Performance and Innovation of Storage Advances through SCSI Express

Marty Czekalski
President, SCSI Trade Association - Emerging Interface and Architecture
Program Manager, Seagate Technology

Greg McSorley
Vice President, SCSI Trade Association
Technical Business Development Manager, Amphenol

SNIA Legal Notice

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

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

Abstract

- Performance and Innovation of Storage Advances through SCSI Express
  - SCSI Express represents the natural evolution of enterprise storage technology building upon decades of customer and industry experience. SCSI Express is based on two of the highest volume and widely deployed, interoperable technologies in the world – SCSI and PCI Express. These two technologies enable unprecedented performance gains while maintaining the enterprise storage experience. This presentation will provide an in-depth overview of SCSI Express including what it is, potential markets, where it is being developed, why it is important to the enterprise computing platform, how it is implemented.

SCSI Logical Abstraction Layer: A Foundation for Innovation

- Preserves Hardened SCSI Command Set
  - Successive Product Generations
  - Accommodates Frequent Technology Shifts
  - Multiple Vendors
  - Multiple Interconnects
- Reduces Time to Market and Integration Costs
- Delivers Enterprise Attributes and Features
  - End-to-End Data Protection
  - Atomic Writes
  - Hinting
  - Task Management
  - Power Management
  - And more on the way

SCSI Logical Abstraction Layer: A foundation for Innovation

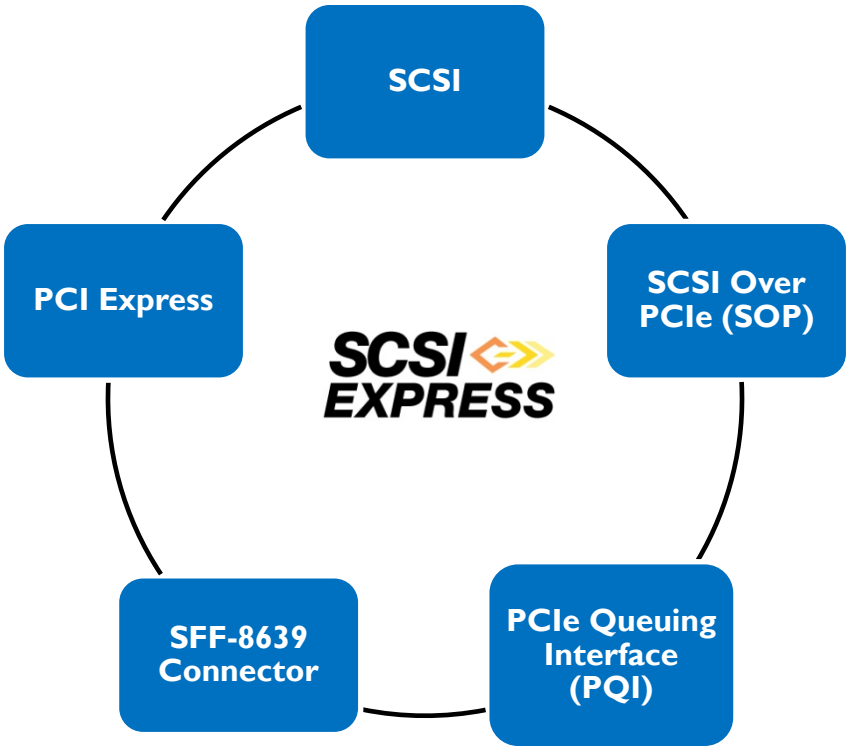
- Operates Over Numerous Transport Layers
  - ATAPI (ATA, SATA)
  - USB
  - Memory sticks
  - Firewire
  - Infiniband
  - iSCSI
  - FC & FCoE
  - Parallel SCSI
  - SAS
- And now, SCSI Over PCIe (SOP, PQI)

SCSI:
The Most Widely Implemented Logical Storage Protocol

SCSI Express Overview

- What is SCSI Express?
  - Proven SCSI protocol combined with PCIe creating an industry standard path to PCIe-based storage
- Why do we need SCSI Express?
  - Deliver proven enterprise storage for PCIe-based storage devices in a standardized ecosystem
  - Take advantage of lower latency PCIe to improve performance
  - Unified management and programming interface

SCSI Express Components



- The SCSI storage command set
- Packages SCSI for a PQI queuing layer
- Flexible, high-performance queuing layer
- Accommodates PCIe, SAS, and SATA drives
- Leading server I/O interconnect

SCSI Express Value Proposition

Performance and Innovation

- Increased performance through lower latency for emerging advanced technologies
- Enables new storage architectures

Reliability

- Proven enterprise SCSI ecosystem
- Architected for nonstop availability

Investment Protection

- Coexistence with SAS via Express Bay and common command set
- Leveraging robust middleware ecosystem

SCSI Express Hardware/Software

SCSI Express Controllers

- Supports SOP-PQI driver functionality on the controller to the target device on the PCIe lanes

SCSI Express Drive/Device

- SOP-PQI protocol
- Connects to SFF-8639
- PCIe up to x4 interface

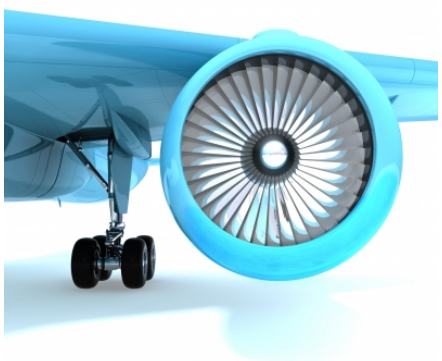
SCSI Express Driver

- Driver supplied by storage OEMs, IHVs or OSVs
- Open Source Linux driver and IHV drivers available

