SNIA SSSI - PCIe Round Table

- **Standards**
- **Technology / Architecture**
- **Deployment Strategies**

Presentations by:

Fusion-io - Intel - Micron - Sata-IO - Seagate - Tailwind

SNIA

Solid State Storage Initiative

SNIA Winter Symposium
SSSI Face to Face
Monday 23 January 2012
10:00 AM - 1:00 PM
St. Claire Hotel, San Jose CA

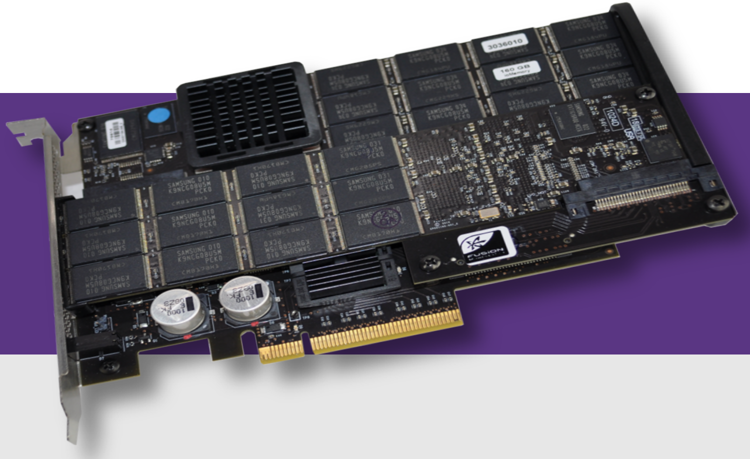
Webex: https//snia.webex.com
Meeting No. 795 947 658 Password: sssi2012
Telecon: 1-877-270-2716 ID: 0021 Password: 8520

Agenda

1.	**10:15 AM - 10:30 AM**	**Introduction - SSS Performance**	**Eden Kim, Chair SNIA SSS TWG**
2.	10:30 AM - 10:45 AM	PCIe SSD Form Factor	Mark Meyers, Intel
3.	10:45 AM - 11:00 AM	Standards & Deployment Models	Marty Czekalski, Seagate
4.	11:00 AM - 11:15 AM	SATA-IO & SATA Express - PCIe for Client Storage	Paul Wassenberg, Sata-IO
5.	11:30 AM - 11:45 AM	PCIe 2.5” Form Factor	Janene Ellefson, Micron
6.	11:45 AM - 12:00 PM	Convergence of Memory & Storage IO Architecture	Moon Kim, Tailwind
7.	12:15 PM - 12:30 PM	Lessons from the Front Lines & Lessons for the Future	Gary Orenstein, Fusion-io
8.	12:30 PM - 1:00 PM	Panel Question & Answers / Working Lunch	

Solid State Storage PCIe . . .

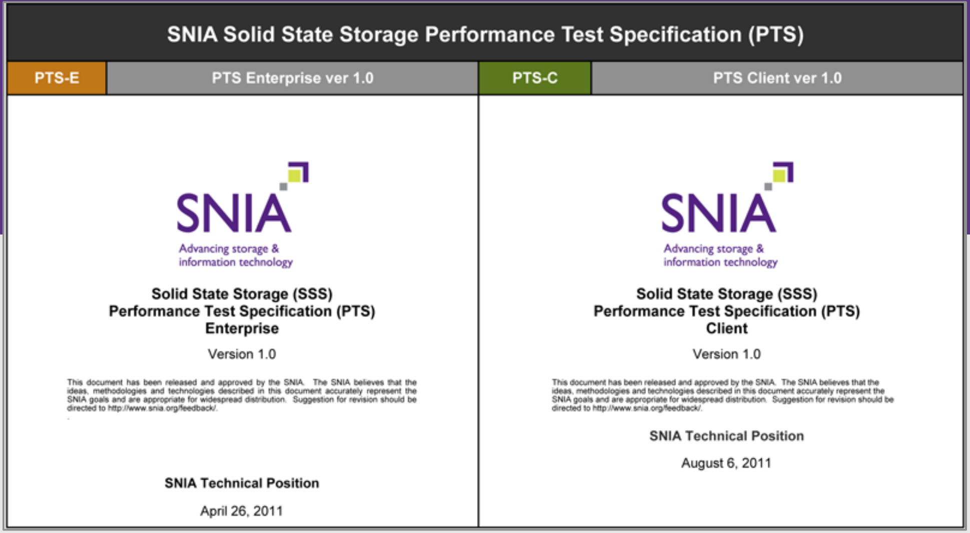
a Round Table



SERIAL ATA · FUSION-iO · TAILWIND STORAGE Semiconductor Storage Device Company · intel · Seagate · Micron Focused on Memory | Engineered for Innovation

What are issues facing Adoption of PCIe Solid State Storage devices?

- Standards for PCIe Attached Storage
- Technology & Architectural Issues
- Mass Storage Ecosystem Adoption & Optimization
- Market & Product Positioning
- Deployment Strategies


SNIA Solid State Storage Performance Test Specification (PTS)

PTS-E	PTS Enterprise ver 1.0	PTS-C	PTS Client ver 1.0

SNIA
Advancing storage & information technology

Solid State Storage (SSS)
Performance Test Specification (PTS)
Enterprise

Version 1.0

This document has been released and approved by the SNIA. The SNIA believes that the ideas, methodologies and technologies described in this document accurately represent the SNIA goals and are appropriate for widespread distribution. Suggestion for revision should be directed to http://www.snia.org/feedback/.

SNIA Technical Position

April 26, 2011

SNIA
Advancing storage & information technology

Solid State Storage (SSS)
Performance Test Specification (PTS)
Client

Version 1.0

This document has been released and approved by the SNIA. The SNIA believes that the ideas, methodologies and technologies described in this document accurately represent the SNIA goals and are appropriate for widespread distribution. Suggestion for revision should be directed to http://www.snia.org/feedback/.

SNIA Technical Position

August 6, 2011

SNIA PTS-C & PTS-E Specifications: Standardizing SSD Performance Test

SNIA SSSI Solid State ***Performance Test Spec*** link:

www.snia.org/tech_activities/standards/curr_standards/pts

Understanding SSD Performance Project link:

www.snia.org/forums/sssi/pts

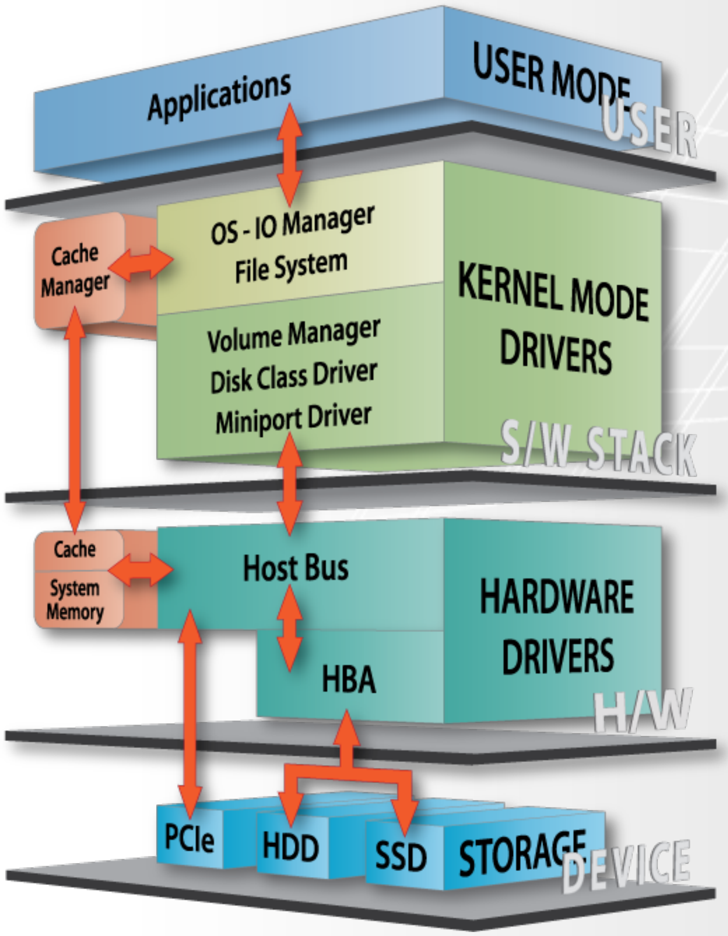
Understanding SSD Performance ***White Paper & Powerpoint*** link:

www.snia.org/forums/sssi/knowledge/education

Understanding SSD Performance ***Webcast*** link:

www.brighttalk.com/webcast/663/40549

PTS Provides a Standardized Methodology to Compare SSD Performance



IOs Traverse the SW / HW Stack

- Storage IOs Must Traverse the SW/HW Stack
- IOs are subject to cache, OS task switching & timing, driver fragmentation & coalescing
- IO can be different at the Device & System level
- Can lose 1:1 correspondence original IO & Physical Device IO
- Performance is Heavily influenced by SW / HW Stack

Solid State Performance Issues

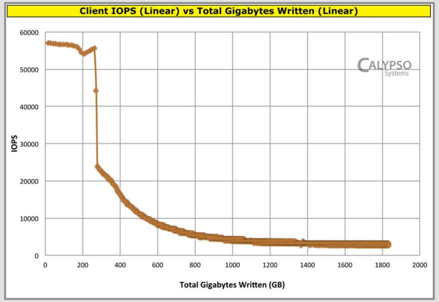
- Solid State Performance is MUCH Faster than HDD Storage
- SSDs must be optimized to Storage Ecosystem
- Solid State Storage employ Virtual Mapping of PBA to LBA
- Asymmetric Read / Write Response Times for Flash
- Response Time & Cost varies for DRAM, PCIe, SLC, MLC, HDD

Reference Test Platform (RTP 2.0)			
Hardware		**Software**	
Processor	Single Intel Xeon 5580W 3.2 Ghz 4 core	**Operating System - Back End**	CentOS 5.6
Motherboard	Intel 5520 HC	**Test Software - Back End**	CTS 6.5
RAM	12 GB ECC DDR3	**Front End - GUI**	Chrome Browser
HBA	6 Gb/s LSI 9212-4e-4i	**Front End: OS, Database**	Windows 7 / MySQL

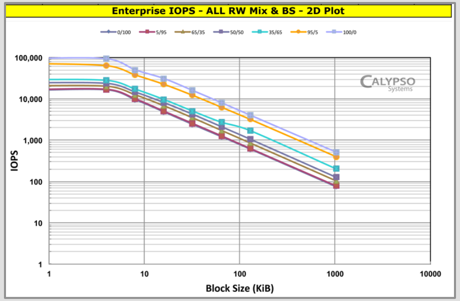
PTS Reference Test Platform - Allows Comparison of PCIe, SAS, SATA, HDD Performance

PTS rev 1.0 Performance Tests

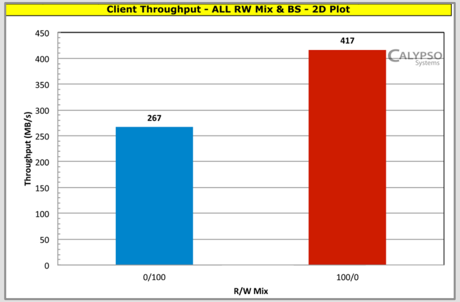
Test	Test Description	Purpose	Metric
WSAT	Continuous RND 4KiB W from FOB, No PC	FOB Performance Evolution over Time	IOPS
IOPS	Large & Small Block RND IOs at Steady State	Steady State IO Transfer Rate per second	IOPS
Throughput	Large Block SEQ R/W Data Transfer at Steady State	Steady State Bandwidth Speed	MB/Sec
Latency	AVE & MAX Response Times measured at a single OIO	Steady State IO Response Time Latency	mSec



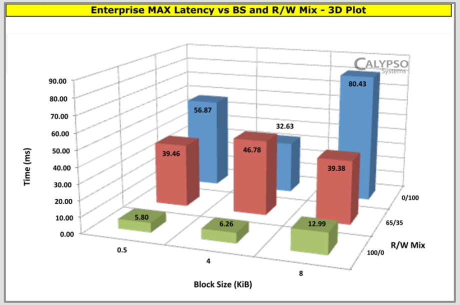
WSAT



IOPS



TP

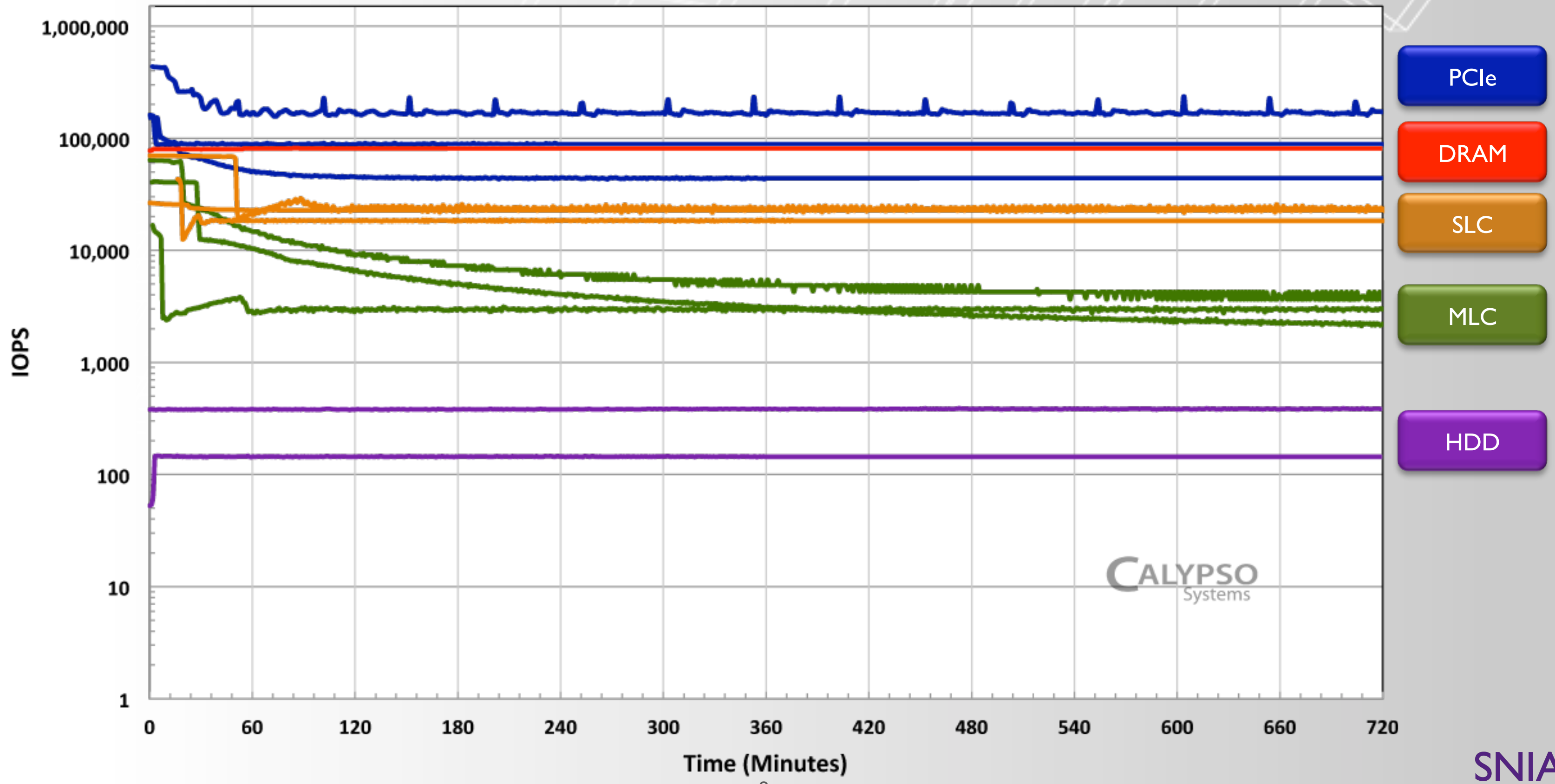


LAT

WSAT Test is useful to Evaluate Solid State Small Block RND Write Behavior

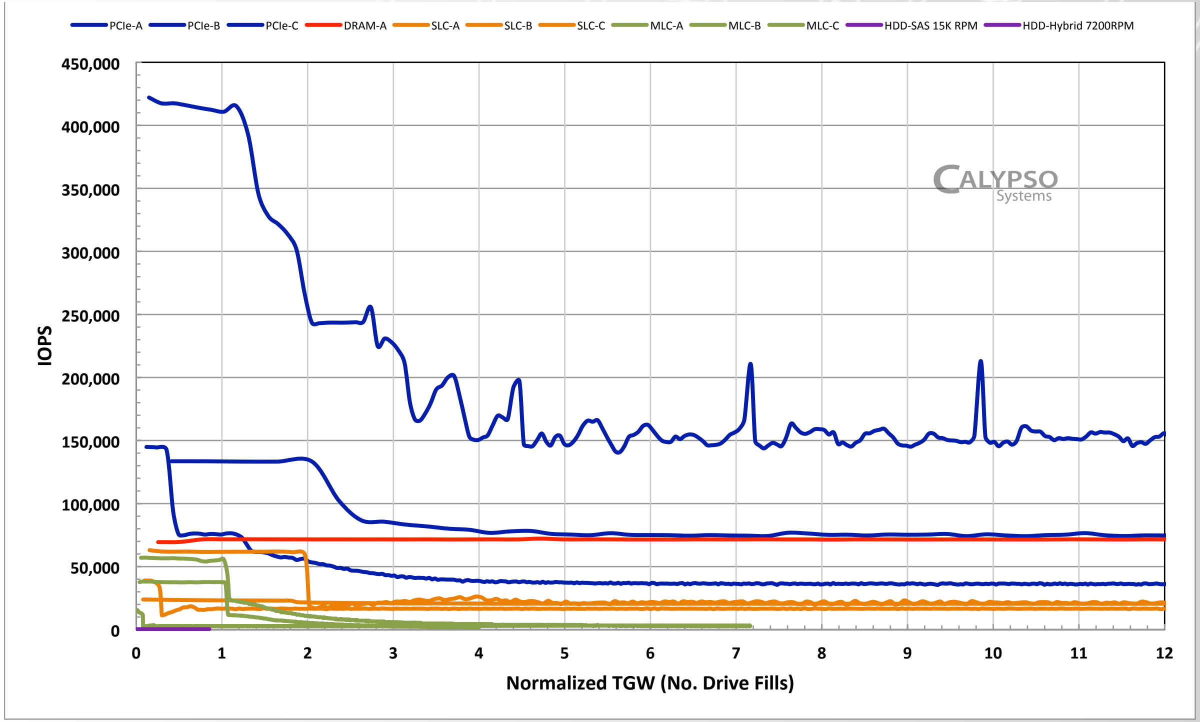
Solid State Storage Technology - RND 4KiB Write Performance*

