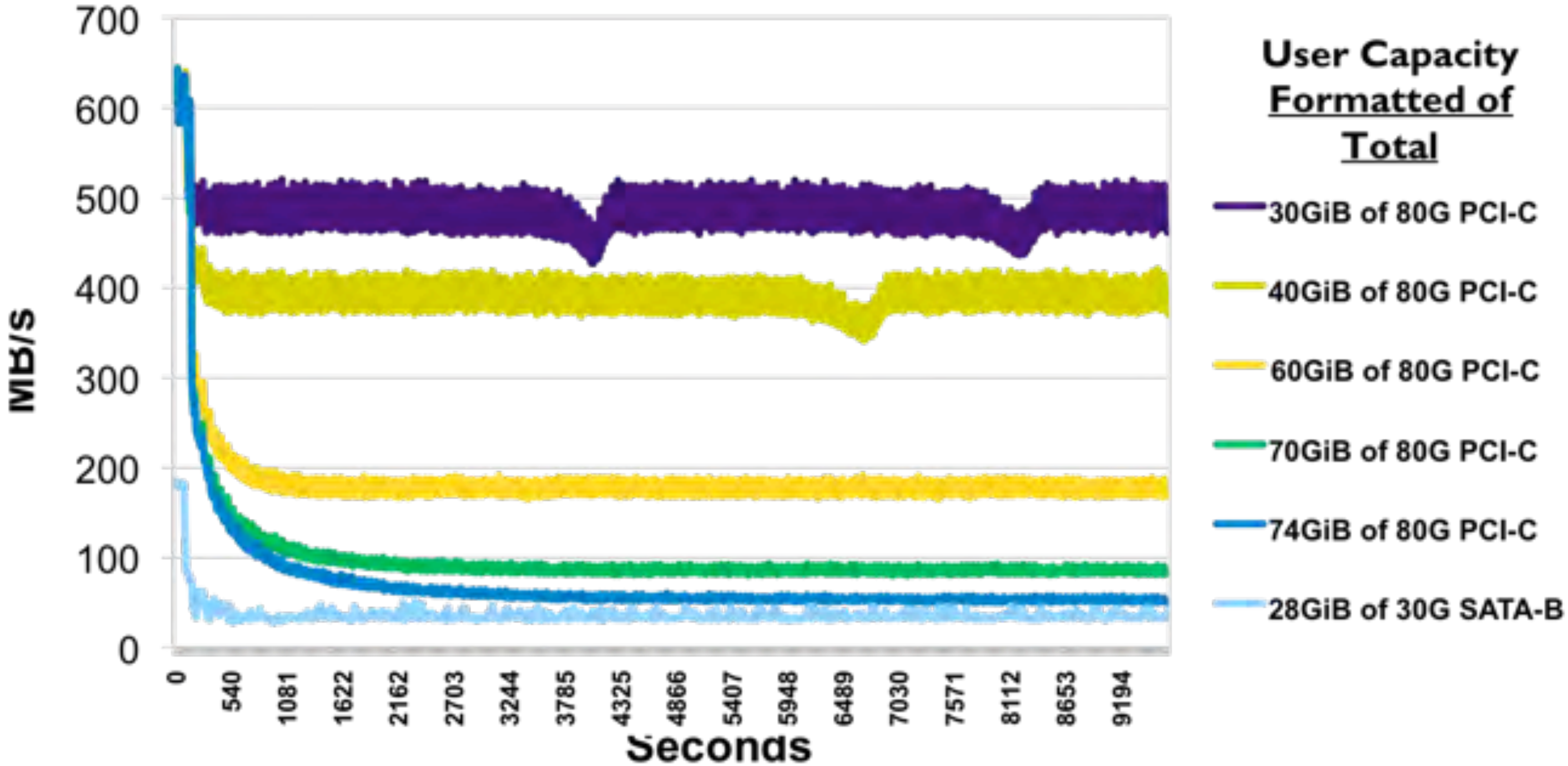
- Continuous writing (no recovery time for GC)

