SNIA Solid State Storage Initiative
PCIe SSD Task Force Meeting No. 6

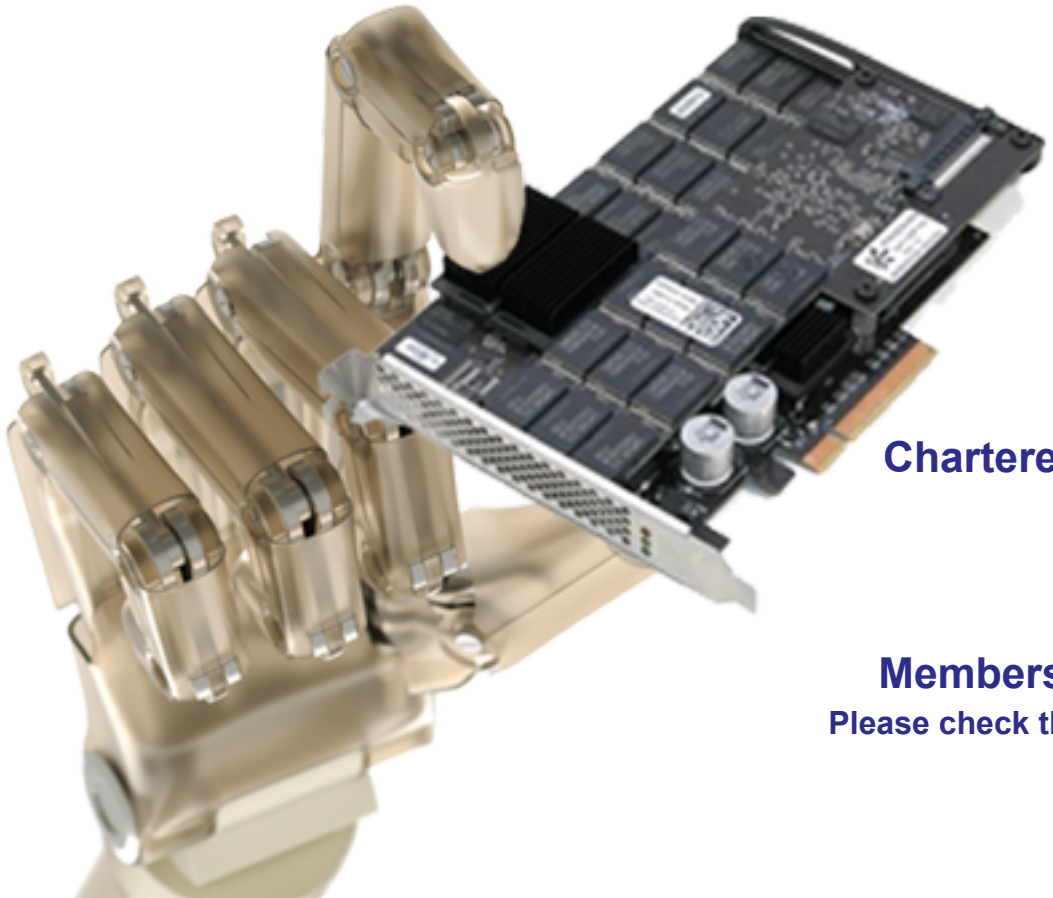
Monday 18 June 2012

SNIA
Solid State Storage Initiative

WELCOME!

SNIA
Advancing storage & information technology

Meeting No. 6
Monday 18JUN2012
4:00 PM - 5:30 PM PST



Welcome to the SNIA
Solid State Storage Initiative

PCIe SSD Task Force

This is an Industry Task Force
Chartered to investigate, discuss & educate
All things PCIe SSD

Membership is Complimentary for (90) Days
Please check the homepage at www.snia.org/forums/sssi/pcie

Task Force Participants

SNIA
Advancing storage & information technology

Concall Guidelines:

- Use your mute button
- ***Sign in webex w/ company name john smith (ABC Co.)***
- Be on time for roll call
- Un Mute when talking
- Use webex chat to ask questions

- Respond to Feedback Requests
- Email String Topic Discussions
- Email comments to reflector pciessd@snia.org
- Send Questions to pciechair@snia.org

(8) OPEN Meetings - Apr - Jul
SSSI Committee Aug - Dec 2012

SNIA
Advancing storage & information technology

Topics	09APR12 Mtg No. 1	23APR12 Mtg No. 2	07MAY12 Mtg No. 3	21MAY12 Mtg No. 4	04JUN12 Mtg No. 5	18JUN12 Mtg No. 6	16JUL12 Mtg No. 7	30JUL12 Mtg No. 8	20JUL12 SSSI FMS Reception
Kick-Off Mtg Issue Identification	X								
Standards		X							
Test Platforms		X							
Performance			X						
System Integration			X						
System Arch Form Factors				X					

Goals: Issue Identification & Committee 2012 Roadmap

Time: 4:00 – 4:05

(8) OPEN Meetings - Apr - Jul
SSSI Mtg & FMS Round Table

SNIA
Advancing storage & information technology

Topics	04JUN12 Mtg No. 5	18JUN12 Mtg No. 6	02JUL12 Holiday	16JUL12 Mtg No. 7	30JUL12 Mtg No. 8	20AUG12 SSSI FMS Reception
Big Picture What's it all Mean	X					
Deployment Strategies Market Development		X				
Where do we go from here?			Holiday	X		
SNIA / SSSI Org, Initiatives & TWGs					X	X
SSSI Task Force F2F FMS Round Table						X SNIA / SSSI FMS Reception

Tentative Topics for Meetings

Time: 4:00 – 4:05

AGENDA – 18 JUNE 12

SNIA
Advancing storage & information technology

I. a. b.	**Administrative** Roll Call; Call Schedule Announcements	4:00 – 4:05 4:05 - 4:10
II. 1. 2. 3.	**Business** **PCIe Marketing: An Analysts' View – Jim Handy, Tom Coughlin** **PCIe Primitives & Persistent Memory – Walt Hubis, Fusion-io** **Developing an Open Kernel NVM Programming Model – Andy Rudoff, Intel**	4:10 - 4:30 4:30 - 4:55 4:55 - 5:25
III. a. b.	**Wrap Up** Discussion Close	5:25 – 5:30

Time: 4:00 – 4:05

Attendance
1 of 4

SNIA
Advancing storage & information technology

Company	No	09APR12	23APR12	07MAY12	21MAY12	04JUN12	18JUN12	16JUL12	30JUL12
Agilent	1	x	x	x	x				
Allion	2	x	x		x	x	x		
AMD	3	x	x						
Apacer	4								
BitSprings	5				x	x	x		
Cadence	6	x							
Calypso	7	x	x	x	x	x	x		
Cisco	8		x	x	x		x		
CLabs	9								
Corsair	10			x					
Coughlin Associates	11	x	x	x			x		
Dell	12	x							
eAsic	13	x			x				
EMC	14	x	x	x	x	x	x		
Enmotus	15	x			x	x			
eTron	16		x						
Fusion-io	17	x	x		x	x	x		
Greenliant	18		x	x	x	x	x		

Time: 4:00 – 4:05

Attendance
2 of 4

SNIA
Advancing storage & information technology

Company	No	09APR12	23APR12	07MAY12	21MAY12	04JUN12	18JUN12	16JUL12	30JUL12
HDS	19	x		x	x				
HP	20	x	x	x	x	x	x		
HGST	21	x	x	x	x	x	x		
Huawei	22	x	x	x	x	x	x		
Hynix (SK Hynix)	23		x	x	x		x		
HyperIO	24	x	x	x		x	x		
IBM	25	x		x					
Intel	26	x	x	x	x	x	x		
Kingston	27								
Lecroy	28		x	x	x		x		
Lenovo	29	x		x		x			
LiteOnIT	30								
Link-A-Media	31								
Lotes	32								
LSI	33				x	x	x		
Lunastar	34	x	x	x	x				
Marvell	35		x		x	x	x		
Micron	36	x	x	x	x	x	x		
Microsoft	37	x	x	x					

