SNIA

Education

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

jonathan thatcher, Fusion-io

SNIA Legal Notice

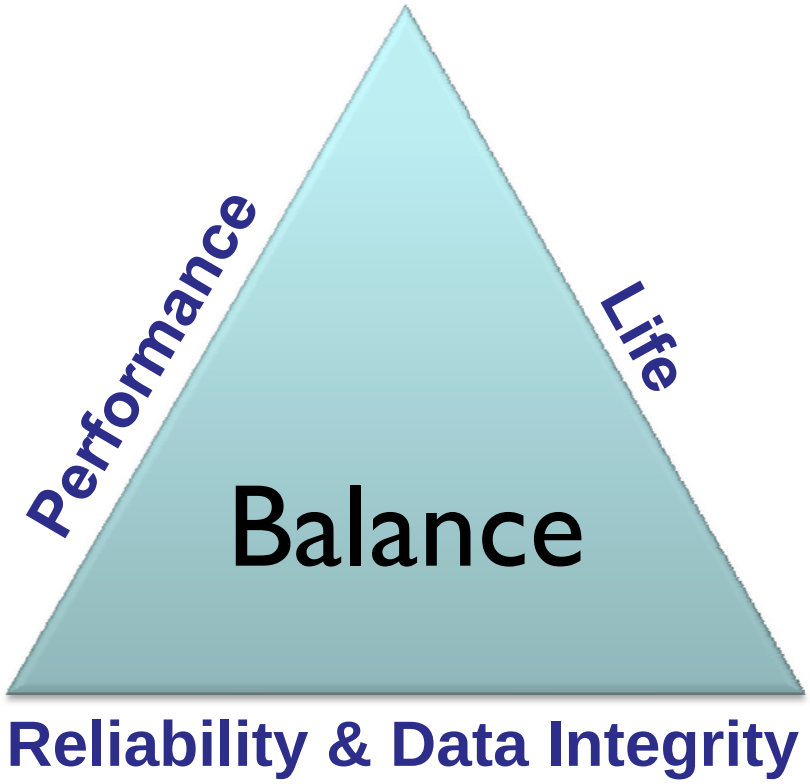
- The material contained in this tutorial is copyrighted by the SNIA.
- Member companies and individual members may use this material in presentations and literature under the following conditions:
  - Any slide or slides used must be reproduced in their entirety without modification
  - The SNIA must 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 tutorial 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. Several examples of successful enterprise deployments with comparative performance and cost-performance will be presented."

Data Integrity + Performance



There can be no data integrity trade-off for performance

Media Reliability / Availability

- The GOOD:
  - No moving parts
  - Catastrophic device failures are rare (post infant mortality)
- The BAD:
  - Relatively high bit error rate, increasing with wear
    - MLC wear rate (higher capacity density) worse than SLC
    - Higher density NAND Flash will increase bit error rate
  - Program and Read Disturbs
- The UGLY:
  - Partial Page Programming
  - Data retention is poor at high temperature
  - Infant mortality is high (large number of parts…)

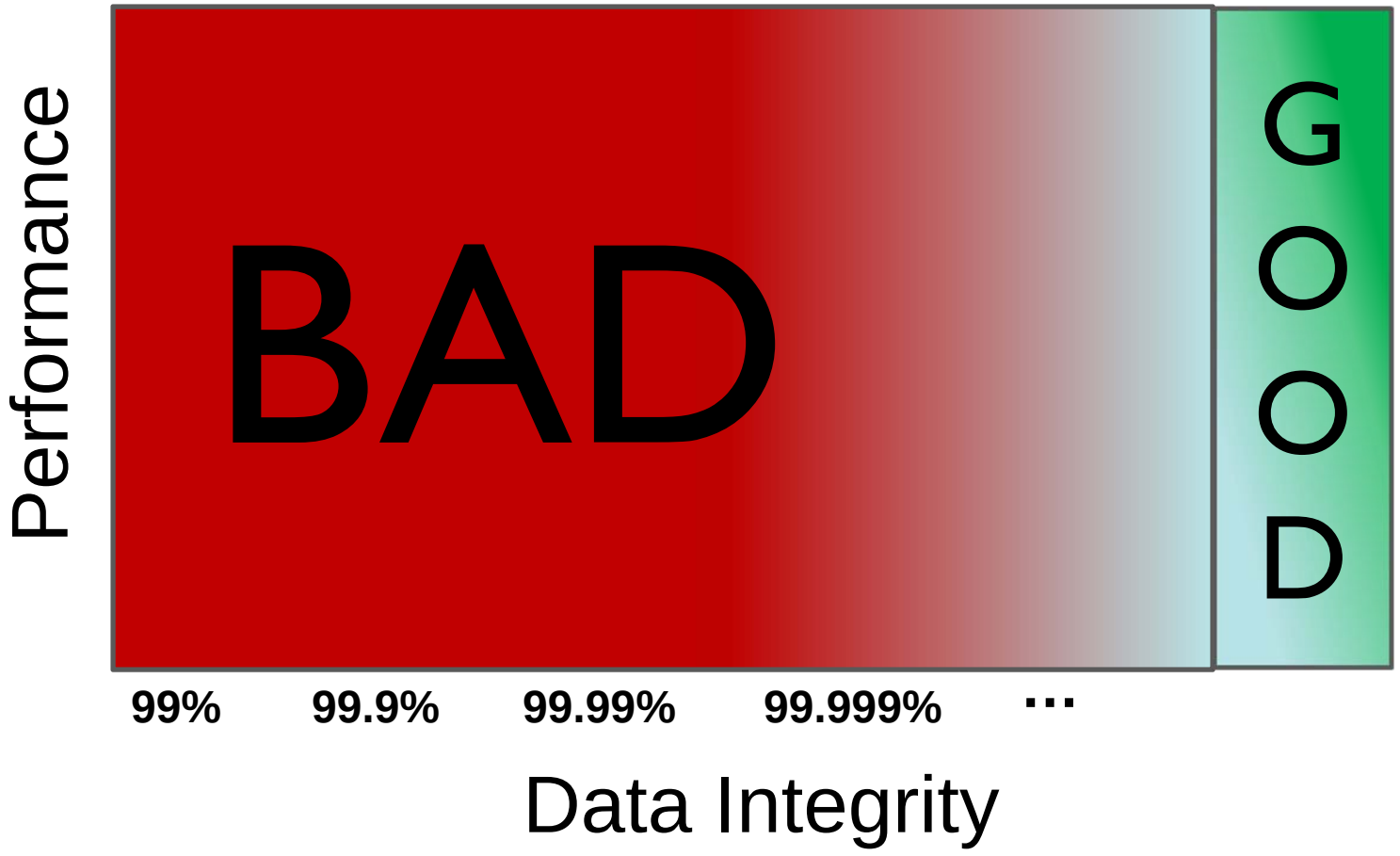
Controller Reliability Management

- Wear leveling & Spare Capacity (e.g. Spare Blocks)
- Read & Program Disturb Controls
- Data & Index Protection
  - ECC Correction
  - Internal RAID
  - Data Integrity Field (DIF)
- Management

Poor Media + Great Controller → Great SSS Solution

Note: Multipage Programming Should Not Be Done

Data Integrity .V. Performance



Performance is about ROI

Education SNIA

ROI

Lower CapEx

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

Lower OpEx

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

Higher Productivity

SSS ROI: Three Quick Examples

Case Studies

Wine.com

Cloudmark

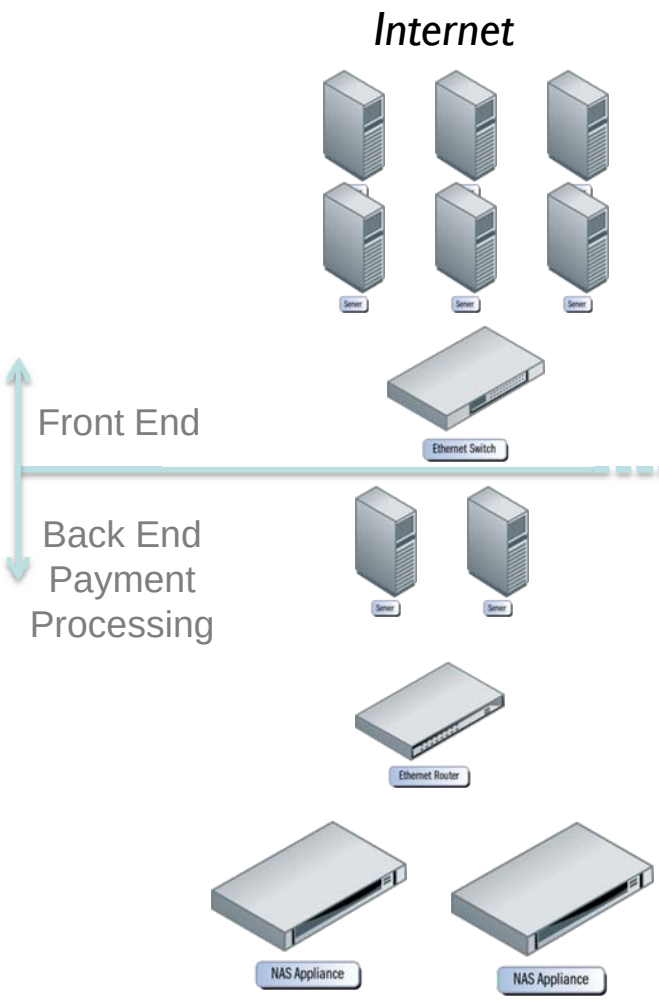
O&G

Company: Wine.com

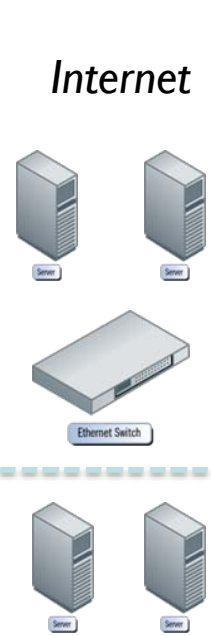
- On-line internet Retail
- Representative Markets:
  - Transaction Processing
  - Data Mining
- Problem
  - Systems not capable → Contract out data mining
  - Performance hitting 100%→ Loss of revenue (4Q07)
  - Cost to meet growth: prohibitive

