

Enterprise NVMHCI

Enabling Enterprise Class PCIe SSDs with Unmatched Performance

Amber Huffman
Principal Engineer
Intel Corporation



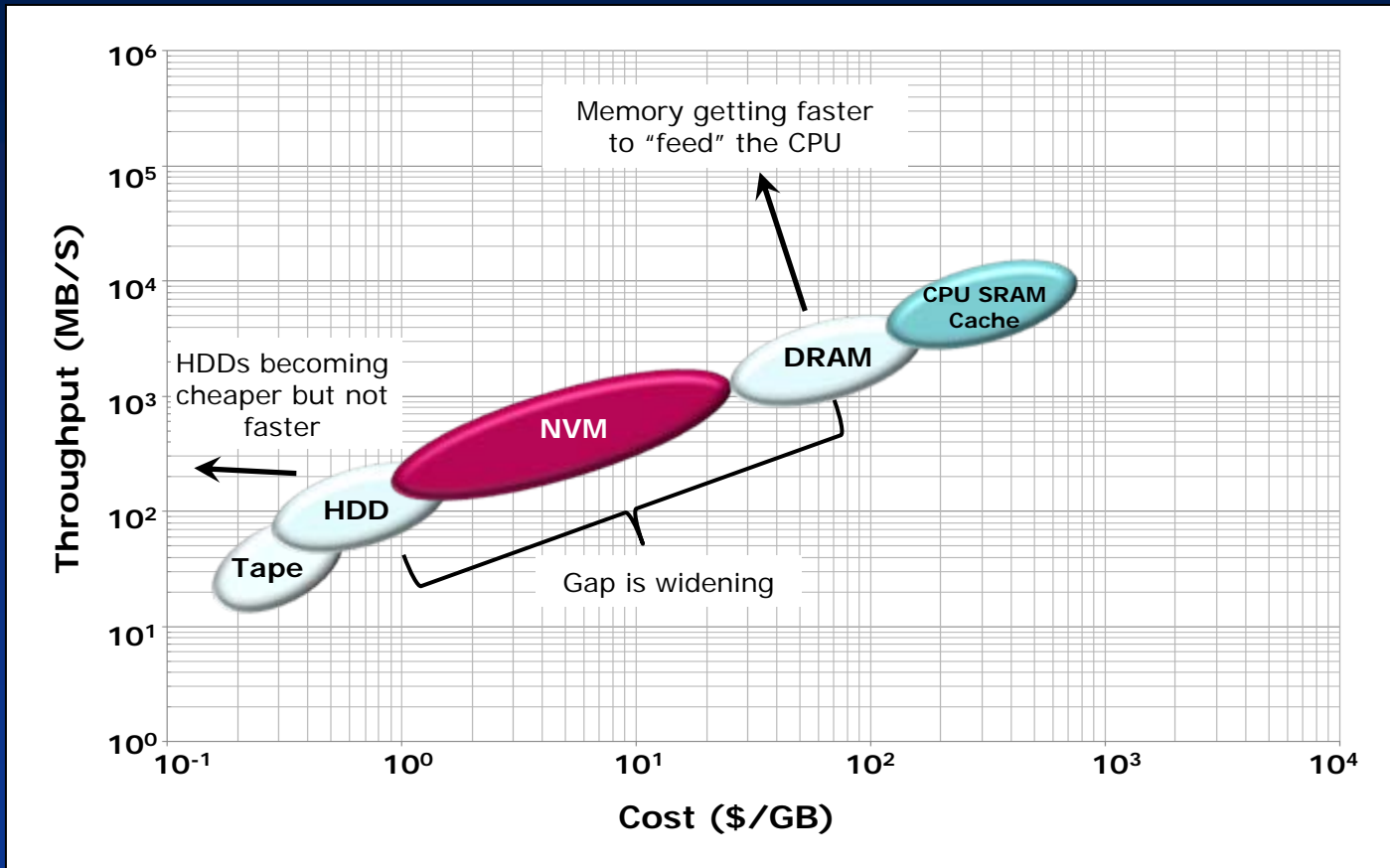
Peter Onufryk
Director of Engineering
IDT



Agenda

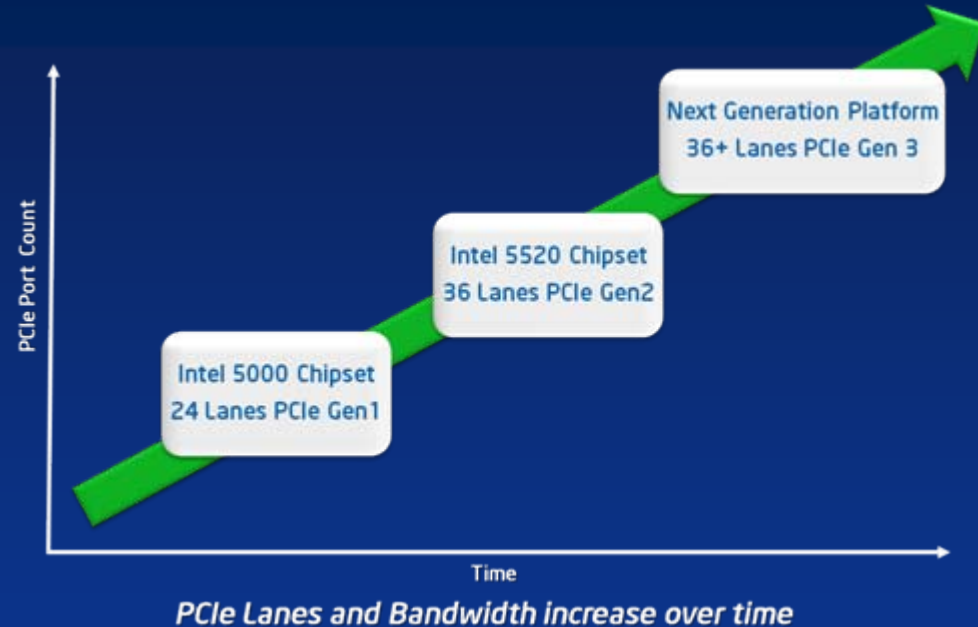
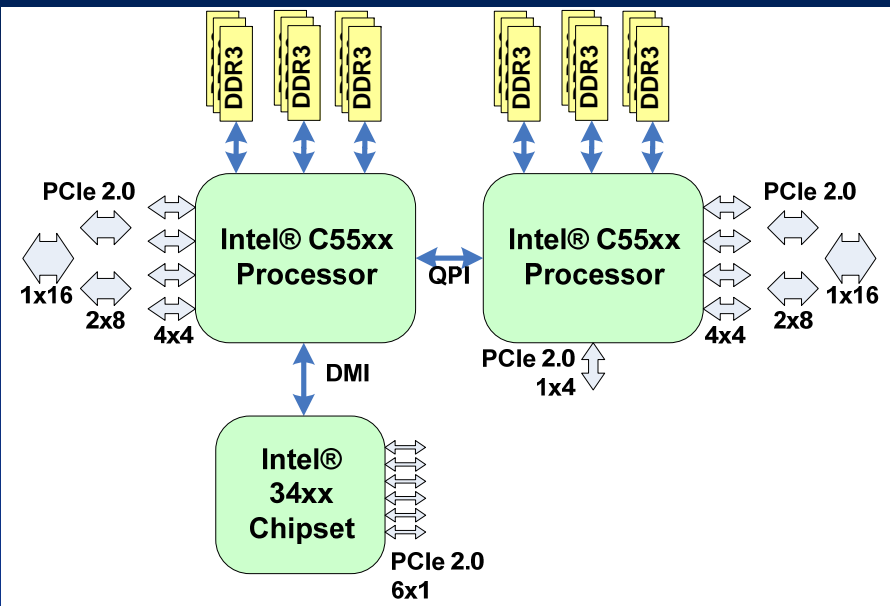
- PCIe SSD Opportunity & Value Proposition
- Why Enterprise NVMHCI
- Interface Attributes
- Queue Mechanism & Command Issue/Completion Path
- Commands & Arbitration
- Out of Order Data Delivery
- Firmware Update
- Security
- End-to-end Data Protection
- Summary

Gap in the Storage/Memory Hierarchy is Growing



NVM is filling the price/performance gap between DRAM and HDD, thereby creating the "I/O Memory Tier"

Platform PCIe Connectivity Continues to Rise



Platform native PCIe connectivity continues to rise.
Enables PCIe SSDs to effectively fill the gap in I/O Memory Tier.