Time: 4:00 – 4:05

Attendance
3 of 4

SNIA
Advancing storage & information technology

Company	No	09APR12	23APR12	07MAY12	21MAY12	04JUN12	18JUN12	16JUL12	30JUL12
Molex	38	x	x	x	x	x	x		
Mushkin	39								
NetApp	40				x		x		
Objective Analysis	41	x			x	x	x		
OCZ	42	x							
Oracle	43	x	x	x			x		
PC Perspectives (Allyn M.)	44								
PLX Technology	45			x	x				
Phison	46	x	x	x	x	x	x		
Renesas	47	x	x	x	x	x			
Samsung	48	x	x	x	x	x	x		
Sandisk	49	x	x						
Seagate	50	x			x	x	x		
Semtech Snowbush IP	51								
Smart Storage	52		x	x	x	x			
SNIA	53	x	x	x	x		x		
STEC	54	x	x	x					

Time: 4:00 – 4:05

Attendance
4 of 4 - TOTAL (65) Companies

SNIA
Advancing storage & information technology

Company	No	09APR12	23APR12	07MAY12	21MAY12	04JUN12	18JUN12	16JUL12	30JUL12
Taejin	55	x	x	x					
Tektronix	56								
TMS	57	x							
Toshiba	58		x	x	x		x		
Tyco Electronics	59	x	x						
Unigen	60	x	x	x					
Viking	61								
Violin Memory	62						x		
Virident	63	x	x	x	x				
WDC	64	x	x	x					
Paul Mitchell (Ind)	65								
		39	37	34	33	23	30		

Time: 4:00 – 4:05

Announcements - Eden Kim

SNIA
Advancing storage & information technology

1. Announcements / Other
   a. Minutes Meeting No. 5 - next slide
   b. Task Force Charter / Structure - Standing Slides
   c. Announcements: NO MTG 02JUL12
      1. Meeting No. 7 – Where do we go from Here? – 16JUL12
         a. PCIe Controller Issues – Narinder Lall, eASIC
         b. Tbd – Gilda Foss, NetApp
         c. Tbd – Contact Chair if you want to present

      2. Meeting No. 8 – SNIA / SSSI presentations to Task Force – 30JUL12
         a. SNIA Board – SNIA Organization –
         b. Initiatives – Paul Wassenberg SSSI
         c. SNIA Tech Council – Arnold Jones SNIA, Don Deel (EMC)
         d. TWGs – Tom West IOTTA, Gilda Foss ABDC, Eden Kim SSS

      3. SNIA/SSSI Flash Memory Summit Reception: email to pciechair@snia.org if you plan to attend

         What: Reception : SSSI PCIe Committee - Roadmap 2012
         Why: New Member Introduction; SSSI PCIe Committee Inauguration
         Who: Invitation only : sponsored by SNIA and the SSSI
         When: Monday 20AUG12 5:30 - 7:00 pm
         Where: Santa Clara Convention Center 2d floor

Time: 4:05 – 4:10

Minutes Meeting No. 5
02JUN12

SNIA
Advancing storage & information technology

1. **Attendance:**
   a. **(23) Companies present of (65)**
   b. **Presentations / Speakers - HyperIO, SSSI, Fusion-io**
2. **Administrative:**
   a. **Task Force Charter** - General Survey of PCIe SSD issues; Recommendations for SSSI Committee 2d half 2012
   b. **Task Force Structure -**
      1. **(3) Open Mtgs left:** 18JUN12, 16JUL12, 30JUL12
      2. **Meeting Topics:**
         Meeting No. 6 - Marketing Issues & Deployment Strategies
         Meeting No. 7 - "Where do we go from here?"
         Meeting No. 8 - Presentation of SNIA / SSSI to Task Force
         RECEPTION - SNIA / SSSI FMS Reception 20AUG12 at Santa Clara Convention Center
         Links: Email Reflector: pciessd@snia.org PCIe Task Force Homepage: www.snia.org/forums/sssi/pts
3. **How IOs Traverse the SW/HW Stack & Implications for PCIe – Tom West, HyperIO**
4. **PCIe Standards Building Blocks, How they fit together – Paul Wassenberg, Chair SSSI, Marvell**
5. **PCIe, SCSI Express & a World beyond disk drives – Gary Orenstein, Fusion-io**
6. **Next Actions -**
   a. **RSVP Flash Memory Summit Reception**
   b. **Agenda / Topics - Meeting No. 6**

Close 5:30PM

Time: 4:00 – 4:10

Task Force Charter: GENERAL SURVEY OF PCIE ISSUES

1. Provide Guidance to Marketplace about PCIe SSDs
    1. Educational Materials
    2. Best Practices Documents
    3. Industry Standards Work
2. Coordinate w/ other Industry Organizations
    1. Complement other groups
    2. Avoid Overlap
    3. Fill Voids
3. Open Industry Forum to SSSI Committee
    1. (90) Day Free Trial Membership
    2. SNIA SSSI Membership Required Aug 2012
    3. No IP/NDA - No Confidential Information may be discussed
    4. Identify Issues & Define Roadmap for Committee

Time: 4:05 – 4:20

Structure
Standing Slide

SNIA
Advancing storage & information technology

Task Force Structure:

1. Webex Meetings - Every other Monday
    1. Starting Monday 09APR12 and every two weeks thereafter
    2. 4:00 PM - 5:30 PM PST
    3. (8) Open Calls prior to SNIA/SSSI Membership Requirement

2. Email Reflector - pciessd@snia.org
    1. Agenda, Minutes & Discussion via reflector until 30JUL12
    2. Post Meeting Survey's for feedback and agenda preparation
    3. Email reflector becomes SSSI member only starting 30JUL12

3. Target Objectives for (90) Day Public Forum Period
    1. Table of Standards Groups
    2. Recommendation on PCIe Hardware Test Platform Standard
    3. Identification of PCIe SSD Performance Issues
    4. Hosting of PCIe Round Table Panel
    5. Other Objectives defined by Task Force
    6. Identity Issues & Recommend SSSI PCIe Committee Roadmap for 2012

Time: 4:05 – 4:20

Links
Standing slide

SNIA
Advancing storage & information technology

SSSI

- SSSI homepage — www.snia.org/forums/sssi
- Understanding SSD Performance Project — www.snia.org/forums/sssi/pts
- SSS Performance Test Specification (PTS) — www.snia.org/pts
- PTS Standard Report Format — www.snia.org/forums/sssi/pts
- SSSI Bright Talk Webcasts — www.snia.org/forums/sssi/knowledge/education
- SSSI White Papers — www.snia.org/forums/sssi/knowledge/education

PCIe Task Force

- PCIe SSD Task Force — www.snia.org/forums/sssi/pcie
- PCIe SSD Task Force reflector — pciessd@snia.org
- PCIe SSD Task Force questions — pciechair@snia.org

Time: 4:05 – 4:20

An Analysts' View of PCIe

Time: 4:10 - 4:30

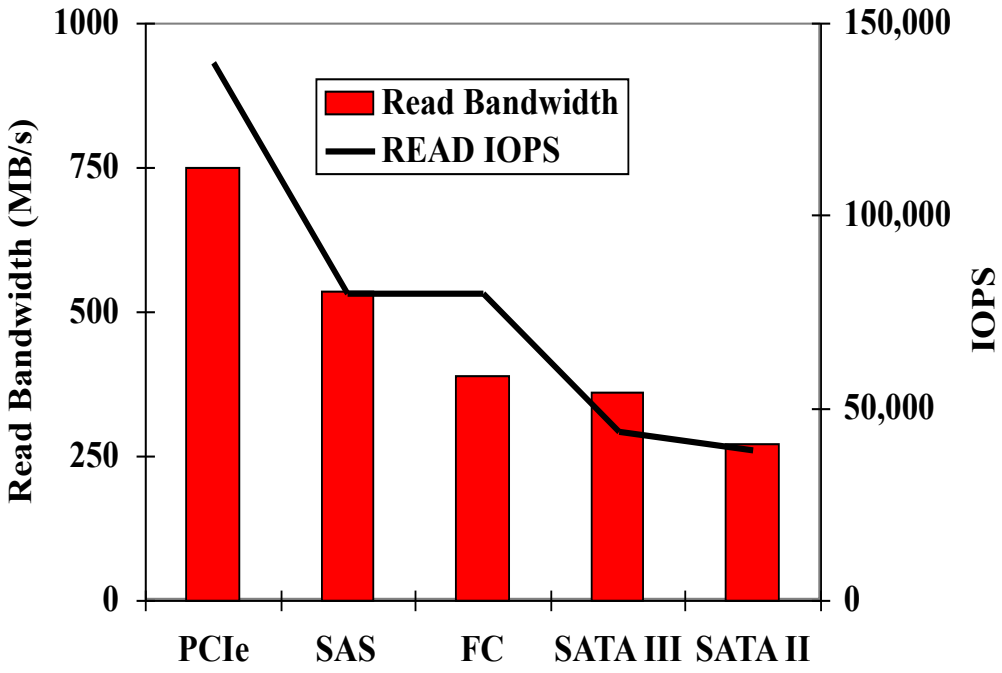
Analysts' View of PCIe

Jim Handy, Objective Analysis
Tom Coughlin, Coughlin Associates

OBJECTIVE ANALYSIS – Semiconductor Market Research

Why PCIe?