Wine.com

Before (3Q2008)



After (1Q2009)



Application level replication for redundancy

33% CapEx avoidance
50% OpEx reduction
33% Footprint reduction

12x improvement on write
- Latency down from 4 to <1ms

14x improvement on read
- Latency down from 12 to <1ms

"Enough capacity to cover 24 months of growth"
Geoffrey Smalling - CTO

Company: Cloudmark

- Email Security – Spam, Phishing, Viruses, etc.
- Protecting 850 Million mailboxes globally
  - 100+ Service Provider customers (mostly Tier-1)
  - High TPS at master – morning bursts are higher
- Problem scaling MySQL & InnoDB worldwide
  - Rapid growth – both customers and new threats
  - I/O limited → hit performance ceiling 50% of the time
  - Slave servers fall behind → increasing risk window
  - High transaction rate → constant disk failures
  - Cost to meet growth: prohibitive

Cloudmark

- Transformed slave clusters to Solid State Storage and retired disk arrays – results:
  - 4 to 1 footprint reduction
  - CapEx avoidance - $2 million
  - 10x performance increase
  - Cluster peak utilization now only 15% of IOPS capability
  - 12 months growth secured with current hardware

"We never thought our problems would ever change from disk I/O problems to CPU bottlenecks!"

Ryan White – Director of Operations

Company: LLOG

- Oil & Gas exploration
- Market:
  - Natural resource discovery using seismic interpretation leading to production
  - Seismic analysis software used for 3D interpretations
    - Manipulation and analysis of large datasets (1Tb+)
  - Quality interpretations result in accurate auction bids
    - Minimize risk and make informed decisions
- Geoscientist productivity challenges
  - Idle time loading data into high-end workstations
  - Lost productivity during basic analysis tasks
  - Time wasted in running jobs in serial

LLOG

- Use of Solid State Storage on workstations
  - Load time reduced from 8 to 2 minutes
  - 3D time slice time reduced from 28 to 18 minutes
- "Supercharged Virtualization" with Solid State Storage
  - Projects took 30+ minutes causing 100% CPU utilization
    - ***Now 10 minutes with projects run in parallel***
  - Increased geoscientist's productivity 5x (projects in parallel)

"We are more competitive because we make better decisions and did it while reducing our costs"

John Hollins – Geophysicist

Performance Matters

Performance

ROI

Media Performance

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

Performance Drivers – SSS Design

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

Performance Drivers - External Cond

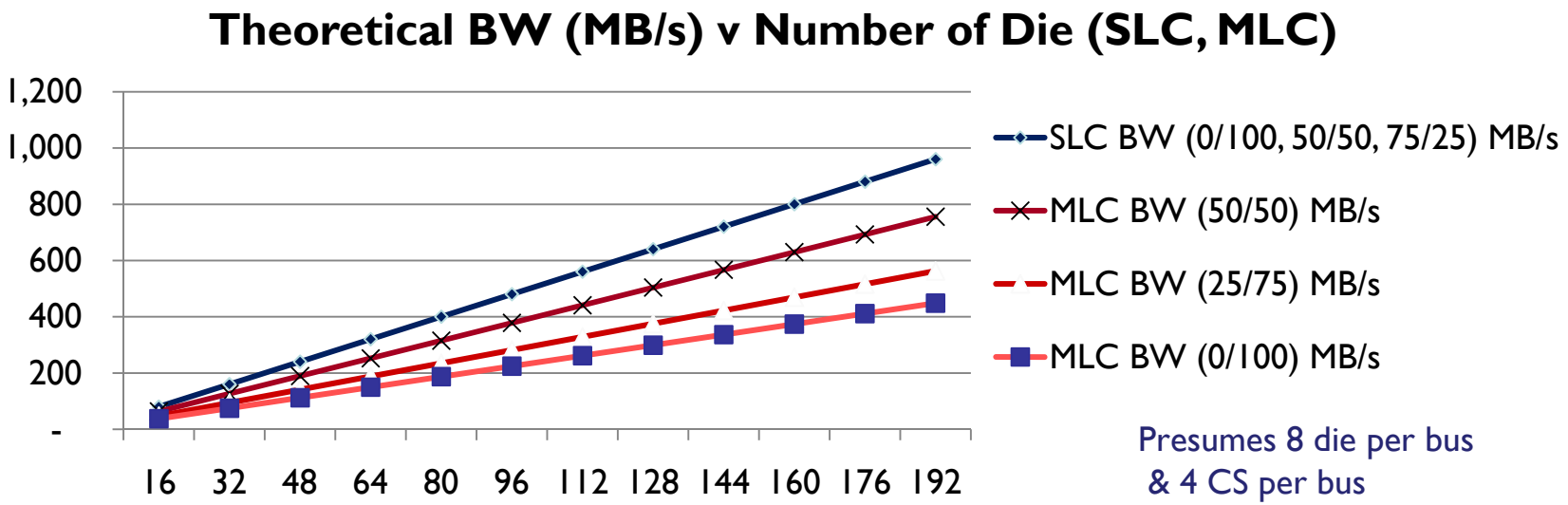
- Transfer Size
- Read/Write Ratios
- Temporal Randomness of Access
- Reserve Capacity Setting (% of used capacity)
- System Limitations (especially wrt scalability)
  - External Controller (#, Type, Performance); # Threads
  - CPU (#Cores, GHz)
  - System Bandwidth
  - Software Stack; Interrupt Handler
- External RAID
- Life of device (change in device affects tErase & tProgram)

Simplified Theoretical Analysis

- Bandwidth Only (Not IOPS)
  - Large Transfers (Data length = Integer * # die)
  - 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
- SLC

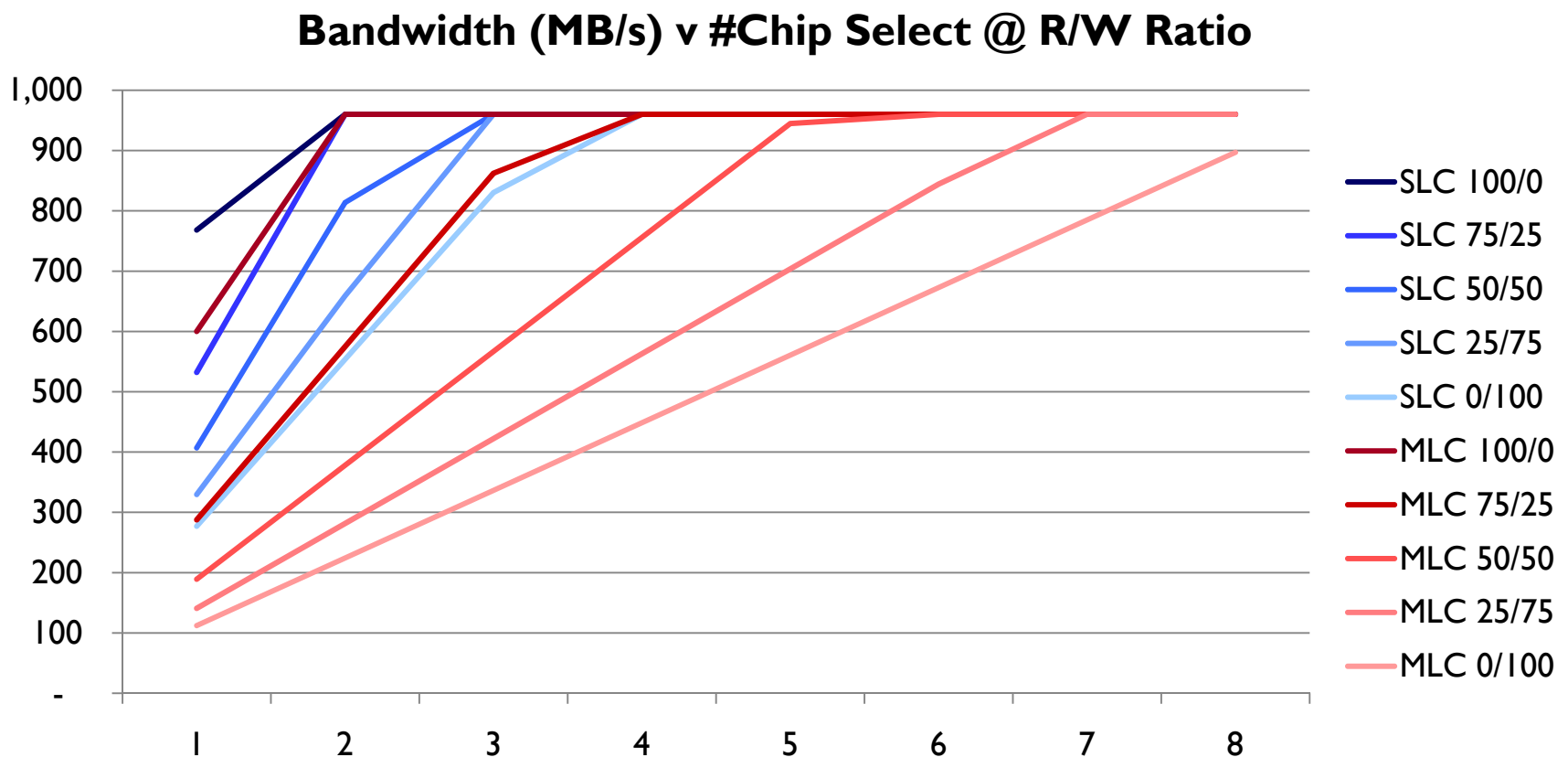
Bandwidth Depends on # Die

		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



Presumes 8 die per bus
& 4 CS per bus

Single Layer Cell v Multi-Layer Cell



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

Performance Data Acquisition

- In this presentation, three commercially available SSS storage solutions were tested under a variety of conditions using a common system:
- Supermicro X7DWE Motherboard
  - Dual, Quad Core Intel X5460: 3.16 GHz
  - 16 GB DRAM
    - 4x4GBKVR800D2D4F5 /4GI, 800MHz FBDIMM DDR2, CL5
  - PCI-E Gen 2 bus
- SATA Controller LSI SAS3081E-R
  - All tests except pathological write, which used on-board SATA
  - Driver version used was: 4.00.43.00; firmware rev: FwRev=011b0000h