THEN…

Efficiency of GC greatly diminished

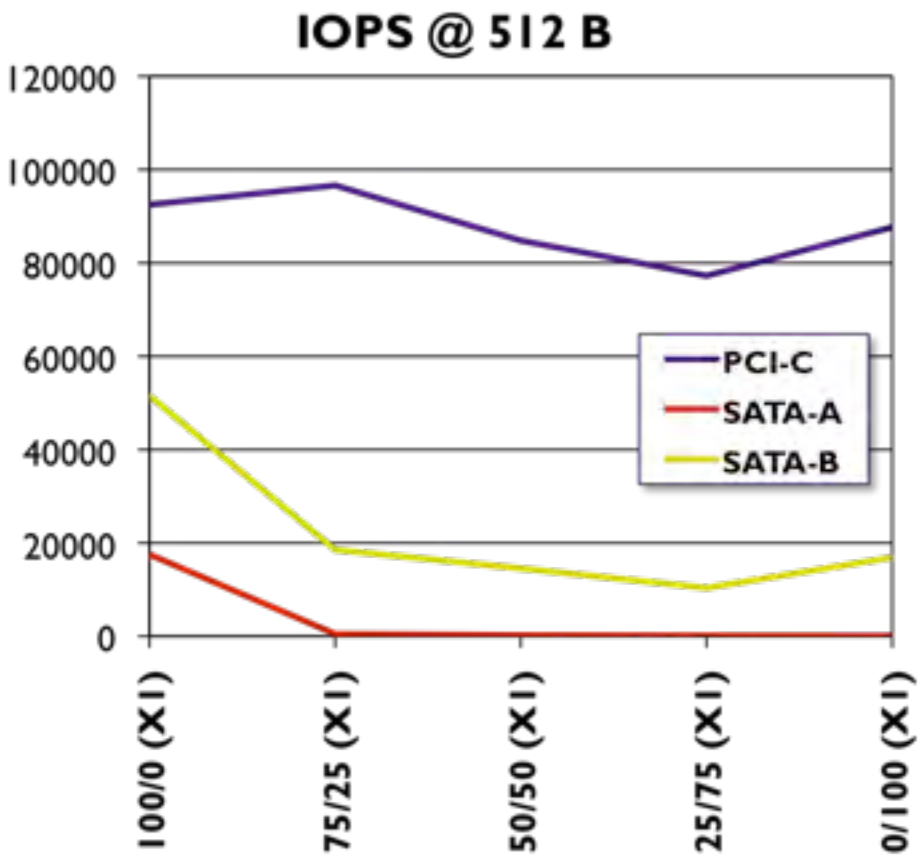
Pathological Write Condition

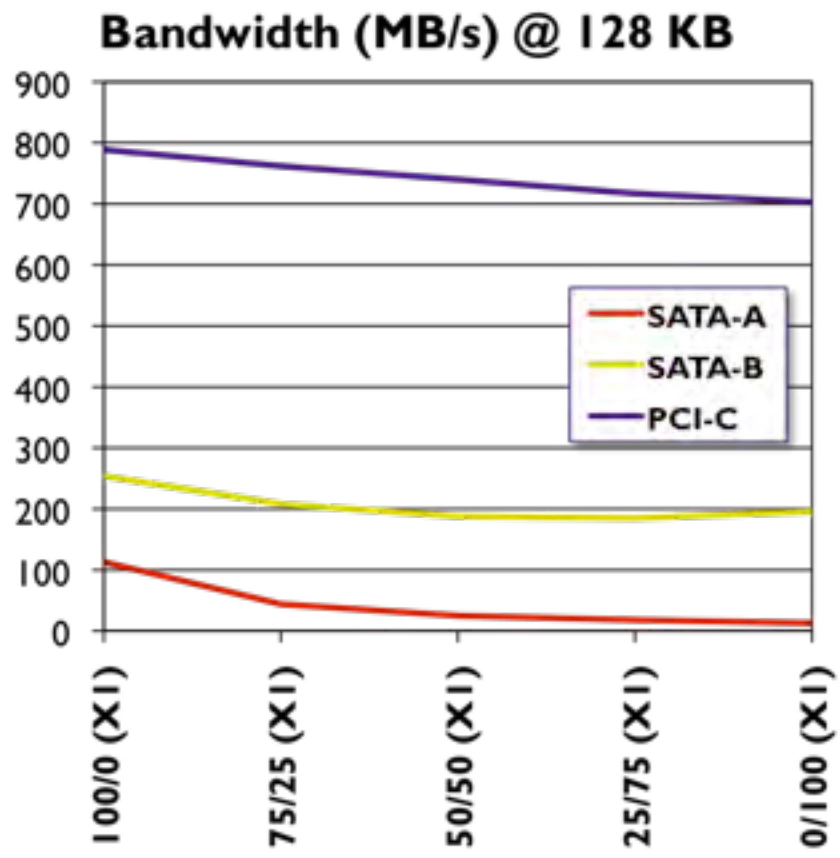
Education SNIA



Performance vs R/W Ratio

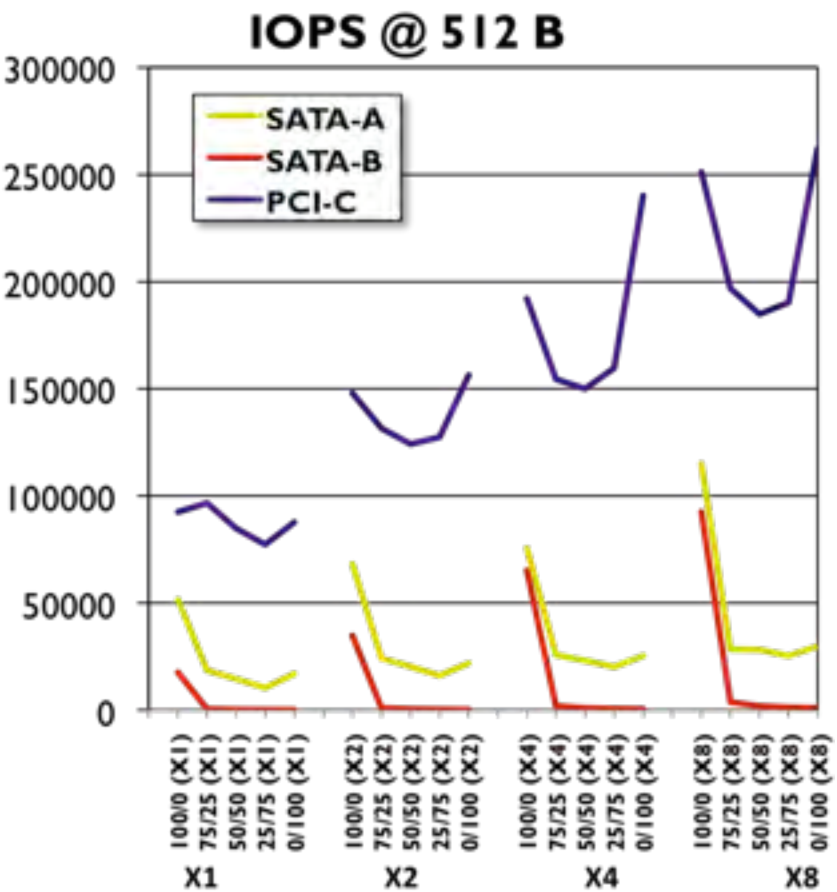
Education SNIA