- Faster than any HDD interface
- Supports multiple channels
- Fewer layers in the protocol



Makes flash more like memory, less like storage

Time: 4:10 - 4:30

What *is* a PCIe SSD?

- One-Hop
  - PCIe to flash through a single translation
  - Generally a faster solution
- Two-Hop
  - Built of a RAID controller or HBA plus SSDs
  - PCIe to SATA/SAS, SATA/SAS to flash
  - Economical implementation, quick time to market

Time: 4:10 - 4:30

PCIe SSD Markets
(An Educated Guess)

SNIA
Advancing storage & information technology

	One-Hop	Two-Hop
Units	Lower	Higher
Prices	Higher	Lower
Competition	Limited	Widespread

Time: 4:10 - 4:30

5-Minute IOPS Survey

- How many IOPS needed?
- Capacity?
- Latency requirements?
- Max system capability?
- Application?

Answers by order of magnitude: 10, 100, 1,000...

Survey is at http://www.SurveyMonkey.com/s/KGKVR6X

Time: 4:10 - 4:30

Caching Will Prevail

SNIA
Advancing storage & information technology

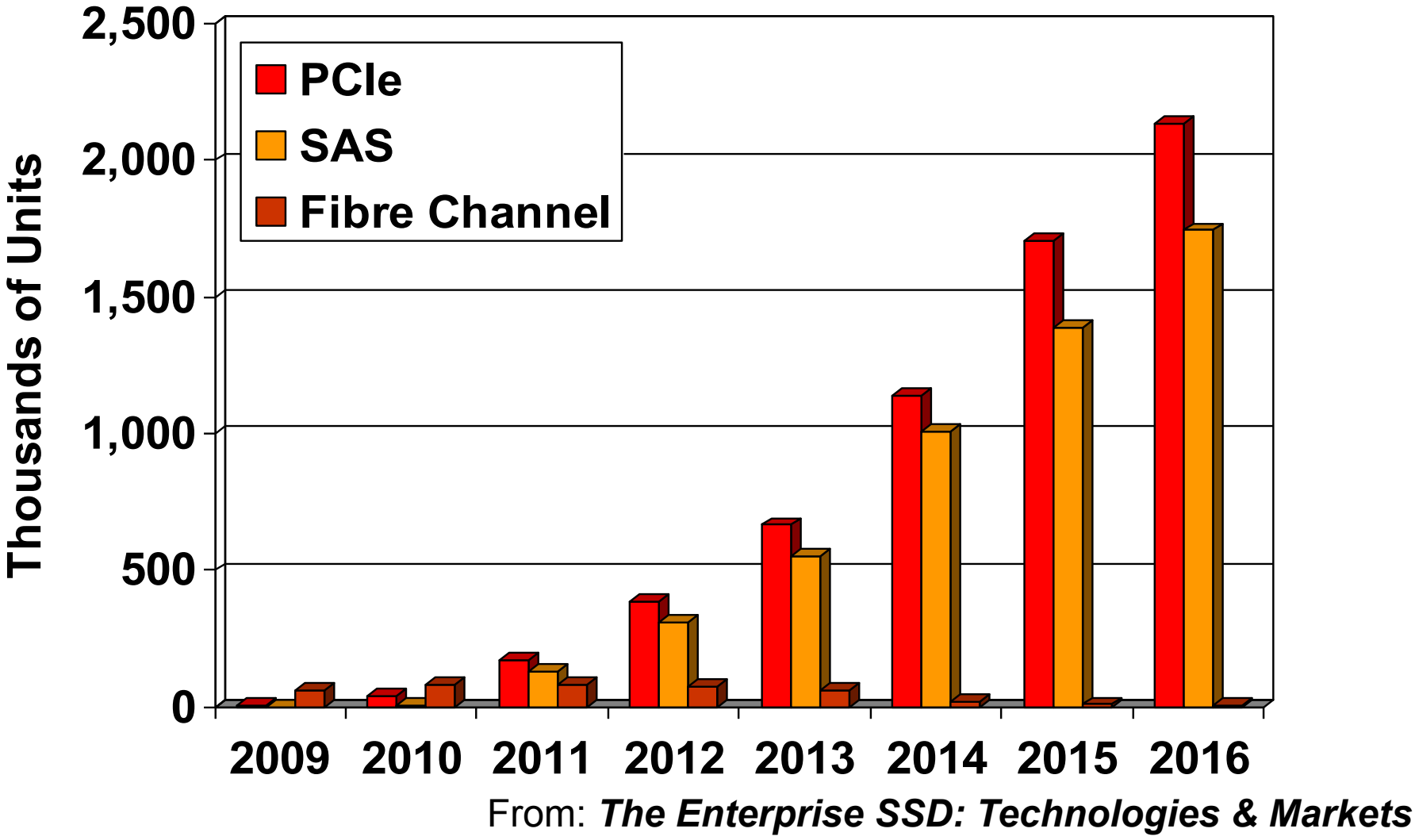
- Manual Data Placement is unsustainable
  - Involves continual hand tuning – labor intensive
  - Inefficient use of flash
- Automatic Data Placement makes sense
  - Autonomous – no user intervention required
    - Based on traffic patterns
  - Uses flash efficiently

Less flash, More speed, No added workload!

Time: 4:10 - 4:30

High-Speed SSD Forecast

SNIA
Advancing storage & information technology



From: ***The Enterprise SSD: Technologies & Markets***

Time: 4:10 - 4:30

Longer-Term Expectations

- SSDs & PCIe will be a stop-gap solution
  - Eventual destination:
    - NAND on the motherboard
    - Hybrid HDDs
- Software will assume a flash layer exists
  - Caching will be incorporated into O/S

All this is 10+ years out!

Time: 4:10 - 4:30

Issues

- One-Hop vs. Two-Hop
- Storage or memory?
  - How to assure coherency?
  - Is PCIe the best interface?
- How long will a cache S/W market exist?
- What happens when flash is superseded?
  - How is this made invisible to the user?

Time: 4:10 - 4:30

Thank You!