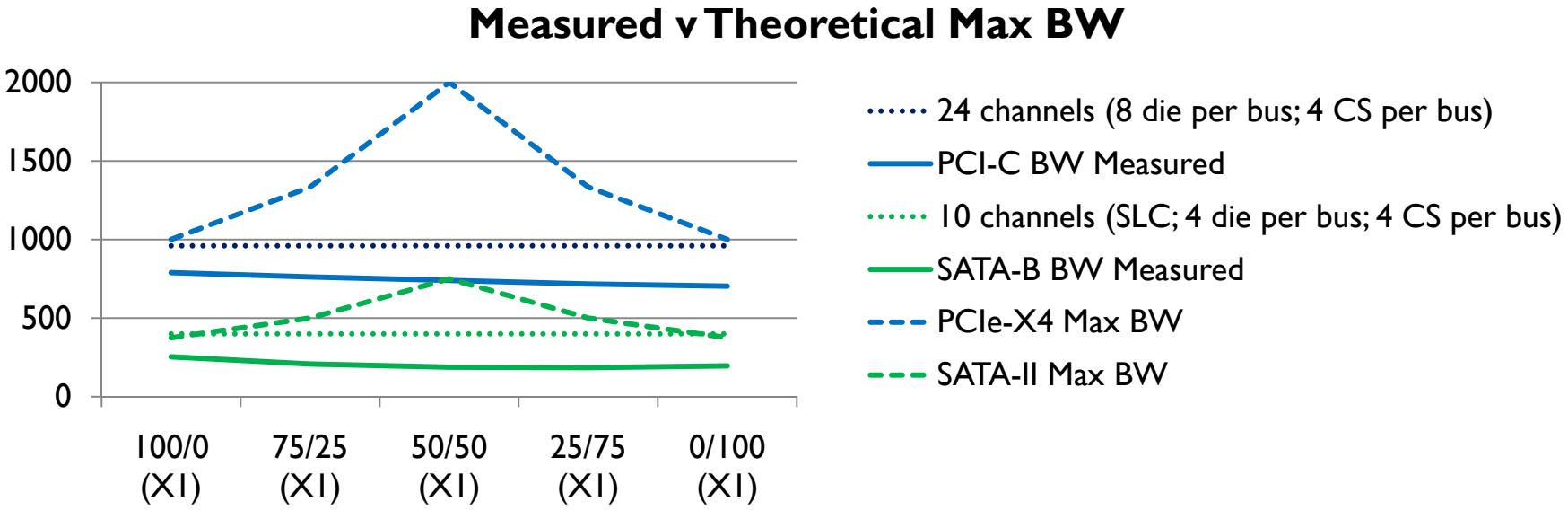
Note: all data collected on devices at beginning of life

Performance Data Acquisition

- Software / OS
  - Linux/CentOS5/RHEL5 2.6.18-92.1.10.el5.
  - "fio" program, version 1.21:
    - From http://freshmeat.net/projects/fio)
  - "IOMeter" V 2006.07.27
    - in Windows 2003/2008 for latency only
- Settings
  - Direct I/O: o_direct used to bypass caching & buffering
  - I/O Scheduler (elevator algorithm) set to null

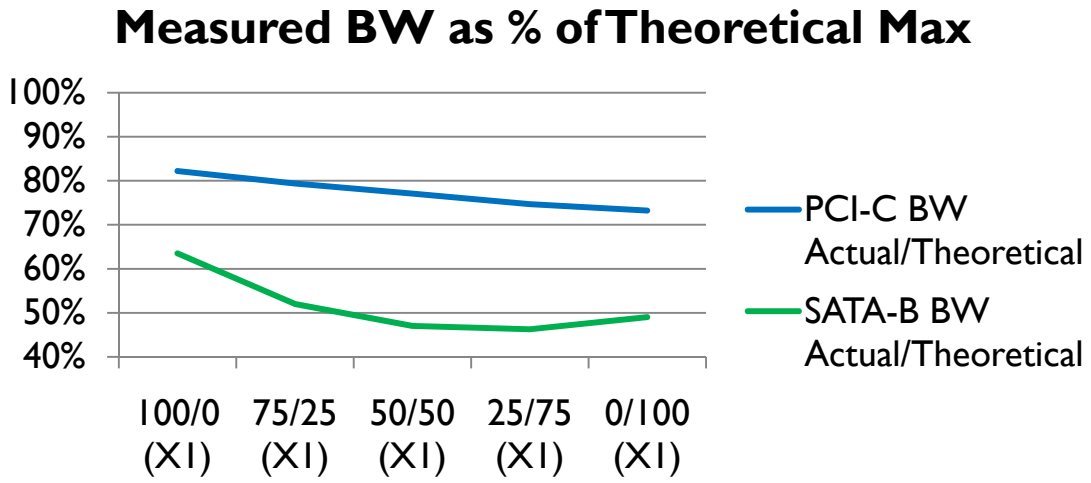
Note: all data collected on devices at beginning of life

Measured v 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



Product Specifications

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	?	11 bits in 240B

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

Buffer holds data while EB-1 Erased

Page	Erase Block 1			
0				
1				
2				
3				

Page	Erase Block 1			
0				
1				
2				
3				

Explanation of Previous Slide

- Assumptions
  - Simplified to show erase blocks with 4 pages, each page having 4 data blocks
  - Invalid (erased or replaced) data is indicated by "—"
  - Old data is indicated by lower case letters
  - New data is indicated by CAPs; Replacement data is indicated by "prime" (e.g. c → C')
- Detail T = t1 to T = t2 transition
  - Data is read from EB-1
  - EB-1 is erased
  - New data {W,X,Y} modifies previous invalid data
  - Data is written back to EB-1
- Detail T = t2 to T = t3 transition
  - Data is read from EB-1 into data buffer
  - EB-1 is erased
  - New data {B', C', Z, A, R'} modifies previous data in data buffer
  - Data is written back to EB-1

Note: backup material for those reviewing or looking at presentation without audio/video

Example 2: Read/Modify/Write

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

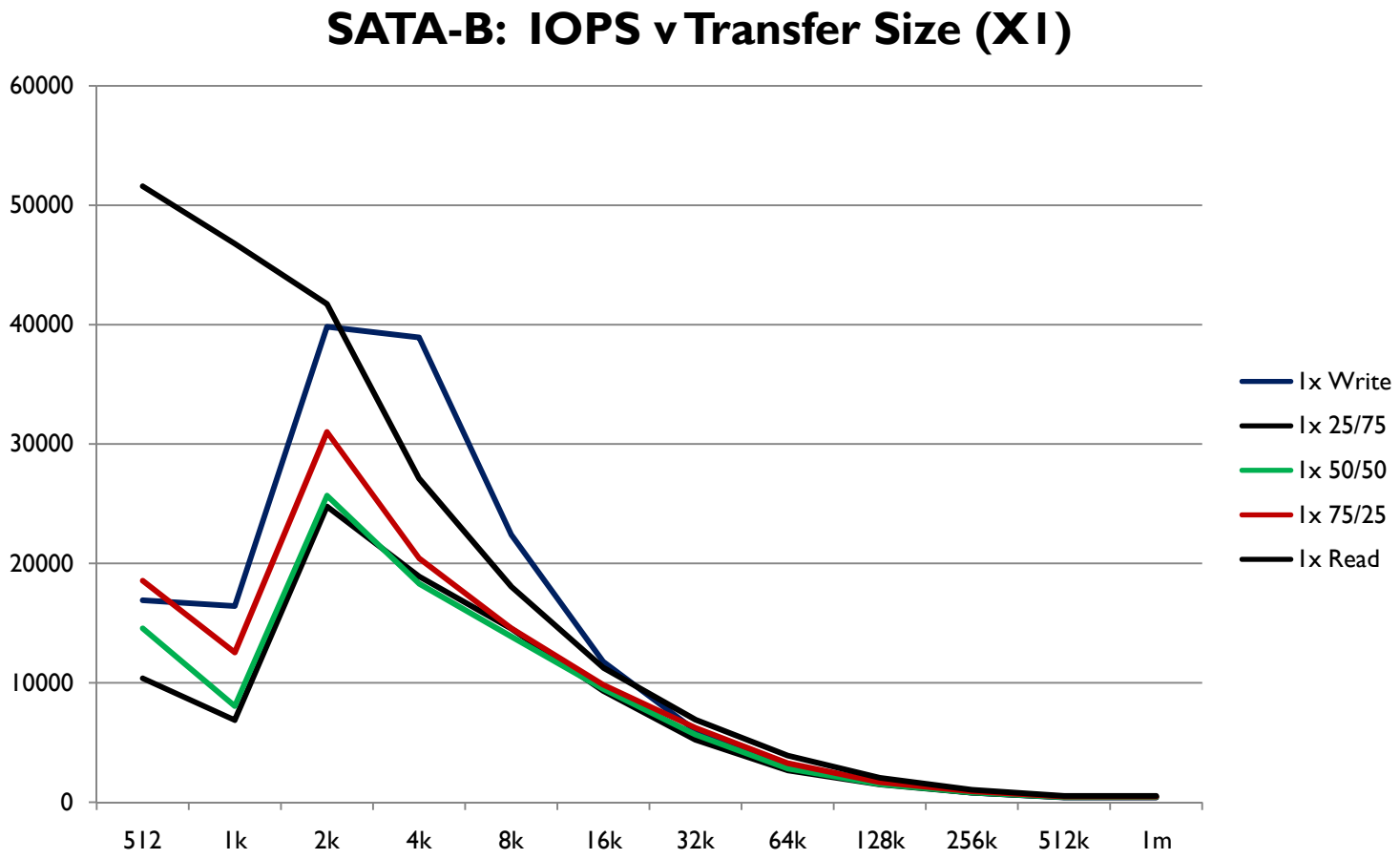
Page	Erase Block 3			
0	B'	C'	w	x
1	j	y	k	l
2	m	Z	A	
3			q	R'