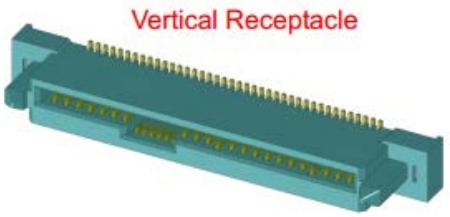
See STA website for additional SCSI Express nomenclature (www.scsita.org)

Express Bay Components

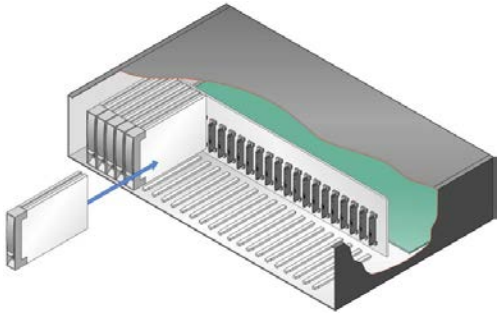
25W Power



Cooling



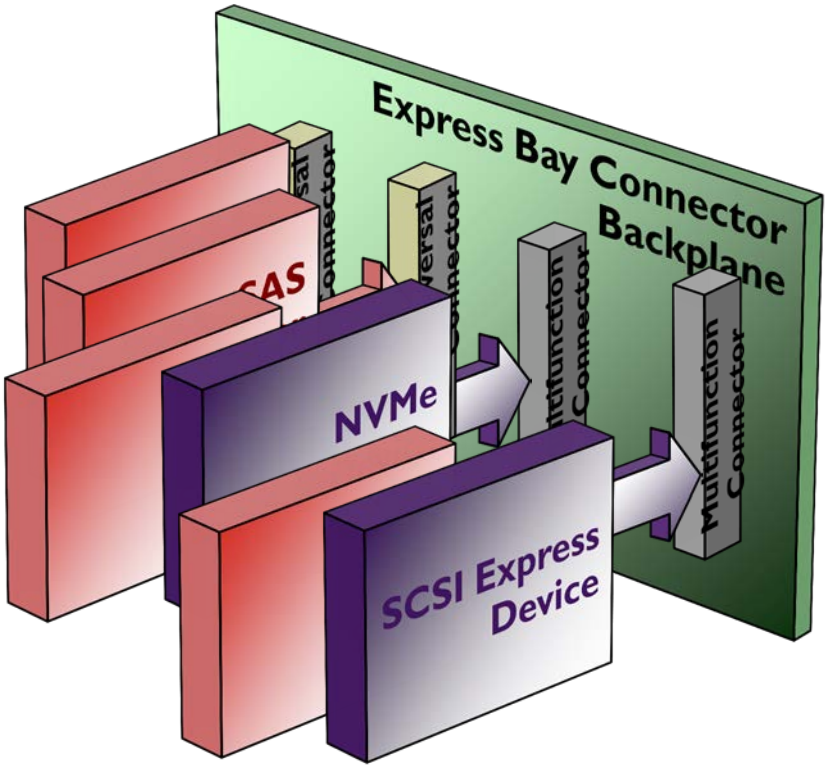
Multifunction Connector



Accessibility / Serviceability

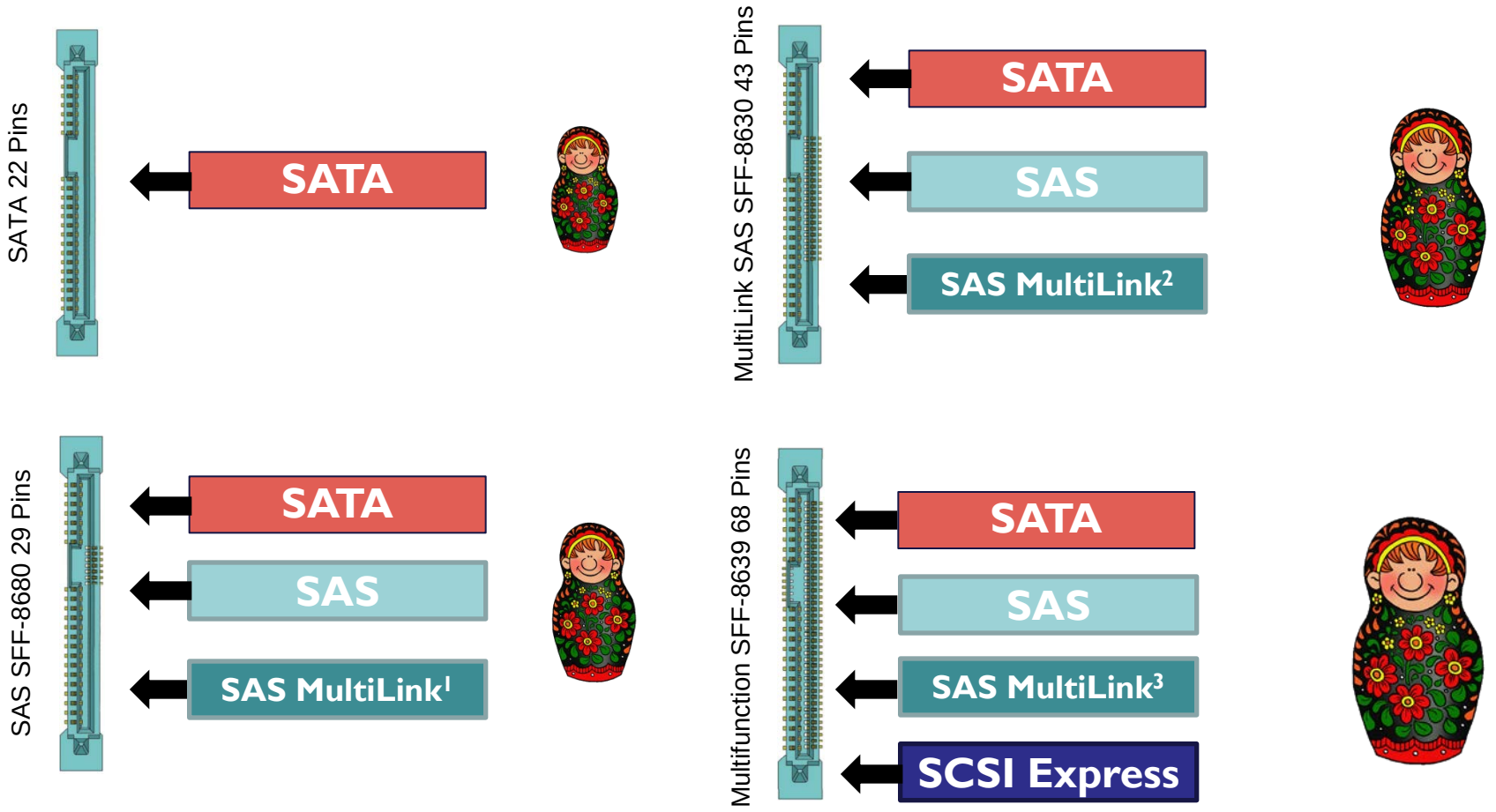
Traditional Drive Form Factor

Express Bay



- Express Bay
  - Up to 25 Watts
    - For both SAS and PCIe
  - SFF-8639 connector
  - PCI-SIG electrical specification
- Objectives
  - Preserve the enterprise storage experience for PCI Express storage
  - Meet SSD performance demands
  - Serviceable, hot-pluggable Express Bay opens up new possibilities…