PCIe SSD Value Proposition

- The market is delivering PCIe SSDs to deliver unmatched performance
 - Plentiful PCIe lanes 36+ lanes
 - Stunning performance opportunity > 3 GB/s (PCIe Gen2 x8)
 - With PCIe scalability > 6 GB/s (PCIe Gen3 x8)
 - Lower latency µsec matter
 - Lower cost with direct attach Eliminate HBA cost

OCZ Launches 4th Generation PCIe SSDs

6:00 PM - April 6, 2010 by Kevin Parrish - source: Tom's Hardware US

OCZ's new Z-Drive PCIe-based SSDs feature removable NAND modules.

[OCZ Technology](#) Group announced today its move into mass production with the fourth generation of PCIe-based solid state drives, the new Z-Drive R2 SSD series. This will actually be the second rendition of the original Z-Drive drives, adding "greater performance and design flexibility" thanks to optimized, interchangeable NAND modules--this will allow for



Seagate teams with LSI to enter PCIe-based SSD game

By Darren Murph posted Jan 26th 2010 1:26PM

Seagate didn't bother serving up a gaggle of new wares at CES this year, but judging by its release shot out today, it's hoping to make a serious splash in the SSD market a bit later on. Thanks to collaboration from LSI, the outfit is expected to deliver its own line of PCI Express-based solid state storage solutions. We're guessing these devices will be similar in scope to the PCIe SSDs already outed by [Fusion-io](#) and [OCZ Technol](#)



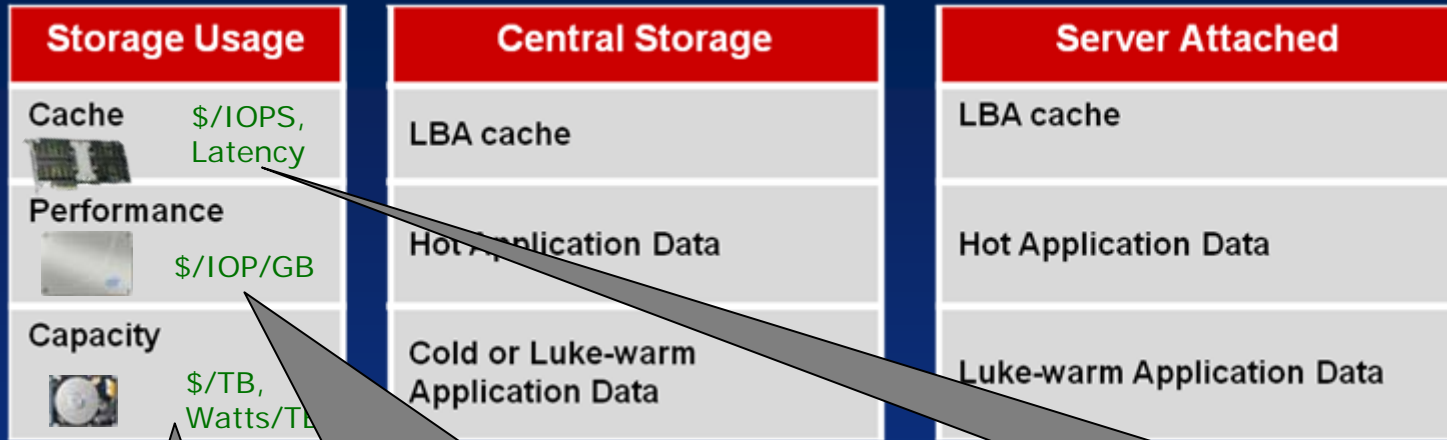
Fusion-io ioMemory VSL Treats Flash Storage As A "New Memory Tier"

Wednesday, July 21, 2010 - by Ray Willington

It's difficult to say if or when [Fusion-io](#)'s newest... average consumer, but as with many things in the technology field, what starts at the highest levels of enterprise eventually filters down to the consumer once kinks have been worked out, prices have adjusted downward and more partner companies have had time to adopt the new process. We are guessing that's exactly the path that ioMemory will take, which is a new flash-optimized OS subsystem.



Enterprise Storage Tiers



SSDs of many flavors replace HDDs for high perf. storage for some apps (e.g. financial, DB)

HDDs best for "data at rest"
- Consistent \$s/GB champ

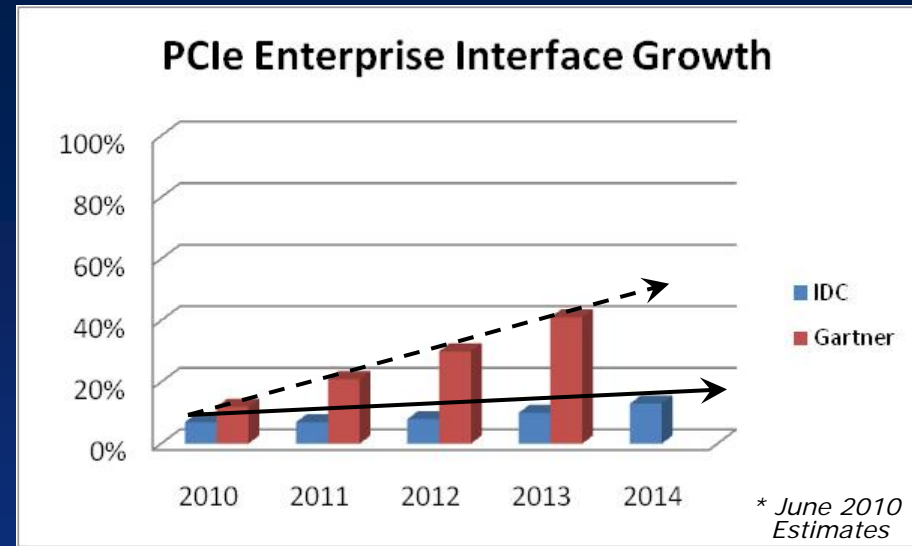
SSD performance enables new storage caching "IOPS tier"

- Many companies testing usage
- Application managed caching

PCIe SSD attributes of high IOPs, high bandwidth, low latency, and lower cost are a great match to the emerging Cache Tier.

Enabling Faster Adoption

- Analysts see a great opportunity for PCIe SSDs in Enterprise
 - Forecasts are from 10% to 40% of Enterprise segment in 2013
- A standard driver and consistent feature set will help place growth for PCIe SSDs on the faster curve
- Adoption inhibitors: Different implementation & unique drivers
 - Today, SSD vendors provide drivers for each OS that OEMs must validate
 - Today, SSDs implement different subsets of features in different ways
- To enable faster adoption and interoperability of PCIe SSDs, industry leaders are defining the Enterprise NVMHCI standard
 - Standard register programming interface & feature set definition
 - Enables standard drivers to be written for each OS
 - Enables interoperability between implementations shortening validation cycles



Companies Driving Enterprise NVMHCI Specification



The NVMHCI Workgroup includes 55+ members, focused on delivering streamlined NVM solutions.

The Value of Enterprise NVMHCI

Microsoft

“A standardized interface functions as a foundation, enabling a volume market for technology innovation while avoiding the compatibility issues that arise from multiple, proprietary interfaces. Enterprise customers are requesting standard interfaces be used on non-volatile-memory products as an enabler to assist broad adoption.”

*Steve Olsson
Lead Program Manager, Storage and File Systems
Microsoft*

The Value of Enterprise NVMHCI

Microsoft

"A standardized interface functions as a foundation, enabling a volume of compatible components to be developed and tested against a single interface, enabling..."

FUJITSU

"The lack of a standard register level interface presents numerous problems when integrating PCIe SSDs into our products, including longer qualification times and functionality that is not uniformly implemented across vendors. Fujitsu Technology Solutions sees Enterprise NVMHCI as an important part of enabling broad adoption in PCIe SSDs emerging in the Enterprise space by resolving these concerns. Joining the working group was a natural choice to foster this industry leading standardization effort."

*Jens-Peter Seick
Senior Vice President x86 Server Product Unit
Fujitsu Technology Solutions*

The Value of Enterprise NVMHCI

Microsoft

“A standardized interface functions as a foundation, enabling a volume of companies to develop a common interface. This interface enables numerous other functions, such as vendor interoperability, NVMe, and additional space growth, leading to a more efficient and cost-effective storage solution.”

FUJITSU

“The lack of a standard register level interface presents a significant barrier to the adoption of new flash based storage devices. We are working with other industry technology leaders to make Enterprise NVMHCI that interface.”



“New flash based storage devices are pushing the limits of traditional storage interfaces. The industry needs a new standard interface, to allow for multi-vendor innovation and take advantage of evolving flash technology and associated storage and platform architecture changes. We are working with other industry technology leaders to make Enterprise NVMHCI that interface.”