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)

"Write Amplification Impact"

- In this example,
  - Data written t1 to t2: 16 blocks
    - NEW DATA {W, X, Y} 3 blocks; Copied Data {b, c, j, k, l, m, q, r} 8 blocks
    - Null Data: 5 blocks
  - Data written t2 to t3: 16 blocks
    - NEW DATA {B', C', Z, A, R'} 5 blocks; Copied Data {w, x, j, y, k, l, m, q} 8 blocks
    - Null Data: 3 blocks
  - (2) EB erasures
  - 25% (8 of 32) writes are user initiated
  - 75% (24 of 32) writes are internal data movement (overhead)
- Important:
  - Amount of valid or invalid data in EB-1 is irrelevant to performance impact
  - "Write Amplification" is workload (access pattern) dependent (e.g., what if the write of R' above was not coincident with B' & C')

SATA-B: IOPS vs Transfer Size



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	--	B'
3	C'	Z	A	R'

Explanation of Previous Slide

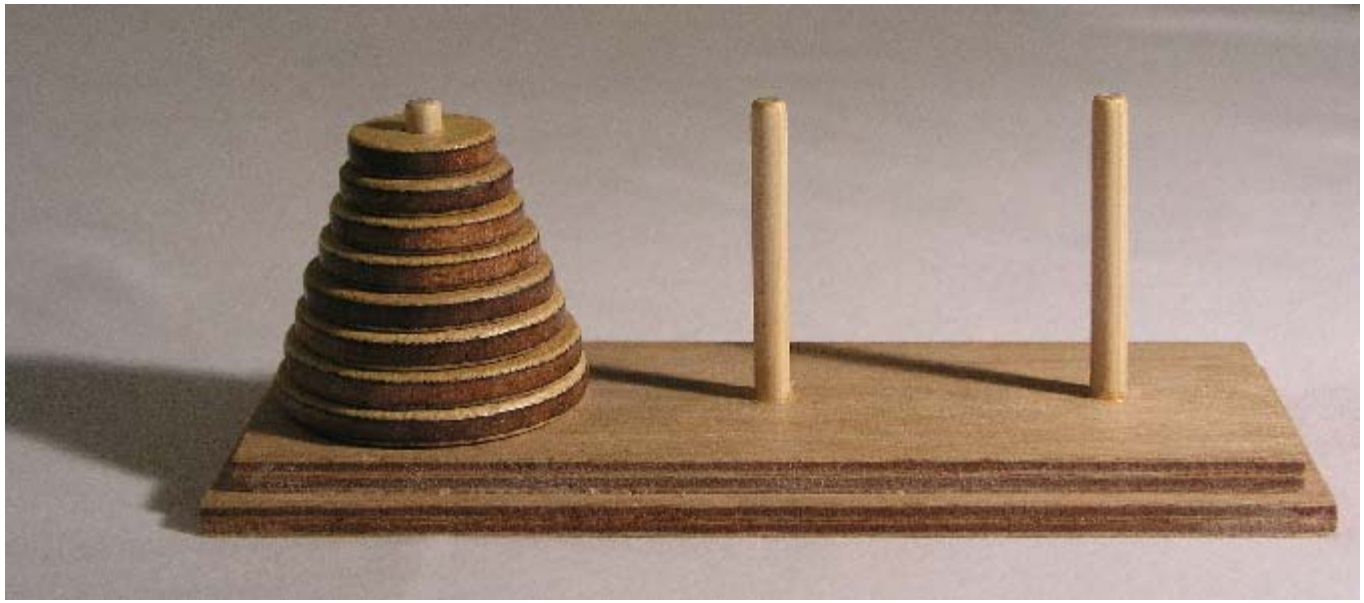
- Assumptions
  - New data blocks and data blocks being garbage collected are interleaved
- Details
  - At Time = t0, erase block 1 (EB-1) is identified for GC
  - At Time = t1, good data is moved from EB-1 to EB-2 (it is implicit that an index is updated accordingly); New data W, X, and Y are added while the GC is taking place. EB-1 is then ready to be erased
  - At Time = t2, EB-2 is erased; Data for b, c, & r have been updated with B', C' & R'; b, c & r are indicated as "Invalid."

Note: backup material for those reviewing or looking at presentation without audio/video

GC Performance Impact

- 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 high percentage of total storage capacity utilized

AND

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

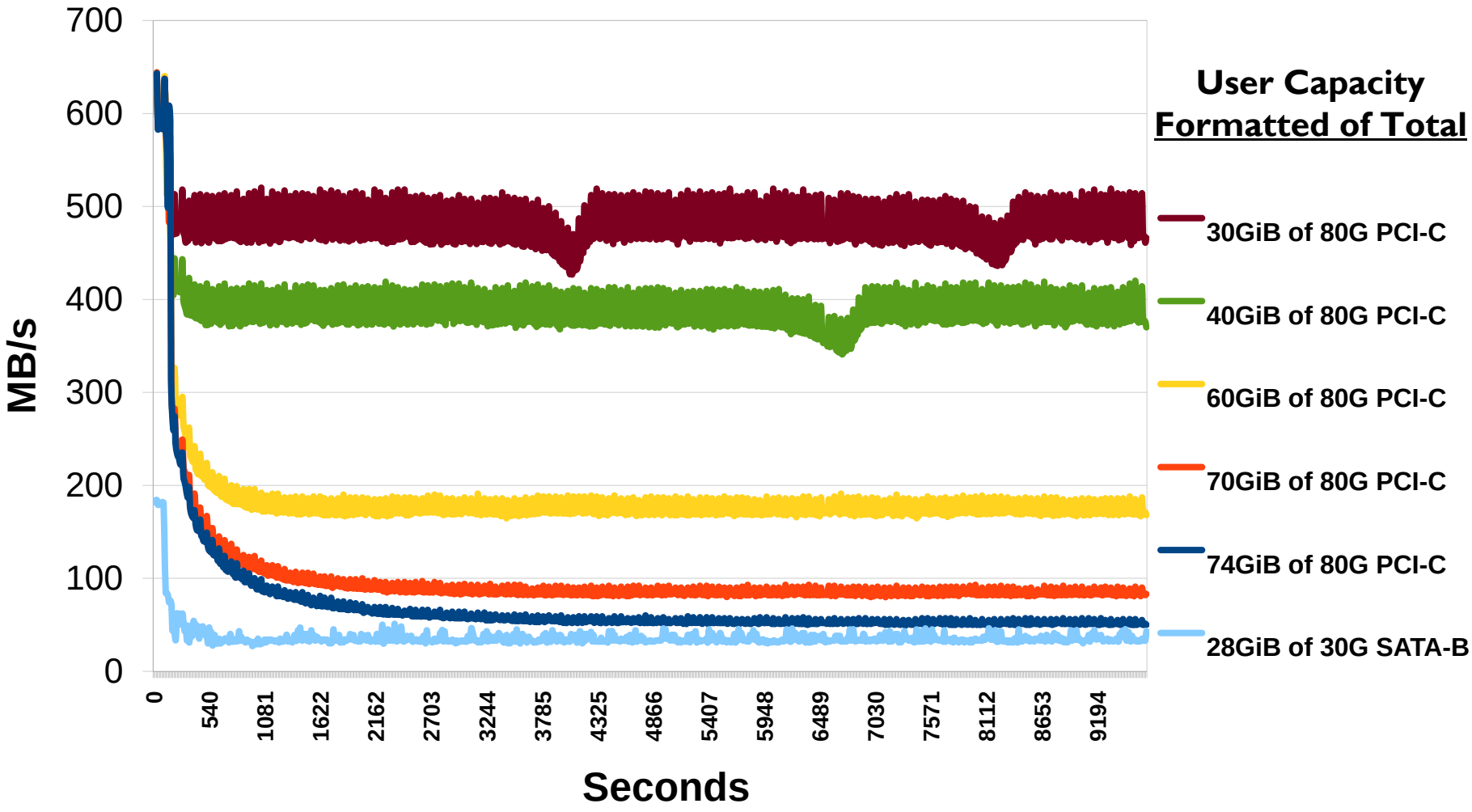
AND

- Continuous writing (no recovery time for GC)