SAS Connector Compatibility

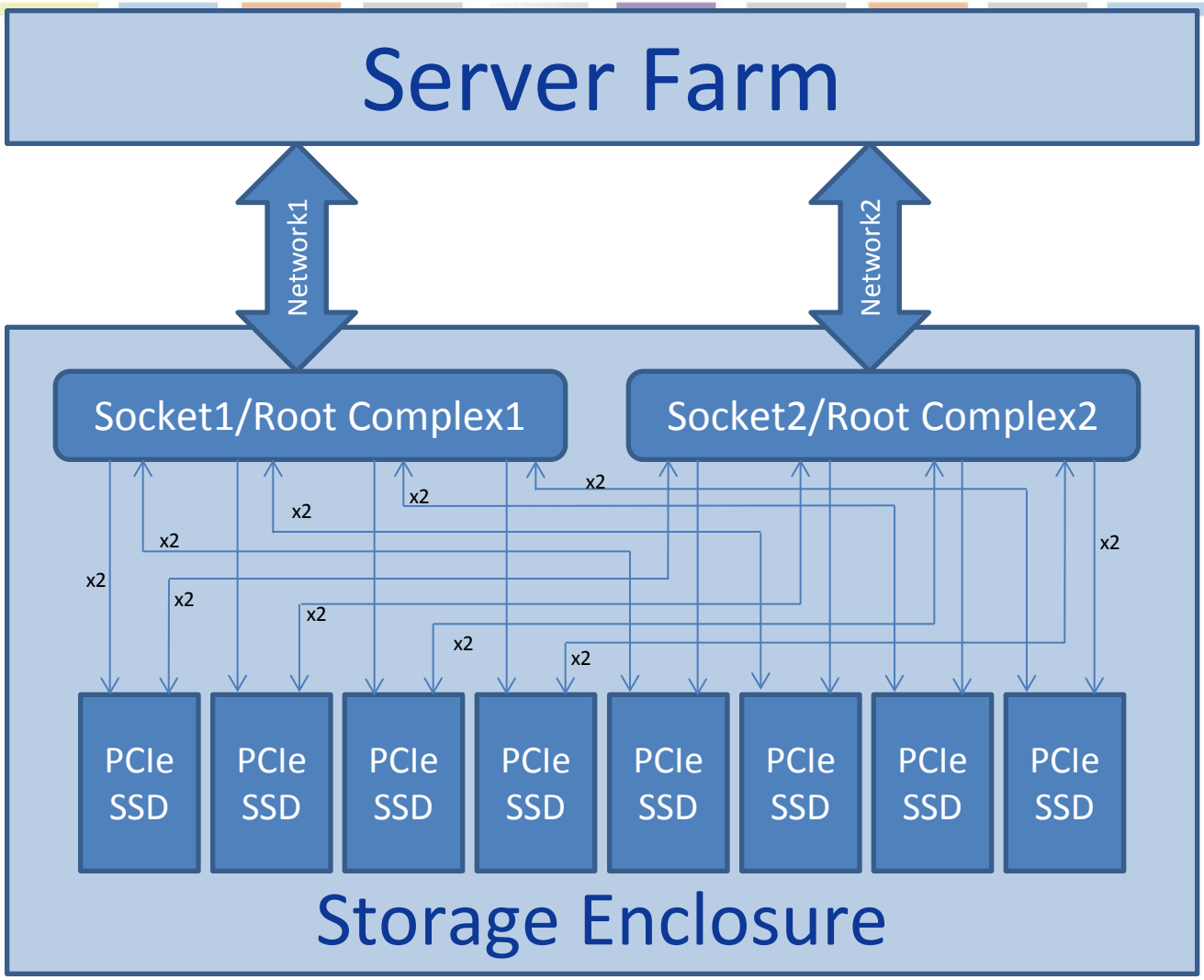


[1] Max two links operate

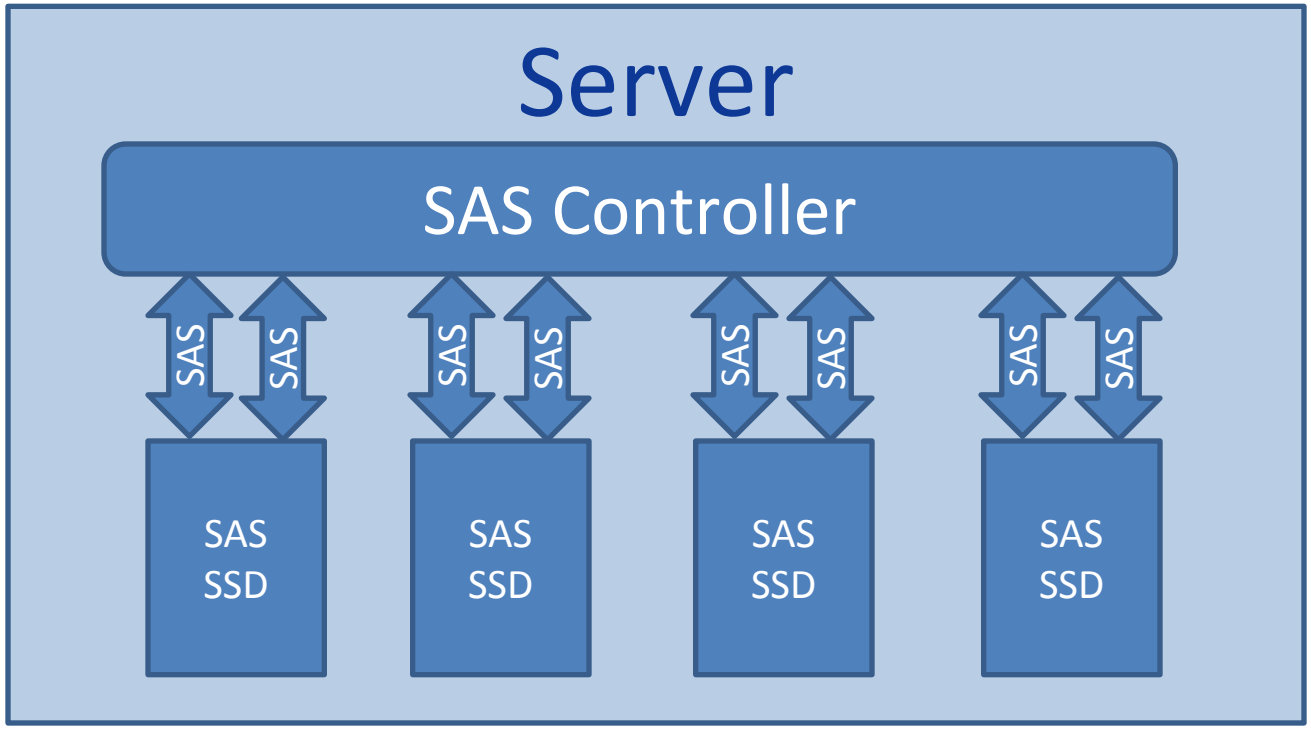
[2] Four links operational

[3] Two or four links operation depending on host provisioning

Dual Port With PCIe for High Availability



Wide Port SAS for Increased Throughput



2.4 GB/s full-duplex per SSD

Express Bay Summary

- Preserve the enterprise storage experience for PCI Express storage
- Meet SSD performance demands with PCIe, SAS, or SATA
- Serviceable, accessible bay offers configurability

SCSI Express Naming

- **SCSI Express Drive**: SCSI Express SSD, SCSI Express HDD (SCSIe SSD or HDD)
  - A PCIe data storage device using solid-state drive technology in a 2.5” Small Form Factor (SFF) or 3.5” Large Form Factor (LFF) that utilizes SCSI over PCIe (SOP) and PCIe Queuing Interface (PCIe) to communicate with the PCIe bus on the host.

- **SCSI Express RAID Controller**:
  - A PCIe controller with Redundant Array of Independent Disks (RAID) fault tolerance capability that communicates SCSI over PCIe (SOP) and PCIe Queuing Interface (PCIe) commands to the PCIe bus on the host.