Jim Handy

PCIe Marketing: An Analysts' View
Jim Handy, Tom Coughlin

SNIA
Advancing storage & information technology

- **Discussion / Questions & Answers**

Time: 4:10 – 4:30

PCIe Primitives & Persistent Memory

Walt Hubis, Fusion-io

1. Exploiting Native Characteristics of Flash with PCI
    1. Storage Primitives
    2. Standardization
    3. Sparse Address Space
    4. Persistent Memory

Time: 4:30 – 4:55

Exploit the Opportunities of Flash and PCI Storage

- Flash + PCIe provides new storage opportunities
- How to take advantage of native Flash capabilities
  - Native log-append writes
    - incorporates copy-on-write basics
  - Native block mapping and allocation
    - incorporate file system basics
  - Native large virtual address space
    - incorporates sparse semantics

Time: 4:30 – 4:55

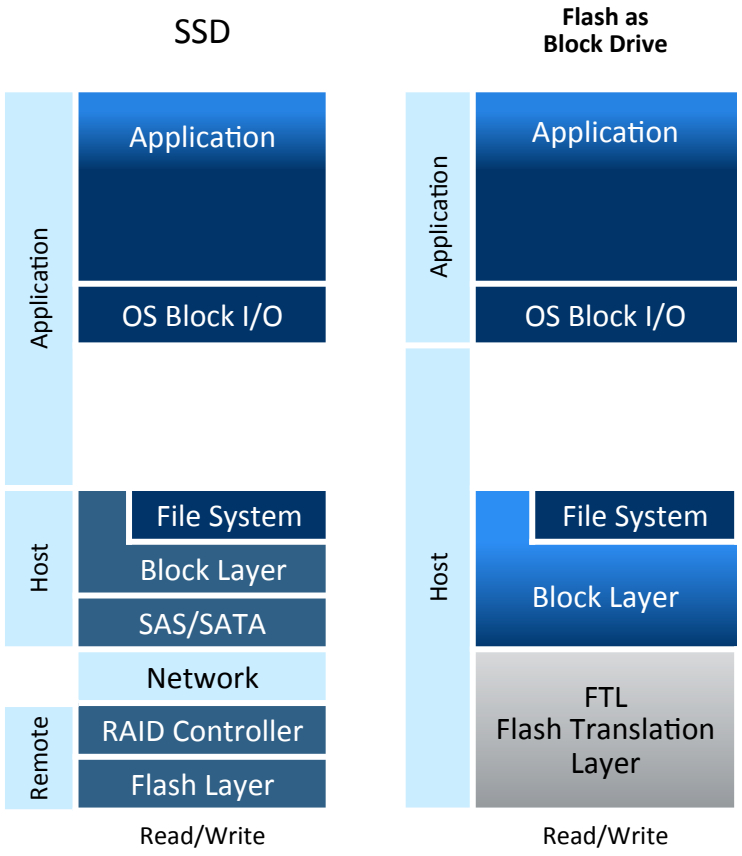
PCIe Flash Storage Overview

SNIA
Advancing storage & information technology

- Background of Storage Architecture
  - How Flash fits in
- Storage Primitives
  - Atomic Writes
  - Sparse Address Space
  - Persistent TRIM and EXISTS
- Memory Access Semantics
- Standardization