* All Data SNIA PTS-E 1.0 WSAT Test Compliant



WSAT: RND 4KiB W - IOPS v TGBW

* All Data SNIA PTS-E 1.0 WSAT Test Compliant



RND 4K W Performance

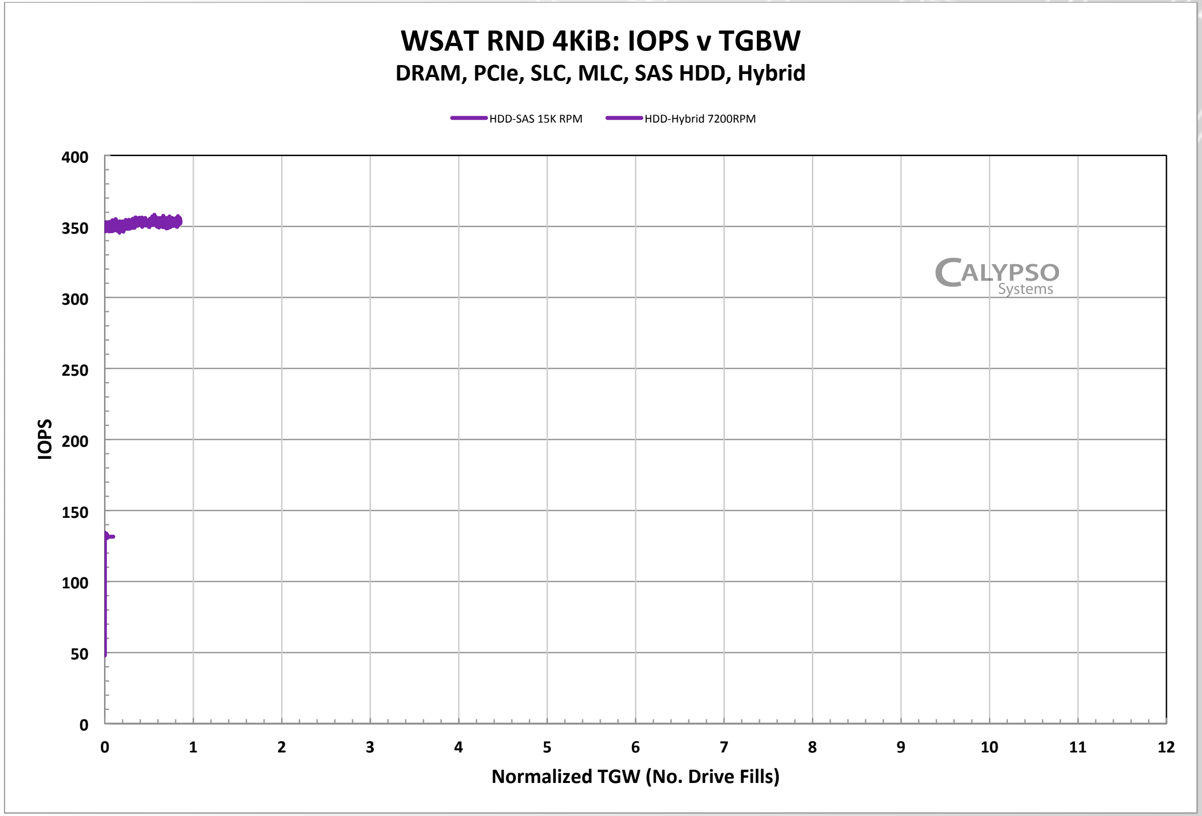
PCIe
150,000 IOPS

DRAM
71,500 IOPS

SLC
20,000 IOPS

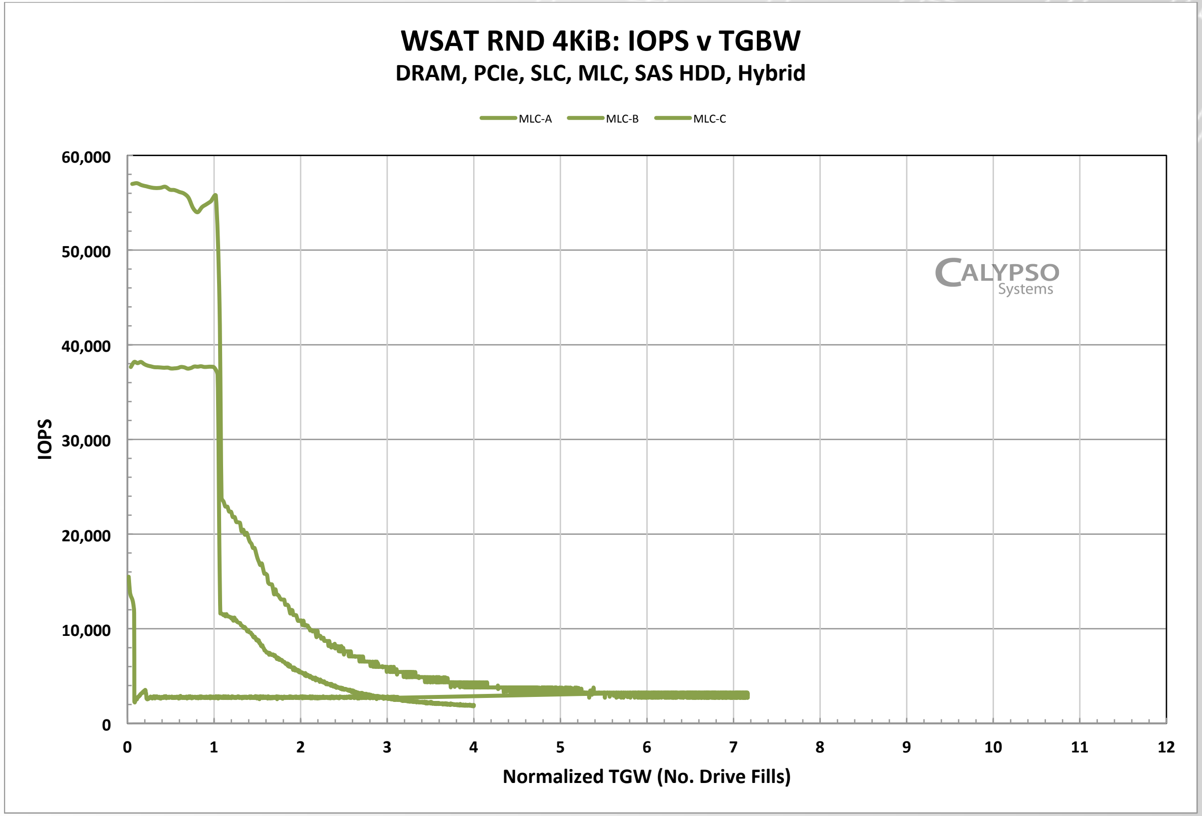
MLC
3,250 IOPS

HDD
350 IOPS

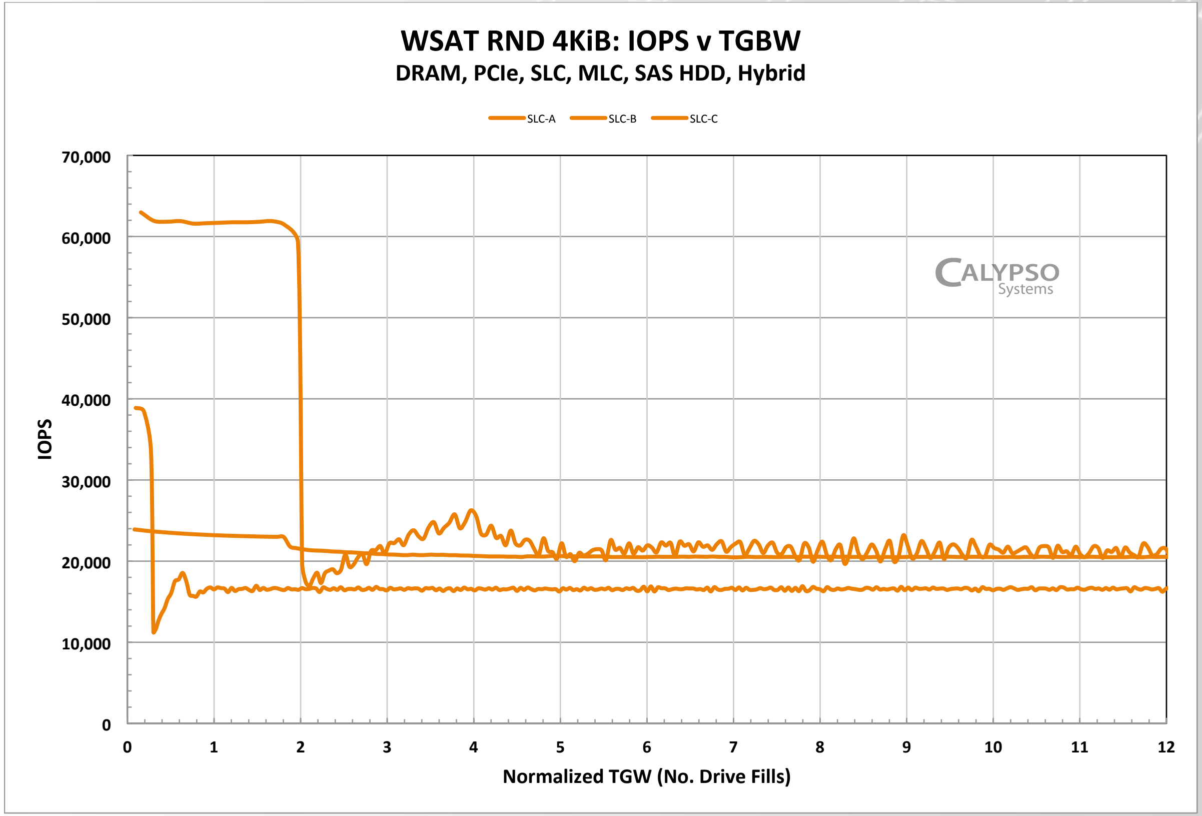


WSAT RND 4KiB: IOPS v TGBW
DRAM, PCIe, SLC, MLC, SAS HDD, Hybrid
HDD-SAS 15K RPM
HDD-Hybrid 7200RPM
IOPS
400
350
300
250
200
150
100
50
0
0
1
2
3
4
5
6
7
8
9
10
11
12
Normalized TGW (No. Drive Fills)
CALYPSO Systems

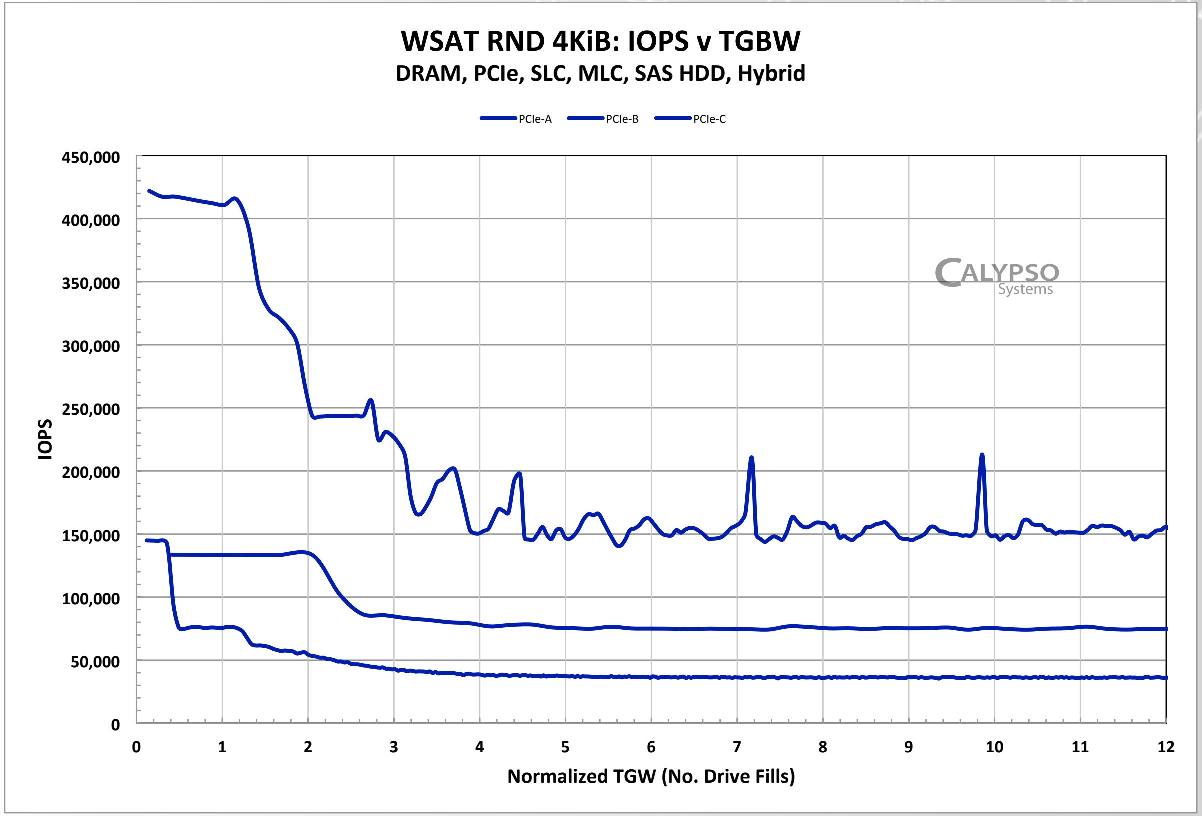
HDD
350 IOPS



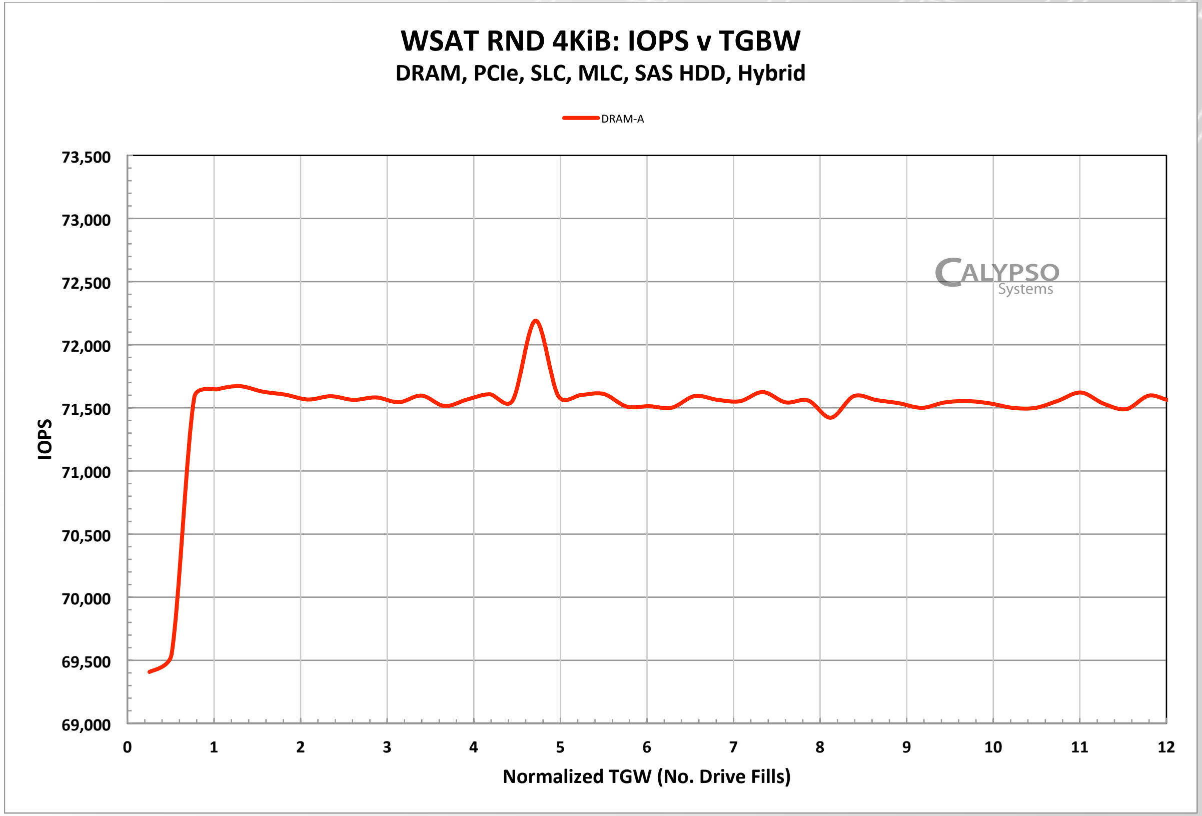
MLC
3,250 IOPS



SLC
20,000 IOPS



PCIe
150,000 IOPS



DRAM
71,500 IOPS

Agenda

1.	**10:15 AM - 10:30 AM**	**Introduction - SSS Performance**	**Eden Kim, Chair SNIA SSS TWG**
2.	10:30 AM - 10:45 AM	PCIe SSD Form Factor	Mark Meyers, Intel
3.	10:45 AM - 11:00 AM	Standards & Deployment Models	Marty Czekalski, Seagate
4.	11:00 AM - 11:15 AM	SATA-IO & SATA Express - PCIe for Client Storage	Paul Wassenberg, Sata-IO
5.	11:30 AM - 11:45 AM	PCIe 2.5” Form Factor	Janene Ellefson, Micron
6.	11:45 AM - 12:00 PM	Convergence of Memory & Storage IO Architecture	Moon Kim, Tailwind
7.	12:15 PM - 12:30 PM	Lessons from the Front Lines & Lessons for the Future	Gary Orenstein, Fusion-io
8.	12:30 PM - 1:00 PM	Panel Question & Answers / Working Lunch	

Mark Meyers, *Intel*

PCIe SSD Form Factor

Abstract

PCIe SSD Form Factor has the attractive attributes that PCIe brings to SSD storage, and adds more capabilities from the existing storage form factors.

Mark is a Server Platform Architect working in Intel’s Datacenter and Connected System group.

Mark is technical chair of the Enterprise SSD Form Factor WG which includes definition of proposed SFF-8639 connector.

Mark has been at Intel for 12 years in various server and IO architecture projects.

Previous employers includes Siemens Nixdorf, Pyramid Technology, and an early stint at Intel.

PCIe SSD Form Factor
for SNIA 2011 Winter Symposium

Mark Myers
Intel Datacenter Platform Architect
January 23, 2012

Introduction

Goal

- Status of the PCIe SSD Form Factor WG summary
- PCIe as a Storage Interface
- Common configurations
- Technical Attributes

Enterprise PCIe SSD Form Factor WG Status

Defined usages and requirement and connector (SFF-8639)

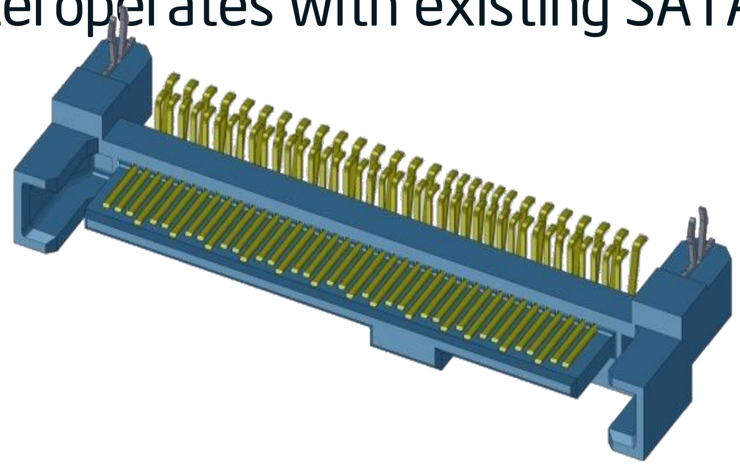
- 5 promoters: Dell, IBM, Fujitsu, EMC, Intel; >50 contributor companies

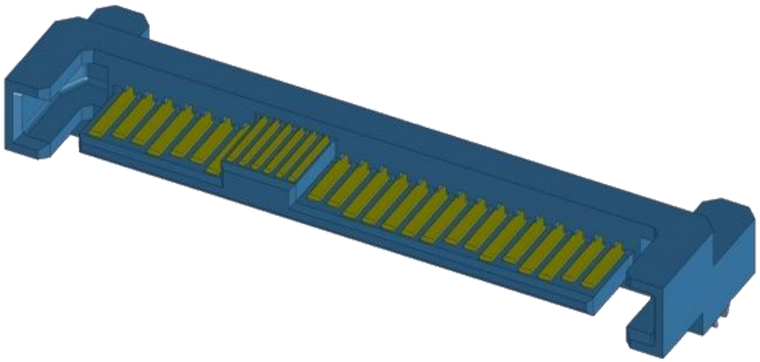
Rev 1.0 Specification Approved http://www.ssdformfactor.org/

- Mechanical piece is SFF-8639 ftp://ftp.seagate.com/sff/SFF-8639.PDF

Looks like existing SAS connector with pins all across both sides

- Interoperates with existing SATA/SAS connector





PCIe as a Storage Interface

PCIe value

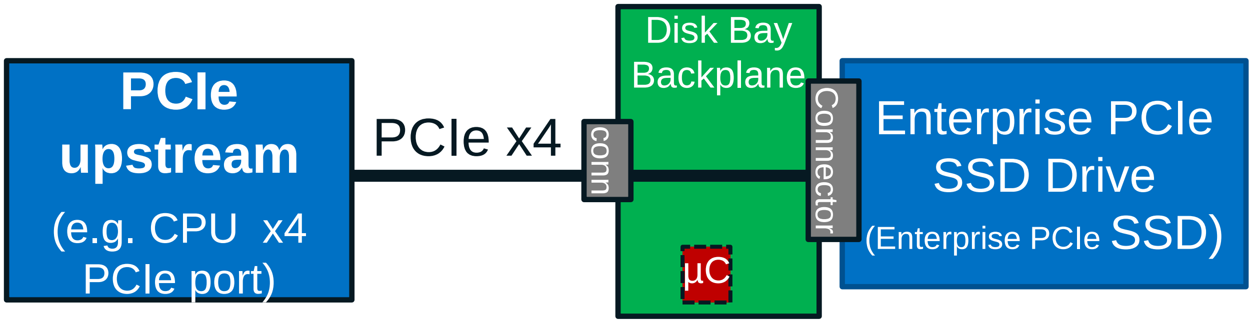
- Industry standard, high BW, multilane, low latency interconnect
- Flexible attach models, discoverable, and supports many form factors ➔ Our work adds a classic 3.5" or 2.5" disk form factor

PCIe as high performance interface; Many storage interfaces;

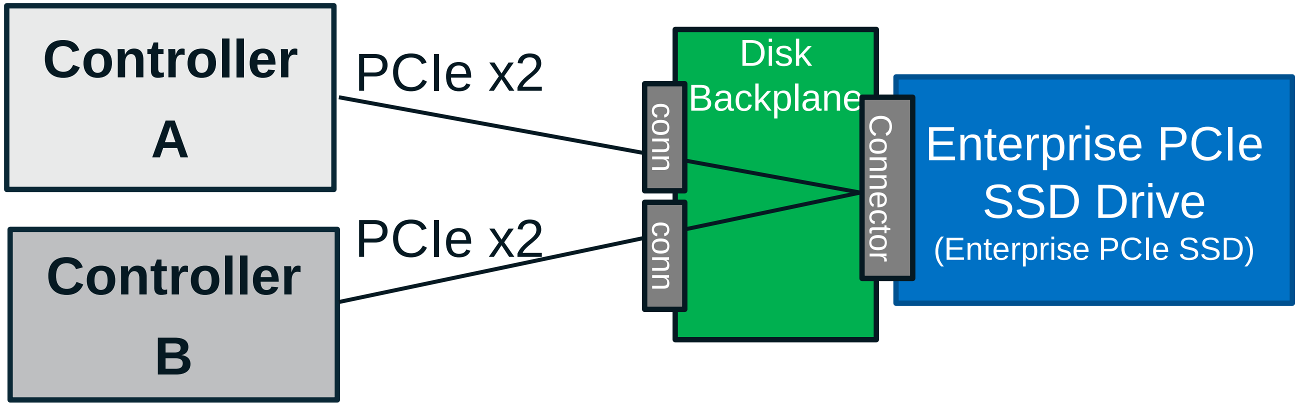
- Hard Disks stay on SATA/SAS for long time, even for many SSDs
- High performance SSD will move to PCIe – higher BW & low latency
- PCIe supports multiple device types: NVM-Express, SOP, proprietary
  - Advocate NVMe as standard block device
  - Expect interface models to evolve as devices improve