THEN

Efficiency of GC greatly diminished

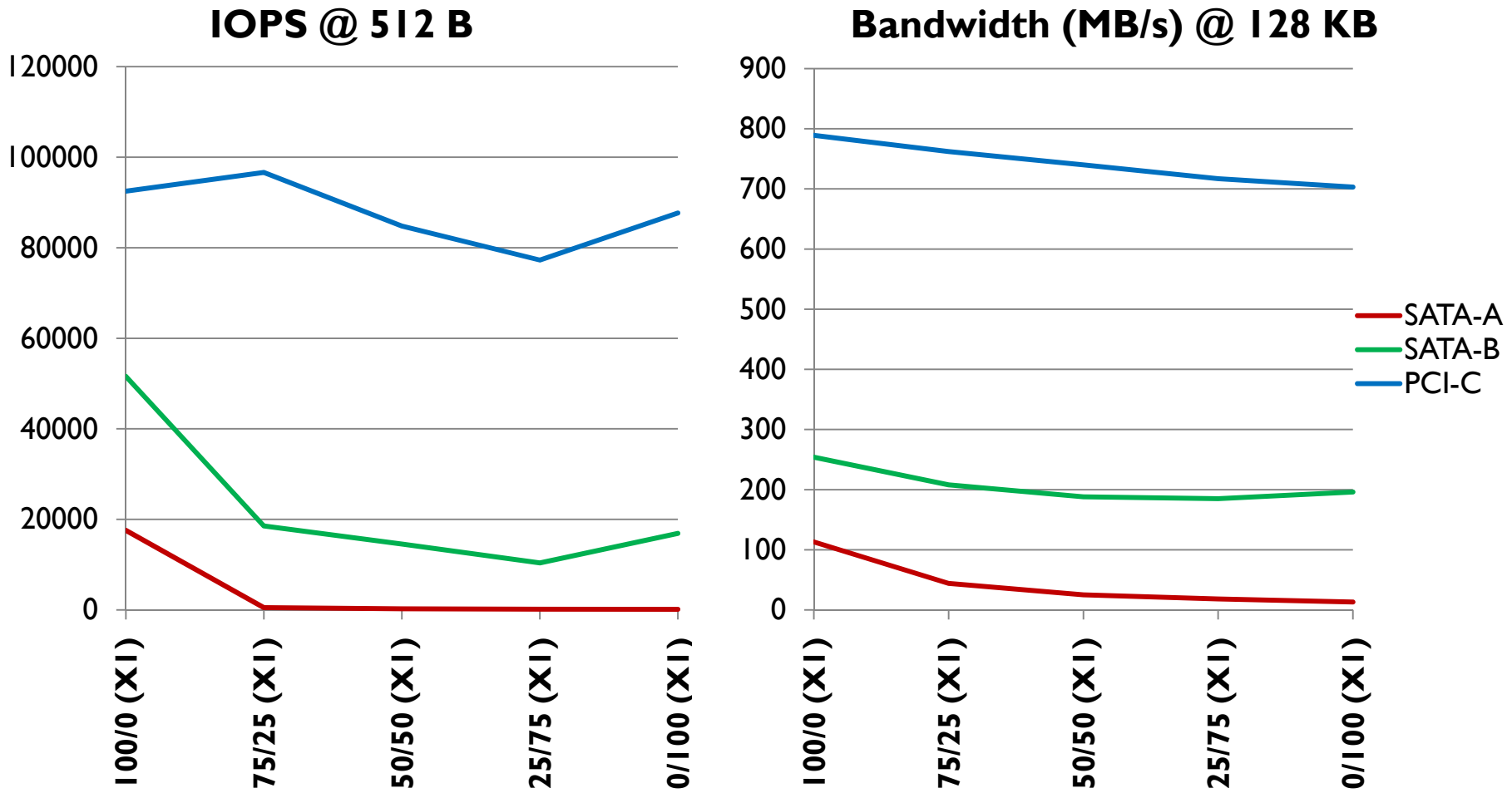
Pathological Write Condition



Scalability

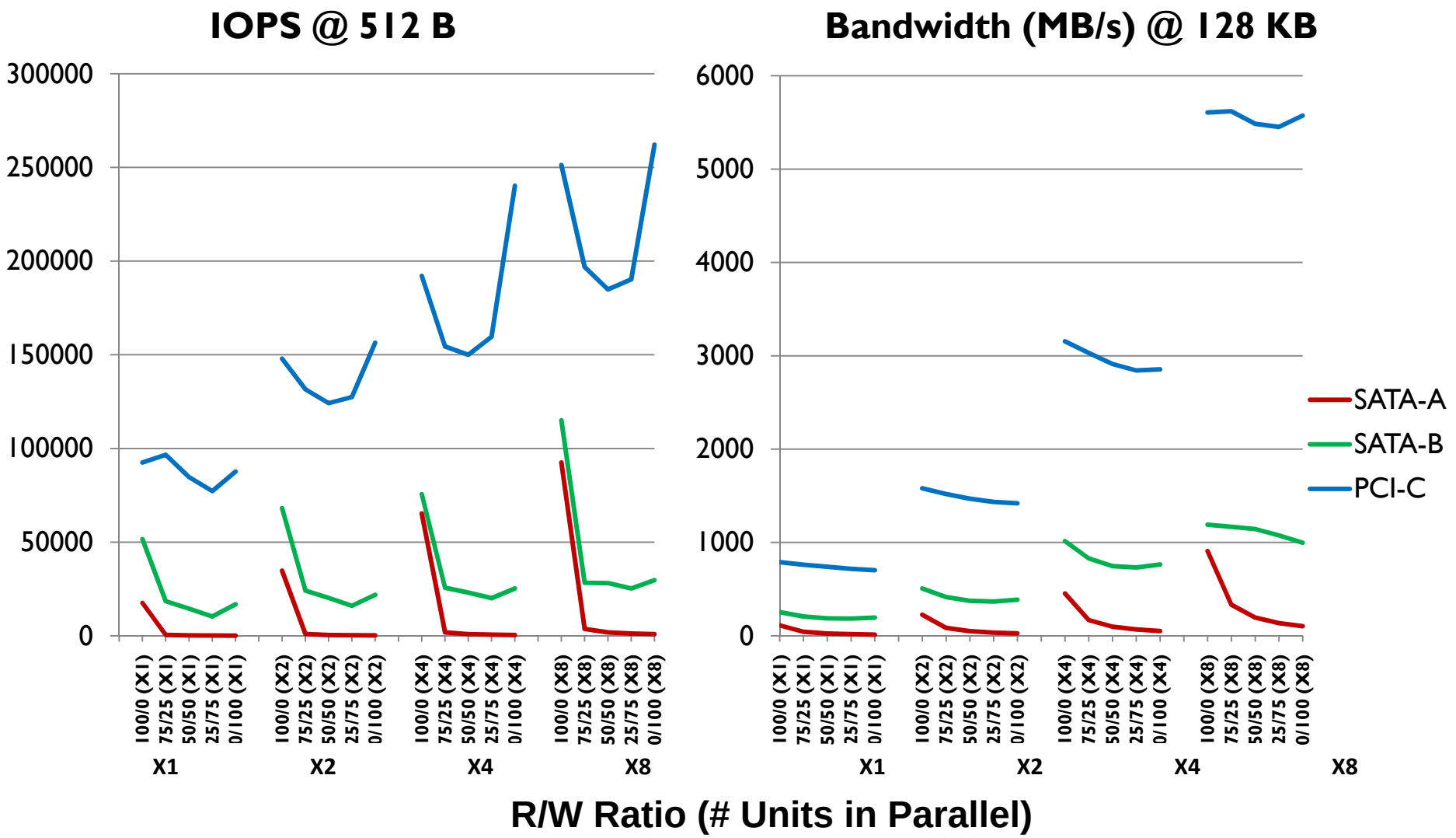
- Following Slides Show Scalability of {1, 2, 4, 8} units
  - Only 1 SATA controller is used – limiting scalability
    - Only 1 thread running
- Measurements taken at Read/Write Ratios of
  - {100/0, 75/25, 50/50, 25/70, 0/100}
  - RMS value is the "root mean square" of these five values
- IOPS measurement taken at 512 Byte Transfers
- Bandwidth taken at 128K Byte Transfers
  - Unless shown differently
  - Linux has a 128K limit

Performance v R/W Ratio

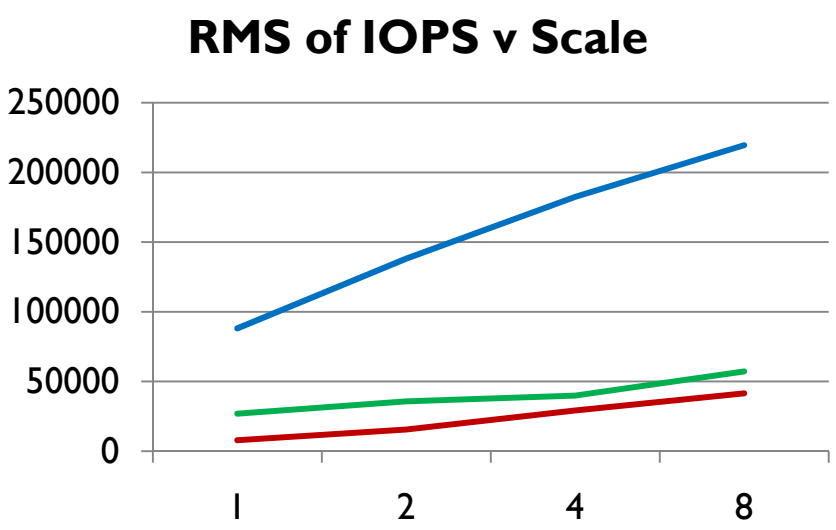


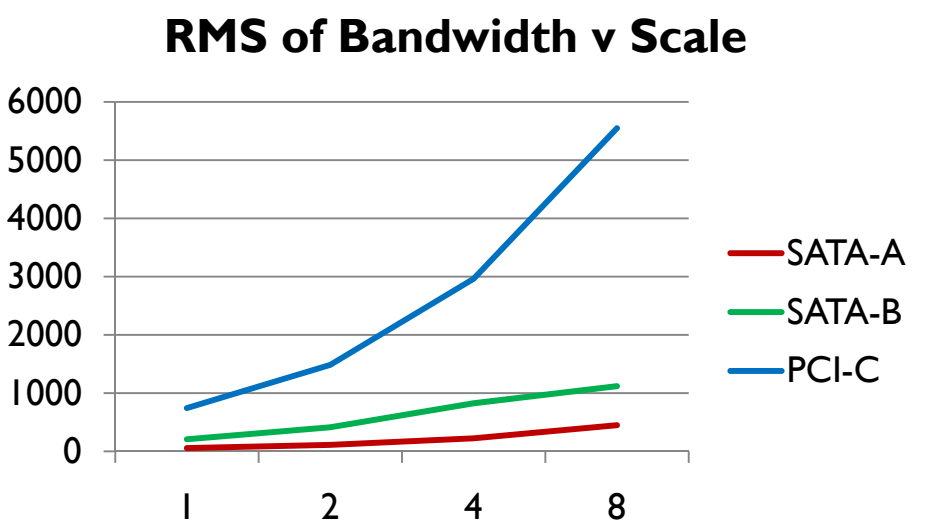
Read/Write Collisions → Drop in Mixed Performance