SCSI Express Naming

- **SCSI Express Card Drive**: SCSI Express SSC, SCSIe SSC (SSC=Solid State Card)
  - A PCIe data storage device using solid-state drive technology in a PCIe card form factor that utilizes SCSI over PCIe (SOP) and PCIe Queuing Interface (PCIe) to communicate with the PCIe bus on the host.

- **SCSI Express M.2 Drive**: SCSI Express SSM, SCSIe SSM (SSM=Solid State Module)
  - A storage device using solid-state drive technology in a M.2 form factor that utilizes SCSI over PCIe (SOP) and PCIe Queuing Interface (PCIe) to communicate with the PCIe bus on the host.

SCSI Express Naming

- **SCSI Express Bridge**:
  - A device using that communicates SCSI over PCIe (SOP) and PCIe Queuing Interface (PCIe) commands downstream to the PCIe bus on the host.

- **SCSI Express Switch**:
  - A device which transmits SCSI over PCIe (SOP) and PCIe Queuing Interface (PCIe) packet commands to and from a host to a SCSI Express device connected to the switch.

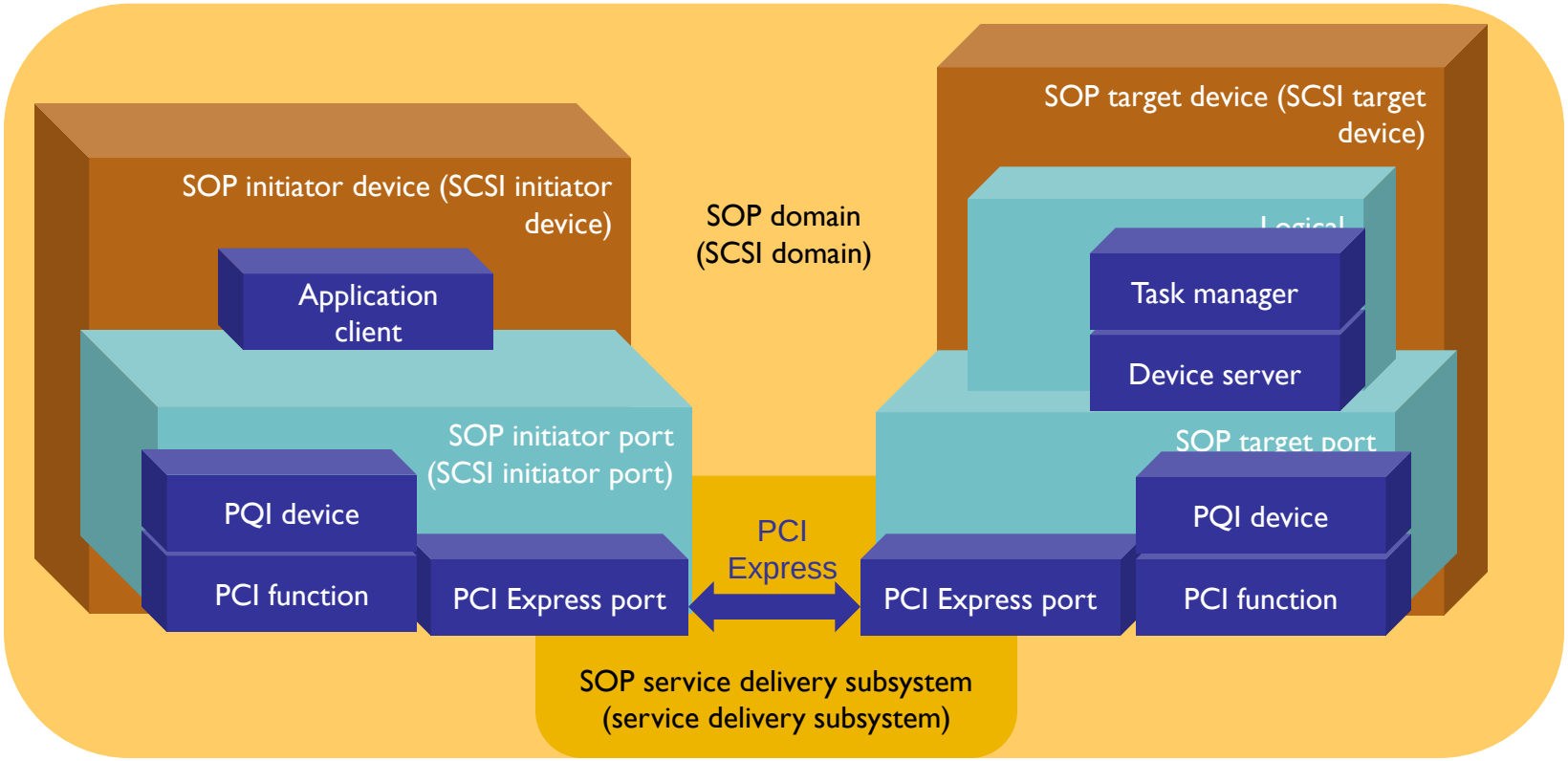
SCSI Express Naming

- **SCSI Express Mezzanine Drive**: SCSI Express SSM (SCSIe SSM)
  - A proprietary PCIe data storage device using solid-state drive technology in a blade mezzanine card form factor that utilizes SCSI over PCIe (SOP) and PCIe Queuing Interface (PCIe) to communicate with the PCIe bus on the host.