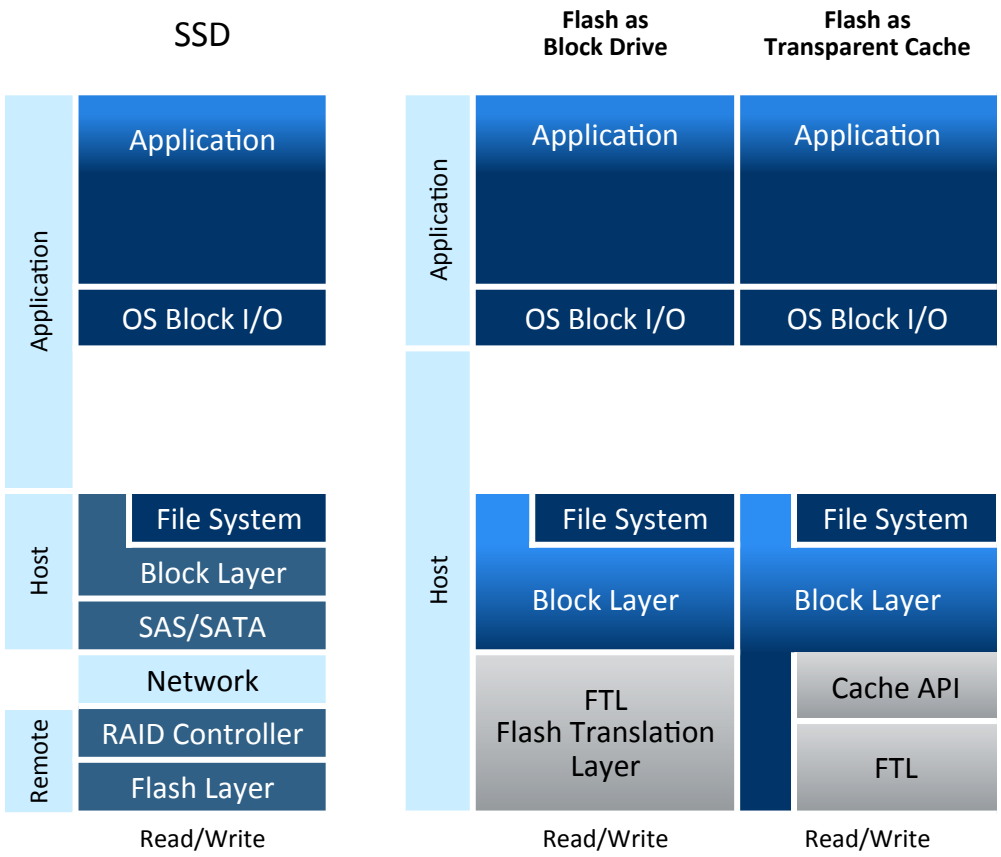
Time: 4:30 – 4:55

Flash Memory Evolution



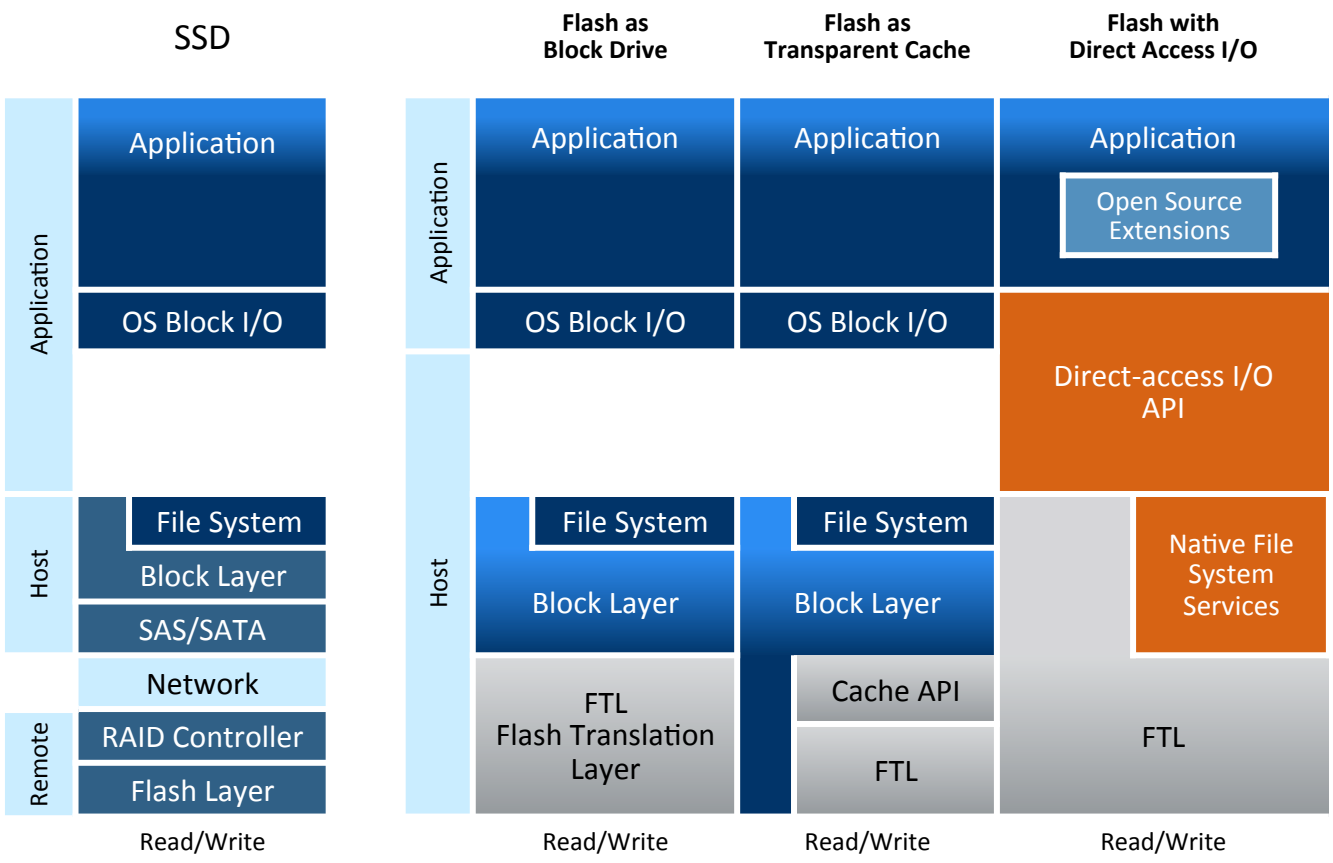
Time: 4:30 – 4:55

Flash Memory Evolution



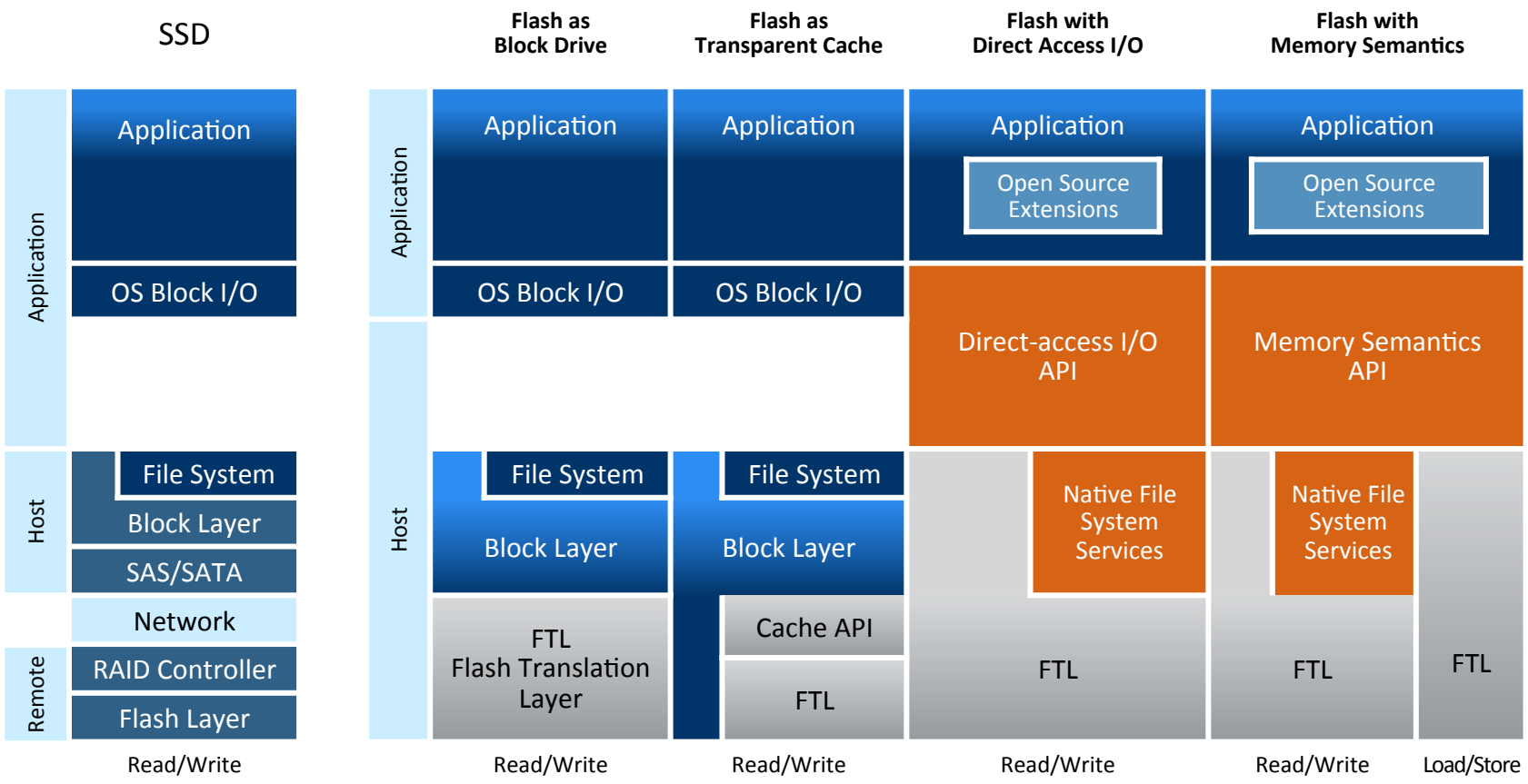
Time: 4:30 – 4:55

Flash Memory Evolution



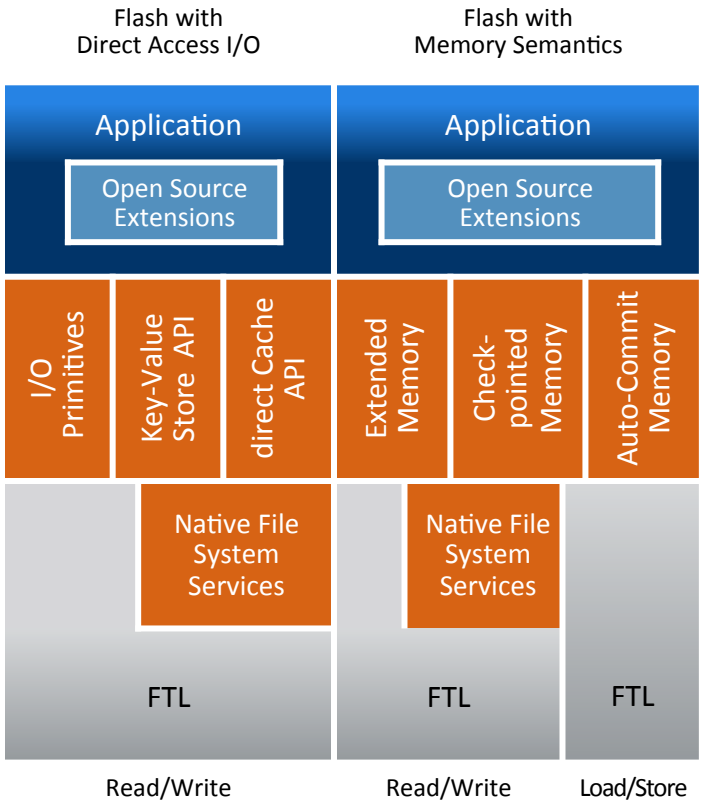
Time: 4:30 – 4:55

Flash Memory Evolution



Time: 4:30 – 4:55

Native API Access



Time: 4:30 – 4:55

Primitives

- How they work
  - Atomic Writes
  - Sparse Address Space
  - Persistent TRIM and EXISTS
- Standardization

Time: 4:30 – 4:55

Primitives

- Used alone or in combination
- Add value to databases, caches, file system and other applications
- Multi-Block Atomic Writes
  - T10 Atomic Write
- Sparse Address Space
  - T10 Mapped, Anchored, and Deallocated LBAs
- Persistent TRIM
  - T10 UNMAP
- EXISTS
  - T10 GET LBA STATUS

ACID Compliant Test	TPS	Data Written
MYSQL	11.7K	24.3GB
MYSQL w/Atomic Write	15.8K	12.15GB

Example performance results: Atomic Writes, October 2011

Time: 4:30 – 4:55

Atomic Write

- Batch multiple I/O operations to multiple independent Flash virtual addresses
- Entire batch written with transactional (ACID) semantics
  - Persisted as a whole or rolled back upon failure
- Application benefits
  - Simplify
  - Less writes

Time: 4:30 – 4:55

Sparse Address Space

- Capacity allocated dynamically on WRITE
  - Similar to thin volume
- LBA address space size far larger than actual capacity
- Capability is extended upward via sparse address space
- High level software usage via primitives
- Support conventional block usages while enabling new usages in cache, file systems, etc.