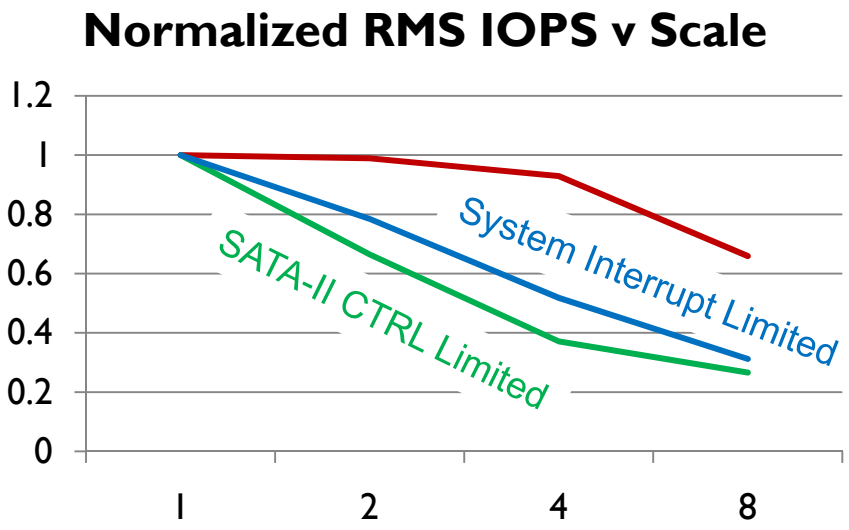
Scalability v R/W Ratio

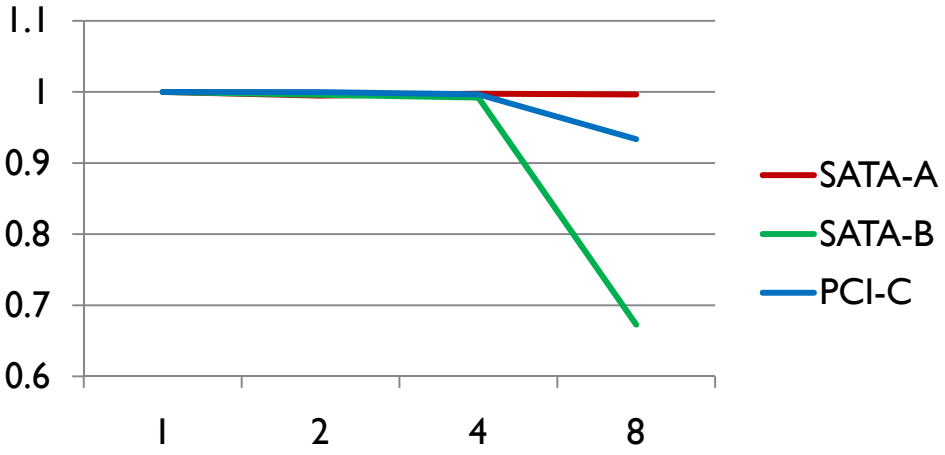


RMS Scalability (# SSS Units)