Paul Prince
CTO, Enterprise Product Group
Dell

Enterprise NVMHCI Goals & Timeline

- Goals for standard:
 - Address Enterprise usage scenarios
 - Enable an efficient & scalable interface, from very high-end to client
 - Ensure no interface impediments to exceeding > 1M IOPs
 - Enable OS vendors to deliver standard high performance drivers
 - Provide a consistent feature set to enable SSD interoperability
 - Reduce TTM for PCIe SSDs by enabling OEMs to validate/qual one PCIe SSD driver for each OS and one consistent feature set

- To get involved, join the NVMHCI Workgroup
 - Details at <http://www.intel.com/standards/nvmhci>

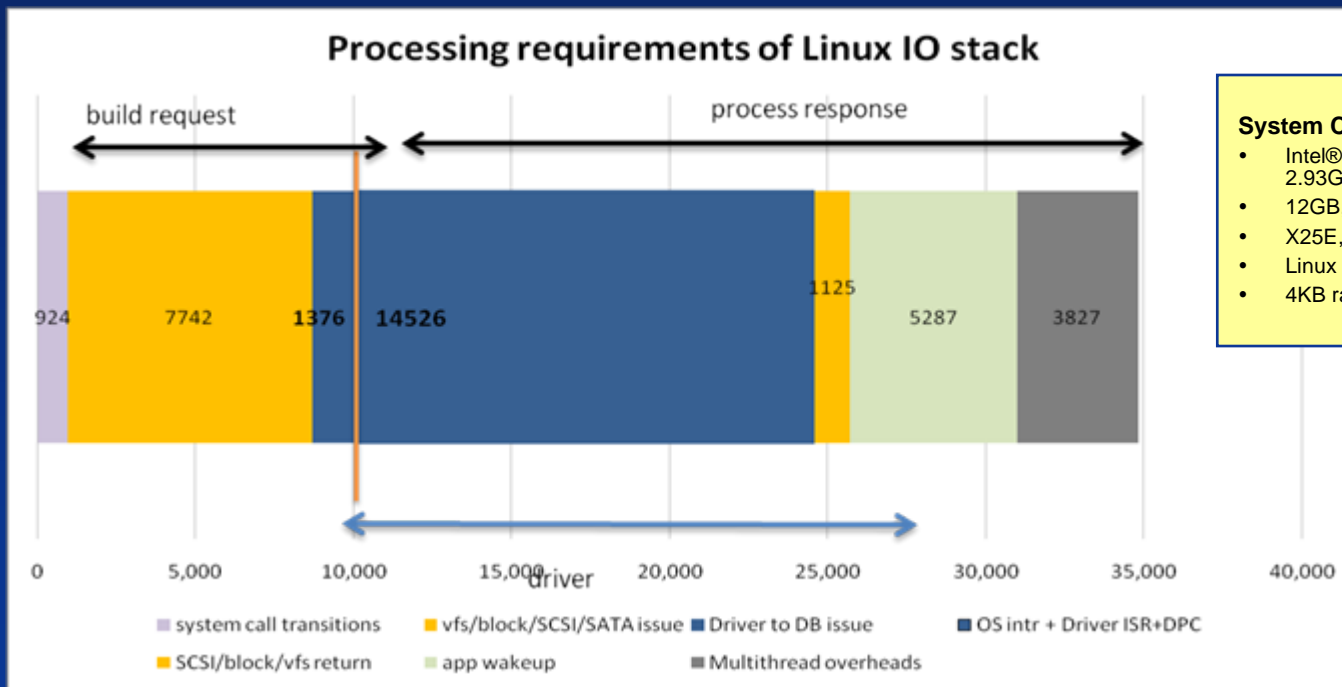
	Apr '10	May '10	Jun '10	Jul '10	Aug '10	Sep '10	Oct '10	Nov '10	Dec '10
Revision	0.5		0.7			0.9		RC	1.0
Definition	0.5: Basic capabilities and approach defined.			0.7: Basic definition complete for all features. Feature freeze.			0.9: Erratum only. RC: Member review. 1.0: Published.		

0.70 revision achieved, available for Contributor review.
Schedule enables product intercept in 2012.


Example Optimization Point

- The Linux* stack using AHCI is ~ 35,000 clocks / IO
- A large impact is uncacheable reads, ~ 2000 clocks each
 - Minimum of 4 uncacheable reads required with AHCI
- Enterprise NVMHCI eliminates uncacheable reads for command issue and completion

Linux AHCI



Enterprise NVMHCI Attributes

- 
- Remove uncacheable reads from command issue/completion
 - Minimize MMIO writes in command issue/completion path
 - Support deep command queues
 - Simplify command decoding and processing
 - Fixed sized (64B) command format
 - Avoid “pointer chasing”
 - Simple DMA scatter/gather list format
 - Provide data usage hints to allow controller optimization of data placement
 - Support MSI-X and flexible interrupt aggregation

Bucket 1: Eliminate performance bottlenecks seen in other interfaces.

Enterprise NVMHCI Attributes



- Do not carry forward HDD command set legacy
- Eight optimized NVM commands
- Efficient *driver level* translation into SCSI management architectures prevalent in Enterprise
- Support for atomic write size, always larger than a sector

Bucket 1: E
s

Bucket 2: Provides an efficient and streamlined command set.

Enterprise NVMHCI Attributes



Bucket 1: E
st

Bucket
stre

Bucket 3:
Provides Enterprise features.

- End-to-end data protection
 - (i.e., T10 DIF / DIX functionality)
- Firmware update
- Encryption
- Comprehensive statistics
- Health status reporting
- Robust error reporting & handling

Enterprise NVMHCI Attributes



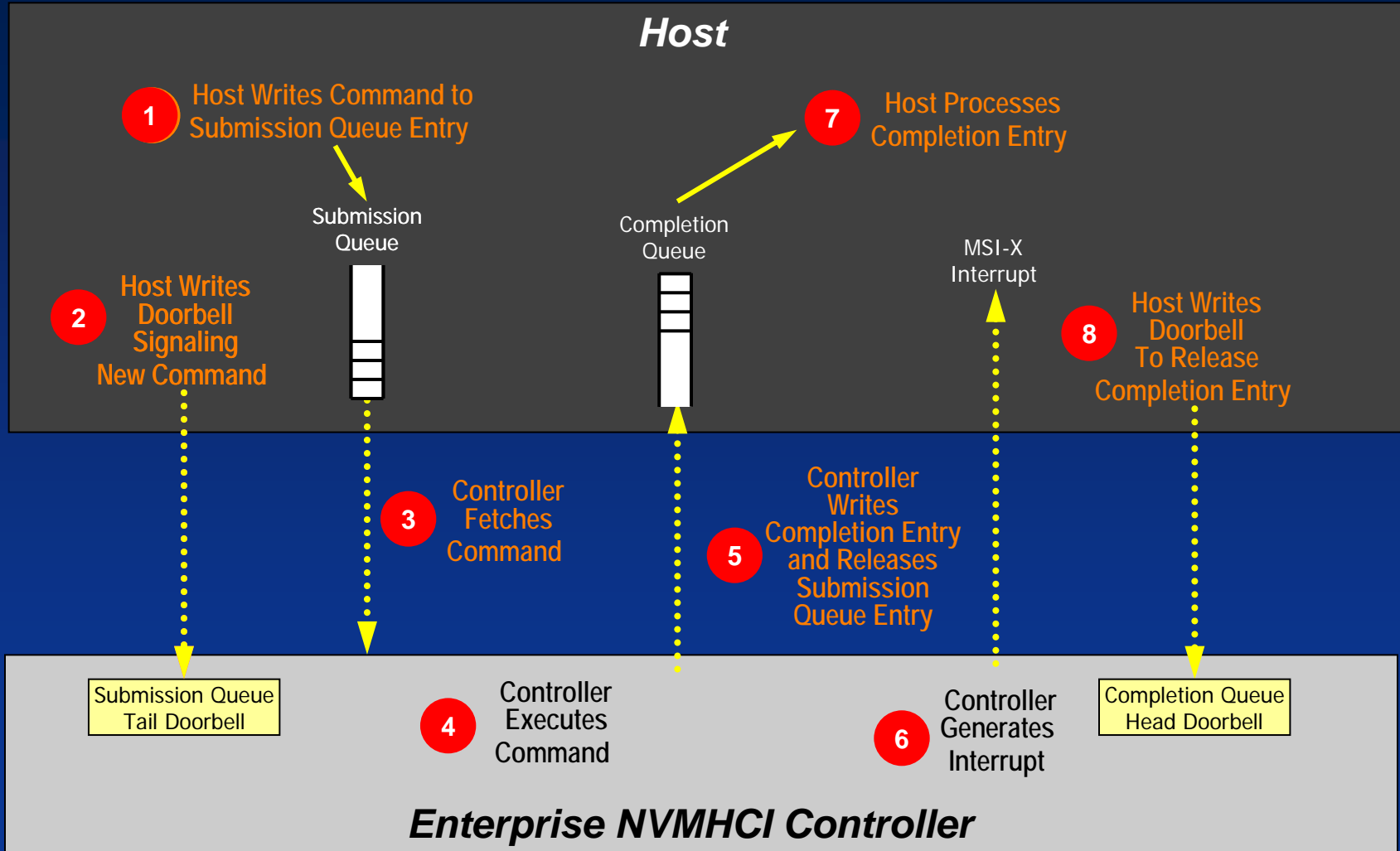
- Support for many core systems
- Supports up to 2K MSI-X vectors
- Support for 64K commands per queue
- Up to 64K Submission & Completion Queues
- Up to 2^{32} outstanding commands to a controller
- Submission & Completion Queues may be mapped on a page basis
- Not tied to any specific NVM technology