Architecture

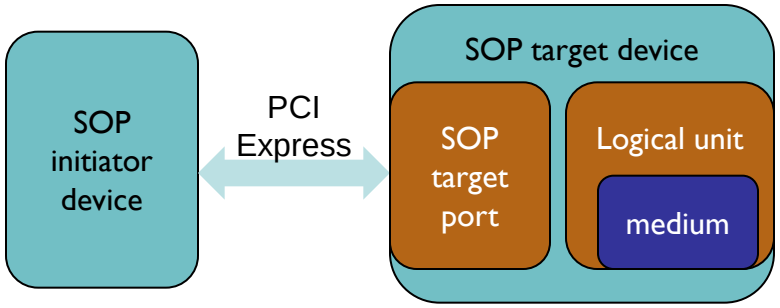
- Key excerpts from SCSI, SOP, and PQI architecture models

SCSI initiator device: a server with a PCI Express Root Port
SCSI target device: an SSD, HDD, HBA, or RAID controller



Simple Devices

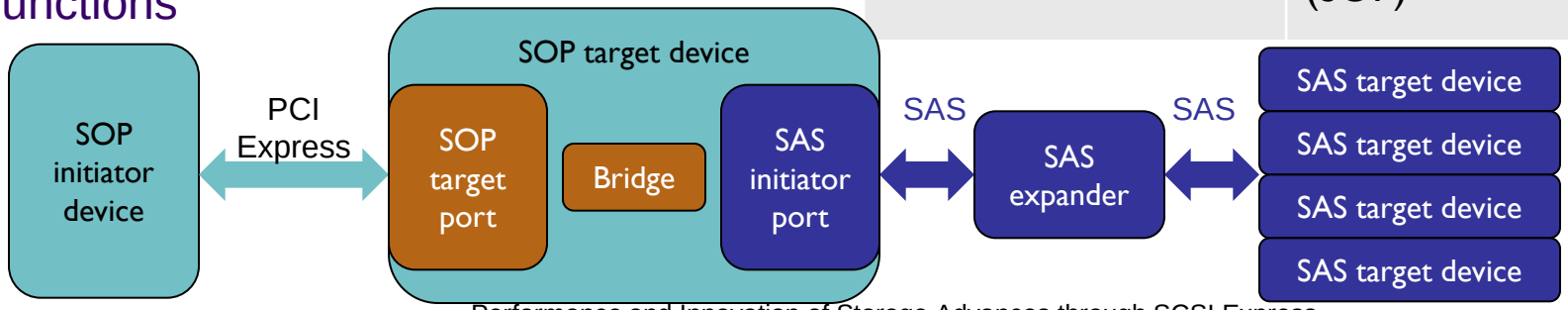
- SSDs, etc
  - Usually just a single logical unit with LUN 0
  - Any SCSI device type is possible
    - SSD, tape drive, optical drive (CD/DVD/BluRay), etc



Bridges

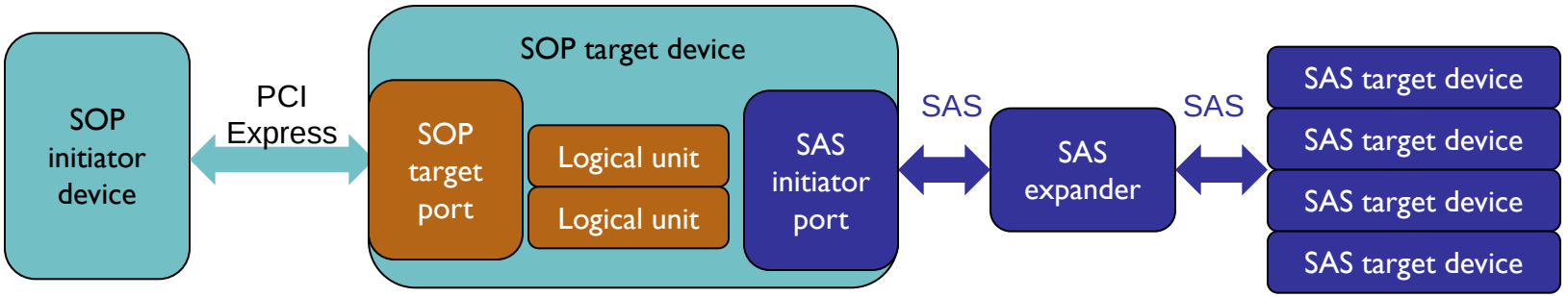
- HBAs
- Bridges from PCI Express to another interconnect supporting SCSI
  - Maps SCSI target devices one-for-one
- Typical terms: host bus adapter (HBA), host controller, host adapter, network interface controller, converged network adapter
- Usually referred to only by the back-end interconnect
  - e.g. "SAS HBA"
- Manage with SOP bridge management functions

Interconnect	SCSI transport protocol
Serial Attached SCSI (SAS)	Serial SCSI Protocol (SSP)
Fibre Channel (FC)	Fibre Channel Protocol (FCP)
Ethernet	Internet SCSI (iSCSI)
Universal Serial Bus (USB)	USB Attached SCSI (UAS)
InfiniBand	SCSI RDMA Protocol (SRP)
PCI Express	SCSI over PCI Express (SOP)