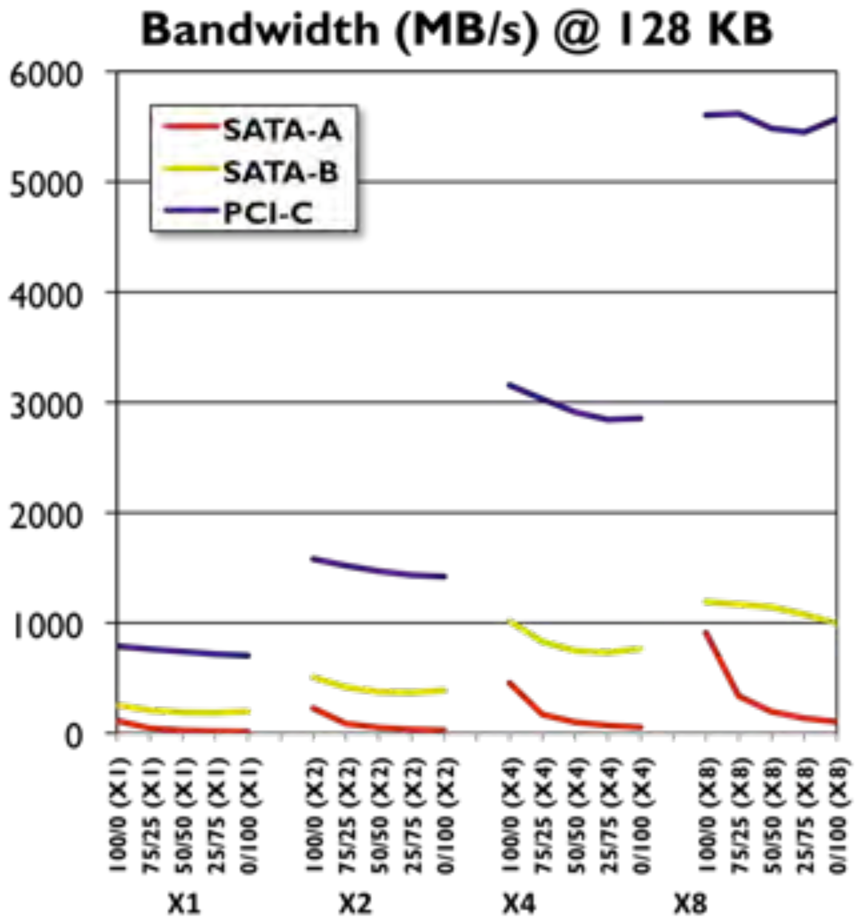




Read/Write Collisions → Drop in Mixed Performance

Scalability versus R/W Ratio

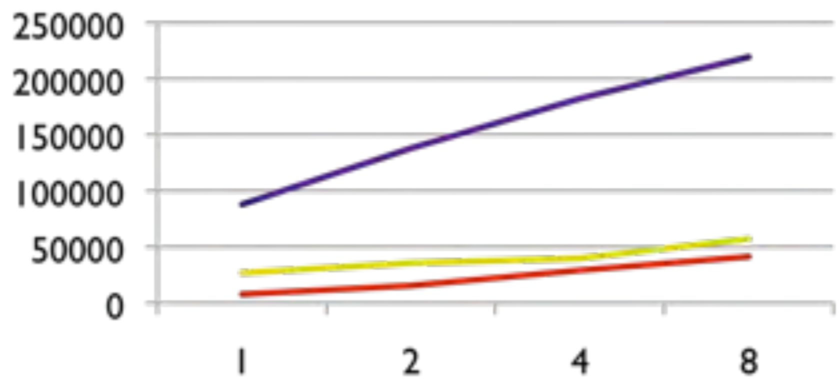




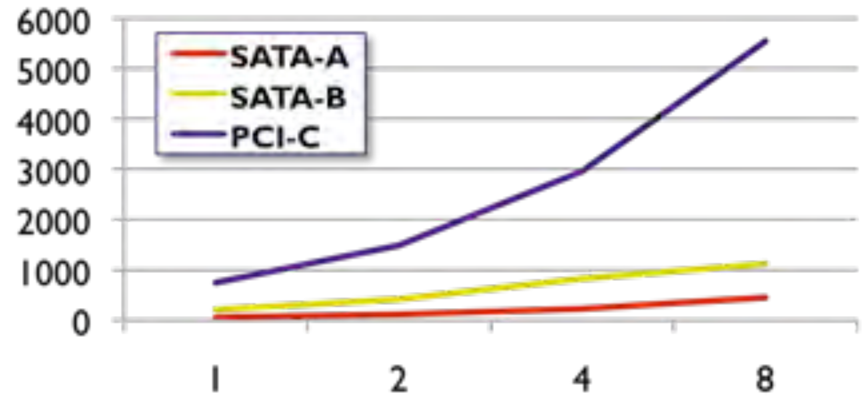
R/W Ratio and Number of Devices in Parallel