Bucket 1: E
s

Bucket
stre

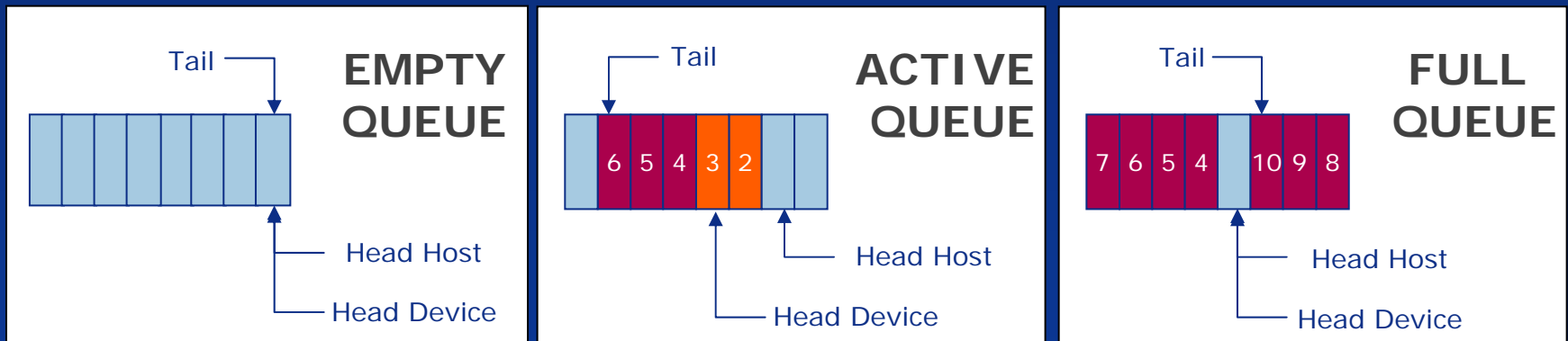
Bucket 4: Provides scalable architecture
for now & the future.

Paired Queue Mechanism



Submission Queue Details

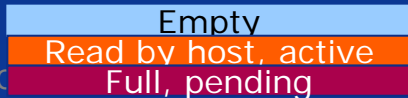
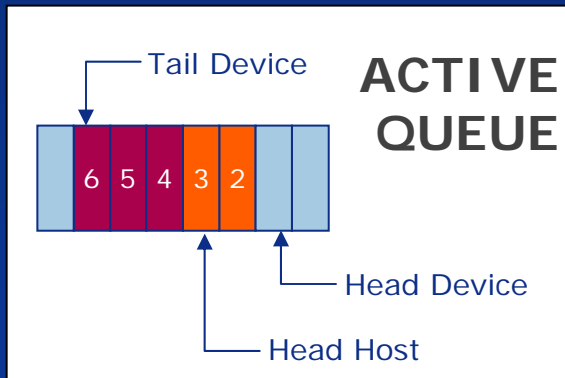
- A submission queue (SQ) is a circular buffer with a fixed slot size of 64B that the host uses to submit commands for execution
- The host updates an SQ Tail doorbell register when there are 1 to n new commands to execute
 - The old SQ Tail value is simply overwritten in the device
- The device reads SQ entries in order and removes them from the SQ, then may execute those commands out of order



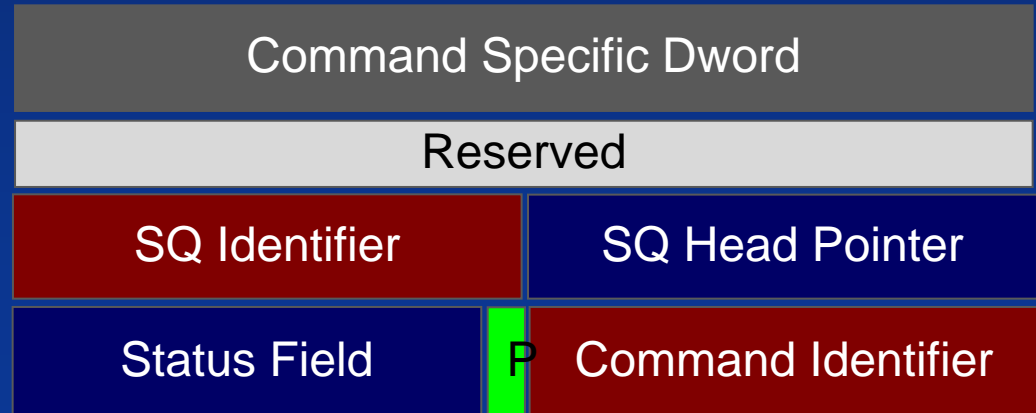
Empty
 Read by device, active
 Full, pending

Completion Queue Details

- A completion queue (CQ) is a circular buffer with a fixed slot size of 16B that the device posts status to for completed commands
- The device identifies the command that completed by the SQ Identifier and the Command Identifier (assigned by software)
- The latest SQ Head pointer is returned in the status to avoid a register read for this information
- The Phase (P) bit indicates whether an entry is new, and inverts each pass through the circular buffer