RAID Controllers

- Less complex than bridges from an SOP perspective
  - Indirectly bridges from PCI Express to another interconnect supporting SCSI
    - Not a one-to-one mapping of SCSI target devices
    - Presents logical drives over PCI Express
      - Created from physical drives
    - Manage with standard SCSI commands
      - REPORT LUNS reports the logical units that have been created
      - Bridge management not involved (unless it's a hybrid HBA + RAID controller



Queuing Layer Services

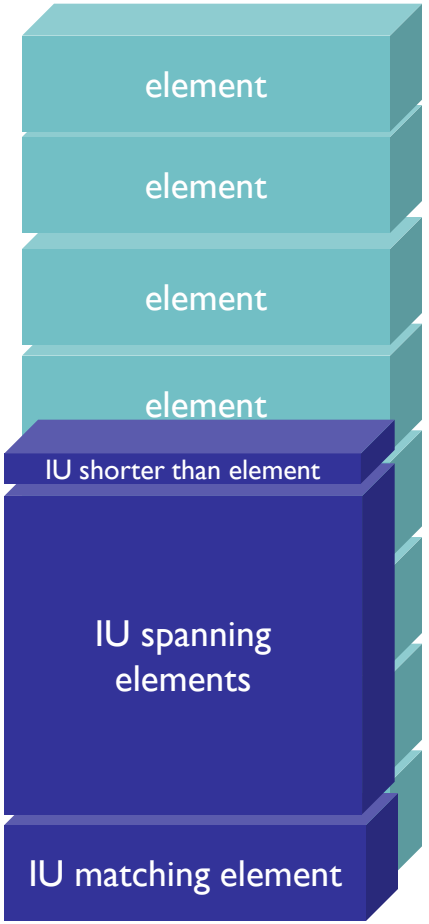
- SOP expects a queuing layer over PCI Express to define
  - inbound queues
    - transfer IUs from SOP initiator port to SOP target port
  - outbound queues
    - transfer IUs from SOP target port to SOP initiator port
- SOP architected to support multiple queuing layers
  - PCI Express Queuing Interface (PQI)
- Information Units (IUs)
  - Messages between a driver or a SCSI Express controller and a SCSI Express device

Information Units (IUs) Types

- SCSI Request/Response IUs
  - Commands, Task Management, Success, Command Response, Task Management Response, etc.
- General Management Request/Response IUs
  - Report General, Report Configuration, Set Configuration, Report Event Configuration, Management Response, Event, Event Acknowledge
- Bridge Management Request/Response IUs
- Administrator Request/Response IUs
- Other
  - Null IU, etc.

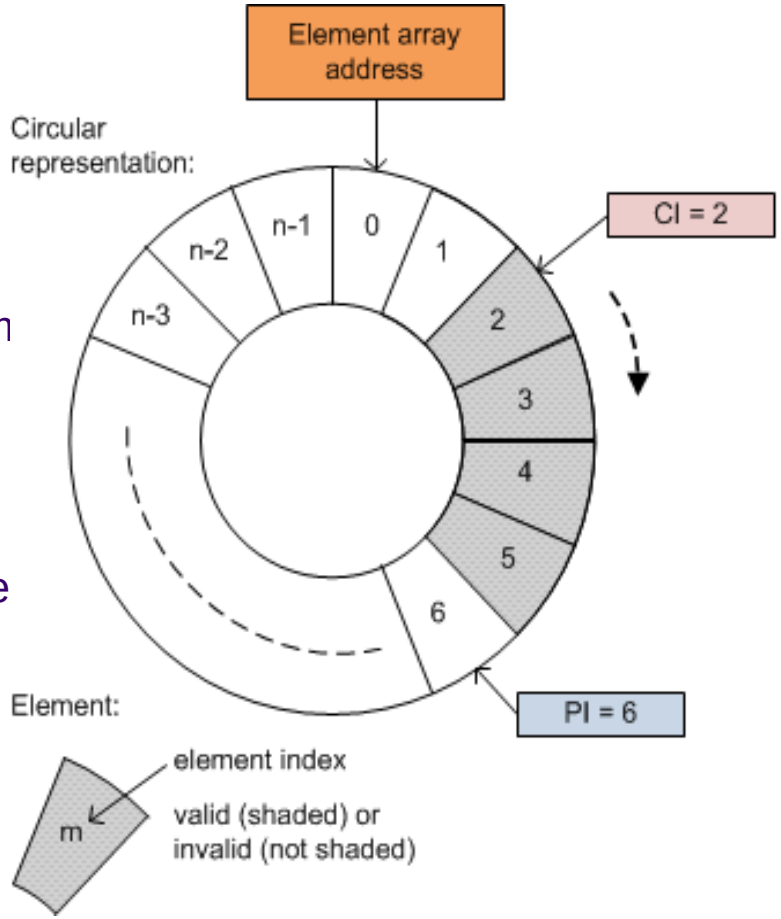
IUs and Queues

- Smaller, equal, or larger than queue element
  - IUs are sent in operational IQs or operational OQs
  - Each IU starts at the beginning of an element
  - No more than one IU in an element
  - Any bytes after the IU are ignored by the recipient up until the end of the element
    - e.g., OQ element size 64 bytes, with a 16 byte SUCCESS IU
  - IU may span multiple elements
    - e.g., OQ element size 16 bytes, with a 64 byte COMMAND RESPONSE IU spanning four elements
    - The IU header is only in the first element, not repeated in each