Write(virtual address)	Allocate on first access
Trim(virtual address)	Hint for block deallocation
Persistent_ trim(virtual address)	Directive for unmapping
Exists(virtual address)	Query state of allocation

Time: 4:30 – 4:55

Trim

- Storage primitive to assist FTLs in garbage collection
- Classic Trim: TRIM(LBA)
  - Hint to FTL to unmap LBA to PBA
  - Improves wear leveling
  - Improves write performance
- Similar to T10 UNMAP command
- Only an advisory command as currently defined
  - A Hint

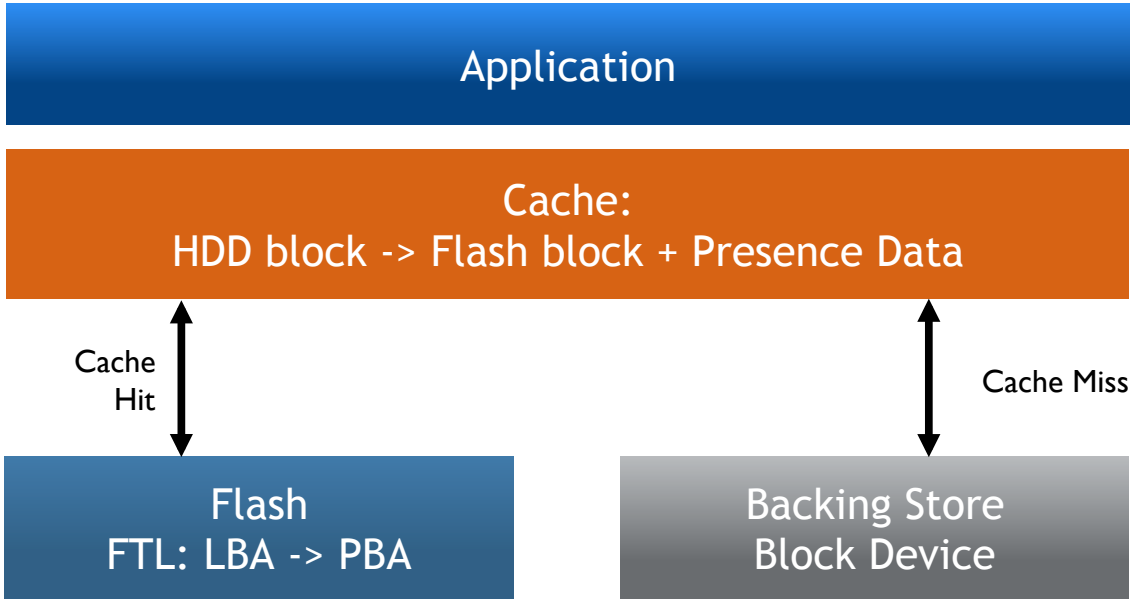
Persistent TRIM and EXISTS

- Persistent Trim PTRIM(LBA)
  - Same properties as TRIM(LBA)
    - Improved wear leveling
    - Improved write performance
  - Well defined with respect to failures
    - Deterministic return of zeros for read
    - Persistent across power failures or errors
- EXISTS (LBA)
  - Query the existence of a particular LBA
  - Enables sparse stores with well defined "presence" semantics

Time: 4:30 – 4:55

Sparse Addressing Example

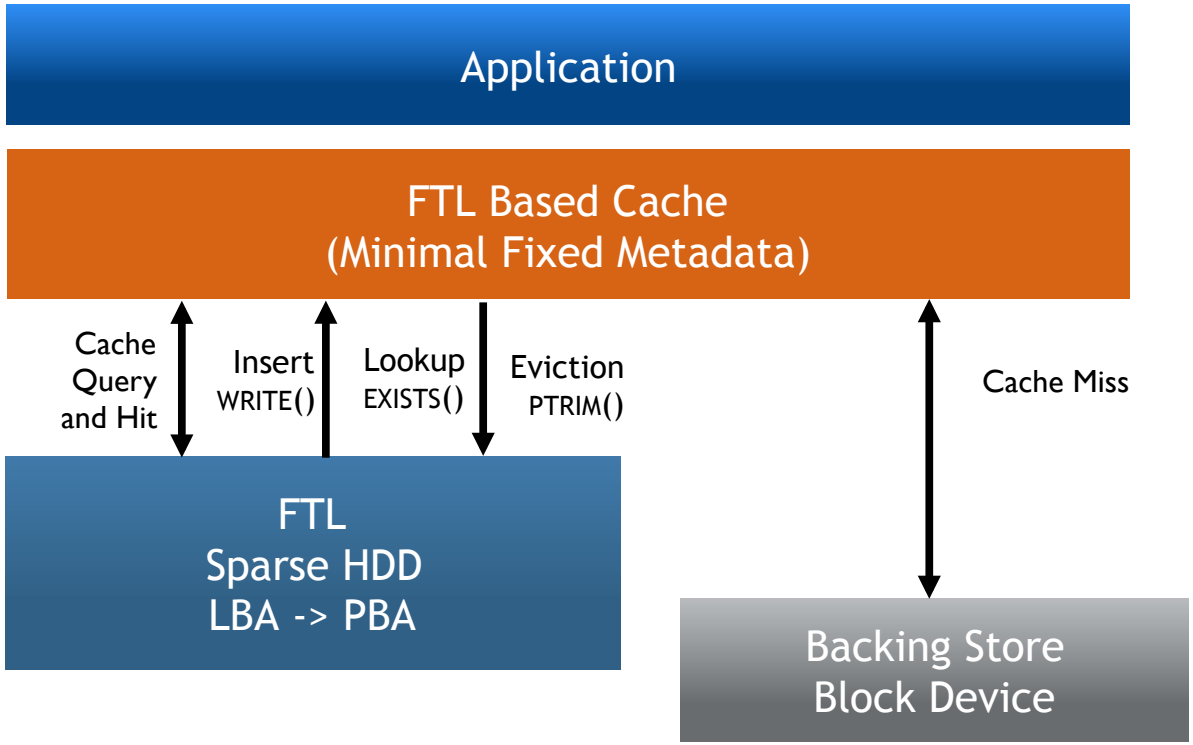
- Conventional Block Cache



Time: 4:30 – 4:55

Sparse Addressing Example

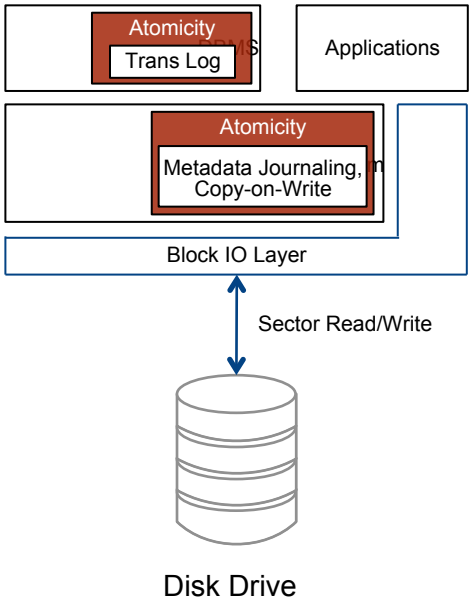
- FTL Based Cache



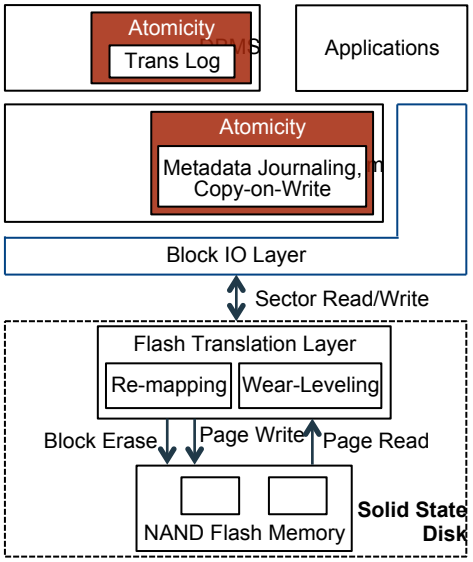
Time: 4:30 – 4:55

Atomic Write

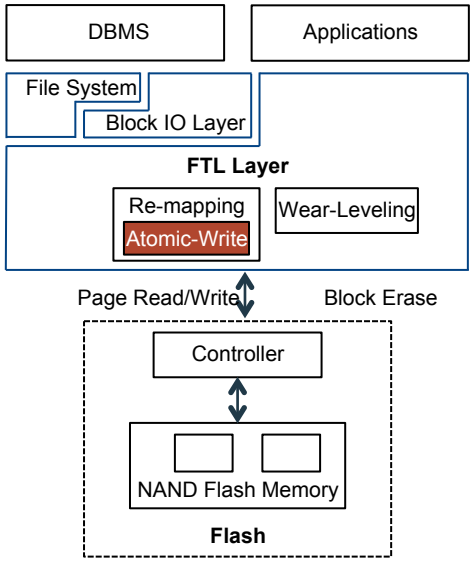
Traditional Atomicity (with Hard Disks)