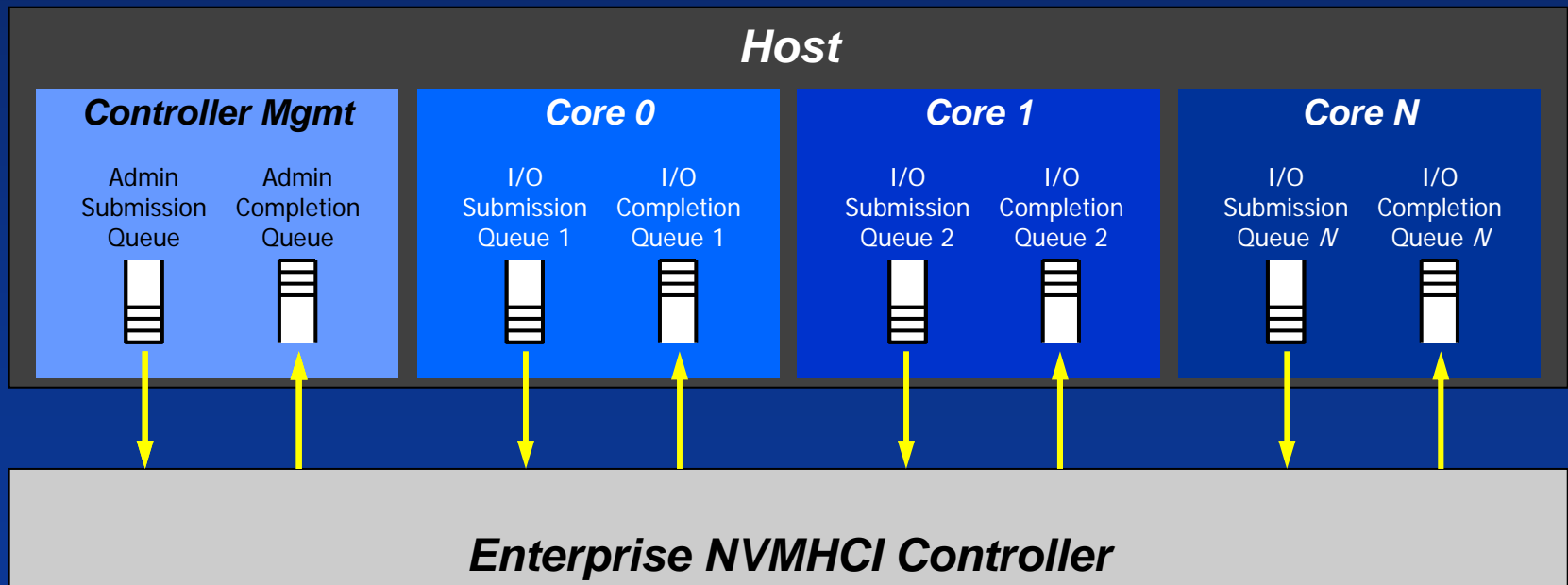


Completion Queue Entry



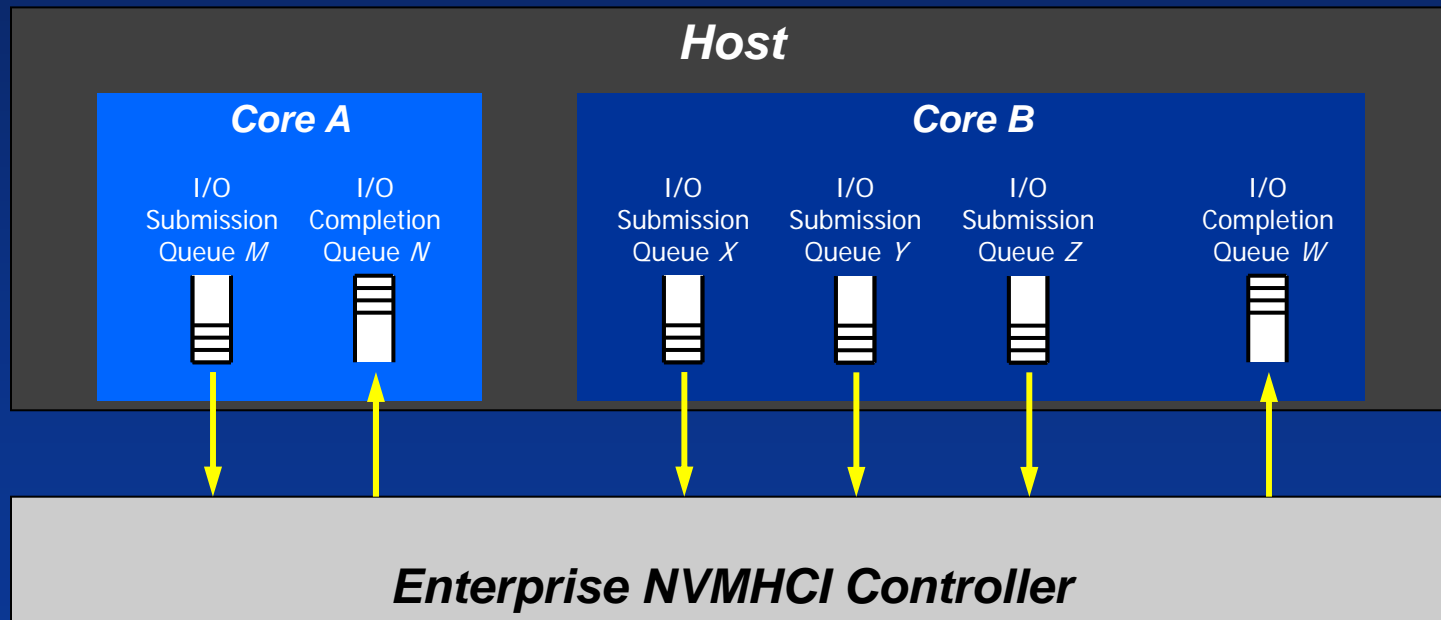
Admin and I/O Queues

- The Admin queue carries out functions that impact the entire device
 - E.g. Queue creation and deletion, command abort
- The driver creates the number of I/O queues that match the system configuration and expected workload
 - E.g. On a four core system, devote a queue pair per core to avoid locking and ensure structures are in the “right” core’s cache
 - Architecturally supports up to 64K I/O submission and completion queues



I/O Queue Mapping Flexibility

- The I/O Submission and Completion Queue mapping is flexible
 - Completion Queue selected when Submission Queue created
- Multiple Submission Queues may be mapped to a single Completion Queue



Command Overview

Management Commands for Queues & Transport

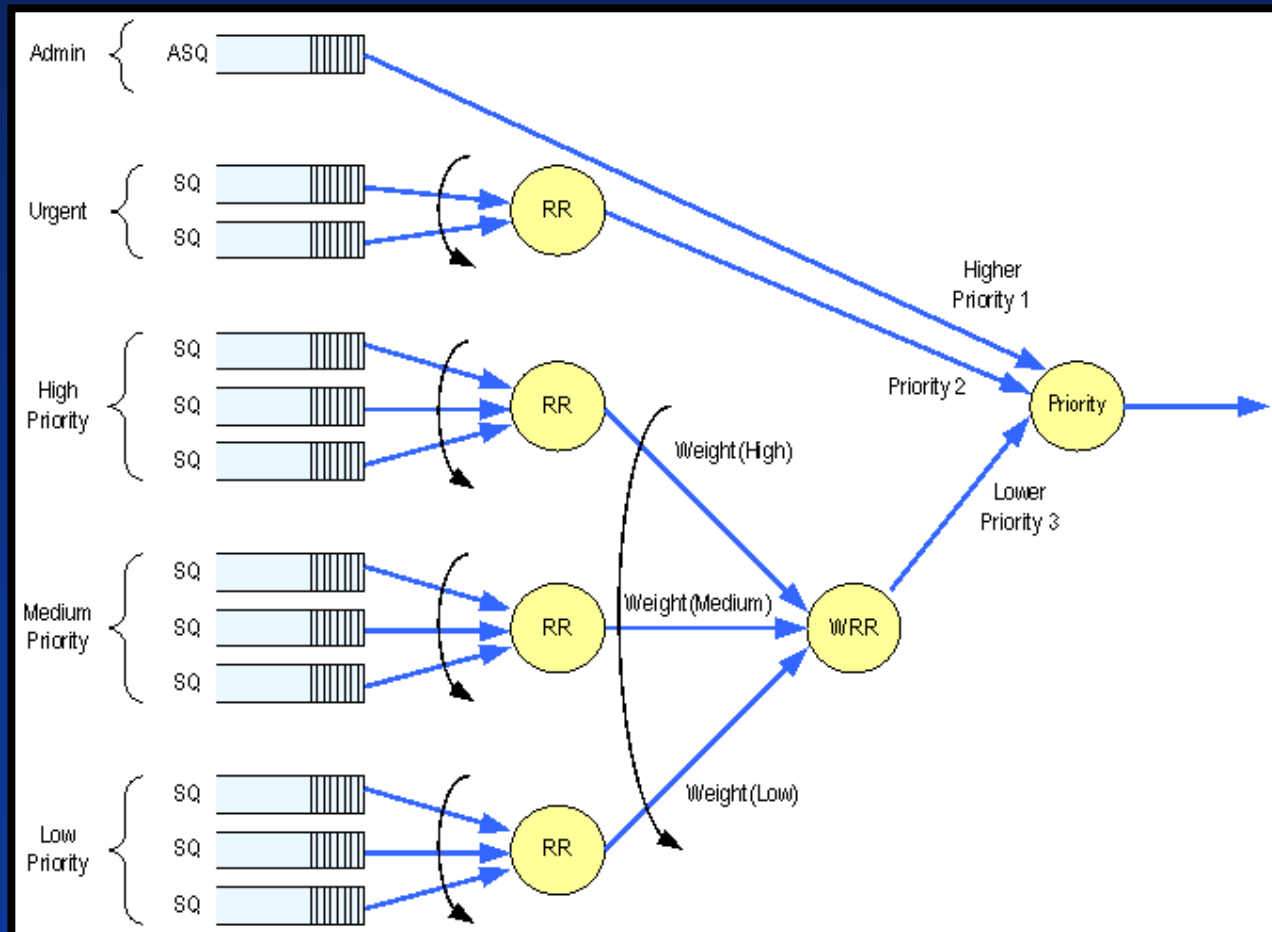
Admin Command	Description
Create I/O Submission Queue	Queue Management
Create I/O Completion Queue	
Delete I/O Submission Queue	
Delete I/O Completion Queue	
Abort Command	
Asynchronous Event Request	Status & Event Reporting
Get Log Page	
Identify	Configuration
Set Feature	
Get Feature	
Firmware Download	Firmware Management
Firmware Activate	
Security Send	Security
Security Receive	

I/O Commands for SSD Functionality

NVM Command	Description
Read	Data Transfer, Including end-to-end data protection & security
Write	
Write Uncorrectable	
Compare	
Compare & Write	
Dataset Management	Data Usage Hints
Flush	Data Ordering
Format NVM	Namespace Management

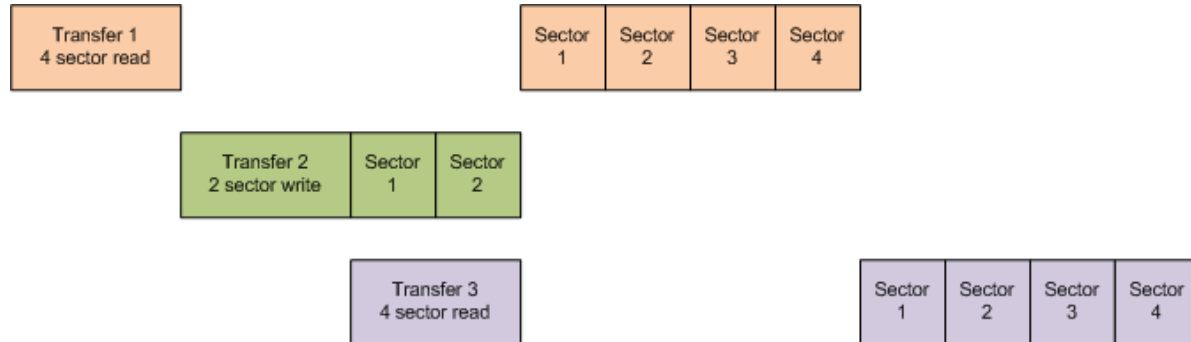
Command Arbitration and Quality of Service