Common Usages: Servers x4, Dual Port storage

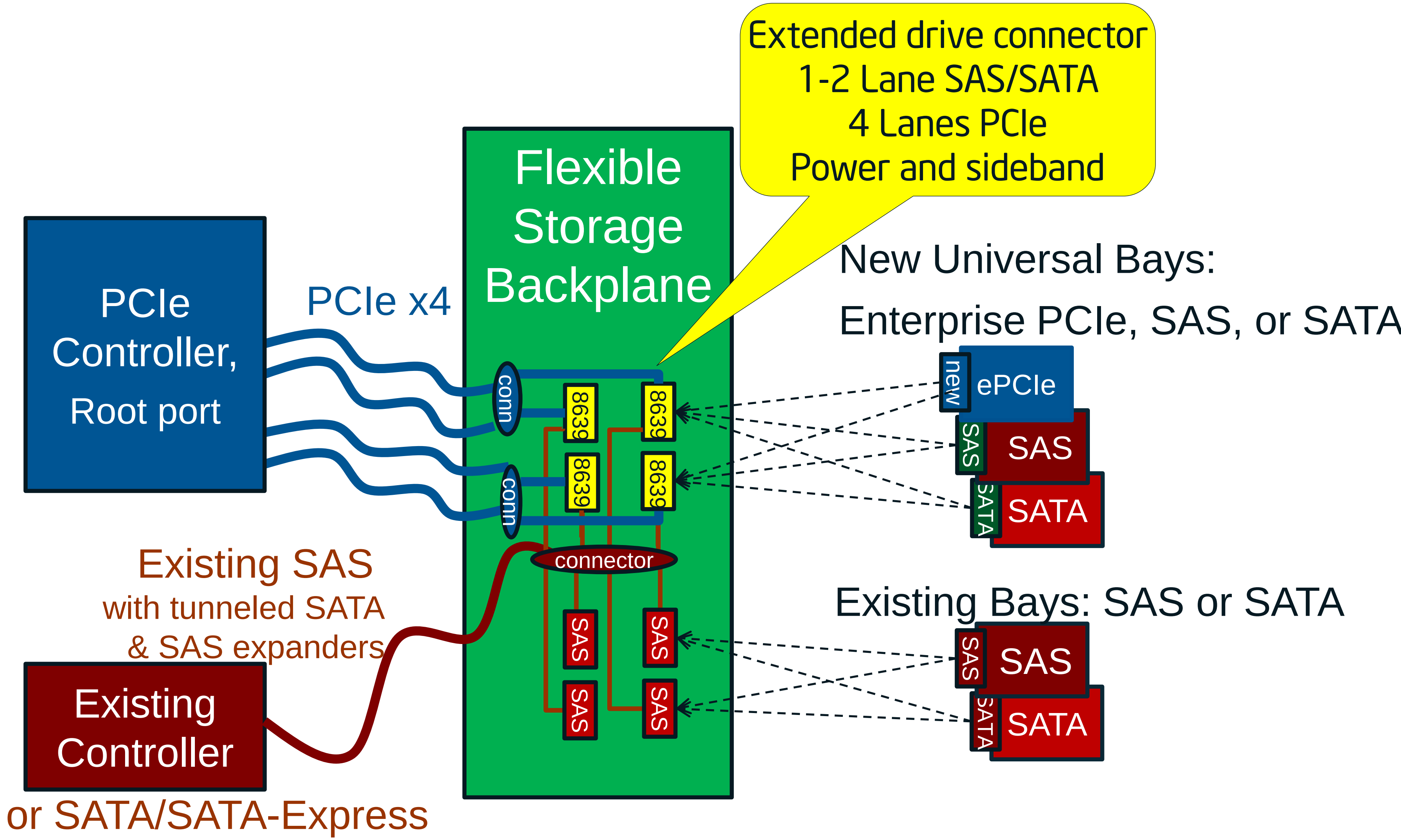
Typical Server configuration



Typical High Availability Storage configuration



Flexible Backplane – Support SAS & PCIe



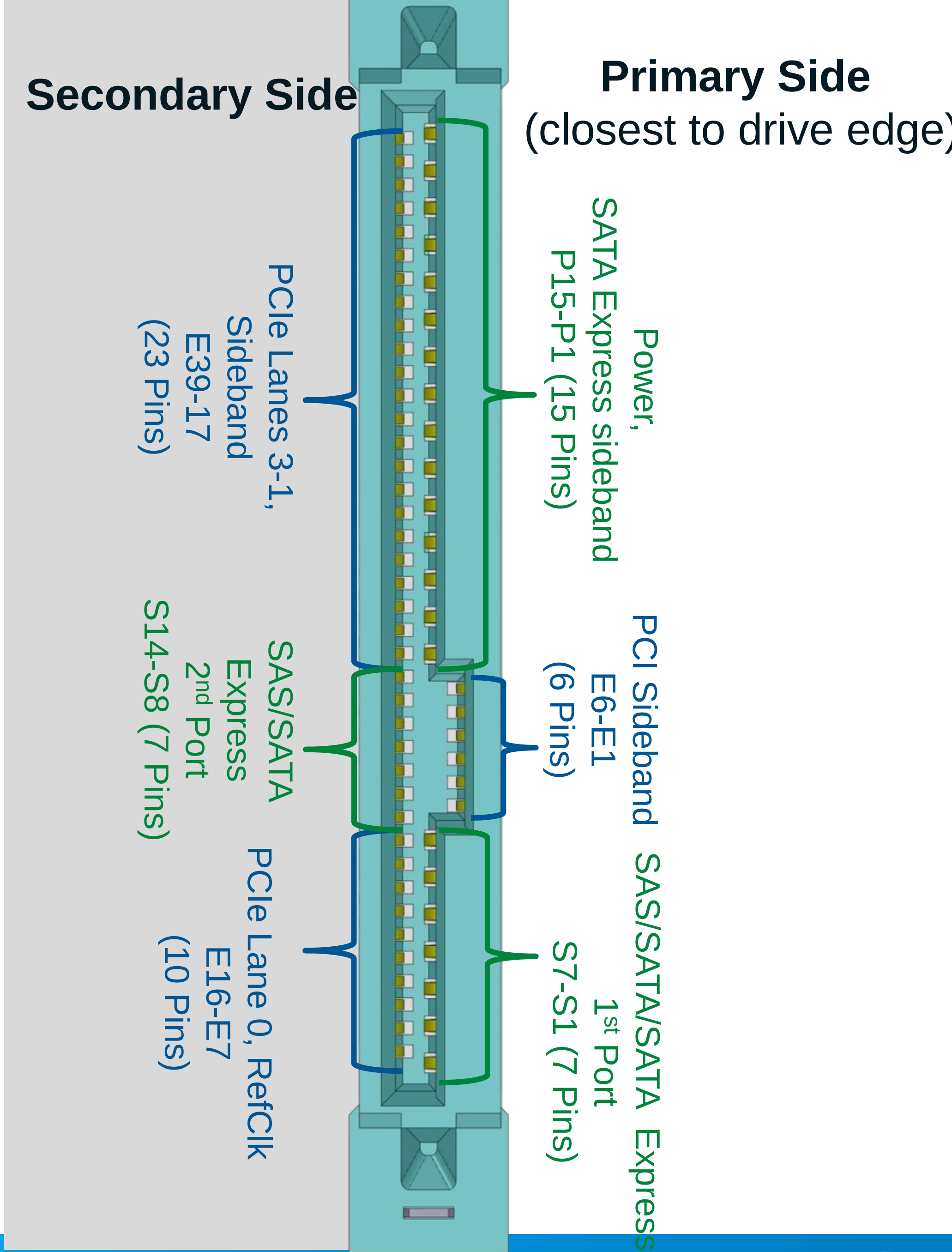
Drives Supported

Support drive types

- Enterprise PCIe x4 SSDs
  - Server x4, Storage Dual Port x2
- Existing SAS drive (dual port)
- Existing SATA drives
- Emerging SATA-Express x1-x2
- Emerging x4 SAS

Support Flexible Backplanes

- Enterprise x4 PCIe SSDs
- SAS/SATA HDDs



Technical Attributes of Specification

- 6 High speed lanes
  - 4 new lanes for Enterprise PCIe
  - 2 existing lanes for SAS/SATA
- Side Band
  - Enterprise: RefClk, ePCIeRst#, SM-Bus, 3.3VAux, DualPort
  - Client/Shared: IfDet#, PRSNT#, cPCIeRst#, Rsvd (pwr mgt)
  - Removed 3.3V, Enterprise SSD supports 12V only
- Keying
  - Support universal receptacle
  - Key to block SATA-express Cable to x4 drive
  - Key to block Enterprise x4 cable to SATA/SAS drive

Conclusion

Enterprise PCIe SSD Form Factor Specification

- Rev 1.0 Approved and Released
- Expect products this year based on standard

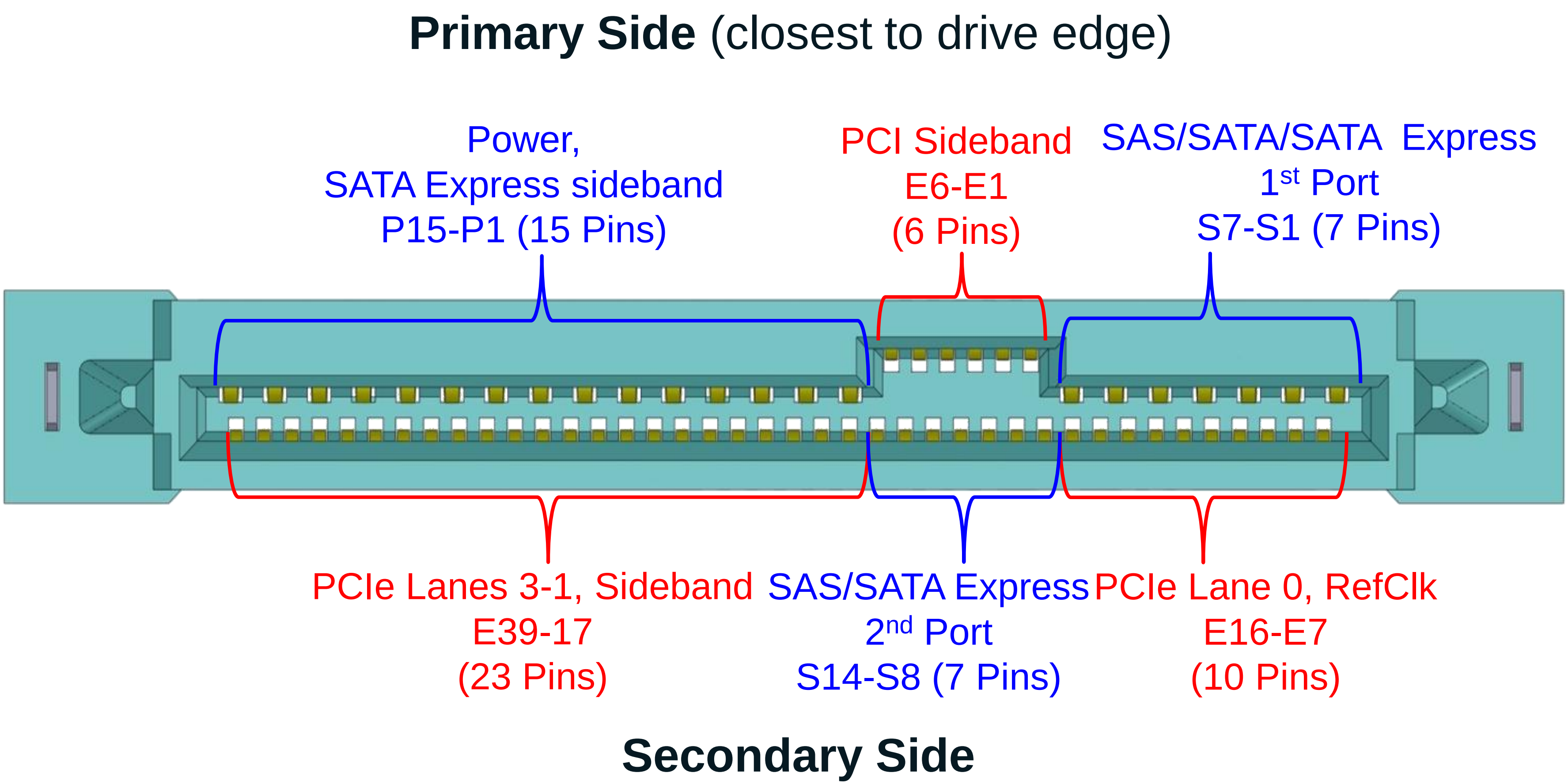
Supports Flexible Storage Backplanes

- High Performance Enterprise x4 PCIe SSDs
  - Using existing PCIe root ports
- Existing SAS/SATA drives
- Emerging SATA-Express and x4 SAS

Thank You

Additional Detail

Overview of Connector Pins



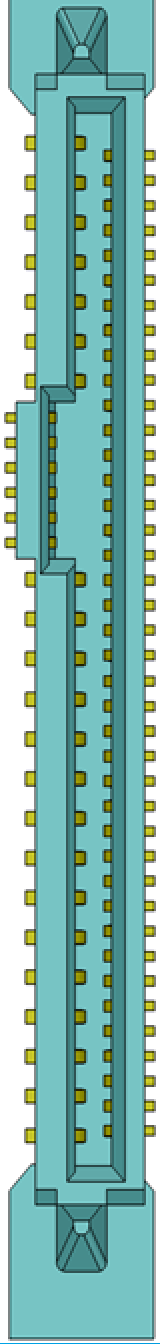
Pin out

Drive	Usage	Signal Description	Name	Mating	Pin #
		Ground	GND	2nd	S1
input	SAS+SATA	SAS/SATA/SATAe 0 Tx+	S0T+ (A+)	3rd	S2
input	SAS+SATA	SAS/SATA/SATAe 0 Tx -	S0T- (A-)	3rd	S3
		Ground	GND	2nd	S4
output	SAS+SATA	SAS/SATA/SATAe 0 Rcv -	S0R- (B-)	3rd	S5
output	SAS+SATA	SAS/SATA/SATAe 0 Rcv +	S0R+ (B+)	3rd	S6
		Ground	GND	2nd	S7
input	Dual Port	ePCIe RefClk + (port B)	RefClk1+	3rd	E1
input	Dual Port	ePCIe RefClk – (port B)	RefClk1-	3rd	E2
input	ePCIe opt	3.3V for SM bus	3.3Vaux	3rd	E3
input	Dual Port	ePCIe Reset (port B)	ePERst1#	3rd	E4
input	ePCIe	ePCIe Reset (port A)	ePERst0#	3rd	E5
		Reserved	RSVD	3rd	E6
input	SATAe +SAS4	Reserved(WAKE#/OBFF), SASAct2	RSVD(Wake#) /SASAct2	3rd	P1
Bi-Dir	SATAe	SATAe Client /SAS reset	sPCIeRst/SAS	3rd	P2
input	SATAe	Reserved (DevSLP#)	RSVD(DevSLP#)	2nd	P3
output	SATAe + ePCIe	Interface Detect (Was GND-precharge)	IfDet#	1st	P4
	all	Ground	GND	2nd	P5
	all			2nd	P6
NC	SAS+SATA	Precharge	5 V	2nd	P7
NC	SAS+SATA	SATA, SATAe, SAS only		3rd	P8
NC	SAS+SATA			3rd	P9
	all	Presence (Drive type)	PRSNT#	2nd	P10
Bi-Dir	all	Activity(output)/Spinup	Activity	3rd	P11
	all	Hot Plug Ground	GND	1st	P12
input	all	Precharge	12 V	2nd	P13
input	all	All – 12V		3rd	P14
input	all	Only power for ePCIe SSD		3rd	P15

ePCIe → Enterprise PCIe (separate from SATA/SAS)

SATAe → SATA Express
(Client PCIe- muxed on SATA/SAS signals)

SAS4 → SAS x4



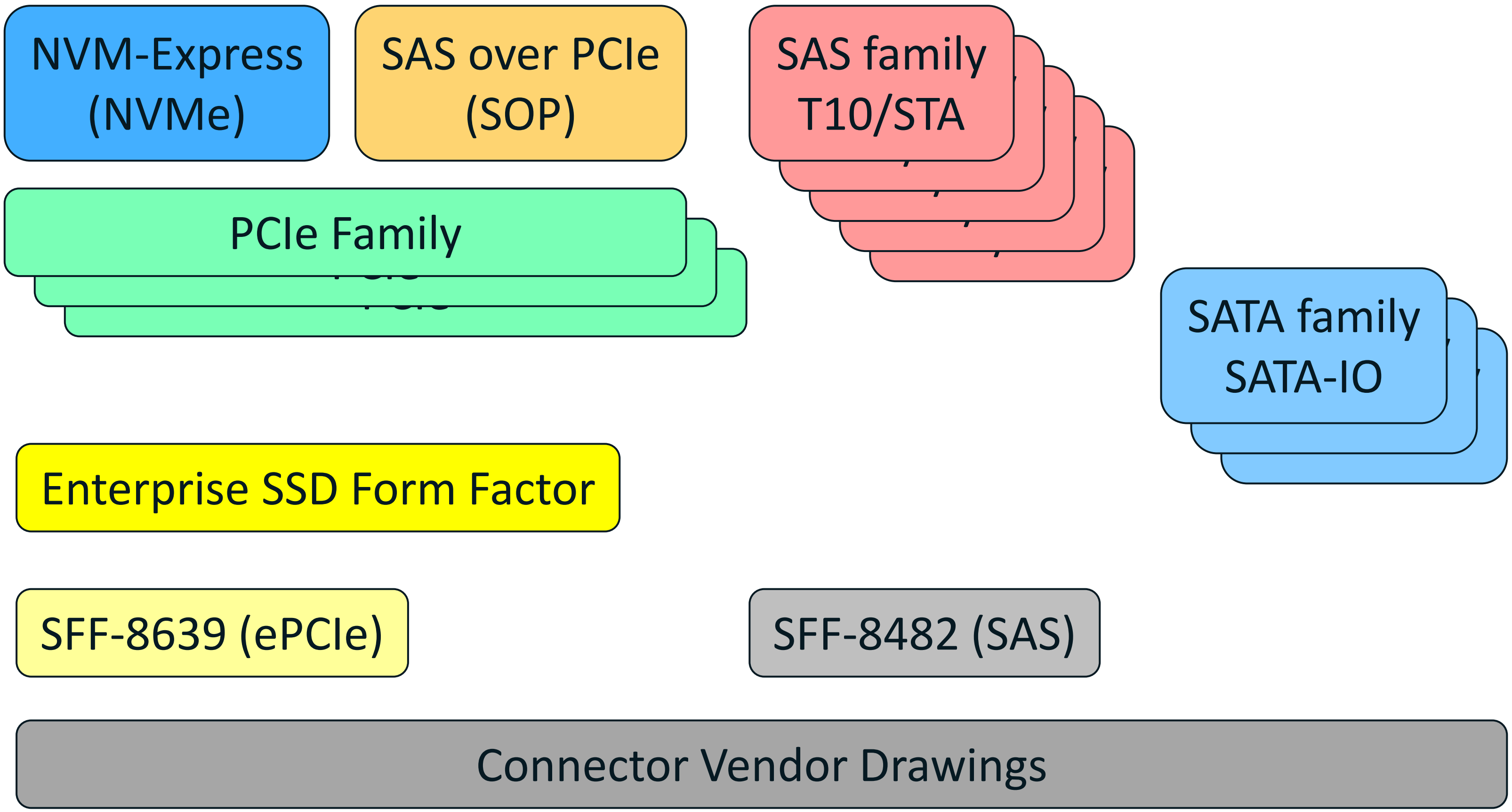
Pin #	Mating	Name	Signal Description	Usage	Drive
E7	3rd	RefClk0+	ePCIe Primary RefClk +	ePCIe	input
E8	3rd	RefClk0-	ePCIe Primary RefClk -	ePCIe	input
E9	2nd	GND	Ground		
E10	3rd	PETp0	ePCIe 0 Transmit +	ePCIe	input
E11	3rd	PETn0	ePCIe 0 Transmit -	ePCIe	input
E12	2nd	GND	Ground		
E13	3rd	PERn0	ePCIe 0 Receive -	ePCIe	output
E14	3rd	PERp0	ePCIe 0 Receive +	ePCIe	output
E15	2nd	GND	Ground		
E16	3rd	RSVD	Reserved		
S8	2nd	GND	Ground		
S9	3rd	S1T+	SAS/SATAe 1 Transmit +	SAS+SATAe	input
S10	3rd	S1T-	SAS/SATAe 1 Transmit -	SAS+SATAe	input
S11	2nd	GND	Ground		
S12	3rd	S1R-	SAS/SATAe 1 Receive -	SAS+SATAe	output
S13	3rd	S1R+	SAS/SATAe 1 Receive +	SAS+SATAe	output
S14	2nd	GND	Ground		
E17	3rd	RSVD	Reserved		
E18	2nd	GND	Ground		
E19	3rd	PETp1/S2T+	ePCIe 1 /SAS 2 Transmit +	ePCIe+SAS4	input
E20	3rd	PETn1/S2T-	ePCIe 1 /SAS 2 Transmit -	ePCIe+SAS4	input
E21	2nd	GND	Ground		
E22	3rd	PERn1/S2R-	ePCIe 1 /SAS 2 Receive -	ePCIe+SAS4	output
E23	3rd	PERp1/S2R+	ePCIe 1 /SAS 2 Receive +	ePCIe+SAS4	output
E24	2nd	GND	Ground		
E25	3rd	PETp2/S3T+	ePCIe2 / SAS 3 Transmit +	ePCIe+SAS4	input
E26	3rd	PETn2/S3T-	ePCIe2 / SAS 3 Transmit -	ePCIe+SAS4	input
E27	2nd	GND	Ground		
E28	3rd	PERn2/S3R-	ePCIe 2 / SAS 3 Receive -	ePCIe+SAS4	output
E29	3rd	PERp2/S3R+	ePCIe 2 / SAS 3 Receive +	ePCIe+SAS4	output
E30	2nd	GND	Ground		
E31	3rd	PETp3	ePCIe 3 Transmit +	ePCIe	input
E32	3rd	PETn3	ePCIe 3 Transmit -	ePCIe	input
E33	2nd	GND	Ground		
E34	3rd	PERn3	ePCIe 3 Receive -	ePCIe	output
E35	3rd	PERp3	ePCIe 3 Receive +	ePCIe	output
E36	2nd	GND	Ground		
E37	3rd	SMClk	SM-Bus Clock	PCIe opt	Bi-Dir
E38	3rd	SMDat	SM-Bus Data	PCIe opt	Bi-Dir
E39	3rd	DualPortEn#	ePCIe 2x2 Select	Dual Port	input

Keying

- Prevent mating if will not work
- Support Universal Receptacle
  - accepts any drive
  - Driver carrier provide keying
- Cable block for client cables
  - Prevent client service calls

	SATA drive	SATA Express drive	SAS drive	Enterprise PCIe drive
Enterprise backplane	Works- system supports (carrier key)	Works- if system supports (carrier key)	Works- if system supports (carrier key)	Works
SAS backplane	Works with STP	Mates-Nonfunctional (requires STP+) (carrier key)	Works	Mates-nonfunctional (carrier key)
SATA Express backplane/laptop	Works	Works	Blocked-Key	Blocked-Key
SATA backplane/laptop		Blocked-Key	Blocked-Key	Blocked-Key
Enterprise cable	Blocked-Key	Blocked-Key	Blocked-Key	Works
SAS cable	Works	Mates-Nonfunctional (requires STP+)	Works	Mates-nonfunctional & no detent retention
SATA Express cable	Works	Blocked-Key	Blocked-Key	Blocked-Key
SATA cable	Works	Blocked-Key	Blocked-Key	Blocked-Key

Layers of Standards



intel®

Agenda

1.	**10:15 AM - 10:30 AM**	**Introduction - SSS Performance**	**Eden Kim, Chair SNIA SSS TWG**
2.	10:30 AM - 10:45 AM	PCIe SSD Form Factor	Mark Meyers, Intel
3.	10:45 AM - 11:00 AM	Standards & Deployment Models	Marty Czekalski, Seagate
4.	11:00 AM - 11:15 AM	SATA-IO & SATA Express - PCIe for Client Storage	Paul Wassenberg, Sata-IO
5.	11:30 AM - 11:45 AM	PCIe 2.5" Form Factor	Janene Ellefson, Micron
6.	11:45 AM - 12:00 PM	Convergence of Memory & Storage IO Architecture	Moon Kim, Tailwind
7.	12:15 PM - 12:30 PM	Lessons from the Front Lines & Lessons for the Future	Gary Orenstein, Fusion-io
8.	12:30 PM - 1:00 PM	Panel Question & Answers / Working Lunch	

Marty Czekalski, *Seagate*

Standards and Deployment Models

Seagate

Abstract:

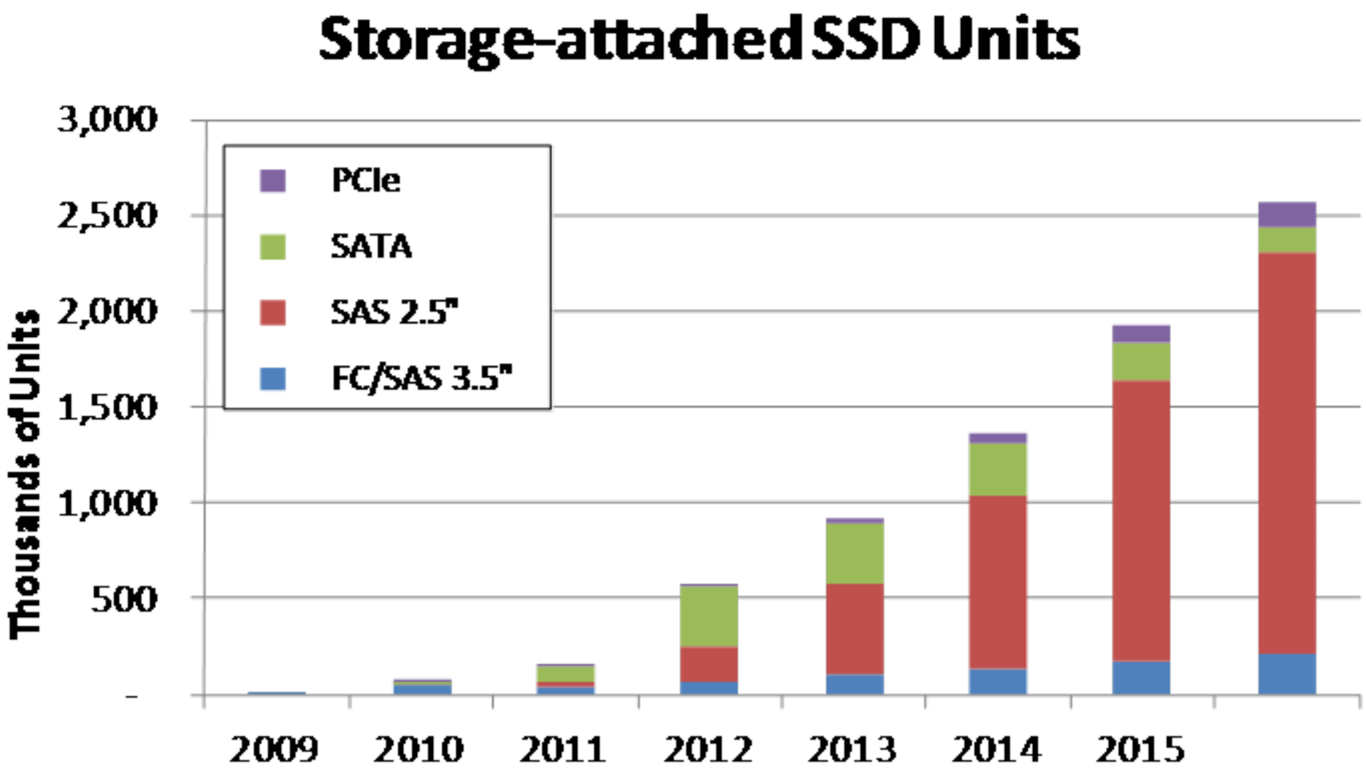
There are multiple standardization activities ongoing for PCIe based storage, some aspects of which overlap. Additionally, there are multiple deployment/provisioning options that will exist in the marketplace. A overview of these activities and issues will be discussed.