RMS Scalability (# SSS Units)

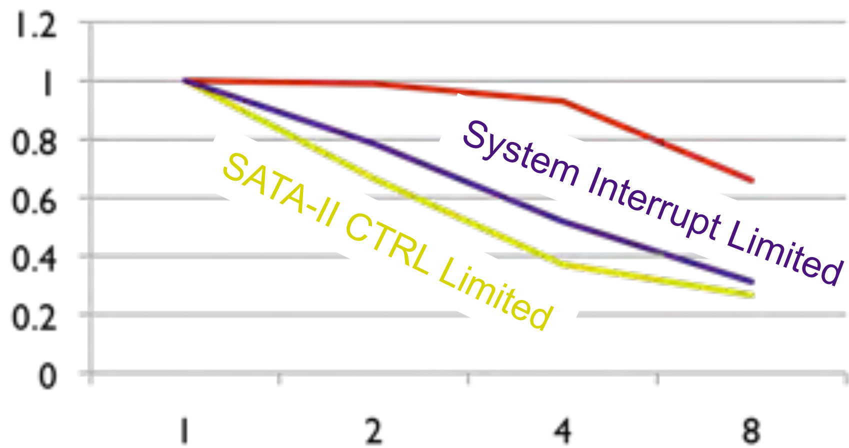
RMS of IOPS v Scale



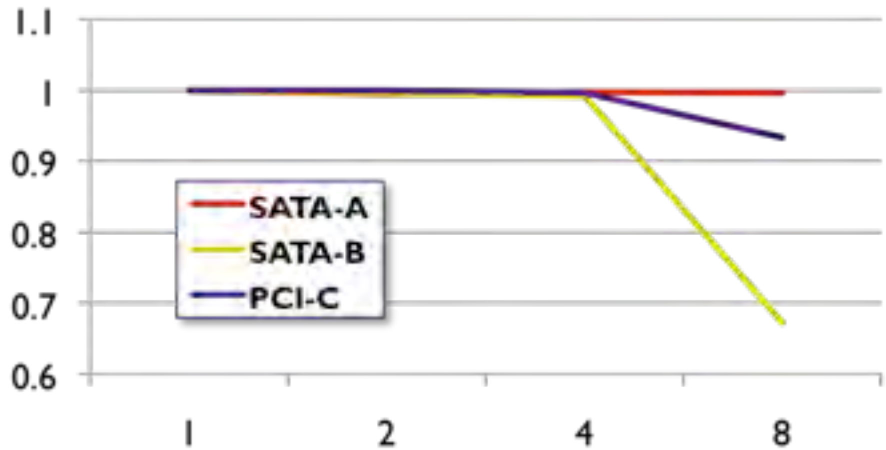