- All Enterprise NVMHCI controllers support round robin command arbitration
- A controller may optionally support weighted round robin with urgent priority class arbitration

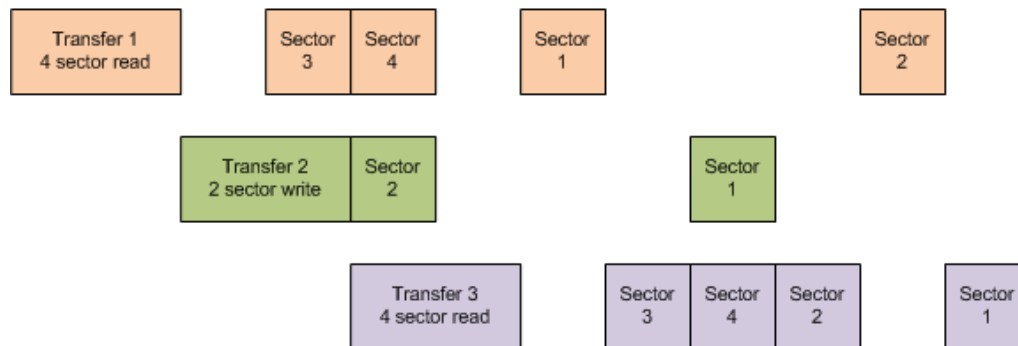


Optimization: Out of Order Data

Traditional storage interfaces



Enterprise NVMHCI



Enterprise NVMHCI reduces latency by allowing more efficient bus ordering and small commands to “slip in”.

Enabling Efficient Out of Order Data Optimization

- Walking a scatter/gather list (SGL) to determine data start locations for portions of a transfer is inefficient
- A fixed size SGL entry enables efficient out of order data & simplifies hardware
- Better approach: Page lists
 - First entry contains an offset
 - “Middle” entries are full page in size
 - Last entry may be less than a page
- The 64B command includes two entries to optimize for 4KB & 8KB I/O
 - For a larger transfer, second entry points to a list of additional entries

1st Entry

Page Base Address	Offset
Page Base Address Upper	

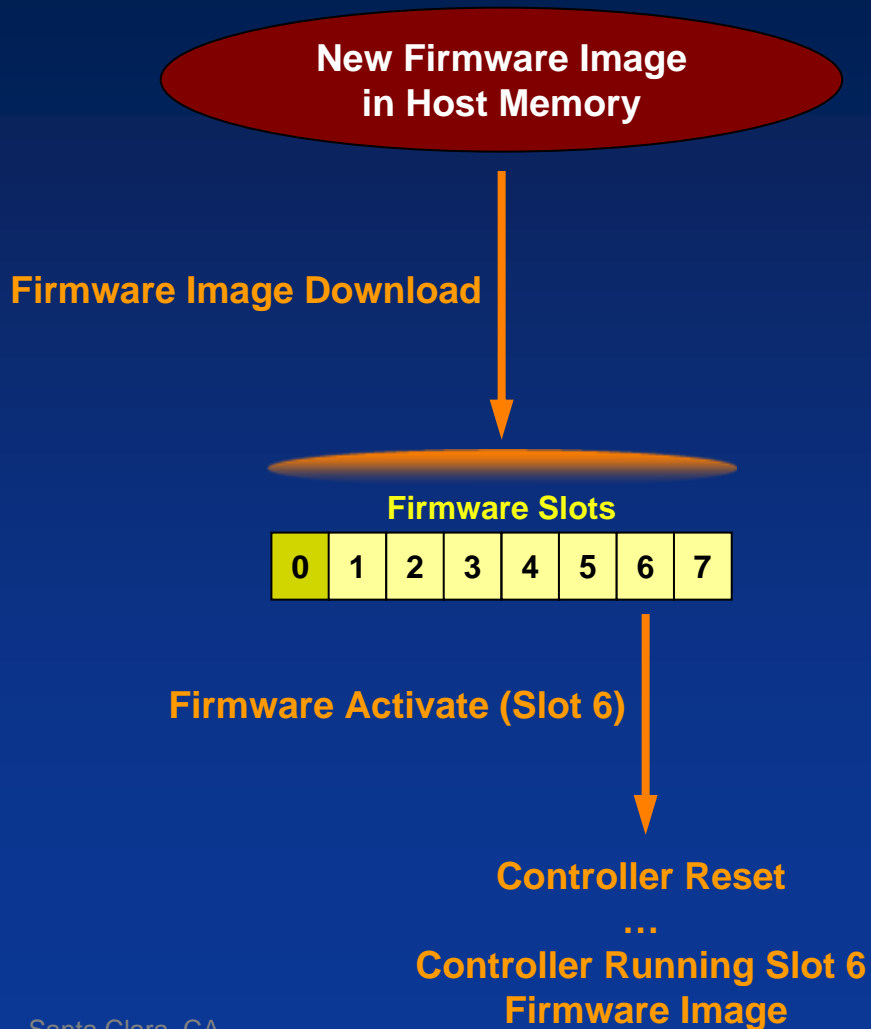
2nd Entry: 4KB unaligned or 8KB

Page Base Address	00h
Page Base Address Upper	

2nd Entry: Pointer to Additional Entries

PRP List Address	00h
Page List Address Upper	

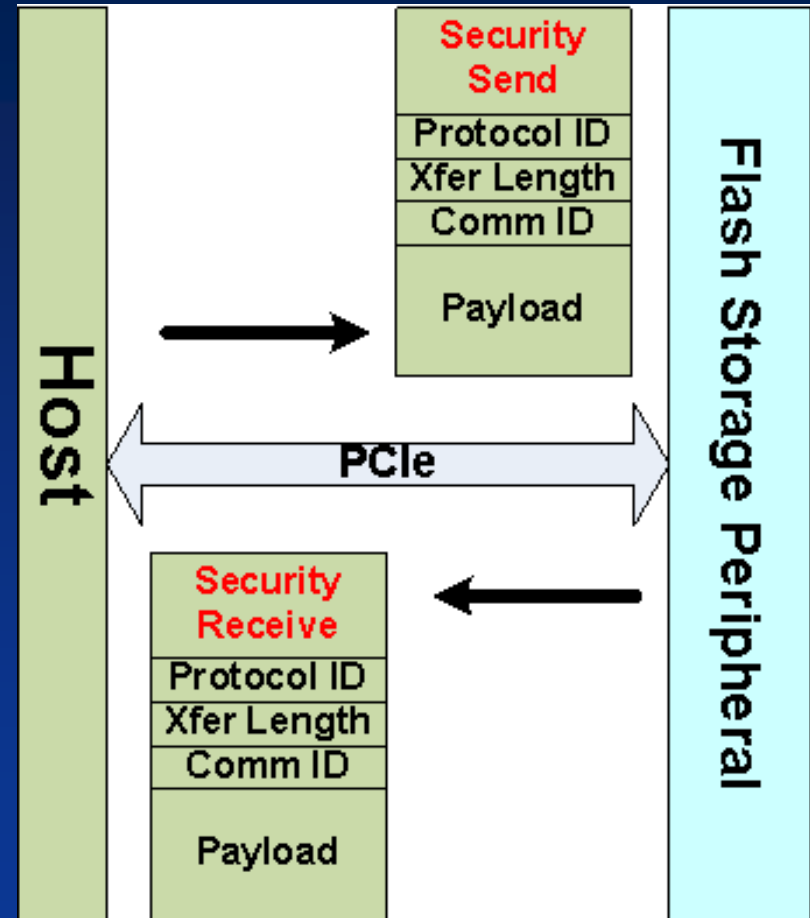
Firmware Update Mechanism



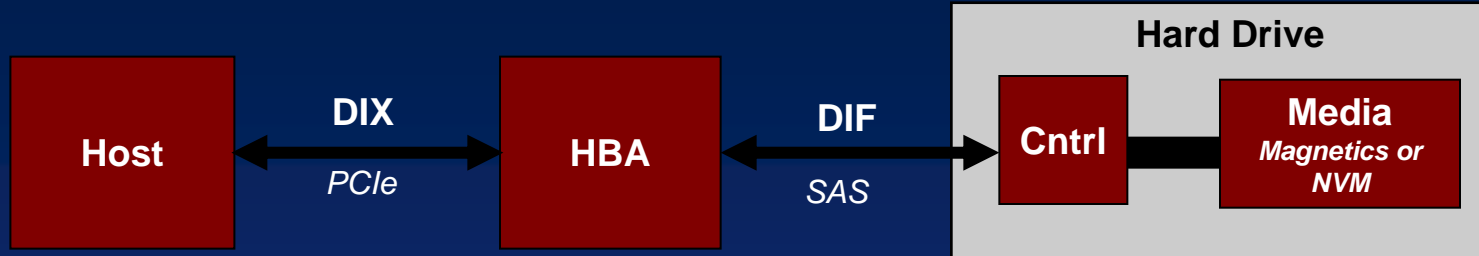
- Firmware slots allows multiple images to be supported
 - Controller supports 1 to 8 slots
- Firmware update process
 - Download Firmware Image: controller transfers image from host
 - Replace Firmware: controller validates image & applies to selected slot
 - Activate Firmware: controller makes selected slot active
 - Firmware update occurs on next reset
- Firmware boot failure
 - Revert to previous active slot or baseline read-only image in slot 0