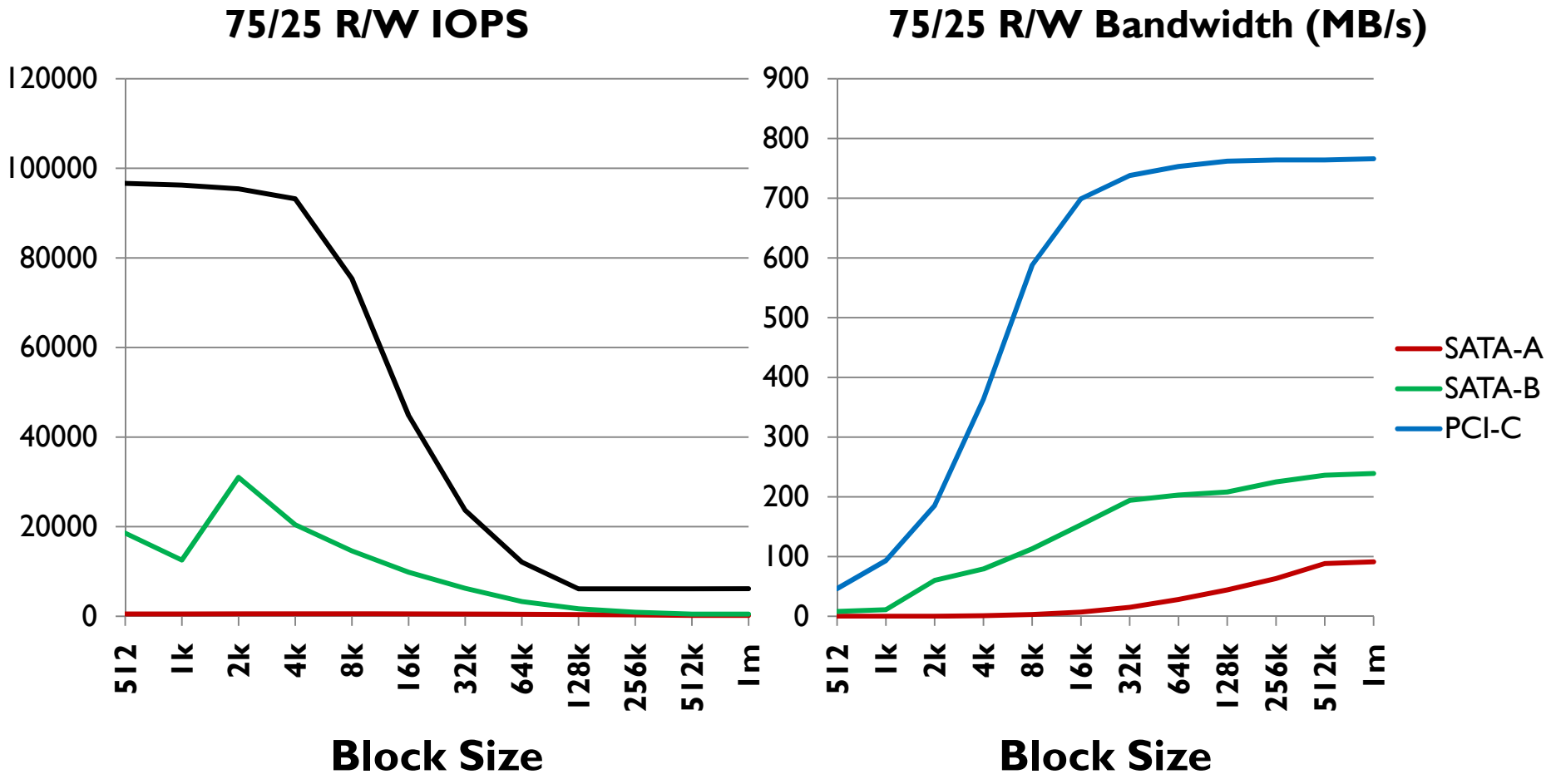




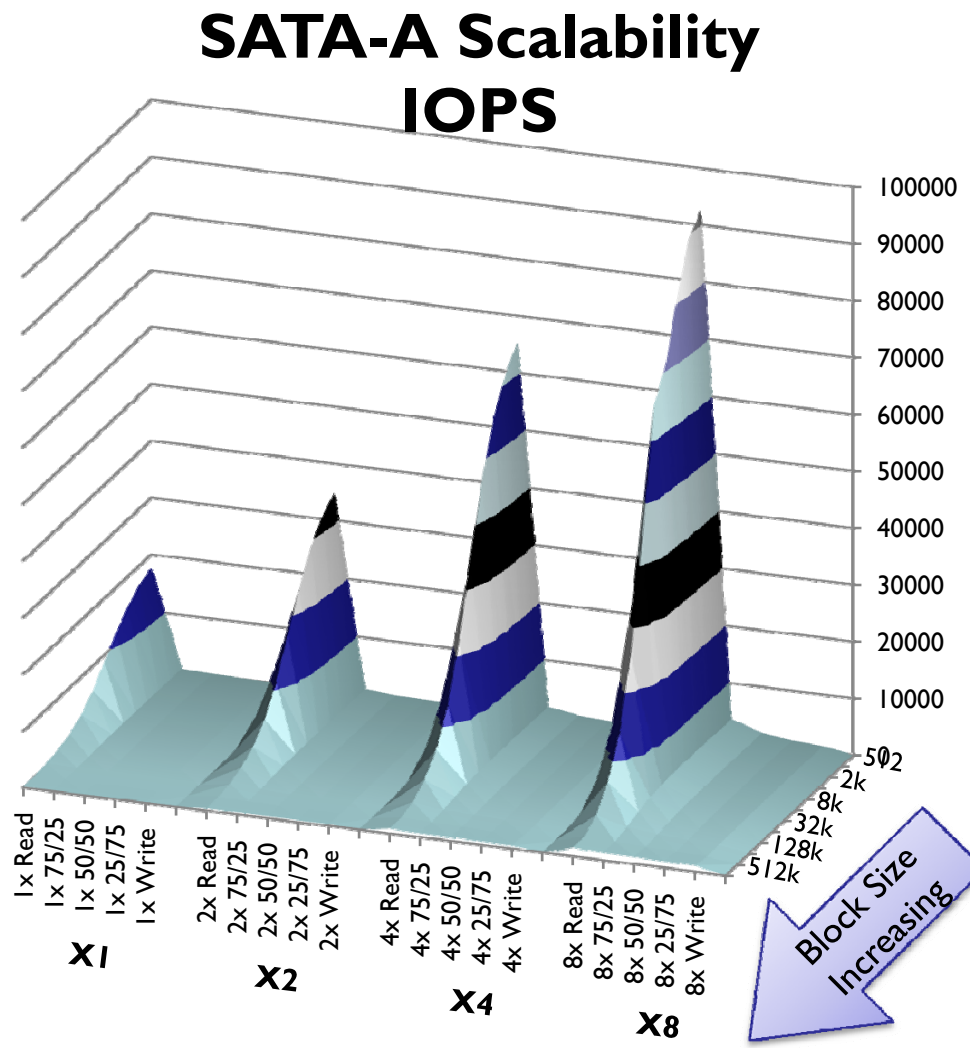


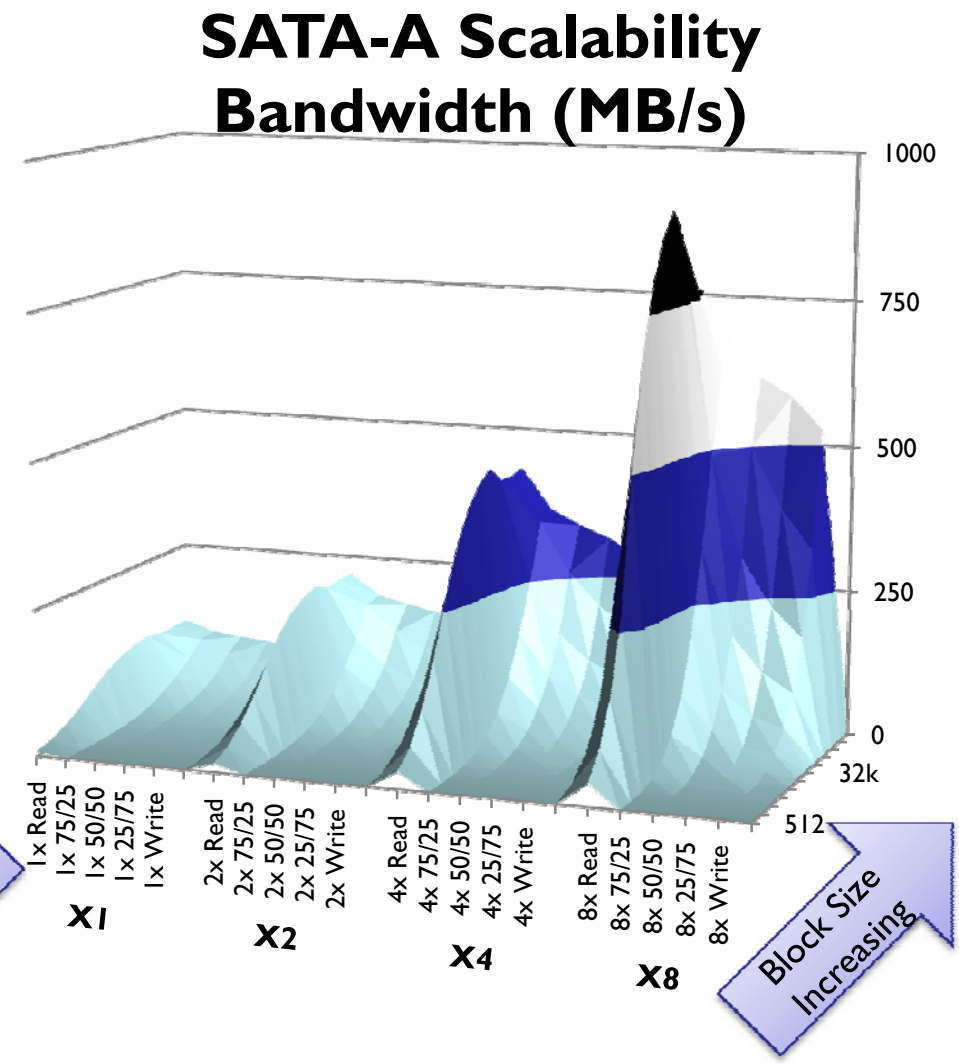


Performance v Block Size (75/25)