Marty Czekalski brings over thirty years of senior engineering management experience in advanced architecture development for Storage and IO subsystem design, ASIC, and Solid State Storage Systems.

He is currently Sr. Staff Program Manager within Seagate's Strategic Planning and Development Group.

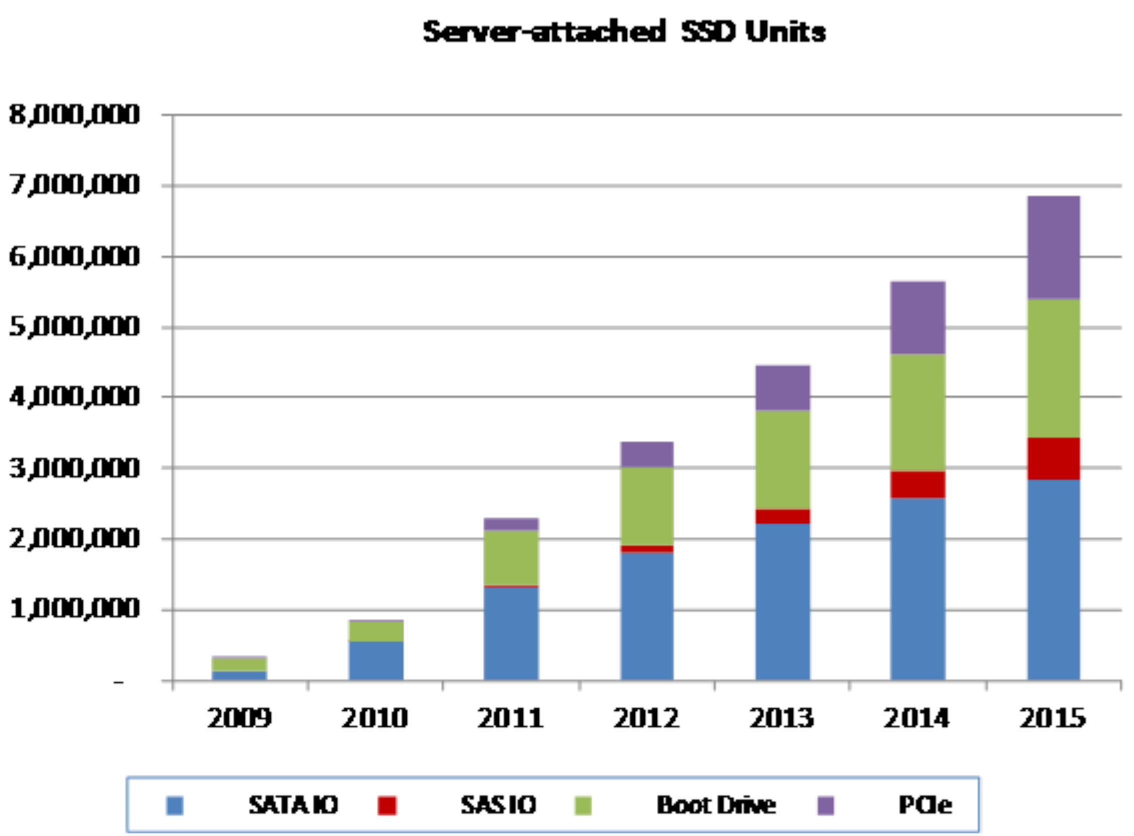
PCIe SSD Alternatives

SAS is the preferred SSD Interface for Storage Systems



Forward-Insights 11-2011

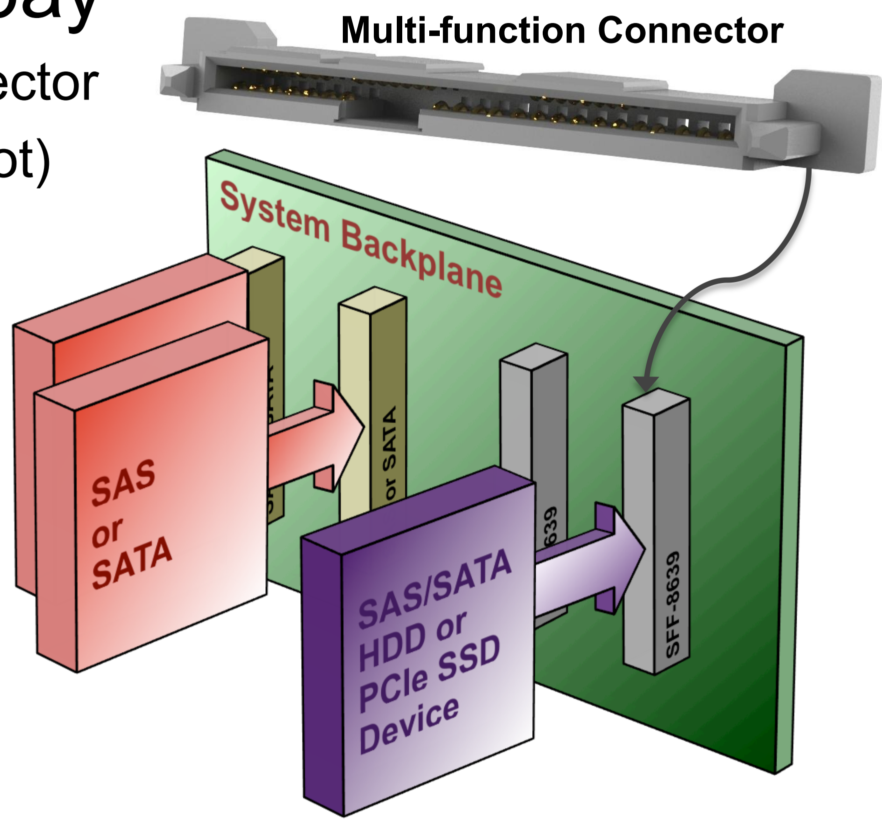
Server Attached SSDs



Forward-Insights 11-2011

Multi-Function Bay

- **Multi-function SAS/PCIe bay**
  - Uses SFF-8639 Multi-function connector
  - High performance (up to 25W per slot)
  - Hot swap, serviceability (SAS)
  - High availability (2 fault domains)
  - Supports a range of devices (system dependent)
    - 12Gb/s SAS
    - 6Gb/s SATA
    - MultiLink SAS (4 SAS Ports)
    - PCIe SSDs (emerging)
      - NVMe, SOP-PQI, Proprietary
    - SATA Express

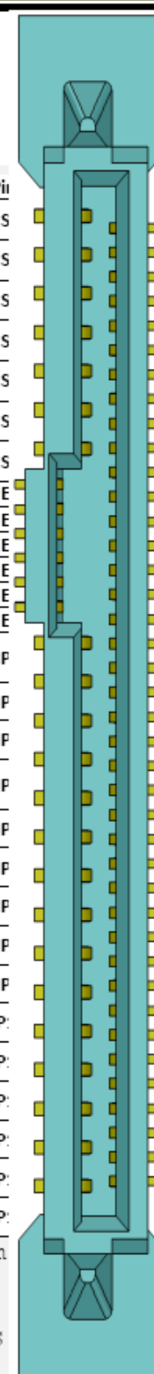


SFF-8639 Signals

Drive	Usage	Signal Description	Name	Mating	Pi
		Ground	GND	2nd	S
input	SAS+SATA	SAS/SATA/SATAe 0 Tx+	S0T+ (A+)	3rd	S
input	SAS+SATA	SAS/SATA/SATAe 0 Tx -	S0T- (A-)	3rd	S
		Ground	GND	2nd	S
output	SAS+SATA	SAS/SATA/SATAe 0 Rcv -	S0R- (B-)	3rd	S
output	SAS+SATA	SAS/SATA/SATAe 0 Rcv +	S0R+ (B+)	3rd	S
		Ground	GND	2nd	S
input	Dual Port	ePCIe RefClk + (port B)	RefClk1+	3rd	E
input	Dual Port	ePCIe RefClk – (port B)	RefClk1-	3rd	E
input	ePCIe opt	3.3V for SM bus	3.3Vaux	3rd	E
input	Dual Port	ePCIe Reset (port B)	ePERst1#	3rd	E
input	ePCIe	ePCIe Reset (port A)	ePERst0#	3rd	E
		Reserved	RSVD	3rd	E
input	SATAe +SAS4	Reserved(WAKE#/OBFF), SASAct2	RSVD(Wake#) /SASAct2	3rd	P
Bi-Dir	SATAe	SATAe Client /SAS reset	sPCIeRst/SAS	3rd	P
input	SATAe	Reserved (DevSLP#)	RSVD(DevSLP#)	2nd	P
output	SATAe + ePCIe	Interface Detect (Was GND-precharge)	IfDet#	1st	P
	all	Ground	GND	2nd	P
	all			2nd	P
NC	SAS+SATA	Precharge	5 V	2nd	P
NC	SAS+SATA	SATA, SATAe, SAS only		3rd	P
NC	SAS+SATA			3rd	P
	all	Presence (Drive type)	PRSNT#	2nd	P1
Bi-Dir	all	Activity(output)/Spinup	Activity	3rd	P1
	all	Hot Plug Ground	GND	1st	P1
input	all	Precharge	12 V	2nd	P1
input	all	All – 12V		3rd	P1
input	all	Only power for ePCIe SSD		3rd	P1

ePCIe → Enterprise PCIe (separate from SATA/SAS)

SATAe → SATA Express (Client PCIe- muxed on SATA/SAS signals)



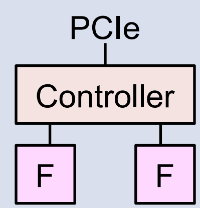
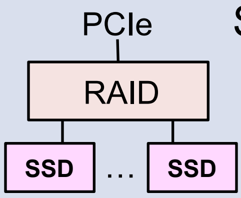
Pin #	Mating	Name	Signal Description	Usage	Drive
E7	3rd	RefClk0+	ePCIe Primary RefClk +	ePCIe	input
E8	3rd	RefClk0-	ePCIe Primary RefClk -	ePCIe	input
E9	2nd	GND	Ground		
E10	3rd	PETp0	ePCIe 0 Transmit +	ePCIe	input
E11	3rd	PETn0	ePCIe 0 Transmit -	ePCIe	input
E12	2nd	GND	Ground		
E13	3rd	PERn0	ePCIe 0 Receive -	ePCIe	output
E14	3rd	PERp0	ePCIe 0 Receive +	ePCIe	output
E15	2nd	GND	Ground		
E16	3rd	RSVD	Reserved		
S8	2nd	GND	Ground		
S9	3rd	S1T+	SAS/SATAe 1 Transmit +	SAS+SATAe	input
S10	3rd	S1T-	SAS/SATAe 1 Transmit -	SAS+SATAe	input
S11	2nd	GND	Ground		
S12	3rd	S1R-	SAS/SATAe 1 Receive -	SAS+SATAe	output
S13	3rd	S1R+	SAS/SATAe 1 Receive +	SAS+SATAe	output
S14	2nd	GND	Ground		
E17	3rd	RSVD	Reserved		
E18	2nd	GND	Ground		
E19	3rd	PETp1/S2T+	ePCIe 1 /SAS 2 Transmit +	ePCIe+SAS4	input
E20	3rd	PETn1/S2T-	ePCIe 1 /SAS 2 Transmit -	ePCIe+SAS4	input
E21	2nd	GND	Ground		
E22	3rd	PERn1/S2R-	ePCIe 1 /SAS 2 Receive -	ePCIe+SAS4	output
E23	3rd	PERp1/S2R+	ePCIe 1 /SAS 2 Receive +	ePCIe+SAS4	output
E24	2nd	GND	Ground		
E25	3rd	PETp2/S3T+	ePCIe2 / SAS 3 Transmit +	ePCIe+SAS4	input
E26	3rd	PETn2/S3T-	ePCIe2 / SAS 3 Transmit -	ePCIe+SAS4	input
E27	2nd	GND	Ground		
E28	3rd	PERn2/S3R-	ePCIe 2 / SAS 3 Receive -	ePCIe+SAS4	output
E29	3rd	PERp2/S3R+	ePCIe 2 / SAS 3 Receive +	ePCIe+SAS4	output
E30	2nd	GND	Ground		
E31	3rd	PETp3	ePCIe 3 Transmit +	ePCIe	input
E32	3rd	PETn3	ePCIe 3 Transmit -	ePCIe	input
E33	2nd	GND	Ground		
E34	3rd	PERn3	ePCIe 3 Receive -	ePCIe	output
E35	3rd	PERp3	ePCIe 3 Receive +	ePCIe	output
E36	2nd	GND	Ground		
E37	3rd	SMClk	SM-Bus Clock	PCIe opt	Bi-DIr
E38	3rd	SMDat	SM-Bus Data	PCIe opt	Bi-Dir
E39	3rd	DualPortEn#	ePCIe 2x2 Select	Dual Port	input

From: SFF-8639 Rev. 0.5, January 3, 2012

T10/STA Standards Update

- Performance Enhancements
  - 12Gb/sec SAS (2013 Product Shipments)
  - Copy Offload
- Power management
  - Ability to adjust power consumption vs performance
- Multi-function (SAS/PCIe) serviceable bay
  - SFF-8639 Connector
- SCSI over PCIe (SOP-PQI)
  - Direct attached devices (e.g. SSDs)
  - HBAs, RAID controllers, and Bridge devices
- New device types – SMR, SSD Commands & Hints

Enterprise Interfaces: PCIe SSDs

	Native	Aggregator
Commands/Transport	 Proprietary (FTL[1] in host/ main memory)	 SCSI or SATA (Multiple SSDs & controller on card)
Committee	None	None
Standards Based	No	**Yes**
Performance with Flash	High	High
CPU/Memory Overhead	High-Low	**Low**
Latency with short queue	Very Low	Low
Latency with deep queue	Moderate	**Low**
Use Case Extensibility	No	**Yes** (RAID, HBA, etc)
Maturity	Evolving	**Based on Proven Industry Architectures**
Enterprise feature set (PI, Security, Mgmt, etc.)	No	Depends on implementation

[1] FTL : Flash Translation Layer

LSI WarpDrive SLP-300 PCIe Solid State Storage Acceleration Card

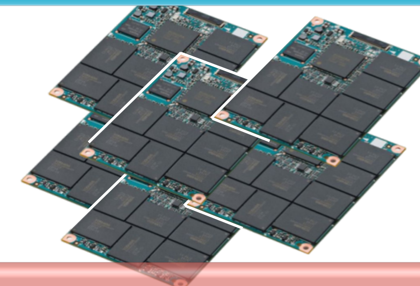
Base LSI Data Protection Layer (DPL) & Storage Management

LSI's robust 6Gb/s SAS Controller

+

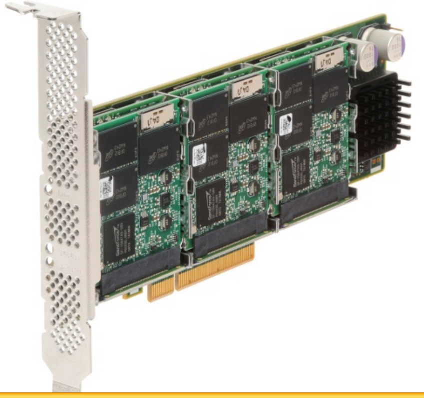
SSD Optimized firmware

+



Up to 6 Custom SSD Modules

=



Enterprise Class

Bootable, ½ height, ½ length HBA

Minimal CPU/RAM overhead

Highest IOPs per $
Highest IOPs per Watt

Application Acceleration for IO Intensive and Latency Sensitive Workloads

Enterprise Interfaces: The Future of PCIe SSDs

SCSI Trade Association

	SOP/PQI[1]	NVMe[2]
Commands/Transport	PCIe Controller F F SOP/PQI[3] (FTL in controller)	PCIe Controller F F NVMe/NVMe (FTL in controller)
Committee	··· **T10/INCITS[4]**	Industry Working Group
Standards Based	**Yes (ANSI/ISO)**	No
Performance with Flash	Very High	Very High
CPU Overhead	Low	Low
Latency with short queue	Very Low	Very Low
Latency with deep queue	Low	Low
Use Case Extensibility	**Yes** (RAID, HBA, etc.)	No (NVM only)
Maturity	**Investment Protection**	TBD
Enterprise feature set (PI, Security, Mgmt, etc.)	**Full Support**	Limited

[1]SOP : SCSI over PCI Express
[2]NVMe : Non- Volatile Memory Express
[3]PCIe Queuing Interface
[4]INCITS : International Committee for Information Technology Standards

SAS/SCSI/SOP Advantages

- Preserves Storage Investment – Logical SCSI
- Broad Open Industry Standards Support
- Dynamic Platform for Storage Innovation
- Enterprise Proven – RAS (Hot Plug)
- Multi-Host, High Queue Depths, Concurrency
- Depth & Breath of Infrastructure
- Ease of integration with existing management infrastructures & features
- Compliments PCIe Attached Storage

So who wins? - TBD

- NVMe has an early lead in development, but not hardened yet
- SOP is behind NVMe in development, but has a more robust ecosystem
- Support across industry is fragmented
- Market is still small, can it sustain the current level of investment?
- SAS controllers > 1 Million IOPS diminish PCIe SSD differentiation
- Once the PCIe capable bays are available, any PCIe device can be packaged in a 2.5” FF and used, in as long as a driver exists.
  - Creates confusion and fragment the market
- Open issues remain
  - Interoperability – Electrical spec for the bay??
  - Hot plug?
  - Compliance testing?
- Will additional form factors emerge and further fragment the market

Agenda

1.	**10:15 AM - 10:30 AM**	**Introduction - SSS Performance**	**Eden Kim, Chair SNIA SSS TWG**
2.	10:30 AM - 10:45 AM	PCIe SSD Form Factor	Mark Meyers, Intel
3.	10:45 AM - 11:00 AM	Standards & Deployment Models	Marty Czekalski, Seagate
4.	11:00 AM - 11:15 AM	SATA-IO & SATA Express - PCIe for Client Storage	Paul Wassenberg, Sata-IO
5.	11:30 AM - 11:45 AM	PCIe 2.5" Form Factor	Janene Ellefson, Micron
6.	11:45 AM - 12:00 PM	Convergence of Memory & Storage IO Architecture	Moon Kim, Tailwind
7.	12:15 PM - 12:30 PM	Lessons from the Front Lines & Lessons for the Future	Gary Orenstein, Fusion-io
8.	12:30 PM - 1:00 PM	Panel Question & Answers / Working Lunch	

Paul Wassenberg, *SATA-IO*

SATA-IO & SATA Express – PCIe for Client Storage

Abstract:

Since its introduction in 2001, SATA technology has evolved from a solely client/server storage interface to provide low-cost, high performance storage solutions for a wide variety of applications. There is an emerging segment of the client storage market, SSDs and hybrid HDDs, that requires higher performance than today's 6Gb/s SATA. To meet the needs of this segment, SATA-IO introduced SATA Express, a new specification that provides higher performance by utilizing readily available, fast, and scalable PCI Express connectivity while preserving established SATA software compatibility. This presentation will describe the details of SATA Express and the implications for devices and systems that will support it.

Paul Wassenberg has over 20 years of experience in data storage and has been deeply involved with storage interface technology, including SATA since its inception. Early in his career, he was a storage controller designer, before moving into Marketing in the HDD industry, and eventually into storage semiconductors.

Paul currently holds the position of Director, Product Marketing with Marvell Semiconductor. In that role, he has responsibility for transceiver technology and HDD/SSD storage standards. He is on the SATA-IO board of directors and chairs the SNIA Solid State Storage Initiative. Paul holds BSEE and MBA degrees from San Jose State University.