Trust and Security Services

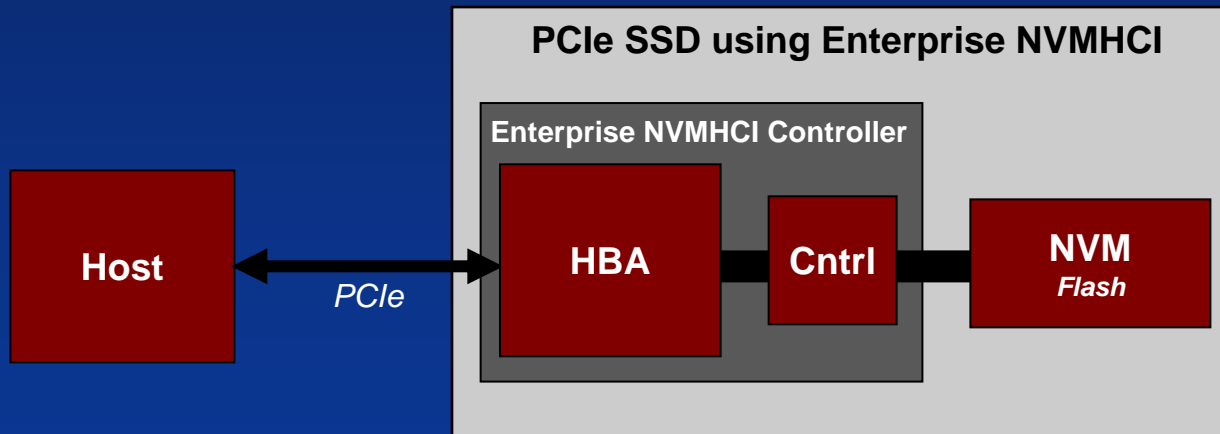
- Security is crucial to NVM as a data-at-rest model
- Encryption and authentication architecture leveraged from existing SSD security concepts such as full drive encryption
 - Simple addition of Security Send and Security Receive to Enterprise NVMHCI command set
- A liaison is being established with the Trusted Computing Group to leverage standard security management
 - Standard architecture for policy-driven access control
 - Includes authentication, encryption, and lifecycle management (deployment to end of life)