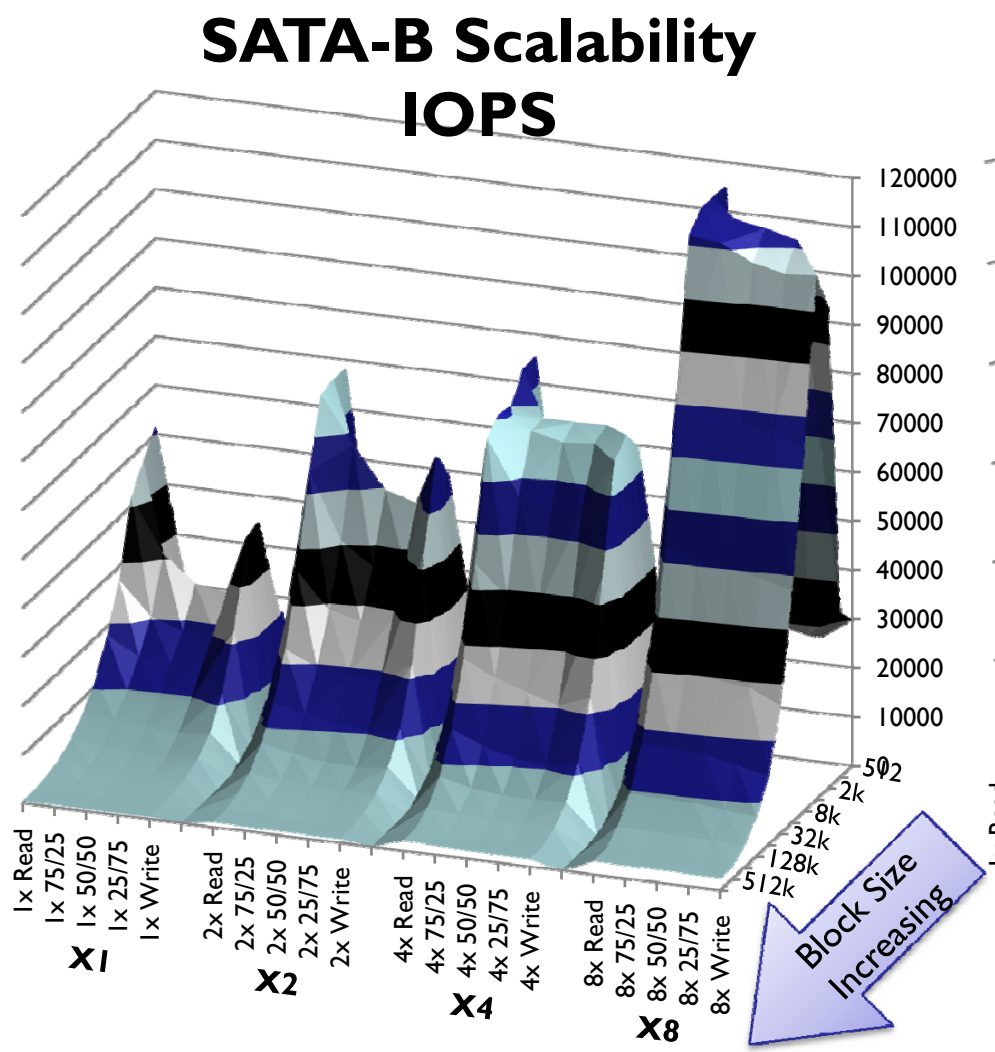


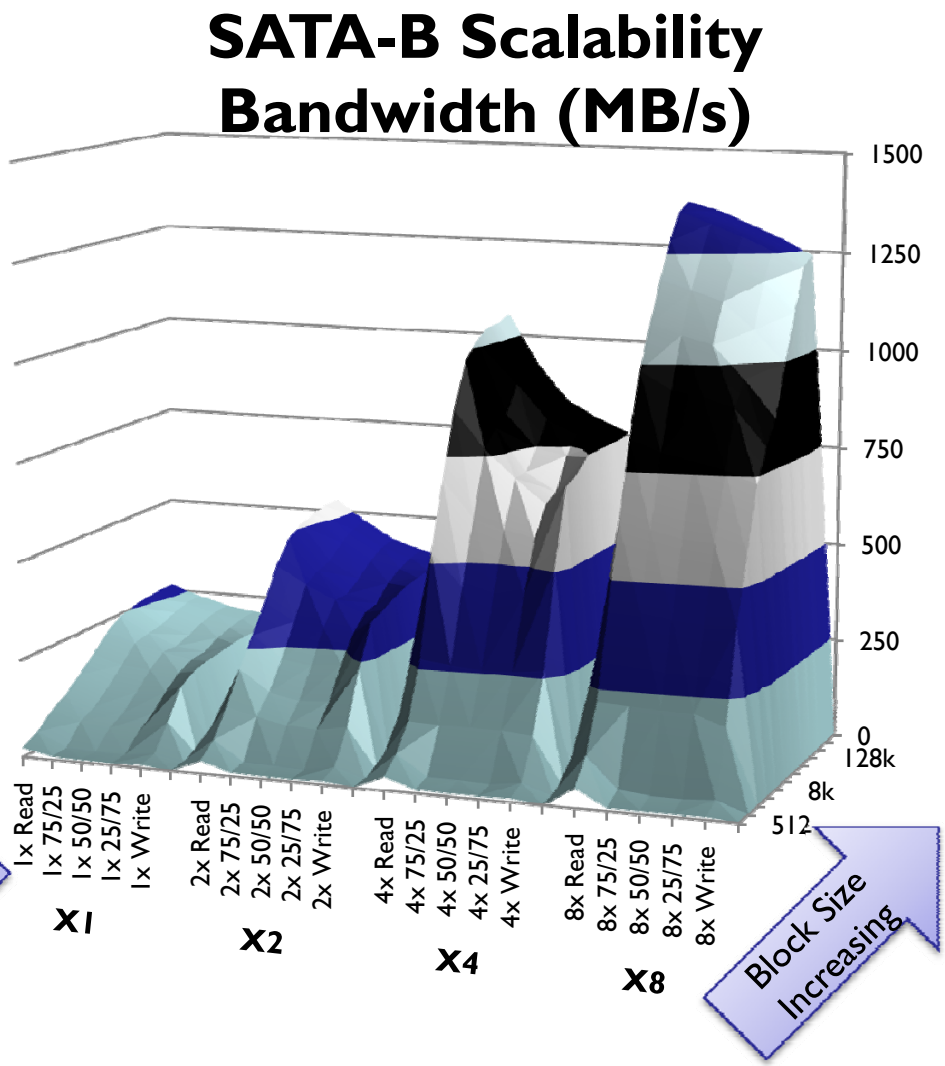
Scalability v RW Ratio v Block Size



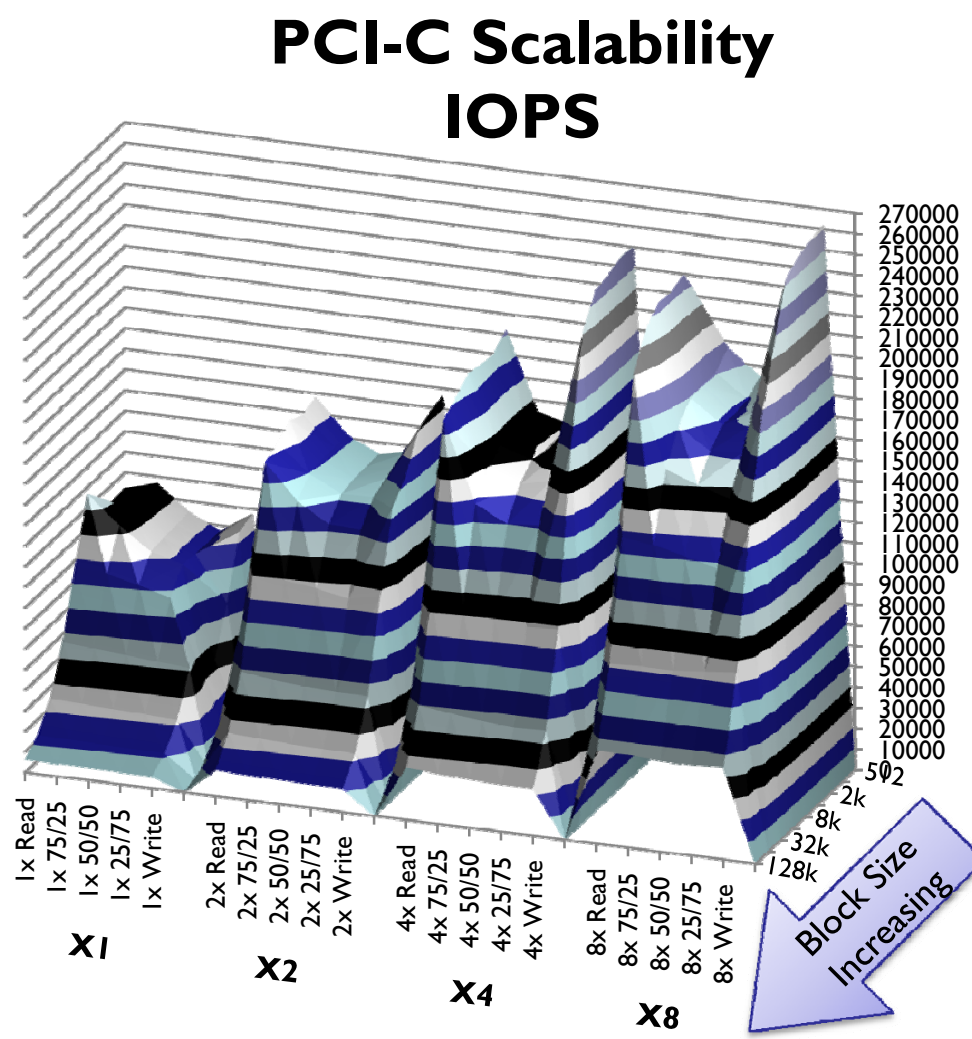


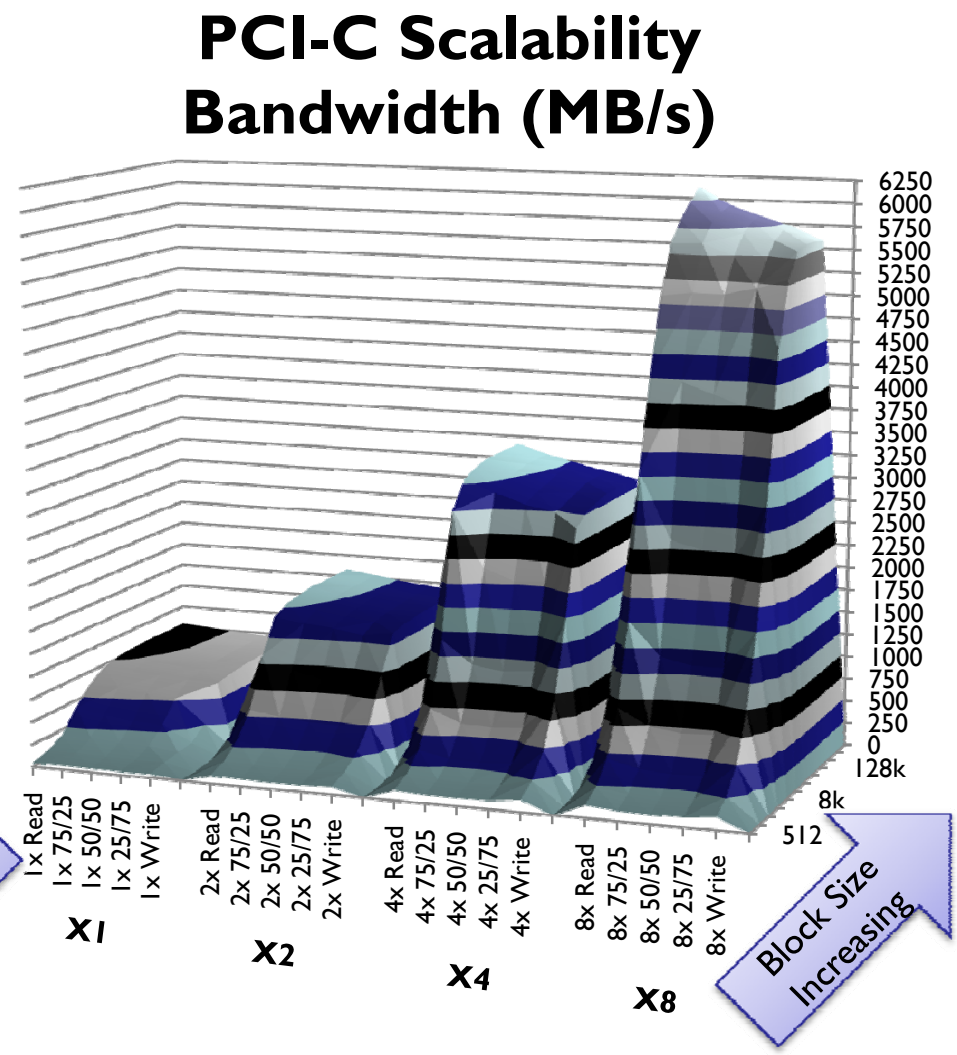
Scalability v RW Ratio v Block Size





Scalability v RW Ratio v Block Size





Media Reliability / Availability

Problem	SSS Solution	System Solution
High Infant Mortality	RAID: NAND Flash	RAID: SSS
High Error Rate & Wearout	RAID: NAND Flash	RAID: SSS
	Robust ECC	No Defragmentation
	DIF	DIF
	Wear Leveling	Access Tuning
	Never Multi-page PGM	
	Thermal Management	Temperature & Air Flow
Non-random	Address Indirection	N/A
R/W not symmetric	More Banks (CS); esp for MLC	Reduce write thrashing

System Level Considerations

- Data / Index Protection (DIF ; RAID)
- Scalability
- Compare system- or data-center-level; not device
- Best case: test on real application, not benchmark
  - Plan to do **tuning** to reach top performance / objectives
  - Applications may have **contra-indicated optimizations**
    - Example: keeping data in close physical proximity (short stroking)
    - Example: caching algorithms

Questions to Ask : Things to Know

- Bandwidth / IOPS at
  - Block size(s) you need
  - R/W ratio you use
  - Steady State / Burst
  - Data's temporal relationship
  - Scalability
  - RAIDing
  - Reserve capacity used
  - BOL / EOL

Questions to Ask : Things to Know

- Design impacts on data integrity; life; failures & performance
  - ECC robustness
  - Write amplification / GC efficiency
  - Internal RAID
  - Bandwidth throttling
  - Partial Page Programming
- Test Conditions
  - RAID On/Off during testing?
  - Caching On/Off during testing?
  - 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

Khaled Amer
Phil Mills
Rob Peglar
Marius Tudor