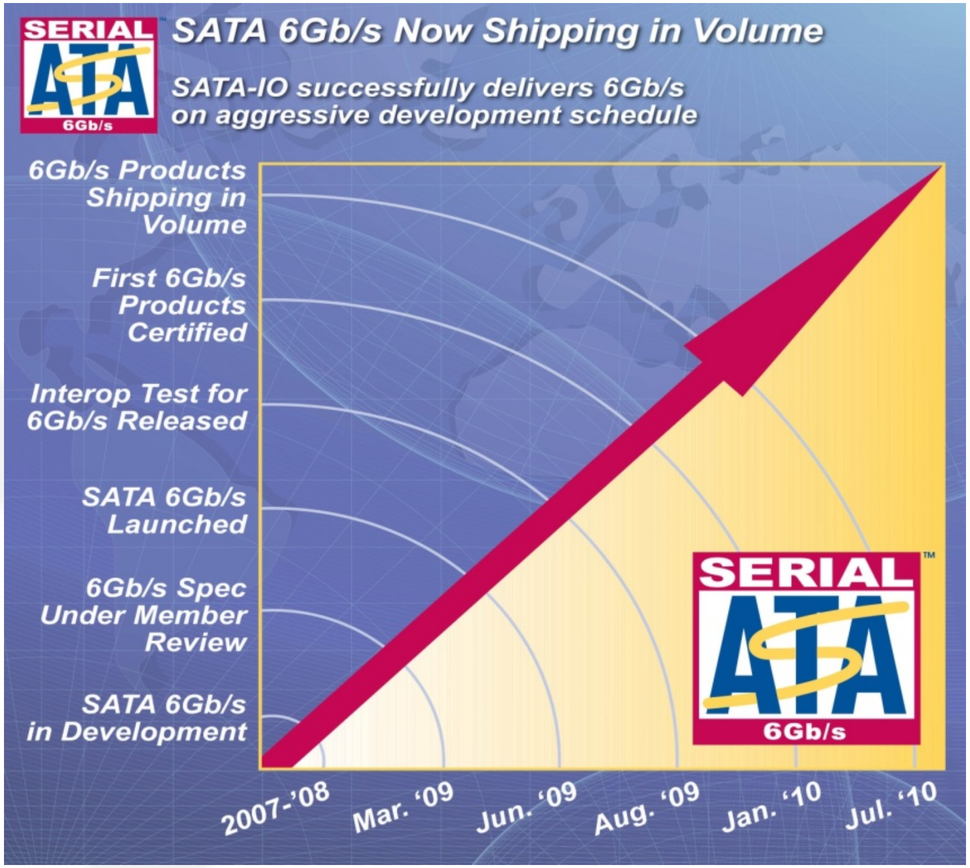
SATA Express

Evolving SATA for High Speed Storage

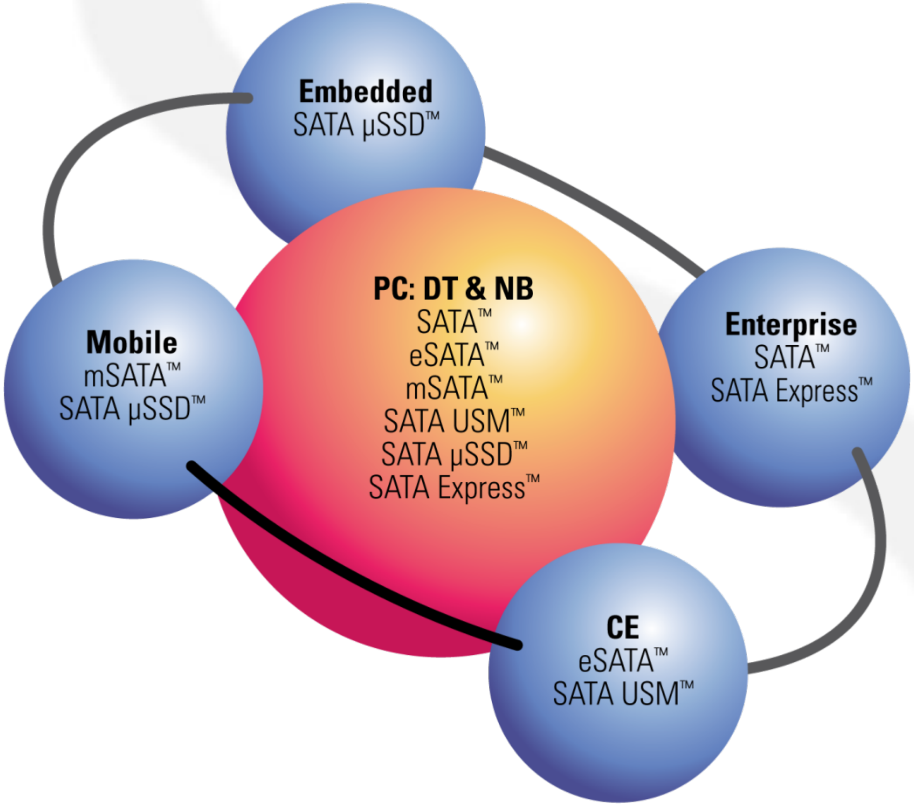
January 23, 2012

SATA for PC Client Storage

- A Mature Interface
  - SATA is the de facto standard for PC storage; also widely implemented in mobile and enterprise applications
  - Adoption of SATA 6Gb/s technology is strong



A Growing Ecosystem



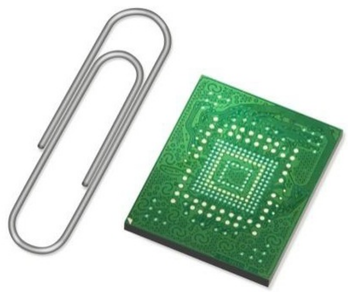
SATA implementations are becoming increasingly application specific

Since its introduction, SATA has evolved into new application spaces and now provides storage interface solutions for HDDs, ODDs, SSDs, and Hybrid HDDs in client PC, mobile, enterprise, CE, and embedded storage markets

Example Application-Specific Implementations

- mSATA (mobile applications)

- SATA µSSD (embedded applications)



- SATA Universal Storage Module (consumer electronics, PC applications)

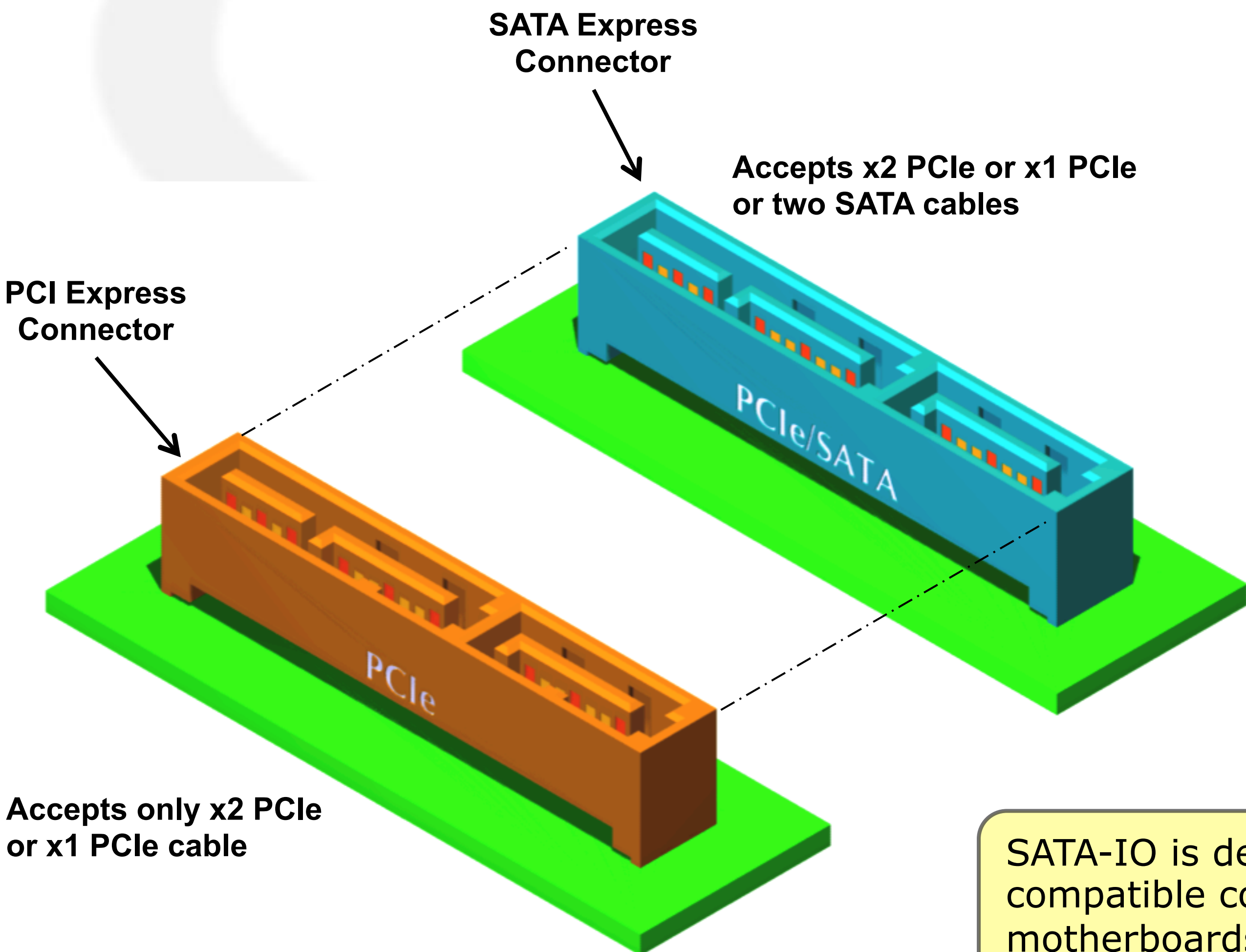
Application Speed Requirements

- Today, most applications are well-served by SATA 6Gb/s and will be for the foreseeable future
- However, SSDs and Hybrid HDDs will soon require greater speeds than those enabled by the current generation of SATA

Introducing SATA Express™

- To meet speed requirements in SSD/hybrid drive applications, SATA-IO is developing SATA Express ™
  - Combines SATA software infrastructure with the PCI Express® (PCIe®) interface
    - Utilizes standard register-level interface such as AHCI
  - Provides up to 8Gb/s and 16Gb/s
    - One lane or two lanes of PCIe
  - Defines new device and motherboard connectors to support both new SATA Express and current SATA devices
  - Will coexist with other application-specific SATA formats

SATA Express Connectors



SATA Express connector supports PCIe and SATA

- Mechanism to detect device interface
- Allows a single motherboard / backplane connector to support both interfaces

SATA Express supports HDD-compatible form factors

- Enables system-level mechanical compatibility

SATA-IO is developing backward compatible connectors for SATA Express motherboards & devices

SATA Express Benefits

- Provides a cost-effective solution for increasing device interface speed
- Specification can be completed and implemented relatively quickly, since both SATA and PCIe are already widely implemented
- Helps ensure seamless coexistence between SATA and PCIe
- Protects developer investments in both interfaces

Next Steps And Timeline

SATA Express is currently under development within the SATA-IO Cable & Connector Work Group

- Completed specification expected within 2012

In the meantime, SATA-IO will continue to optimize the existing SATA infrastructure for a wide variety of applications

- SATA will continue to be the mainstream storage interface for the foreseeable future

Agenda

1.	**10:15 AM - 10:30 AM**	**Introduction - SSS Performance**	**Eden Kim, Chair SNIA SSS TWG**
2.	10:30 AM - 10:45 AM	PCIe SSD Form Factor	Mark Meyers, Intel
3.	10:45 AM - 11:00 AM	Standards & Deployment Models	Marty Czekalski, Seagate
4.	11:00 AM - 11:15 AM	SATA-IO & SATA Express - PCIe for Client Storage	Paul Wassenberg, Sata-IO
5.	11:30 AM - 11:45 AM	PCIe 2.5" Form Factor	Janene Ellefson, Micron
6.	11:45 AM - 12:00 PM	Convergence of Memory & Storage IO Architecture	Moon Kim, Tailwind
7.	12:15 PM - 12:30 PM	Lessons from the Front Lines & Lessons for the Future	Gary Orenstein, Fusion-io
8.	12:30 PM - 1:00 PM	Panel Question & Answers / Working Lunch	

Janene Ellefson, *Micron*

PCIe 2.5" Form Factor

Abstract:

A key factor standing in the way of widespread PCIe SSD adoption is serviceability of the current card form factor. In most hosts, the card form factor requires that the system be powered down and the unit be opened up to remove the existing card and insert a new card. This is not optimal given the widespread adoption of virtualization. Powering down a machine can disrupt overall system efficiency. Providing the industry with a robust form factor that can be serviceable and still provide PCIe high-performance capability will be a game changer and will increase adoption.

The 2.5-inch form factor is an overall industry standard, and when coupled with a PCIe interface and a SATA/SAS combo connector, it becomes a portable, compact, hot-pluggable PCIe device that is very compelling and enables better performance and serviceability in enterprise systems. Enterprise applications everywhere will benefit from the increased performance, lower energy consumption compared to HDDs, and hot plug serviceability.

Janene Ellefson is the Product Marketing Manager for Enterprise PCIe SSDs and is responsible for worldwide PCIe SSD marketing efforts.

She joined Micron in 1989 and has spent the majority of her Micron career in various marketing roles, supporting NOR Flash and NAND Flash products.

Ms. Ellefson holds a BS from the University of Phoenix in business and marketing.

PCIe 2.5” Form Factor

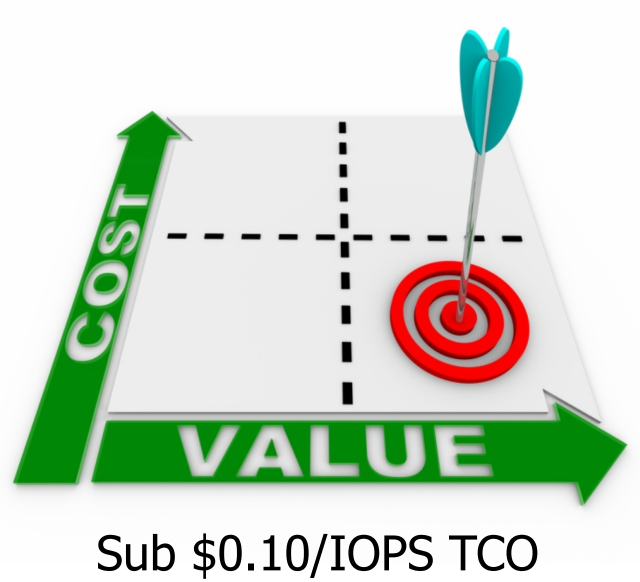
Janene Ellefson

Product Marketing Manager – PCIe SSD

Advantages of PCIe over SATA/SAS SSDs

Higher Performance

Lower power
consumption



Sub $0.10/IOPS TCO

Lots of advantages for PCIe
Enterprise would use it more if they could

What’s Holding it Back?

- No Hot-Swap capability
- Too much space
- Limited PCIe Slots
- Power down required

Todays PCIe form factors are not optimal for Enterprise serviceability

Propose the 2.5" PCIe Form Factor

All the performance of PCIe with the serviceability standards of SATA/SAS

2.5" Advantages

- PCIe performance
- Common Form Factor
- Compactness
- Serviceability
- Lower TCO
- Supports RAID

Summary

- PCIe offers lots of advantages – Adoption rates are low
- Today's PCIe form factors are not optimal for Enterprise serviceability
- 2.5" Form Factor: All the performance of PCIe with the serviceability standards of SATA/SAS
- 2.5" = increase PCIe adoption

Micron®

Focused on Memory | Engineered for Innovation

January 23, 2012

Agenda

1.	**10:15 AM - 10:30 AM**	**Introduction - SSS Performance**	**Eden Kim, Chair SNIA SSS TWG**
2.	10:30 AM - 10:45 AM	PCIe SSD Form Factor	Mark Meyers, Intel
3.	10:45 AM - 11:00 AM	Standards & Deployment Models	Marty Czekalski, Seagate
4.	11:00 AM - 11:15 AM	SATA-IO & SATA Express - PCIe for Client Storage	Paul Wassenberg, Sata-IO
5.	11:30 AM - 11:45 AM	PCIe 2.5” Form Factor	Janene Ellefson, Micron
6.	11:45 AM - 12:00 PM	Convergence of Memory & Storage IO Architecture	Moon Kim, Tailwind
7.	12:15 PM - 12:30 PM	Lessons from the Front Lines & Lessons for the Future	Gary Orenstein, Fusion-io
8.	12:30 PM - 1:00 PM	Panel Question & Answers / Working Lunch	

Dr. Moon Kim, Phd, *Tailwind*

Convergence of Memory and Storage IO Architecture

TAILWIND STORAGE
Semiconductor Storage Device Company

Abstract:

As storage devices are used in memory technologies (e.g., flash and DDR devices) in order to speed up data access, storage system designs have not been changed. That is, convention designs still utilize I/O interfaces such as PCIe.

As such, conventional designs have storage access imbalances Although, memory system technology has been utilizing DRAM-based approaches, many business applications require even larger memory spaces in order to take advantage of more recent CPU technology advancement. In this presentation, the use of the extended memory access architecture will be introduced.

As a venture partner of the Harbor Pacific Capital, Dr. Moon J. Kim serves as the CEO of TailWind Storage company. Most recently, Dr. Kim served as the Vice-Chairman & CEO Technology Advisor of Samsung Electronics Corp., where he led several special projects. He also served the executive technology advisor of LG and the senior managing executive of Exponent, a New York based technology consulting company. Dr. Kim is specialized in IO and memory architecture on HPC and main frame servers. During his 28 years in IBM R&D, he led and managed all aspects of IT technology and server development. He held the prestigious title of **IBM Master Inventor** and has led numerous **Emerging Technology** developments. He has produced over **130 inventions** and has authored several system and IT technology books and published numerous technical papers. He is an expert on the technology industry in Asia. Recently he was awarded twice by the Chinese Academy of Science for this work on the China National Supercomputing Grid and multicore processor development projects. He can be reached at mjkim@harborpac.com and (650) 690-0795, (845) 702-2422.

TAILWIND STORAGE

Semiconductor Storage Device Company

SNIA Presentation

January 2012

Dr. Moon J Kim

A New Era in Storage Architecture

- High IO demand causes IO congestion. DRAM has the highest bandwidth and least latency for CPUs, thus making it reasonable to exploit DRAM as an IO channel.
- Conventional IOs, such as PCIe, demand too many supporting resources for the IO itself, and several CPU cycles are required to move the data.
- New and innovative technologies are needed to bring IOs closer to CPU.

Expanded Storage Architecture

US006026462A

United States Patent [19]
George et al.

[11] **Patent Number: 6,026,462**
[45] **Date of Patent: *Feb. 15, 2000**

[54] MAIN STORAGE AND EXPANDED STORAGE REASSIGNMENT FACILITY

[75] Inventors: **Jonel George**, Pleasant Valley; **Steven Gardner Glassen**, Wallkill; **Matthew Anthony Krygowski**, Hopewell Junction; **Moon Ju Kim**, Wappingers Falls; **Allen Herman Preston**; **David Emmett Stucki**, both of Poughkeepsie, all of N.Y.

[73] Assignee: **International Business Machines Corporation**, Armonk, N.Y.

[*] Notice: This patent is subject to a terminal disclaimer.

[21] Appl. No.: **08/897,449**

[22] Filed: **Jul. 22, 1997**

Related U.S. Application Data

[62] Division of application No. 08/635,537, Apr. 22, 1996, Pat. No. 5,704,055, which is a continuation of application No. 08/070,588, Jun. 1, 1993, abandoned.

[58] **Field of Search** 711/2, 200, 201, 711/206, 170; 74/208, 209, 200, 201, 206

[56] **References Cited**

U.S. PATENT DOCUMENTS

4,926,322	5/1990	Stimac et al.	395/500
5,704,055	12/1997	George et al.	711/2

Primary Examiner—Tuan V. Thai
Attorney, Agent, or Firm—Lane, Aitken & McCann; Lynn L. Augspurger

[57] **ABSTRACT**

A data processing system has a processing unit and a memory which provides a common pool of physical storage. This storage is initially assigned as either main storage or expanded storage during power on. Subsequent to the initial assignment, storage assigned as main storage or expanded storage may be unassigned and thus returned to the common pool. Once returned to the common pool, the storage may be reassigned as either main storage or expanded storage. The storage reassignment is done dynamically without requiring a reset action and transparent to the operating system and any active application programs.

United States Patent [19]
George et al.

[54] **DYNAMIC RECONFIGURATION OF MAI STORAGE AND EXPANDED STORAGE BY MEANS OF A SERVICE CALL LOGICAL PROCESSOR**

[75] Inventors: **Jonel George**, Pleasant Valley; **Stev Gardner Glassen**, Wallkill; **Matthe Anthony Krygowski**, Hopewell Junction; **Moon Ju Kim**, Wappinger Falls; **Allen Herman Preston**; **Davi Emmett Stucki**, both of Poughkeeps all of N.Y.

[73] Assignee: **International Business Machines Corporation**, Armonk, N.Y.

[21] Appl. No.: **635,537**

[22] Filed: **Apr. 22, 1996**

Related U.S. Application Data

[63] Continuation of Ser. No. 70,588, Jun. 1, 1993, abandoned.

5,479,631 12/1995 Manners et al. 395/465

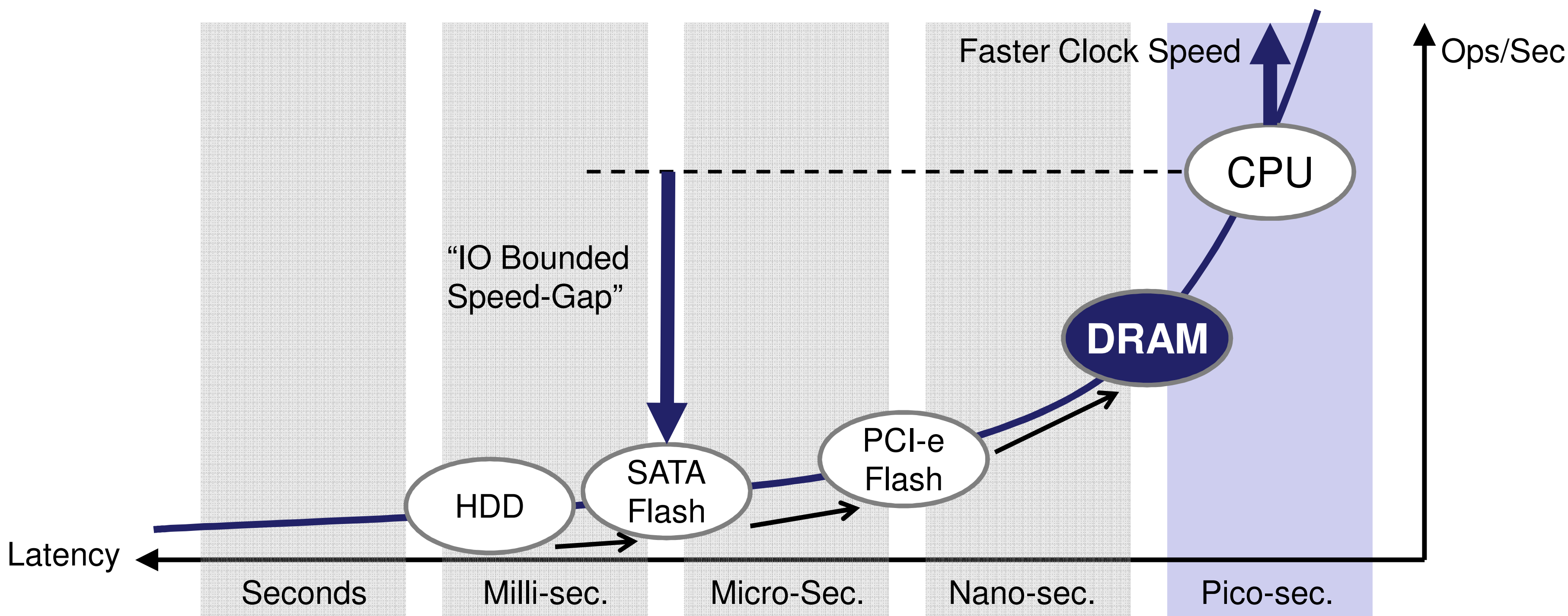
Primary Examiner—Tod R. Swann
Assistant Examiner—Tuan V. Thai
Attorney, Agent, or Firm—Lynn L. Augspurger; Laurence J. Marhoefer

[57] **ABSTRACT**

A data processing system has a processing unit and a

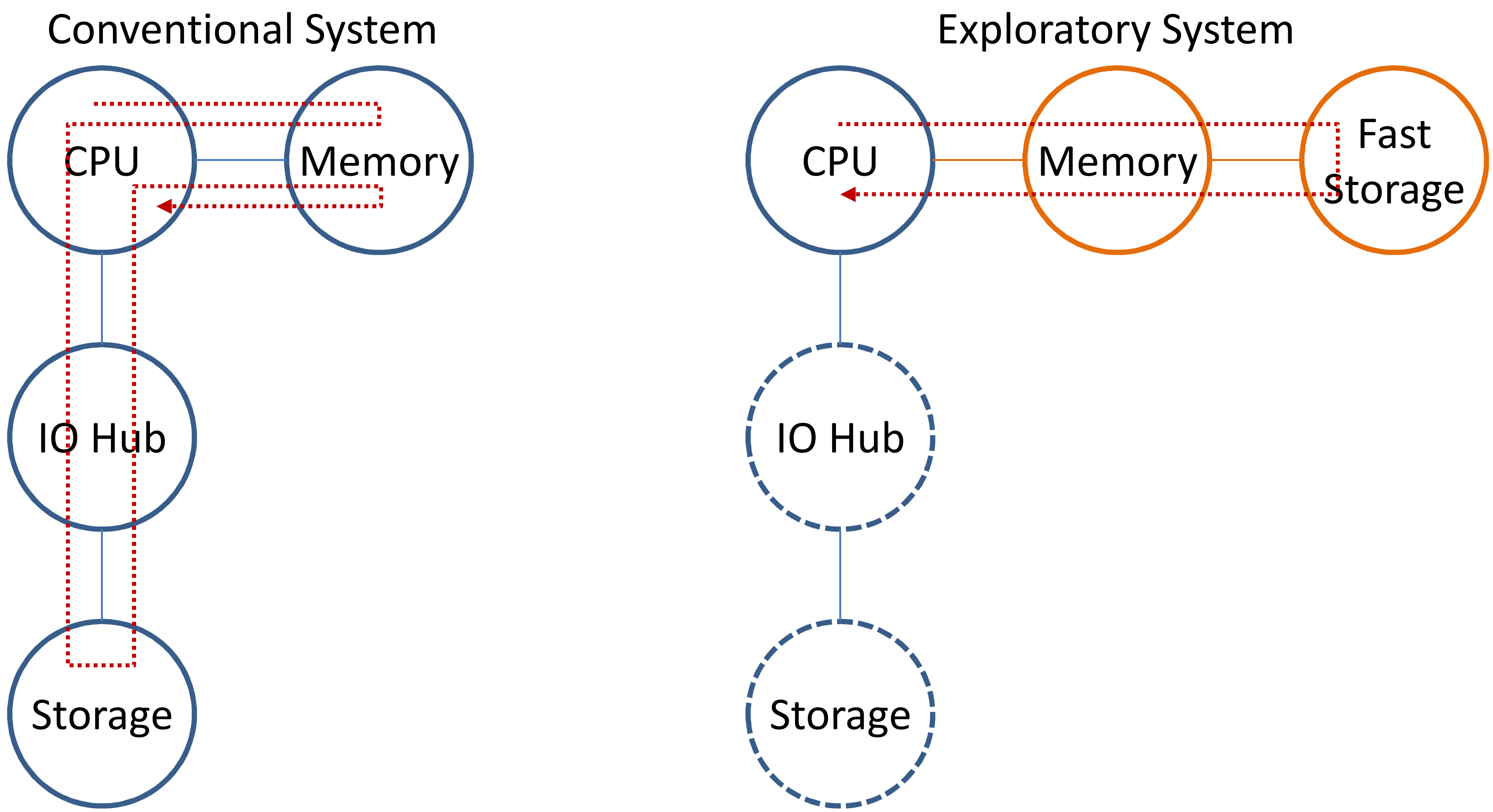
Problem – Increasing need for faster storage

- As CPUs reach faster clock speeds, storage technologies have evolved to reduce the "Speed-Gap" between the CPU and the storage device.