Traditional Atomicity (with SSD)



Atomicity in Flash



Time: 4:30 – 4:55

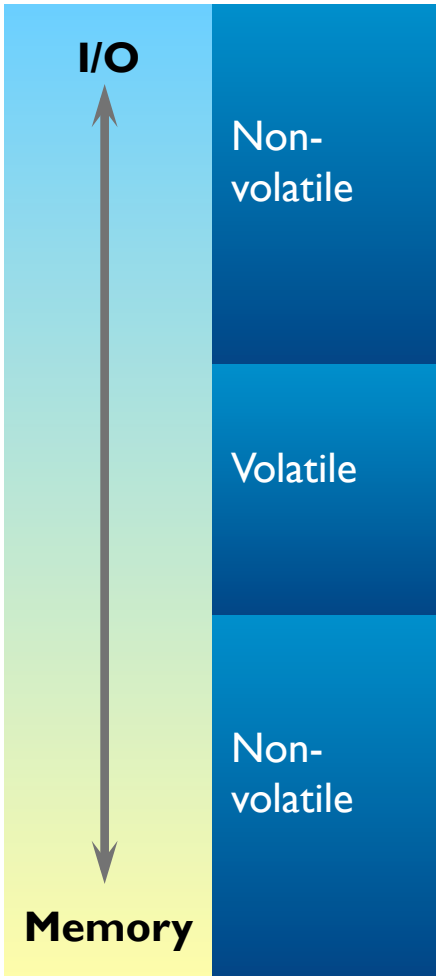
Standardization

- Primitives standards are being proposed in T10
- Atomic Writes
  - Adds atomic write command
  - In T10 CAP committee now
  - SBC-4 SPC-5 Atomic-Writes
    - http://www.t10.org/cgi-bin/ac.pl?t=d&f=11-229r4.pdf
- PTRIM
  - Proposal to add persistence to existing UNMAP command
    - Current UNMAP command is a hint
  - Preliminary proposal is being circulated in T10 now

Standardization

- EXISTS
  - Existing GET LBA STATUS functionality sufficient
  - Proposal to modify command to limit problems with sparse data
  - Preliminary proposal in process
- Scatter/Gather
  - Proposal to allow scattered writes and gathered reads
  - In T10 CAP Committee now
  - SBC-4 SPC-5 Scattered writes, optionally atomic
    - http://www.t10.org/cgi-bin/ac.pl?t=d&f=12-086r2.pdf
  - SBC-4 SPC-5 Gathered reads
    - http://www.t10.org/cgi-bin/ac.pl?t=d&f=12-087r2.pdf

Persistent Memory



I/O-programming semantics
- Open file descriptor (e.g. open())
- write() read() data blocks to file descriptor
- Write multiple data blocks atomically
- kv_put(), kv_get() kev-value pairs to file descriptor

Memory-programming semantics
- Allocate virtual memory (e.g. malloc())
- memcpy()
- dereference pointer to assign/access data to virtual address

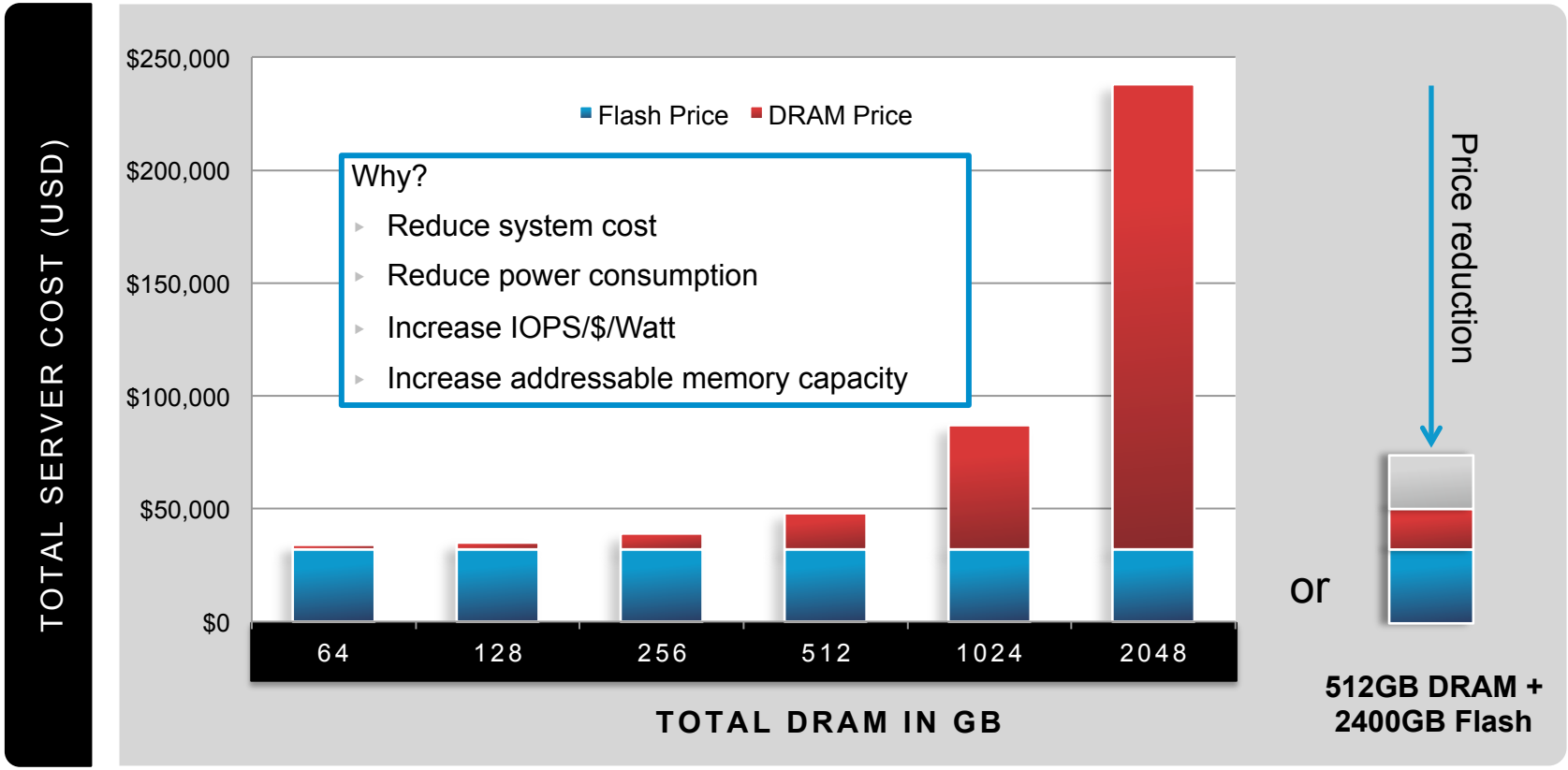
Memory-programming semantics example
- Allocate virtual memory
- memcpy()
- dereference pointer to assign data to virtual address
- issue flush when ready to commit