RMS of Bandwidth v Scale



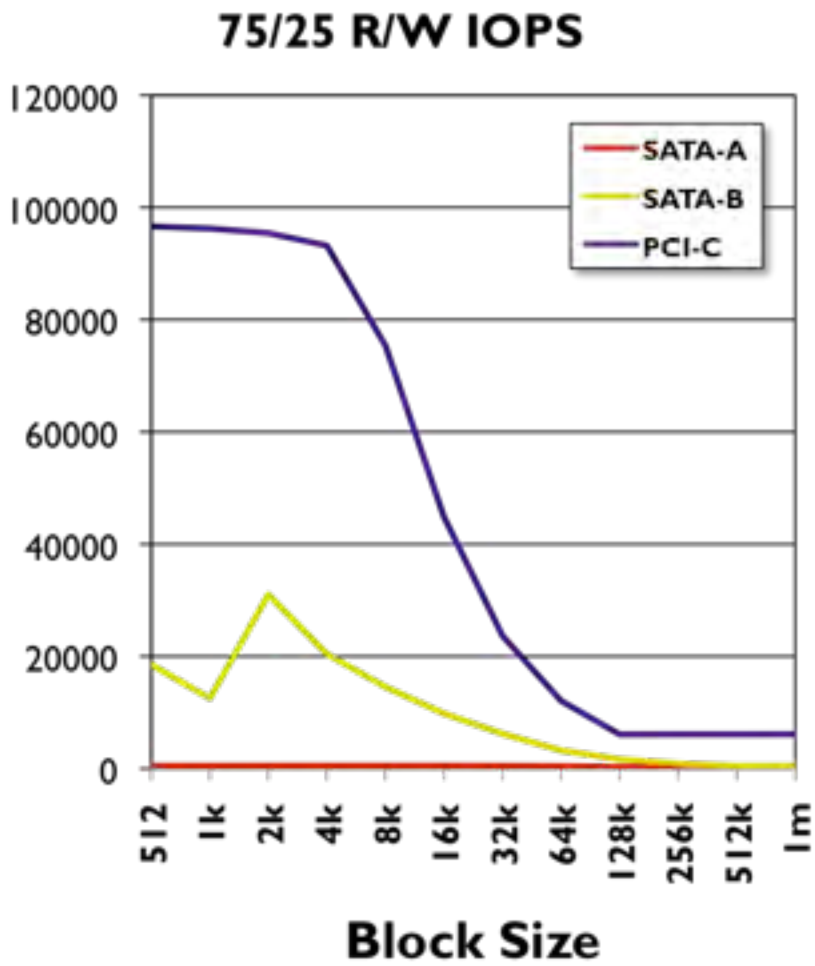
Normalized RMS IOPS v Scale

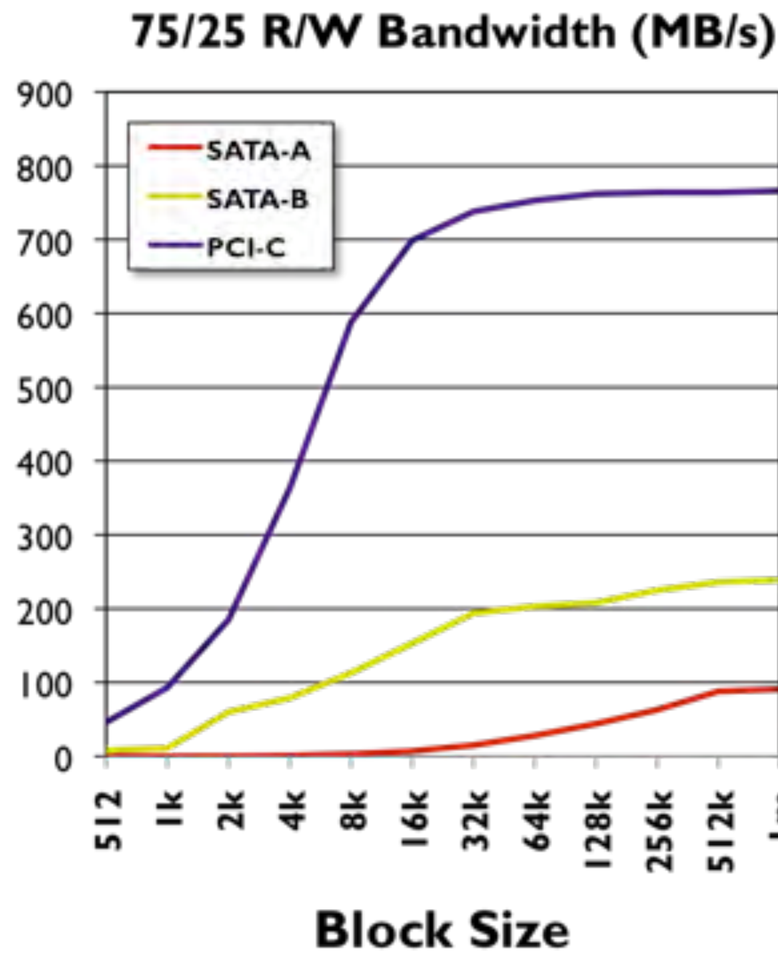