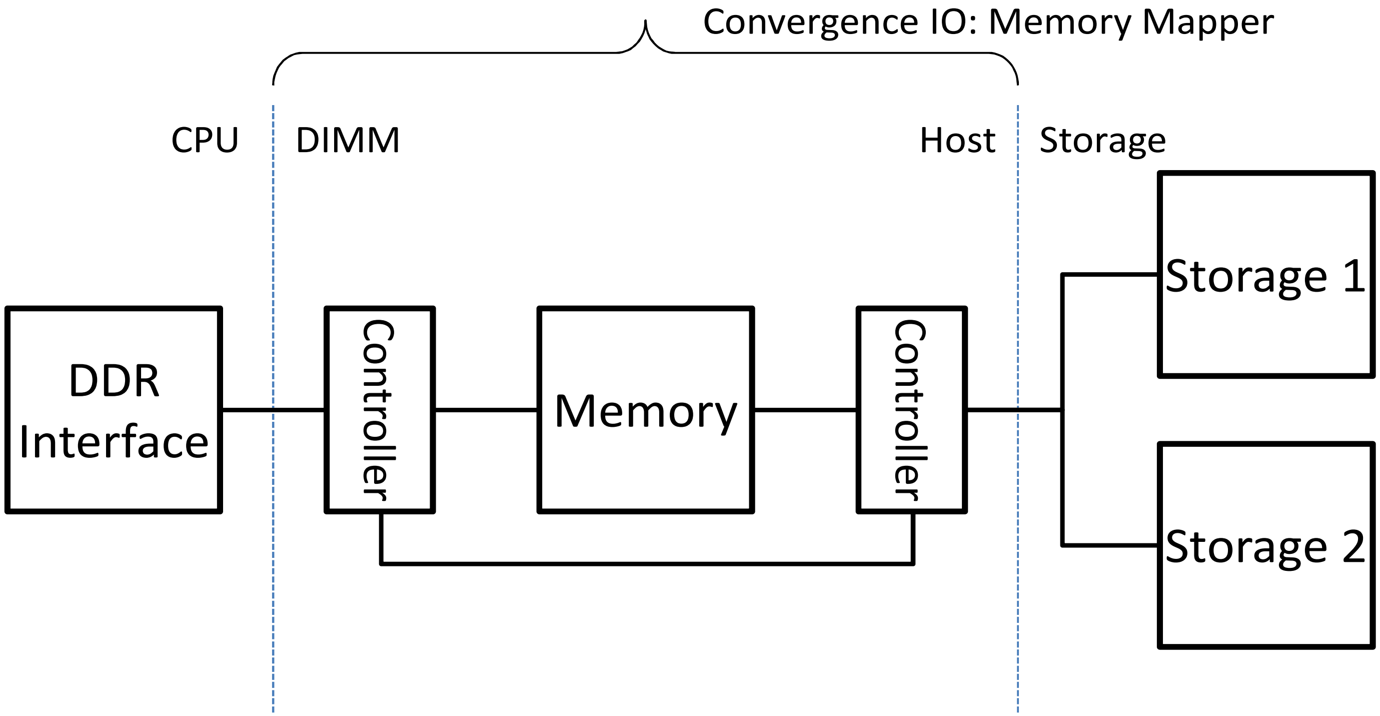


Source : Shirish Jamthe , Director of System Engineering, Virident Systems, Inc., August 2011

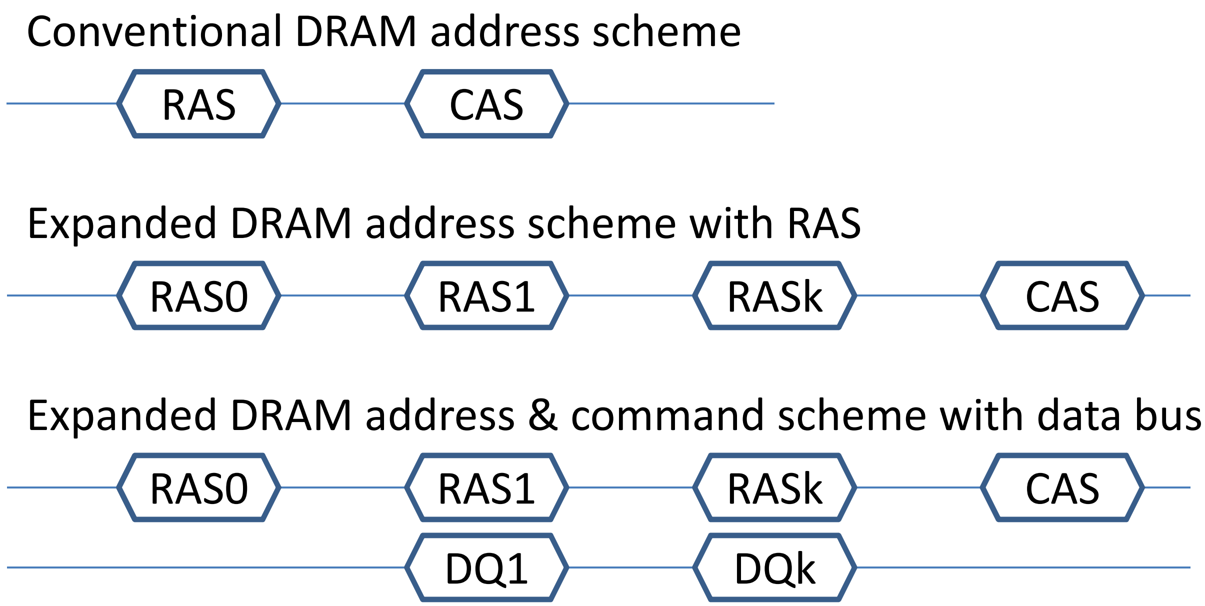
New Architecture Consideration



DDR+ Extension and Memory Mapper



Example Implementation



- Slight modification and expansion of SDRAM address scheme allows infinite address space extension, additional command mode, status register space, etc.

New Architecture

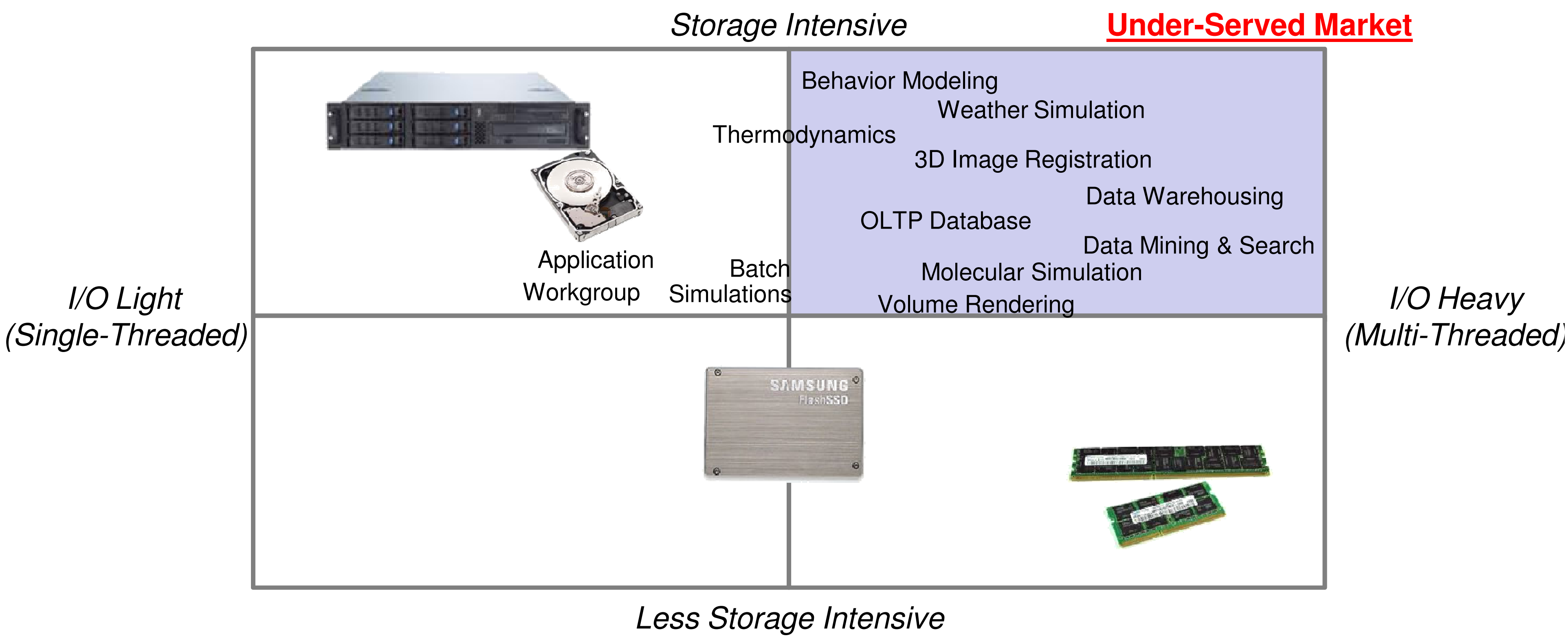
- Memory space can extend additional description tag stored in specified register and memory location.
- Memory space and thread are virtualized within the limited memory space.
- Thread sees physical address space.
- OS maintains virtualization of threads.
- Storage is connected through memory-to-storage mapper.
- Memory can serve as a large off-CPU cache of storage.
- Storage should be fast enough to support memory operation.
- Storage can be accessed directly.

Tailwind Storage Company

- Tailwind's DDR-based storage technology meets the increasing need for ultra-fast storage devices that match faster CPUs.
- Tailwind Storage prototypes have been approved by major OEM partners.
- Tailwind maintains a robust IP portfolio.
- Tailwind's team has over 100 years of IO & Memory experience in storage system technology.

Problem – High Performance Computing Environment

- Existing storage technologies are unable to fully meet demanding performance in <u>multi-threaded, data-heavy computing environments</u>.



NAND Flash Memory Technology

- NAND technology maintains a lower effective capacity.
- IOPS testing: latency is effective and favorable under NAND. It may not reflect the real memory operations.
- NAND scaling usually increases latency.

Solution – Benefits of Our DDR Storage Products

- **Expandable**
- **Unbeatable Speed**
  - Much faster than flash based SSD
  - Access to storage is closer to speed of CPU
- **Sustainable Performance**
  - No performance degradation*
  - Symmetric read/write performance
  - Linear and transparent
  - Consistent performance regardless of workload mix

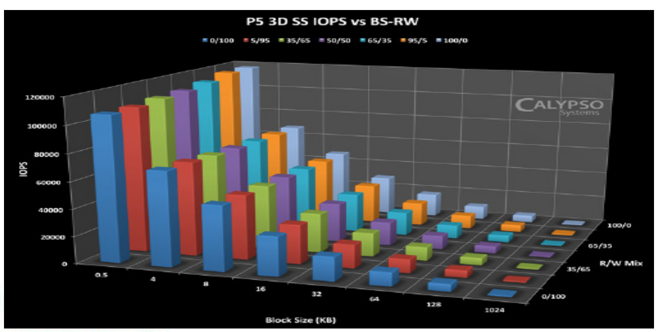
Our Solution – DDR Advantages

Latency

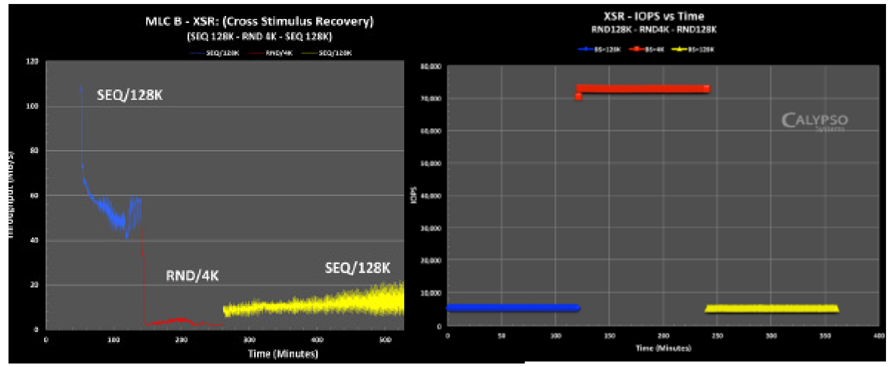
- Faster than Flash SSD and HDD , in the order of nanoseconds instead of milliseconds or microseconds

Sustainability of Performance**

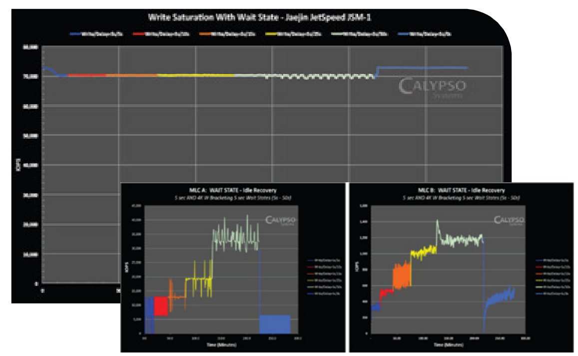
- No performance degradation
- Symmetric, and linear read/write performance
    - Read / Write Parity
    - Consistent performance regardless of work load mix



Fast transition in handling mix block sizes**



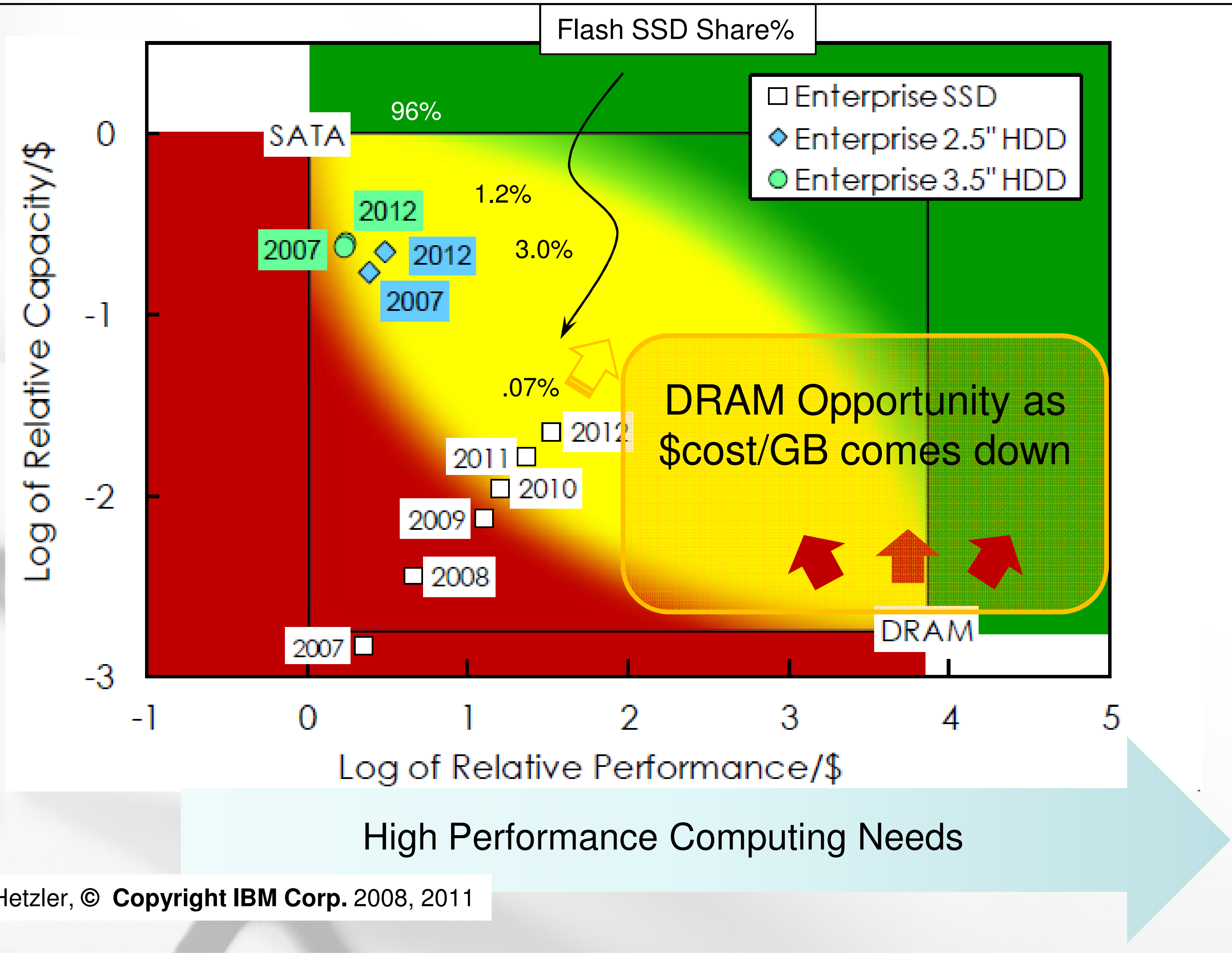
No idle recovery required, minimum background garbage collection**



**Actual test results by an independent test service company with Tailwind's Pro-E

IT Storage Hierarchy, Trends and Opportunity

Flash SSD share % is still small
Opportunity for DRAM as market demands higher performance



Our Solution – Y 2011 Early Adopter products

- Pro Extended
  - 64GB DDR, 700MB/s
  - Initial evaluation completed with prototype from OEM

- Hybrid SSD Storage & Server
  - 8 Core CPU, 512GB DDR, 5GB/s
  - Evaluation approved by major OEM for market development

- Super-Mini
  - 8 Core CPU, 1TB DDR, 16GB/s
  - Customer evaluation in progress

Y2012 TW Product Specification

Feature	*Pro-Extreme Prototype	Backdraft	2nd Backdraft
Memory technology	DDR2 SDRAM	DDR3 SDRAM	DDR3 SDRAM
Capacity	64GB	512GB max.	1024GB max.
Host interface	PCIe Gen. 1, 4x	PCIe Gen. 2, 8x	PCIe Gen. 2, 16x
Host bandwidth	0.8GB/s	4GB/s	8GB/s
Form factor	Full length PCIe	Half, full, dual PCIe	Half, full, dual PCIe

Contact Information

For more information:

Dr. Moon J Kim

- mjkim@tailwindstorage.com
- Tel: 650-690-0795
- 525 University Ave, Suite 100, Palo Alto, CA 94301

Agenda

1.	**10:15 AM - 10:30 AM**	**Introduction - SSS Performance**	**Eden Kim, Chair SNIA SSS TWG**
2.	10:30 AM - 10:45 AM	PCIe SSD Form Factor	Mark Meyers, Intel
3.	10:45 AM - 11:00 AM	Standards & Deployment Models	Marty Czekalski, Seagate
4.	11:00 AM - 11:15 AM	SATA-IO & SATA Express - PCIe for Client Storage	Paul Wassenberg, Sata-IO
5.	11:30 AM - 11:45 AM	PCIe 2.5" Form Factor	Janene Ellefson, Micron
6.	11:45 AM - 12:00 PM	Convergence of Memory & Storage IO Architecture	Moon Kim, Tailwind
7.	12:15 PM - 12:30 PM	Lessons from the Front Lines & Lessons for the Future	Gary Orenstein, Fusion-io
8.	12:30 PM - 1:00 PM	Panel Question & Answers / Working Lunch	

Gary Orenstein, *Fusion-io*

PCIe – Lessons from the Front Lines; and a Look to the Future

Abstract:

In a matter on no time, at least in storage years, NAND flash has emerged in the data center as a force changing the storage landscape. Perhaps no area where this impact has been more visible and more dramatic is in the placement of NAND flash close to the CPUs. By placing process-critical data close to the CPU customers see leap fold performance improvements for their applications and databases.

This talk will explore customer input, reactions, and lessons on new models of deploying NAND flash using PCIe, along with taking a look at the future. Today the industry is on the cusp of a new storage continuum. PCIe as a storage mechanism now spans everything from high end servers like the HP DL 980 with up to 16 PCIe I/O expansion slots, all the way down to Thunderbolt, a consumer focused link based on PCIe. There are also important industry initiatives underway like SCSI Express and activities within T10. This talk will cover some of the latest proposals and how the industry and customers stand to benefit from these developments.

VP of Products, Fusion-io, Gary has served in leadership roles at numerous data center infrastructure companies. Prior to Fusion-io he was the vice president of marketing at MaxiScale, focused on web scale file systems and acquired by Overland Storage.

Prior to MaxiScale, he was the vice president of marketing and business development at Gear6, focusing on storage and web caching. He also served as vice president of marketing at Compellent which went public and 2007, and was a co-founder at Nishan Systems, acquired by McDATA/Brocade.