Time: 4:30 – 4:55

Memory-Access Semantics

- Why 100% in-memory datasets?
  - >10x performance improvement typical
  - Dramatically simplified programming model
    - Faster time-to-market
- Downsides of 100% in-memory datasets
  - Changes to data not persistent
  - DRAM ~10x more expensive than Flash at high densities
  - 100% DRAM uses 10-20x more power than DRAM/Flash systems.
  - Big datasets may not fit in DRAM

Memory Access Semantics Advantages



Time: 4:30 – 4:55

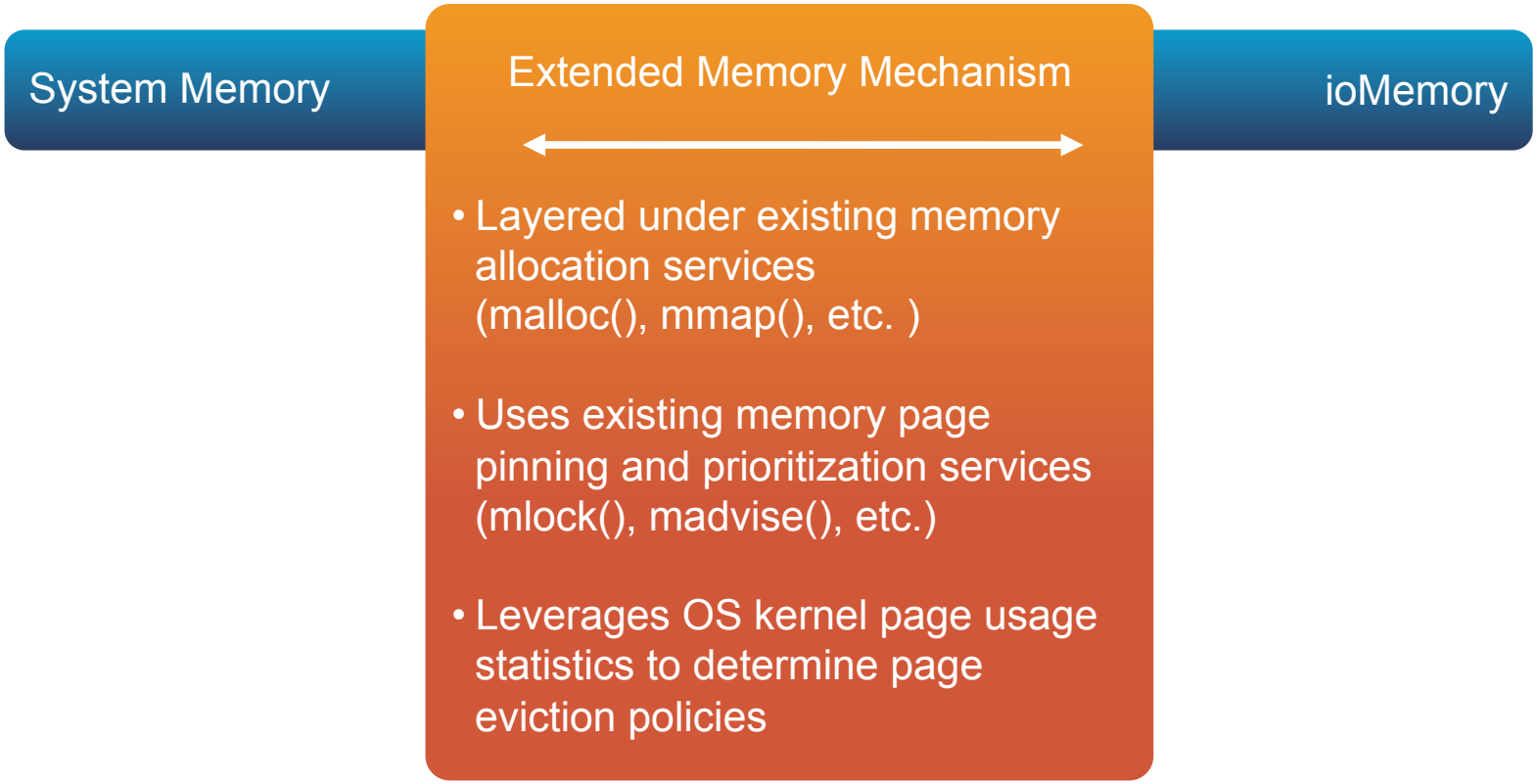
OS Swap vs. Extended Memory

- Originally designed as a last resort to prevent OOM (out-of-memory) failures
- Never tuned for high-performance demand-paging
- Never tuned for multi-threaded apps
- Poor performance, ex. < 30 MB/sec throughput

- No application code changes required
- Designed to migrate hot pages to DRAM and cold pages to Flash
- Tuned to run natively on flash (leverages native characteristics)
- Tuned for multi-threaded apps
- 10-15x throughput improvement over standard OS Swap

Time: 4:30 – 4:55

Extended Memory Overview



Questions

Walt Hubis
Fusion-io Corporation
whubis@fusion-io.com

Time: 4:30 – 4:55

- **Discussion / Questions & Answers**

Time: 4:30 – 4:55

Developing An Open Kernel NVM Programming Model

Andy Rudoff, Intel
andy.rudoff@intel.com

Time: 4:55 - 5:25

Agenda

- Goals – What do we mean by Programming Model?
- Discovery
- Block-Oriented Capabilities
- Memory-Oriented Capabilities
- Next Steps

Time: 4:55 - 5:25

Goals: A Programming Model

- Not an API
  - At least not defined initially by a SNIA TWG
  - OSVs own their kernel APIs
- Components using NVM need:
  - Common ideas they can depend on
  - Evolutionary path
  - Flexibility
- Programming Model
  - Published Spec of capabilities
  - Stop short of defining the API

Time: 4:55 - 5:25

Discovery

- Perhaps the most important idea here
- Components can discover HW capabilities
- Examples:
  - Capacity
  - Performance
  - Write-Atomicity
  - New funky feature X
- Must be extensible, for an evolving world

Time: 4:55 - 5:25

Block I/O

- Traditional block stacks haven't changed much in decades
- A recent example is TRIM
  - Block stack had to be modified to allow this through
  - Typically a way to detect the capability was added too
- Examples:
  - Submitting "fused" block commands
  - I/O barriers

Time: 4:55 - 5:25

Memory-Mapped I/O

- Emerging NVM technologies enable this model
- Is there a "malloc()" for Persistent Memory?
- How is it named, managed?
- What's the permission model?
- Issues start to line up with the file I/O model
- Examples:
  - Open a blob of NVM by name
  - Map a blob of NVM
  - Sync a blob of NVM

Next Steps

- NVM Programming Technical Work Group
- Approved by SNIA TC June 12, 2012
- From the charter: members only site nvmptwg
  - The NVM Programming TWG is created for the purpose of accelerating availability of software enabling NVM (Non-Volatile Memory) hardware. The TWG creates specifications which provide guidance to operating system, device driver, and application developers. These specifications are vendor agnostic and support all the NVM technologies of member companies.
- Open to all SNIA members
- Calls start week of July 9th – tentatively 10JUL 1 PM PST
- For questions: nvmptwg-chair@snia.org

Time: 4:55 - 5:25

Developing an Open Kernel NVM Programming Model – Andy Rudoff, Intel

- **Discussion / Questions & Answers**

Time: 4:55 - 5:25

Open Discussion
Next Actions

SNIA
Advancing storage & information technology

Open Discussion:

Next Actions:

- **Email response to reflector:** ***Would you attend a private reception at FMS if invited?***
- **Agenda & Topics**
  - **Meeting No. 7 – Where do we go from Here? 18JUL12**
    a. **Narinder Lall, eASIC - PCIe Controller Issues**
    b. **Gilda Foss, NetAPP- tbd**
    c. **Tbd –** ***contact chair if you are interested to present***
  - **Meeting No. 8 – SNIA / SSSI Presentation to Task Force 30JUL12**
    a. **SNIA Organization**
    b. **Solid State Storage Initiative**
    c. **SNIA Technical Council**
    d. **SNIA Technical Working Groups**
- **Email poll for Mtg No. 8 – comments / questions on Task Force Presentations**

Time: 5:25 - 5:30

Supplemental Slides