End-to-End Data Protection

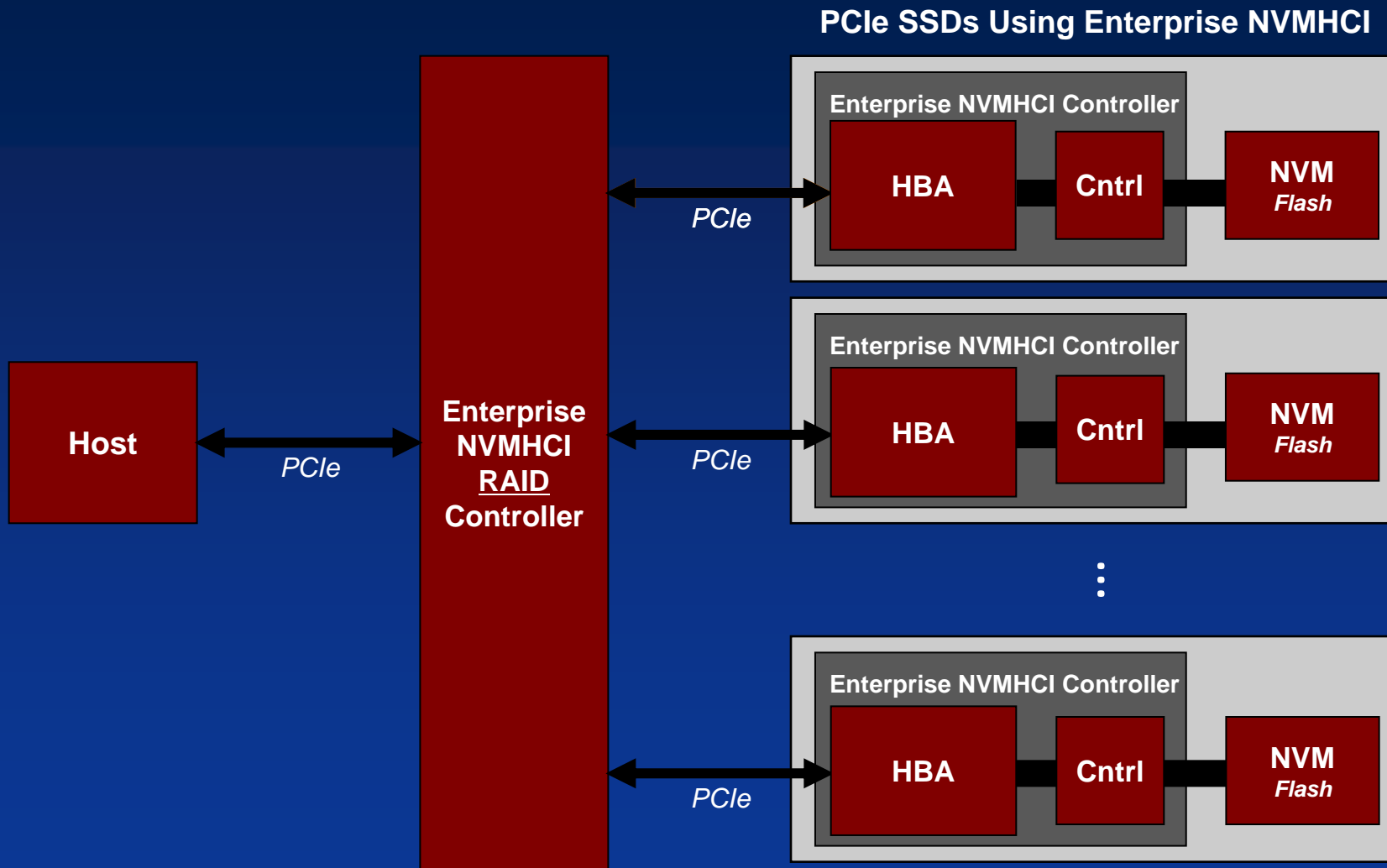


Traditional Storage System

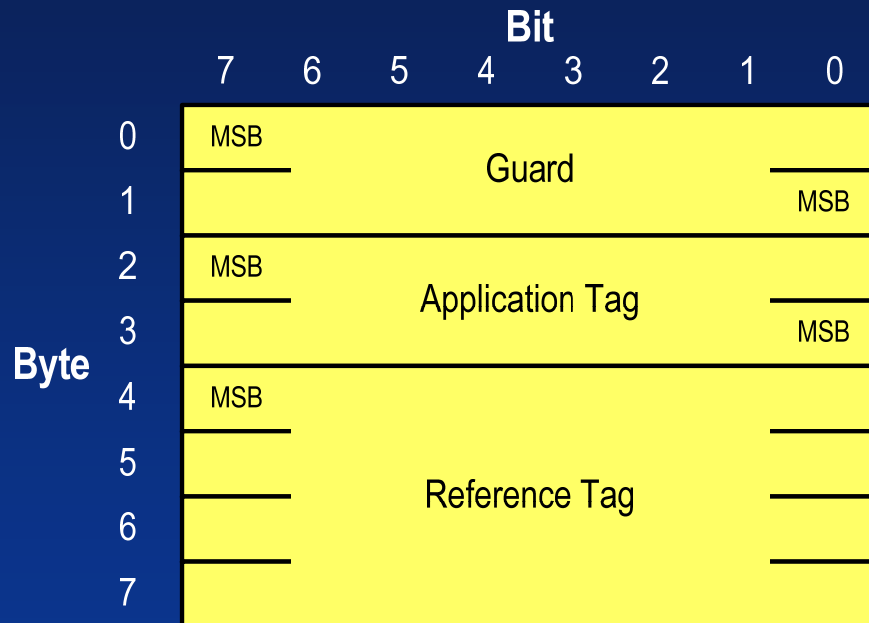


Enterprise NVMHCI Storage System

End-to-End Data Protection with Hardware RAID

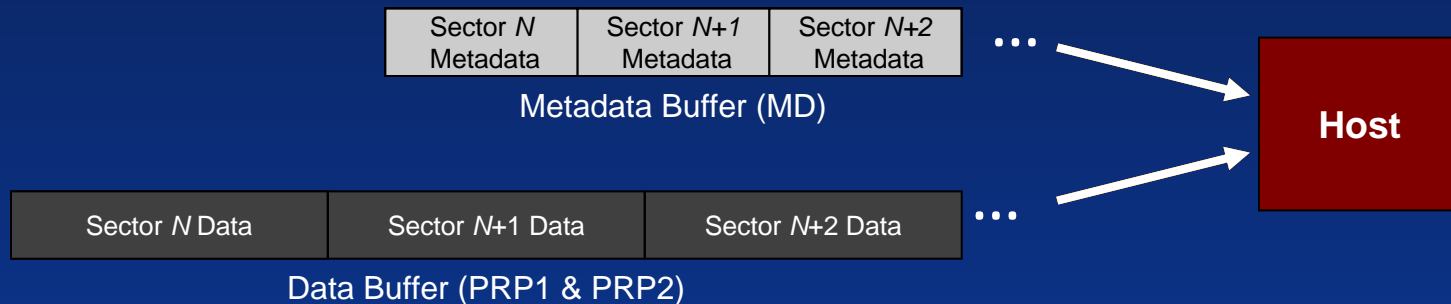
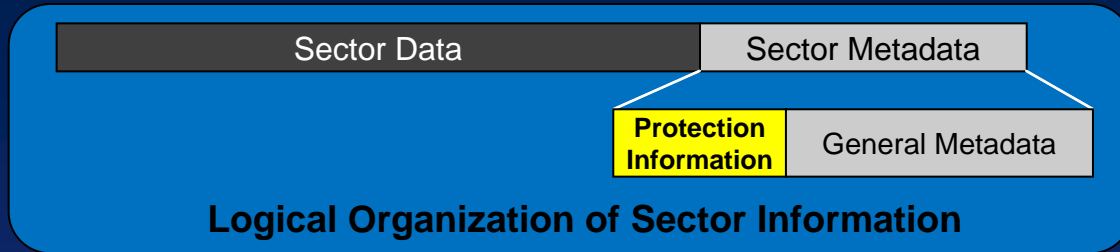


Data Protection Information

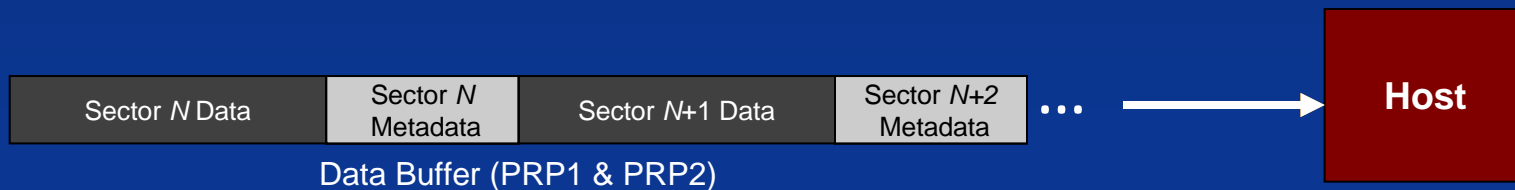


- Data protection information associated with each sector
 - Same format as DIF / DIX
 - Consumes first 8 bytes of metadata
- Guard field
 - CRC-16 as defined by T10 DIF
 - IP Checksum not supported
- Application tag field
 - Same definition as T10 DIF
 - May be used to disable checking of protection information (i.e., 0xFFFF)
 - Generally opaque data not interpreted by controller
- Reference tag field
 - Same definition as T10 DIF
 - May be used to disable checking of protection info (i.e., 0xFFFF_FFFF)
 - Incrementing value associated with sector address or value provided as part of command

Host Metadata Buffer: Organization / Transfer Options

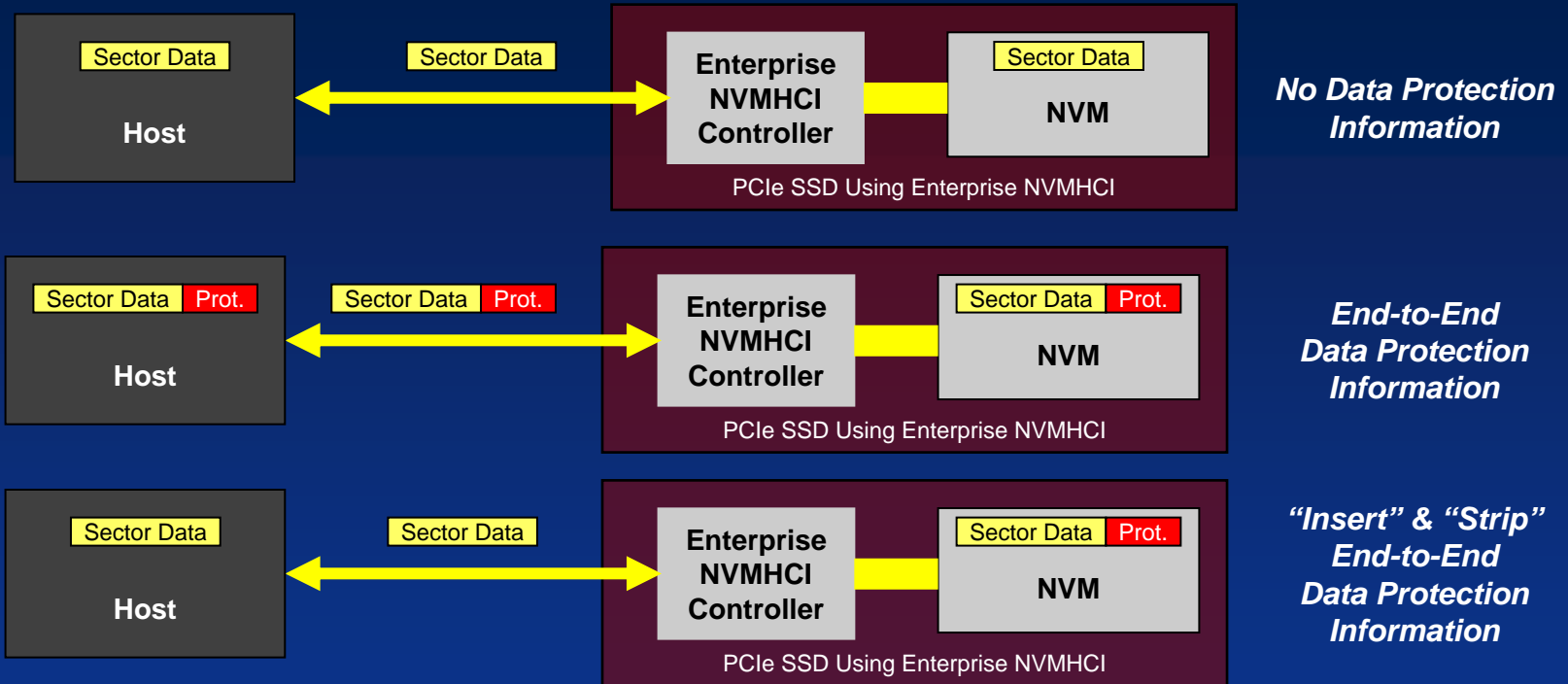


Data with Metadata in Separate Contiguous Buffer, "DIX Like"



Metadata as Part of Sector Data, "DIF like"

End-to-End Data Protection Options



- Functionally compatible with T10 DIF & DIX, including DIF Type 1, 2, and 3
- End-to-end protection configured per namespace with NVM Format command
- Controller may optionally "insert" and "strip" protection information

Summary

- Enterprise NVMeHCI fosters interoperability & faster adoption for PCIe SSDs
 - Standard OS drivers
 - Consistent Enterprise feature set
 - Reduced OEM validation and qualification
- Enterprise NVMeHCI has been optimized for ultra high performance
 - Support for many core systems
 - Normal operation requires no reads from controller
 - Support for a large number of outstanding operations
- Enterprise NVMeHCI is on track for completion this year enabling product intercept in 2012
 - 0.70 specification available now to NVMeHCI members