PCIe Storage – Lessons From the Front Lines and a Look to the Future

Gary Orenstein, VP of Products

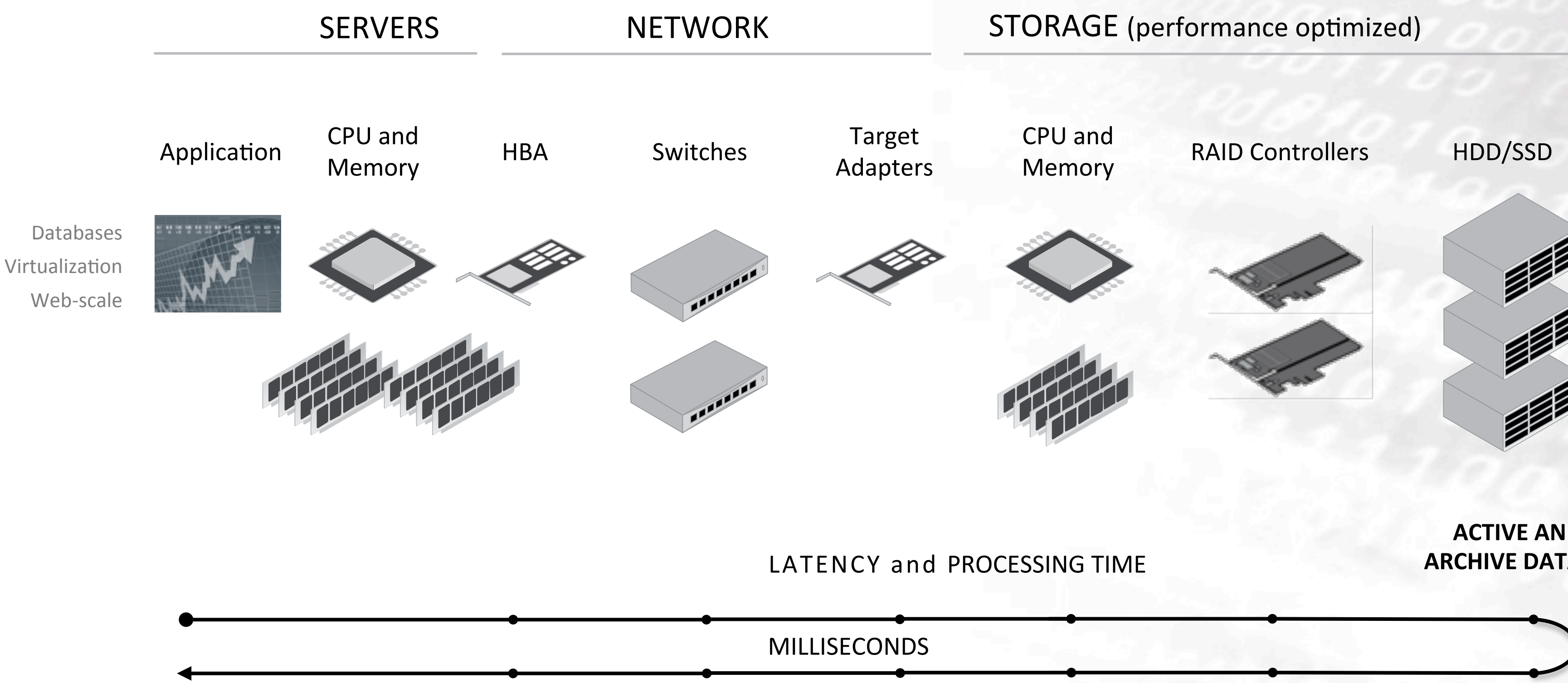
FUSION-iO

SNIA™

Winter Symposium
January 2012

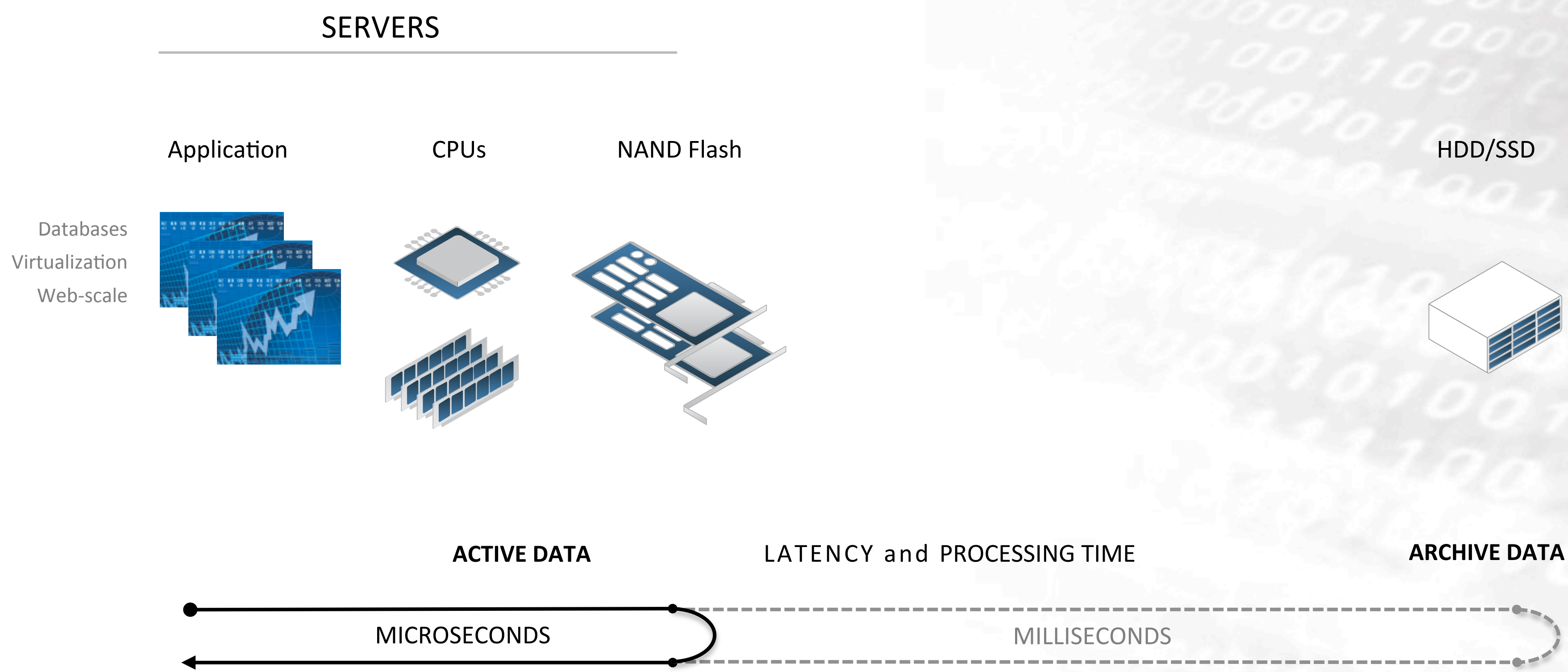

Lessons
from the
front lines

TRADITIONAL ARCHITECTURE



PROCESS CRITICAL DATA CLOSE TO THE CPU

FUSION-IO



APPLICATION ACCELERATION EXAMPLES

FUSION-io

Microsoft SQL Server
Case Study Equinox

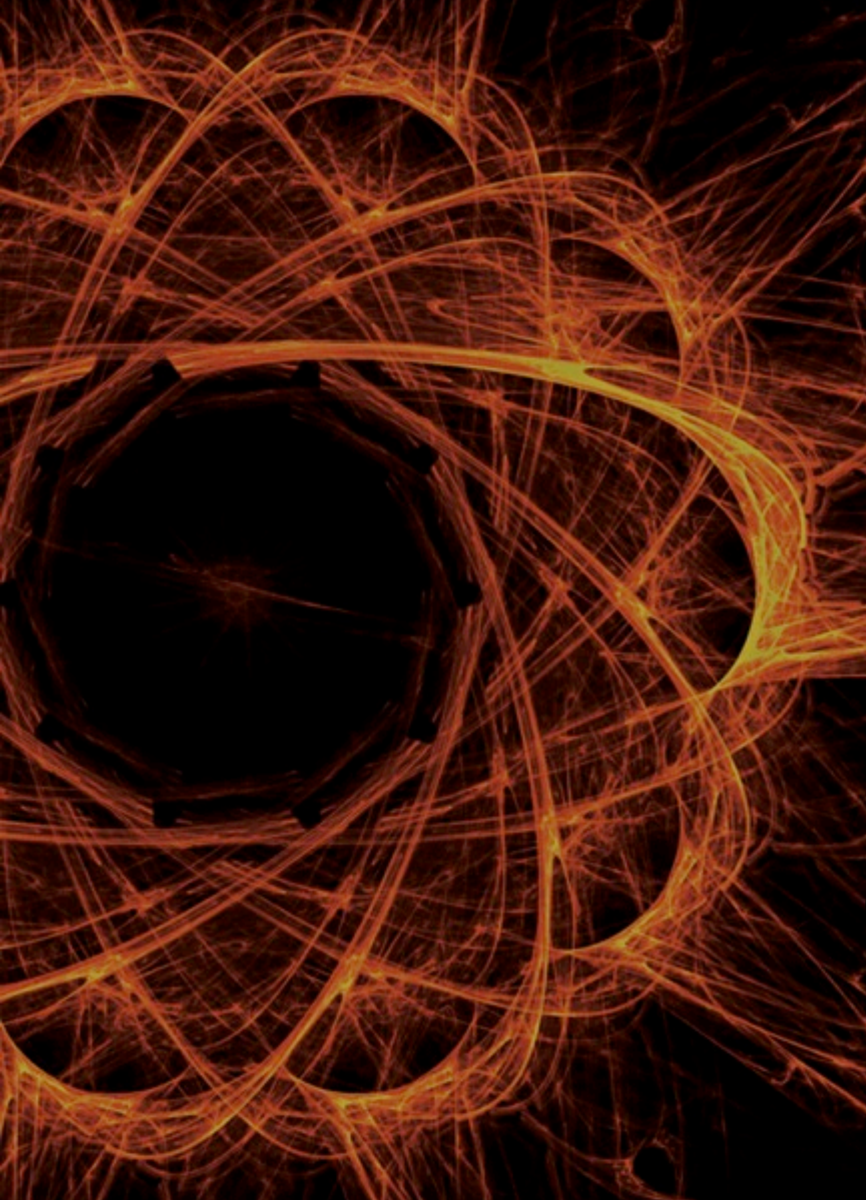
AVERAGE DATABASE THROUGHPUT

VMware View
Case Study V3

PERFORMANCE PER RACK UNIT

ORACLE
Case Study Datalogix

QUERY TIME

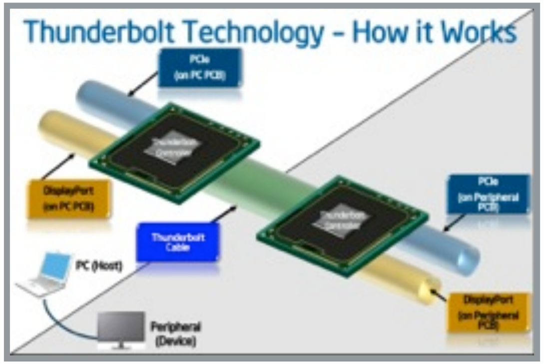


Turning Point for PCIe

THE WAY WE WERE

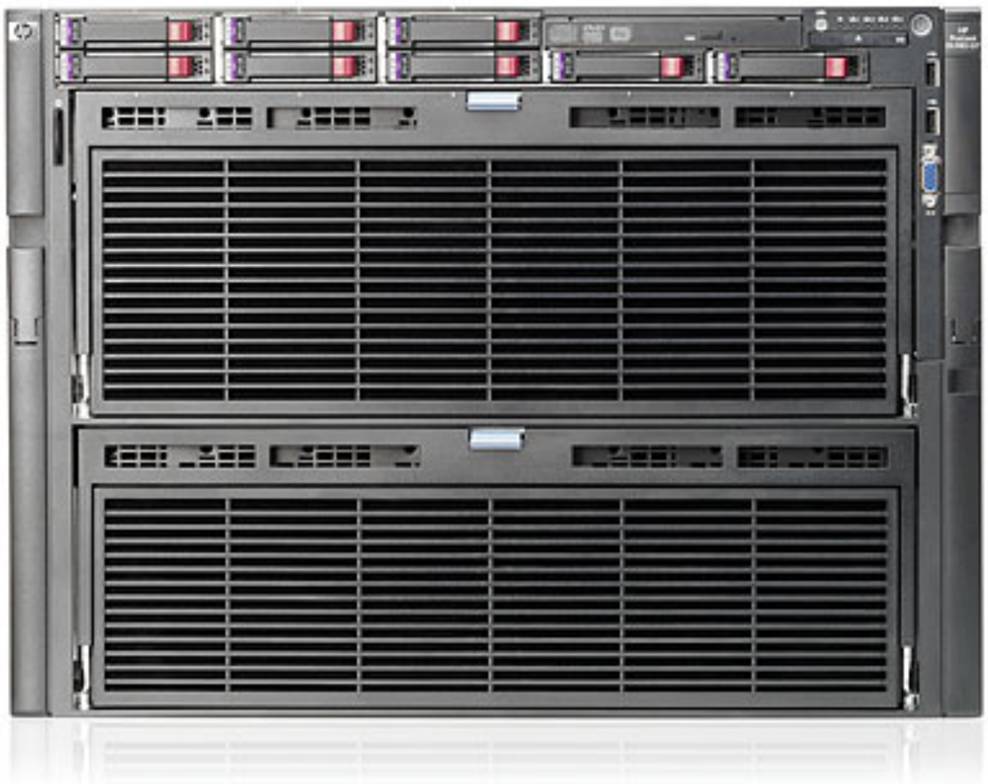
Consumer	SMB/SME	Enterprise
• USB • SATA • IDE	• SATA • SAS • FC	• SAS • FC • IB

THUNDERBOLT PCIE DEVICES



CUBIX
Visual & GPU Compute Solutions

HP DL 980



Up to 11 full height/full length slots supported

2410 GB x 11

26.5TB per server

x11

WHERE WE ARE HEADED

Consumer	SMB/SME	Enterprise
PCIe		

SCSI Express

SCSI EXPRESS

A set of industry initiatives delivering a
PCIe Express based enterprise storage solution

Industry Initiative	Focus
SCSI Over PCIe (SOP)	Streamline SCSI command set optimized for solid state
PCIe Queuing Interface (PQI)	Flexible and extensible transport layer
Universal drive connector	Supporting current and emerging devices
PCIe physical layer	Drive error handling and asynchronous hot add/ remove
Native OS support	Standard drivers to support range of devices

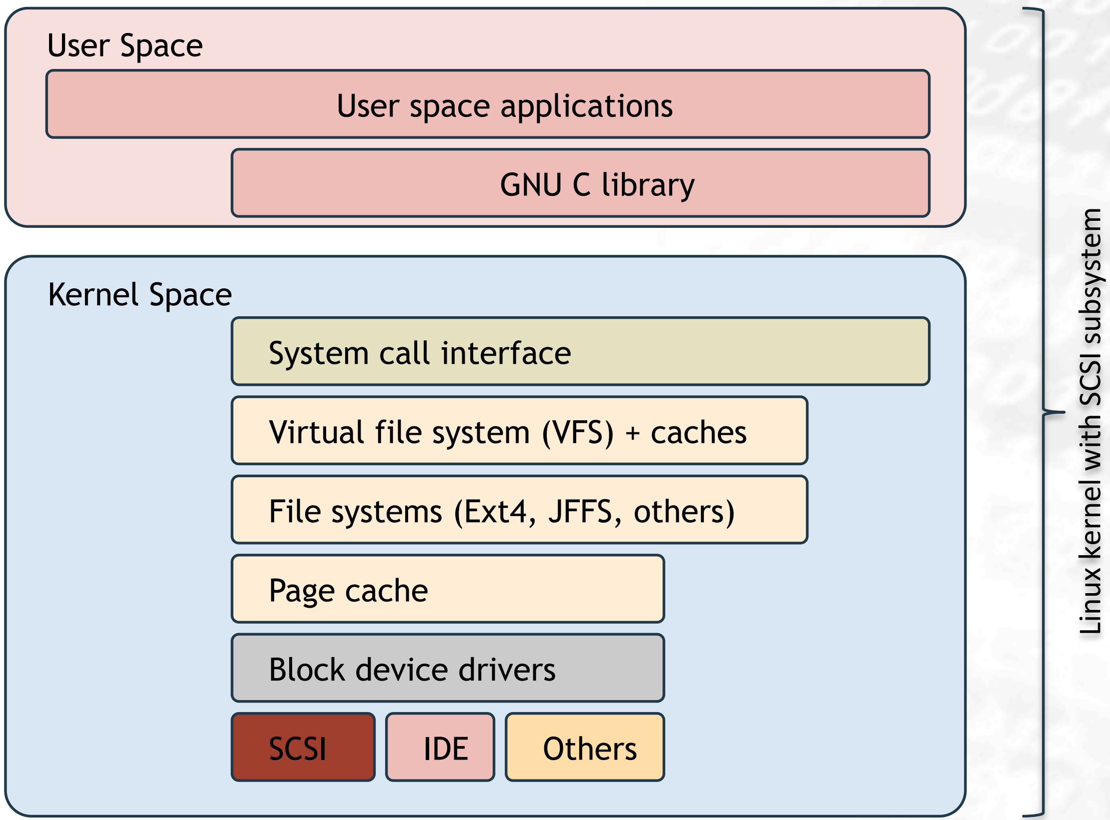
SCSI EXPRESS AND NVM EXPRESS

SCSI Express	NVM Express (NVMe)
A standard to combine SCSI and PCIe	A register level interface for host software to communicate with a non-volatile memory subsystem
Enterprise Roots (SCSI based) SCSI reliability and dependability	**Consumer Roots (ATA based)**
Extensible configurations	Limited configuration support
Driven in Industry Storage Forums - ANSI T10	Proprietary governance / limited expertise

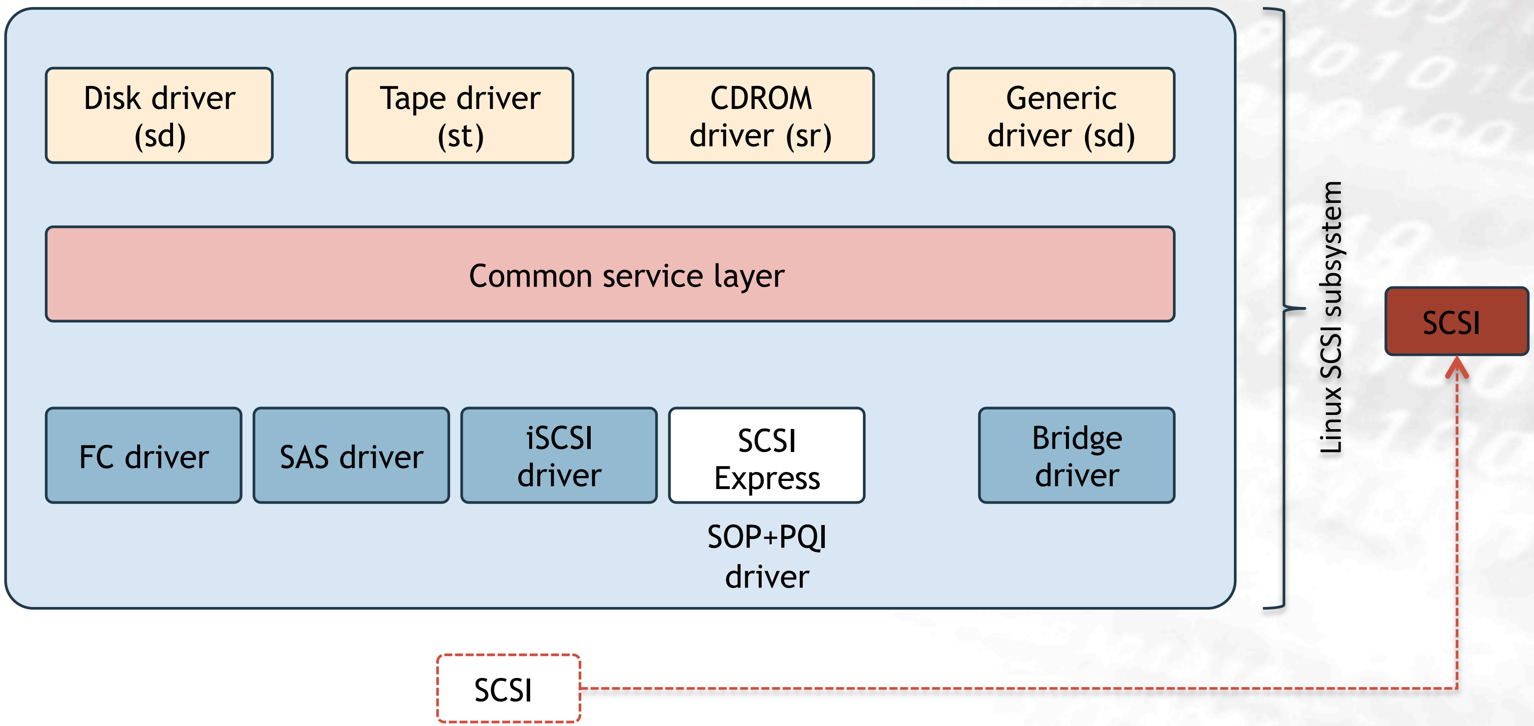
SCSI EXPRESS BENEFITS

- Embrace PCIe
- Fill gap between DRAM and HDD
- Embrace SCSI
- Work together on standards
- Ensure a quality ecosystem

LINUX KERNEL WITH SCSI SUBSYSTEM



LINUX SCSI SUBSYSTEM



Don’t forget software

FLASH: NEW MEDIA OR NEW ARCHITECTURE

FUSiON-iO

Is GPS technology a new map or new architecture?