Normalized RMS Bandwidth v Scale

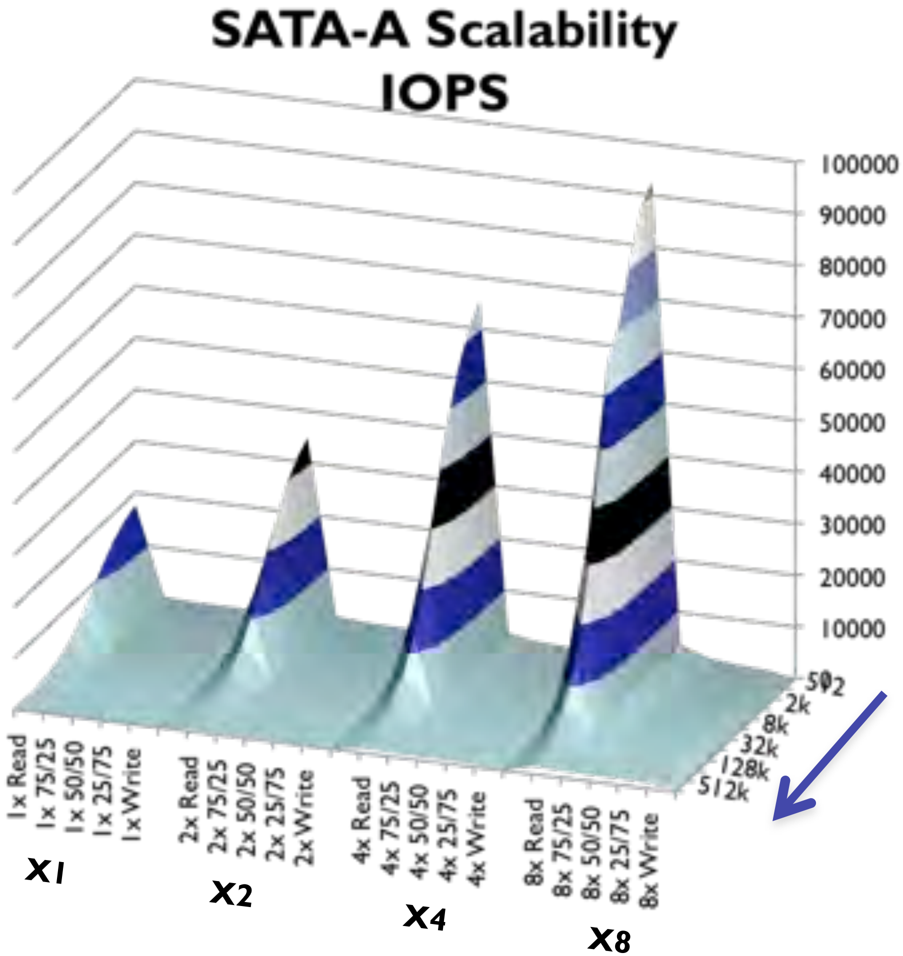


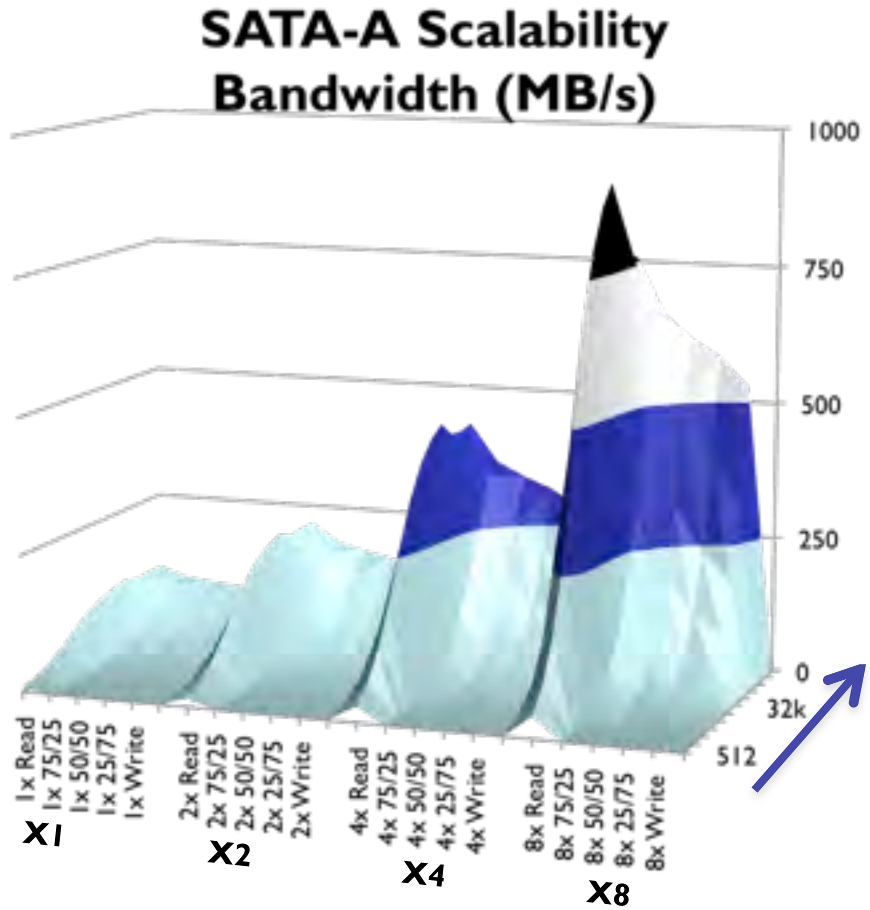
Performance vs Block Size (75/25)