CONVENTIONAL ARCHITECTURE

FUSiON-iO

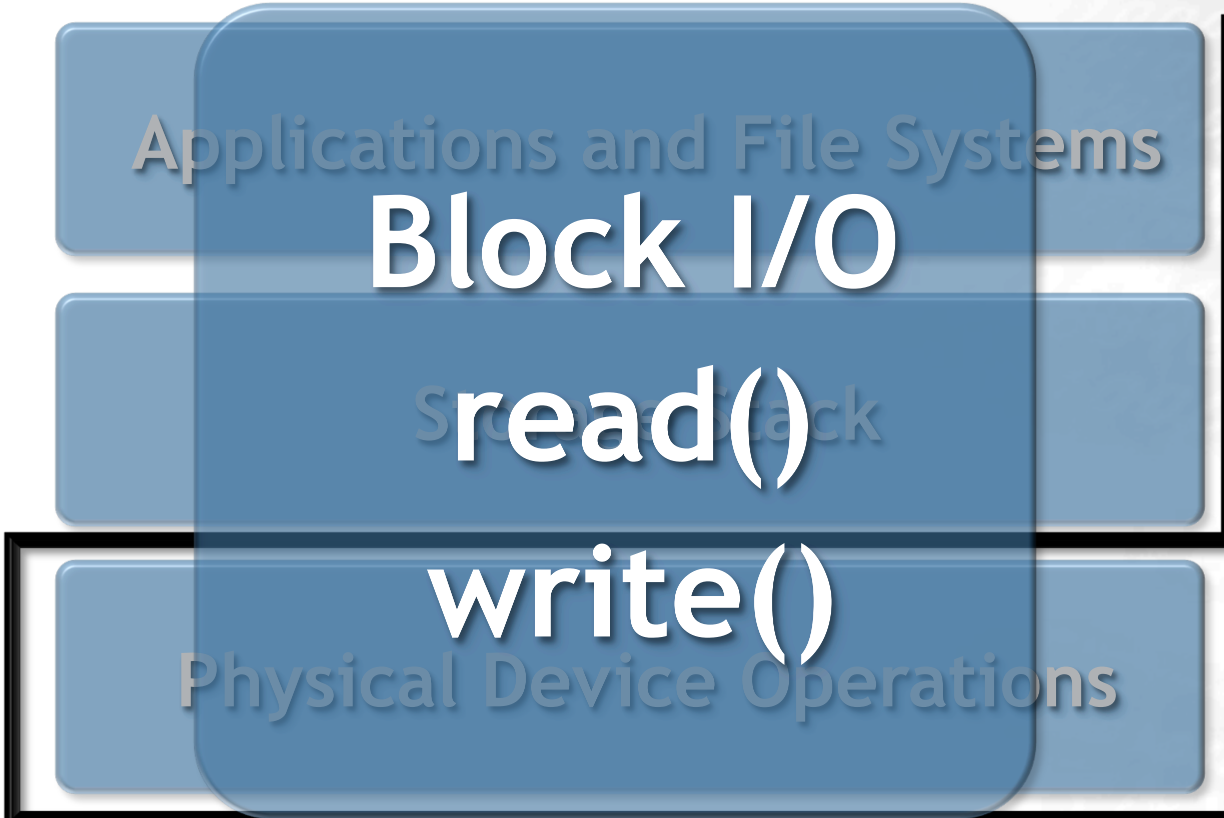
Applications and File Systems

Storage Stack

Physical Device Operations

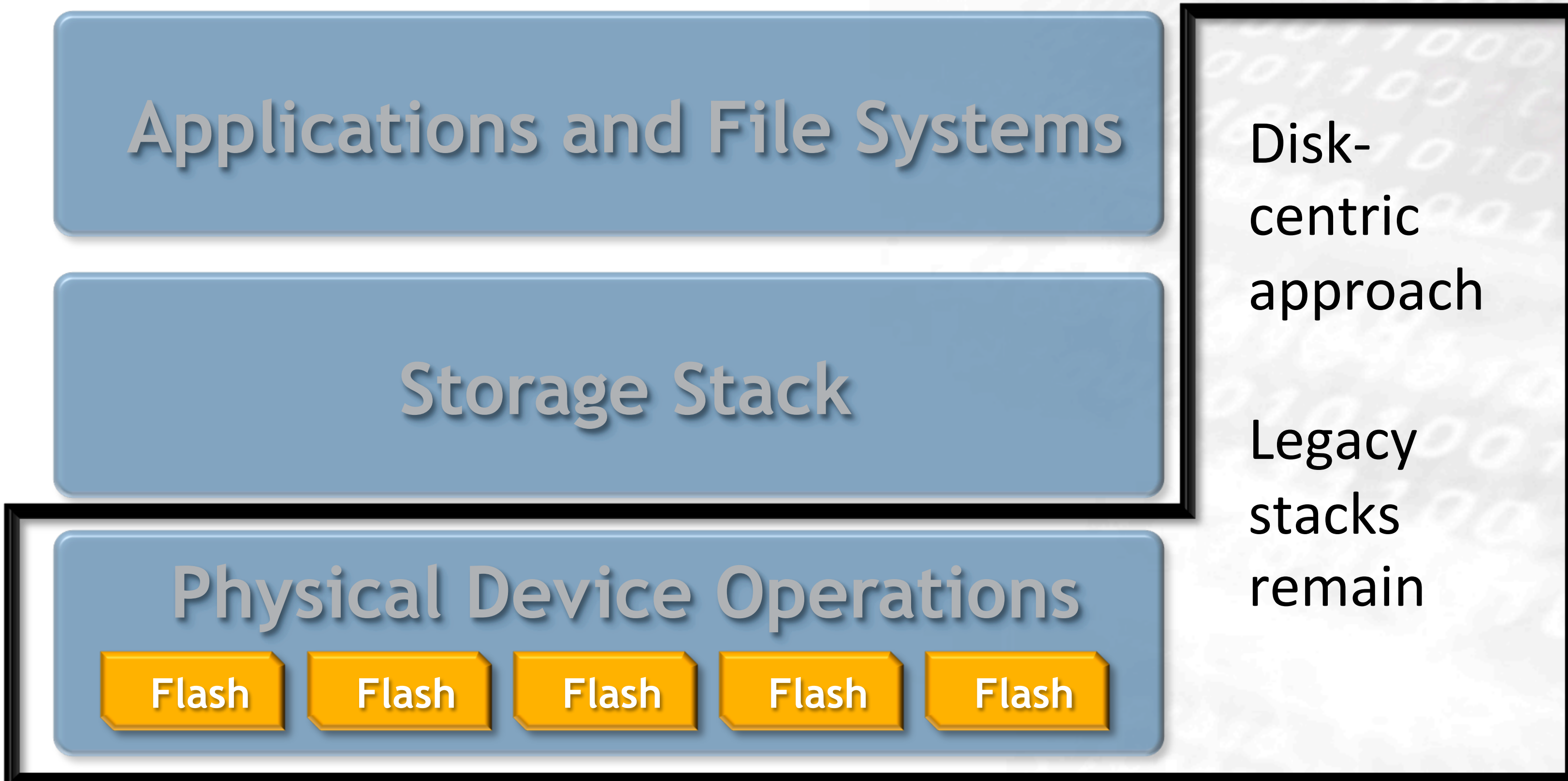
STUCK ON DRIVES AND BLOCK I/O

FUSION-io



Entirety of software and physical stacks optimized for rotating disks

FLASH ONLY AS A FAST DISK

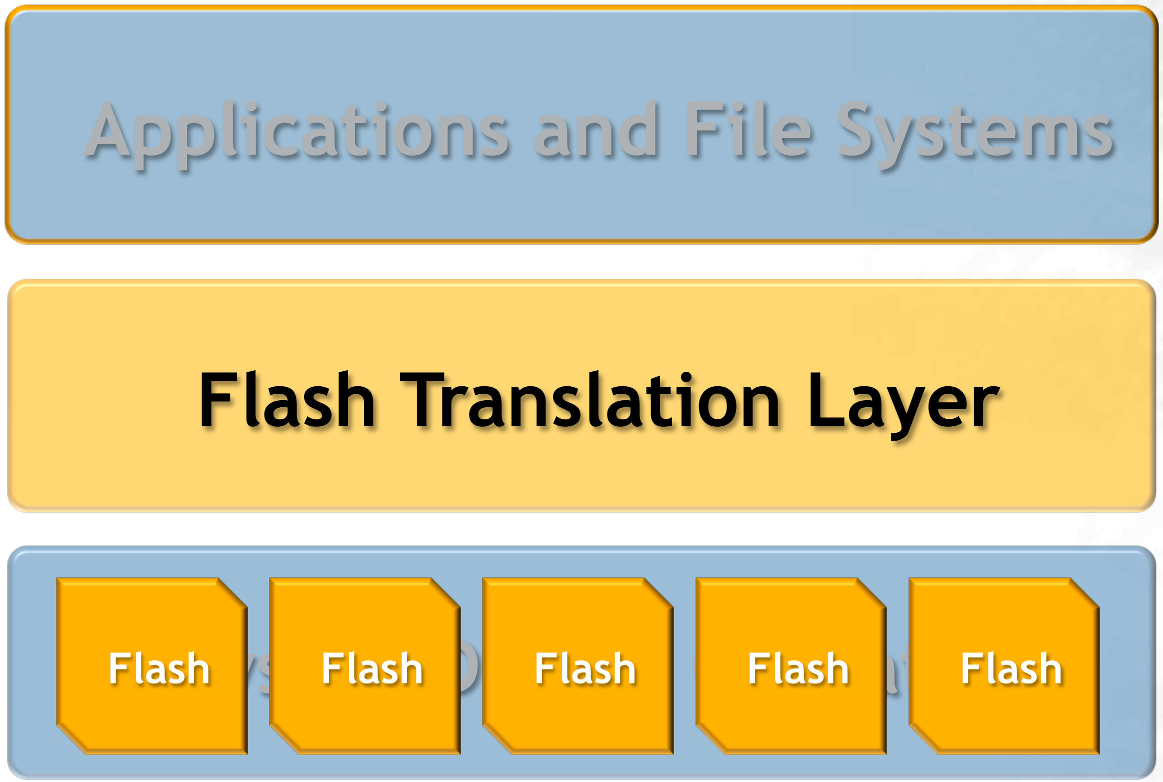


BUT FLASH IS DIFFERENT

- Asymmetric read/write latencies
- Write-impact on durability
- Unique erase characteristics

FLASH AS A NEW ARCHITECTURE

FUSION-io



Flash-centric approach

Retain backwards compatibility with conventional block I/O

FLASH TRANSLATION LAYER 101

Input

Logical Block Address (LBA)

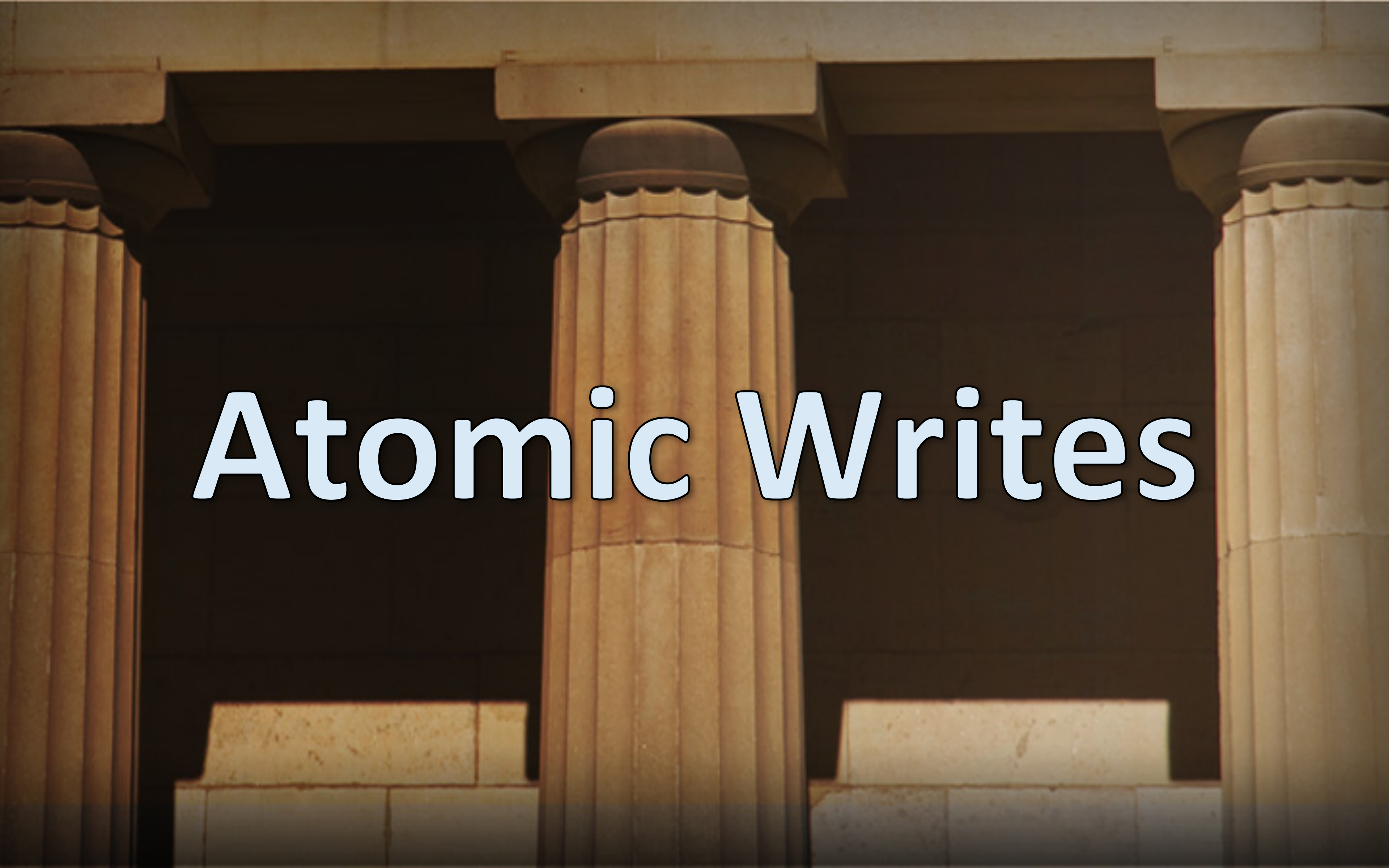
Flash Translation Layer

Output

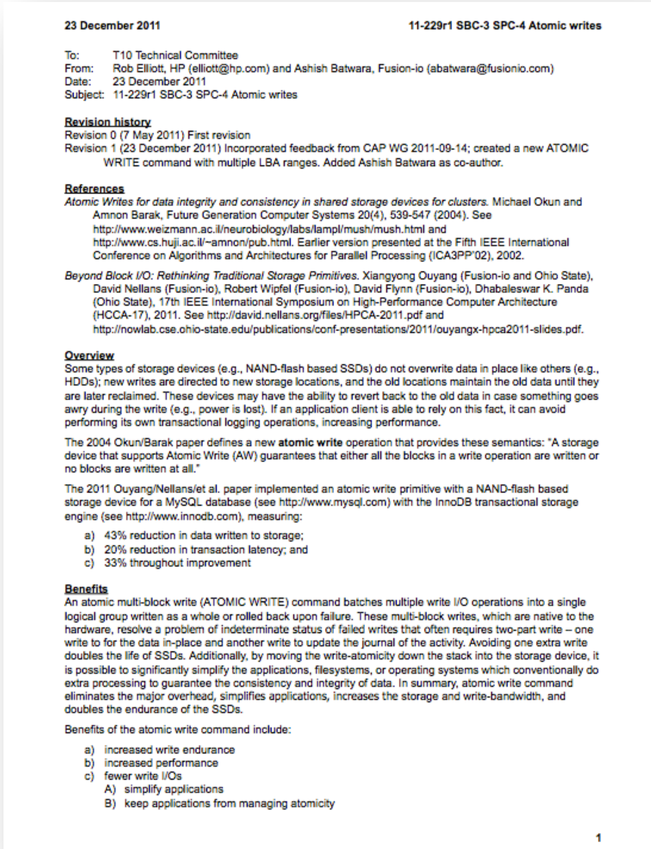
Commands to Physical NAND flash

OPTIMIZING SOFTWARE STACKS FOR FLASH

- Virtualize the storage layer
- Retain compatibility with conventional block I/O
- Deliver new flash-native capabilities


Atomic Writes

ATOMIC WRITES



23 December 2011 11-229r1 SBC-3 SPC-4 Atomic writes

To: T10 Technical Committee
From: Rob Elliott, HP (elliott@hp.com) and Ashish Batwara, Fusion-io (abatwara@fusionio.com)
Date: 23 December 2011
Subject: 11-229r1 SBC-3 SPC-4 Atomic writes

Revision history
Revision 0 (7 May 2011) First revision
Revision 1 (23 December 2011) Incorporated feedback from CAP WG 2011-09-14; created a new ATOMIC WRITE command with multiple LBA ranges. Added Ashish Batwara as co-author.

References
Atomic Writes for data integrity and consistency in shared storage devices for clusters. Michael Okun and Amnon Barak, Future Generation Computer Systems 20(4), 539-547 (2004). See http://www.weizmann.ac.il/neurobiology/labs/lampl/mush/mush.html and http://www.cs.huji.ac.il/~amnon/pub.html. Earlier version presented at the Fifth IEEE International Conference on Algorithms and Architectures for Parallel Processing (ICA3PP'02), 2002.

Beyond Block I/O: Rethinking Traditional Storage Primitives. Xiangyong Ouyang (Fusion-io and Ohio State), David Nellans (Fusion-io), Robert Wipfel (Fusion-io), David Flynn (Fusion-io), Dhabaleswar K. Panda (Ohio State), 17th IEEE International Symposium on High-Performance Computer Architecture (HCCA-17), 2011. See http://david.nellans.org/files/HPCA-2011.pdf and http://nowlab.cse.ohio-state.edu/publications/conf-presentations/2011/ouyangx-hpca2011-slides.pdf.

Overview
Some types of storage devices (e.g., NAND-flash based SSDs) do not overwrite data in place like others (e.g., HDDs); new writes are directed to new storage locations, and the old locations maintain the old data until they are later reclaimed. These devices may have the ability to revert back to the old data in case something goes awry during the write (e.g., power is lost). If an application client is able to rely on this fact, it can avoid performing its own transactional logging operations, increasing performance.

The 2004 Okun/Barak paper defines a new **atomic write** operation that provides these semantics: "A storage device that supports Atomic Write (AW) guarantees that either all the blocks in a write operation are written or no blocks are written at all."

The 2011 Ouyang/Nellans/et al. paper implemented an atomic write primitive with a NAND-flash based storage device for a MySQL database (see http://www.mysql.com) with the InnoDB transactional storage engine (see http://www.innodb.com), measuring:

a) 43% reduction in data written to storage;
b) 20% reduction in transaction latency; and
c) 33% throughout improvement

Benefits
An atomic multi-block write (ATOMIC WRITE) command batches multiple write I/O operations into a single logical group written as a whole or rolled back upon failure. These multi-block writes, which are native to the hardware, resolve a problem of indeterminate status of failed writes that often requires two-part write – one write to for the data in-place and another write to update the journal of the activity. Avoiding one extra write doubles the life of SSDs. Additionally, by moving the write-atomicity down the stack into the storage device, it is possible to significantly simplify the applications, filesystems, or operating systems which conventionally do extra processing to guarantee the consistency and integrity of data. In summary, atomic write command eliminates the major overhead, simplifies applications, increases the storage and write-bandwidth, and doubles the endurance of the SSDs.

Benefits of the atomic write command include:

a) increased write endurance
b) increased performance
c) fewer write I/Os
   A) simplify applications
   B) keep applications from managing atomicity

1

http://www.t10.org/

Doc:11-229R1

IT IS ABOUT TRANSACTIONS

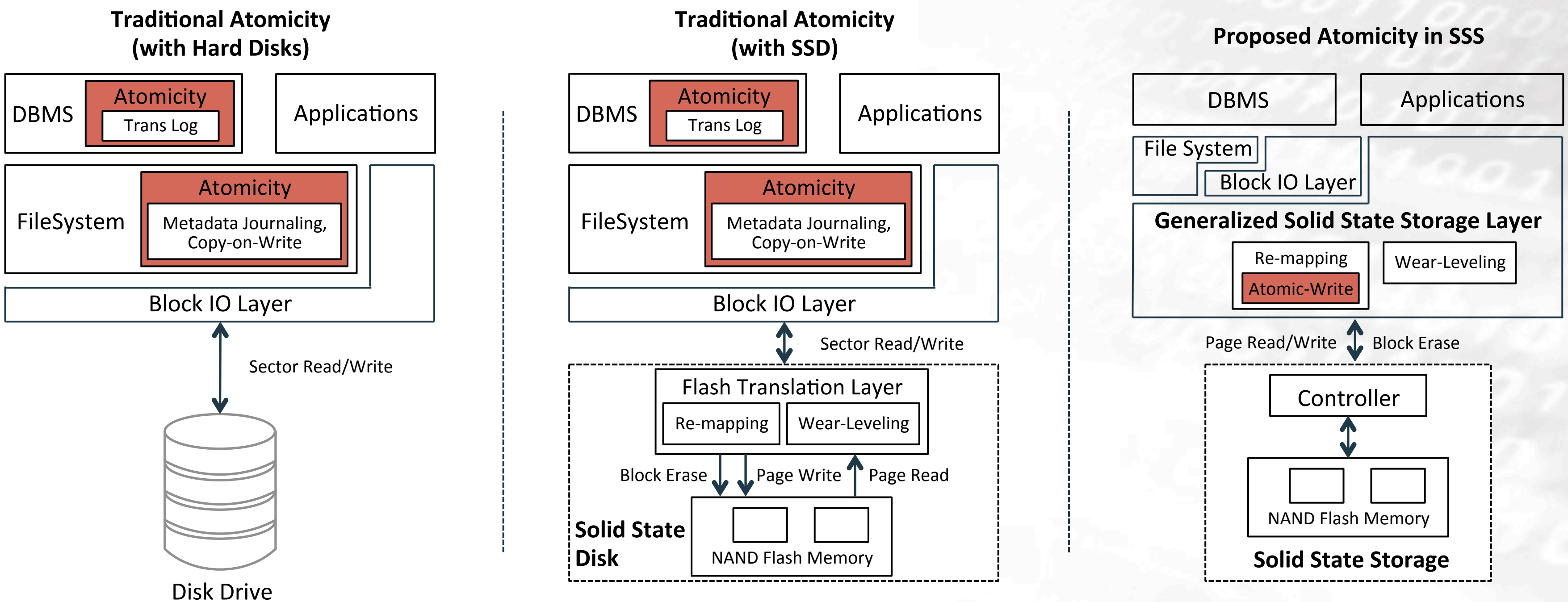
- Building block of applications and databases
- Transactional Semantics
  - Data Integrity
  - Concurrency
  - Crash Recovery

- Applications
- File Systems
- Databases
- Web Services
- Search Engines
- Mission Critical Computing

ATOMIC WRITES

- Batch multiple I/O operations into a single logical group
- Multiple I/Os are persisted as a whole or rolled back upon failure

ATOMIC WRITES – OPTIMIZED



Moving the Atomic-Write Primitive into Storage Stack

OPPORTUNITIES FOR ADVANCEMENT

MySQL Extension for Atomic Writes

35% PERFORMANCE

2X ENDURANCE

½ WRITE WORKLOAD

Gary Orenstein

go@fusionio.com

@garyorenstein

THANK YOU

Agenda

1.	**10:15 AM - 10:30 AM**	**Introduction - SSS Performance**	**Eden Kim, Chair SNIA SSS TWG**
2.	10:30 AM - 10:45 AM	PCIe SSD Form Factor	Mark Meyers, Intel
3.	10:45 AM - 11:00 AM	Standards & Deployment Models	Marty Czekalski, Seagate
4.	11:00 AM - 11:15 AM	SATA-IO & SATA Express - PCIe for Client Storage	Paul Wassenberg, Sata-IO
5.	11:30 AM - 11:45 AM	PCIe 2.5” Form Factor	Janene Ellefson, Micron
6.	11:45 AM - 12:00 PM	Convergence of Memory & Storage IO Architecture	Moon Kim, Tailwind
7.	12:15 PM - 12:30 PM	Lessons from the Front Lines & Lessons for the Future	Gary Orenstien, Fusion-io
8.	12:30 PM - 1:00 PM	Panel Question & Answers / Working Lunch	

PCIe Round Table . .

Questions for the Panel

- Will any one of the competing PCIe interface standards prevail as the Industry Standard and why?
- Is PCIe SSS suitable for both Client and Enterprise Applications?
- How does the higher cost per GB of PCIe Solid State Storage affect adoption?
- Will PCIe SSS become standardized as a Block IO device driver?
- What does one DO with a million IOPS? i.e. limitations of bus, bandwidth, system optimization
- Doesn't the move to virtualization work against the adoption of DAS-oriented PCIe SSD??
- Does PCIe flash make more sense as a memory or as a storage element?

Thank You for your Participation in the PCIe Round Table at the 2012 SNIA Face-to-Face