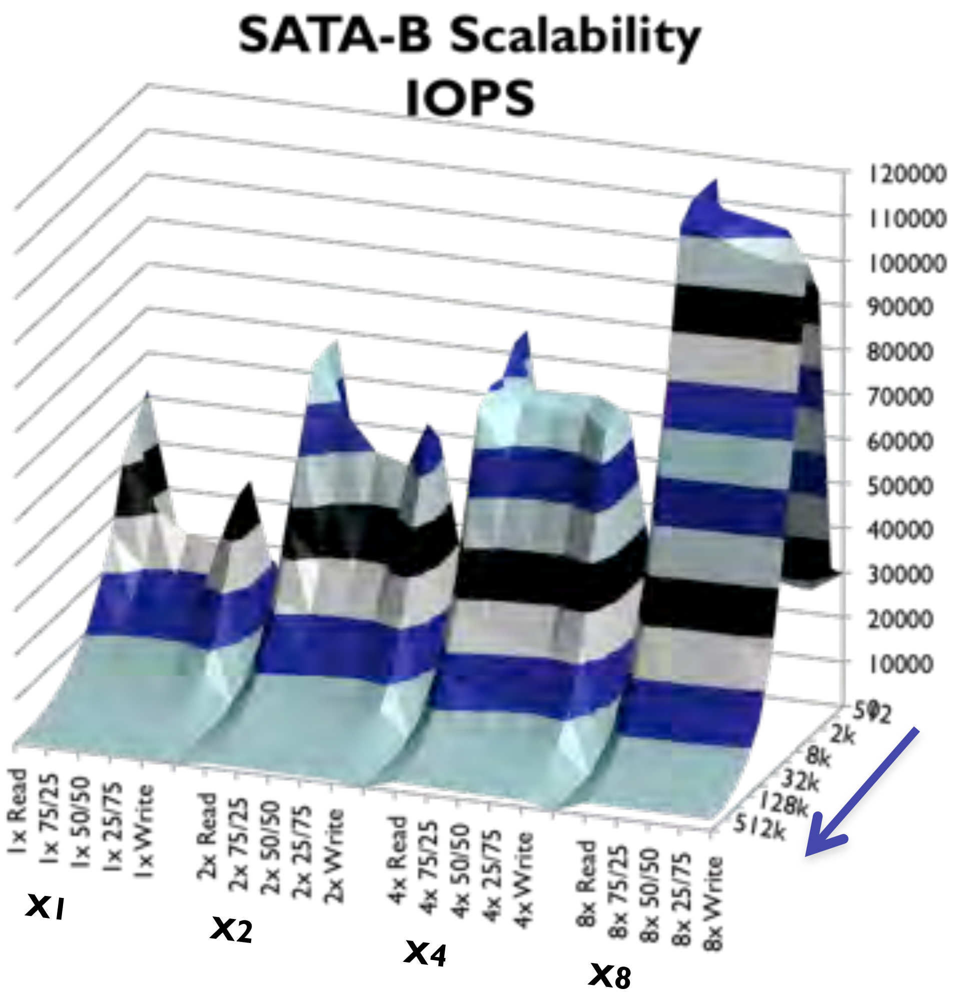


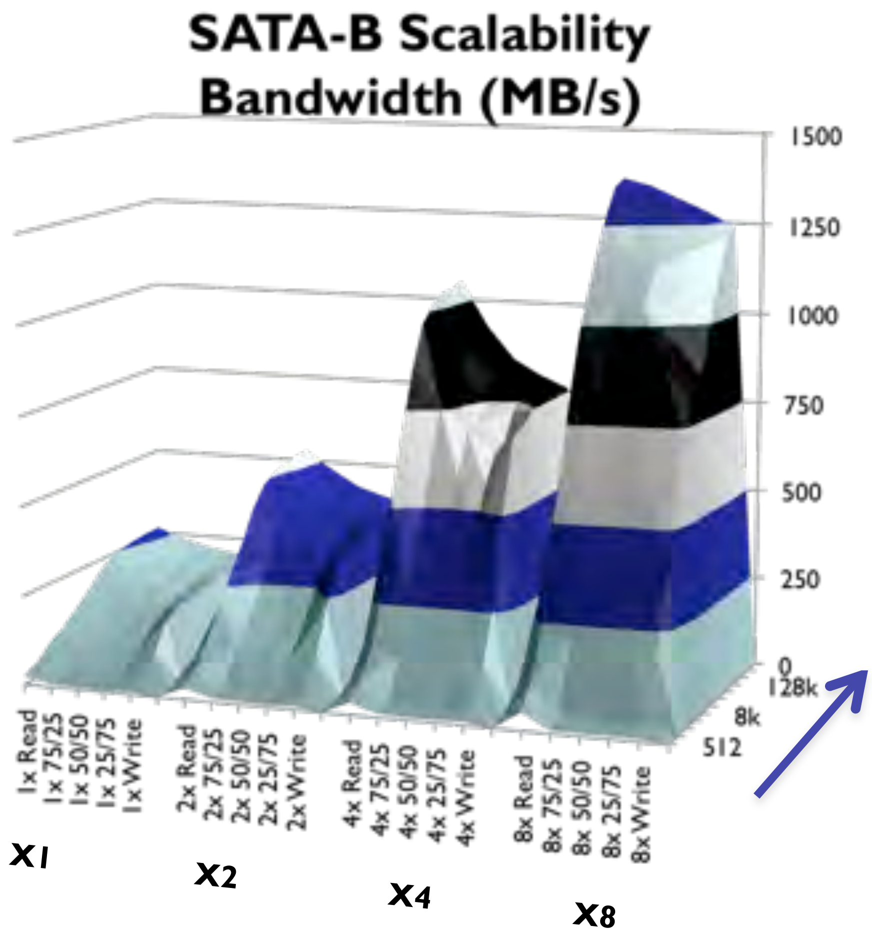
SATA-A Scalability vs R/W vs Block Size





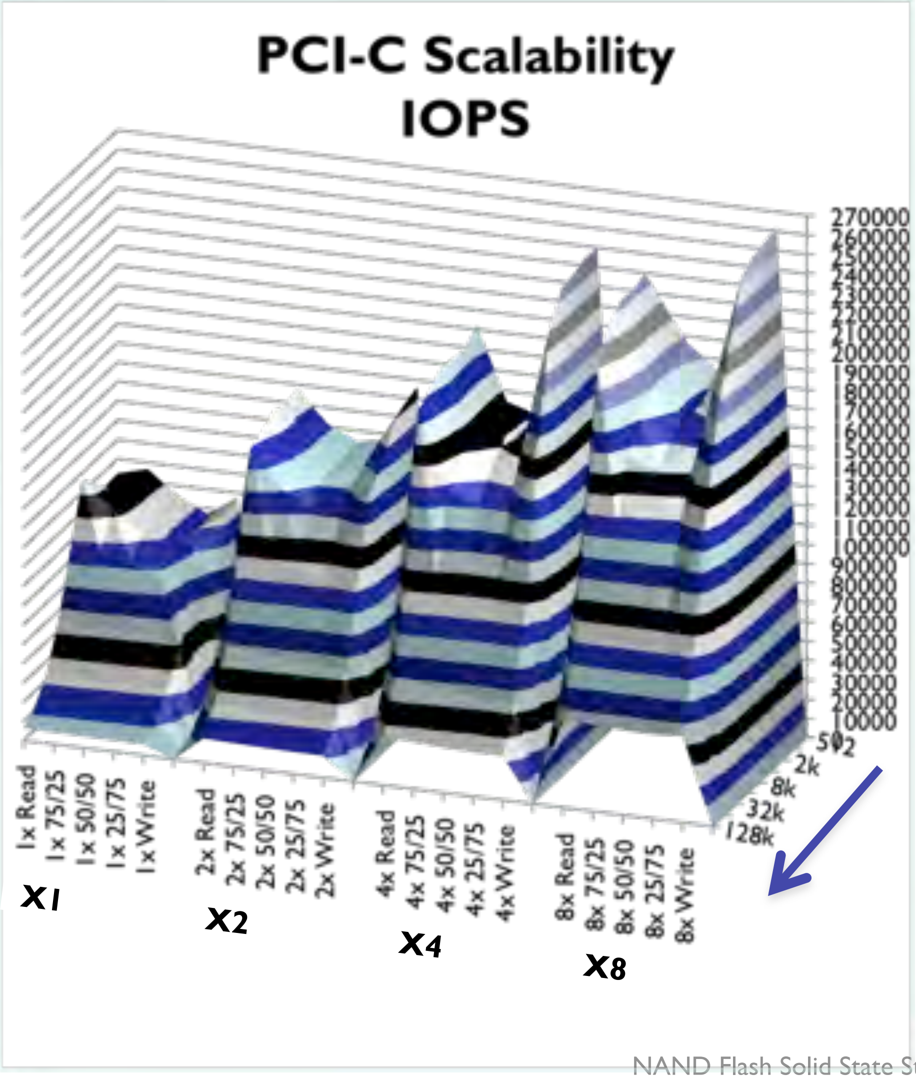
SATA-B Scalability vs R/W vs Block Size

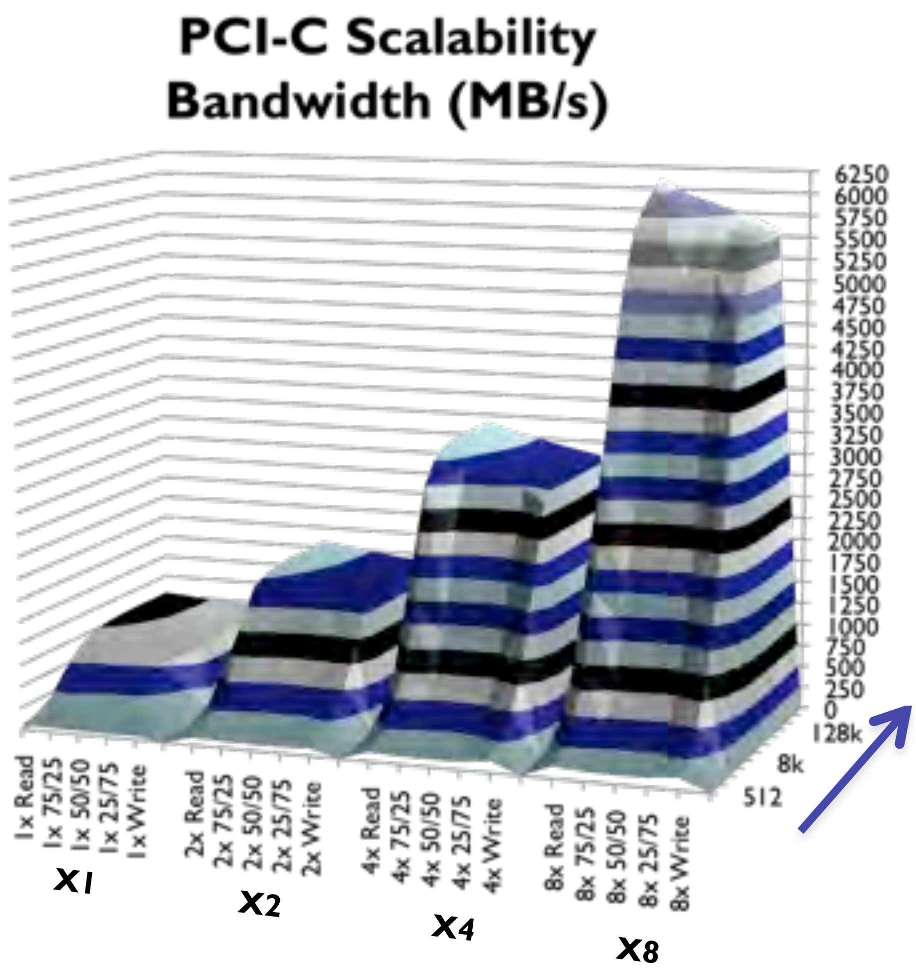




PCI-C Scalability vs R/W vs Block Size

Education SNIA





System Level Considerations

Education SNIA

- Data / Index Protection (RAID and DIF)
- Scalability
- Compare system- or data-center-level
  - Not device
- Best case: test on real application
  - Not benchmark
  - Plan to do tuning to reach top perf. / objectives
  - Applications may have contra-indicated optimizations
    - Keeping data in close physical proximity (short stroking)
    - Caching algorithms

Questions to Ask : Things to Know

Education SNIA

- Bandwidth / IOPS at
  - **Block size(s) you need**
  - **R/W ratio you use**
  - **Steady State / Burst**
  - Reserve capacity used
  - Data's temporal relationship
  - Scalability
  - RAIDing
  - BOL / EOL
- Design impacts on data integrity; life; failures & perf.
  - ECC robustness
  - Write amplification / GC efficiency
  - Internal RAID
  - Bandwidth throttling
  - Partial Page Programming
- Test Conditions
  - Workload
  - Temporal Relationships
  - User capacity / reserve capacity

Q&A / Feedback

- Please send any questions or comments on this presentation to SNIA: : ***tracksolidstate@snia.org***

Many thanks to the following individuals for their contributions to this tutorial.
- SNIA Education Committee

Jonathan Thatcher
Khaled Amer
Phil Mills
Rob Peglar
Marius Tudor