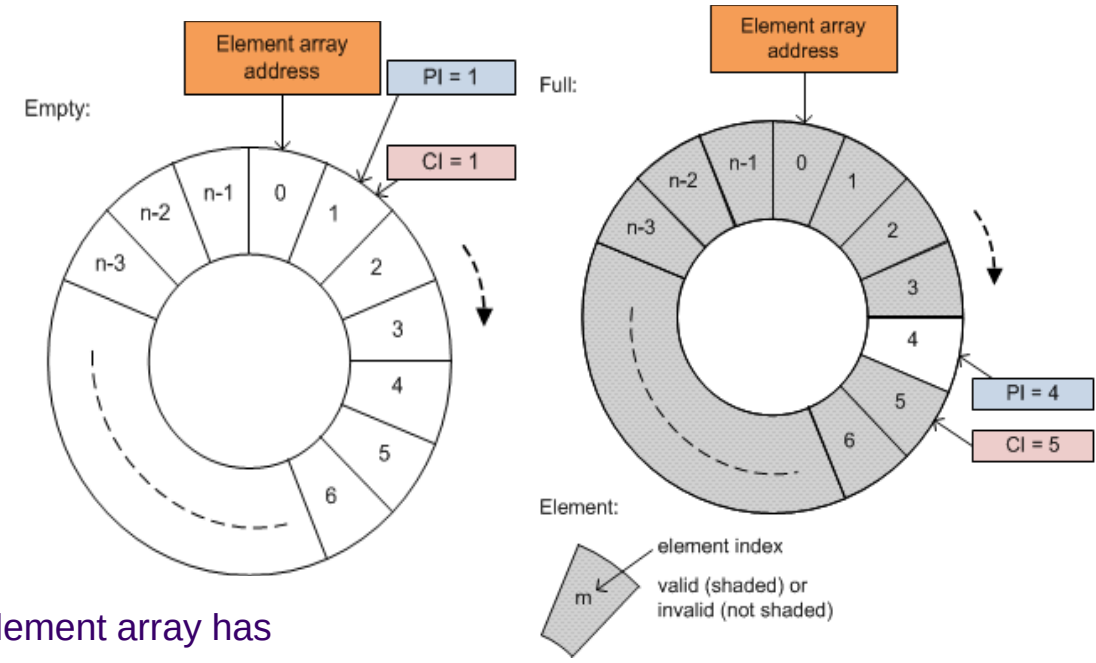
Circular Queue Basics

- Circular queue basics
  - Element array
    - Fixed size elements (e.g., 64 bytes)
  - Producer index (PI)
    - Location to which producer writes elem
    - Write to element array[PI++]
    - Wrap at size of the element array
  - Consumer index (CI)
    - Location from which consumer reads e
    - Read from element array[CI++]
    - Wrap at size of the element array



Empty and Full Circular Queues

- One element is always unused
  - Empty
    - PI equals CI
    - e.g., PI=17 and CI=17
  - Entries to consume
    - e.g., PI=18 and CI=17
  - Full
    - PI is one behind CI
      - e.g., PI=17 and CI=18
      - e.g., PI=62 and CI=63
      - e.g., PI=63 and CI=0 (if element array has 64 elements)
      - e.g., PI=0 and CI=1
  - One queue element is always unused
    - e.g., maximum 63 entries in a queue of 64

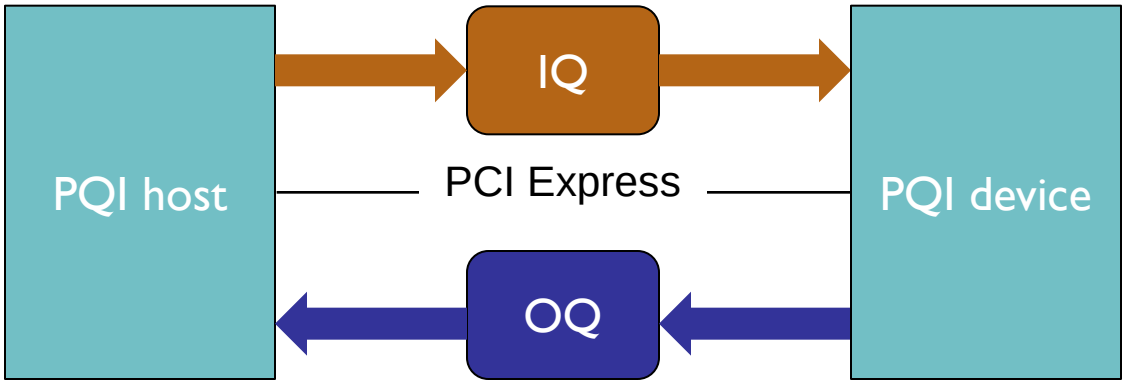


Inbound Queues (IQs) and Outbound Queues (OQs)

- Named from the PQI device's perspective

- **Inbound queues (IQs)**
  - PQI host to PQI device
    - Administrator request IUs
    - SCSI request IUs (in SOP)

- **Outbound queues (OQs)**
  - PQI device to PQI host
    - Administrator response IUs
    - SCSI response IUs (in SOP)



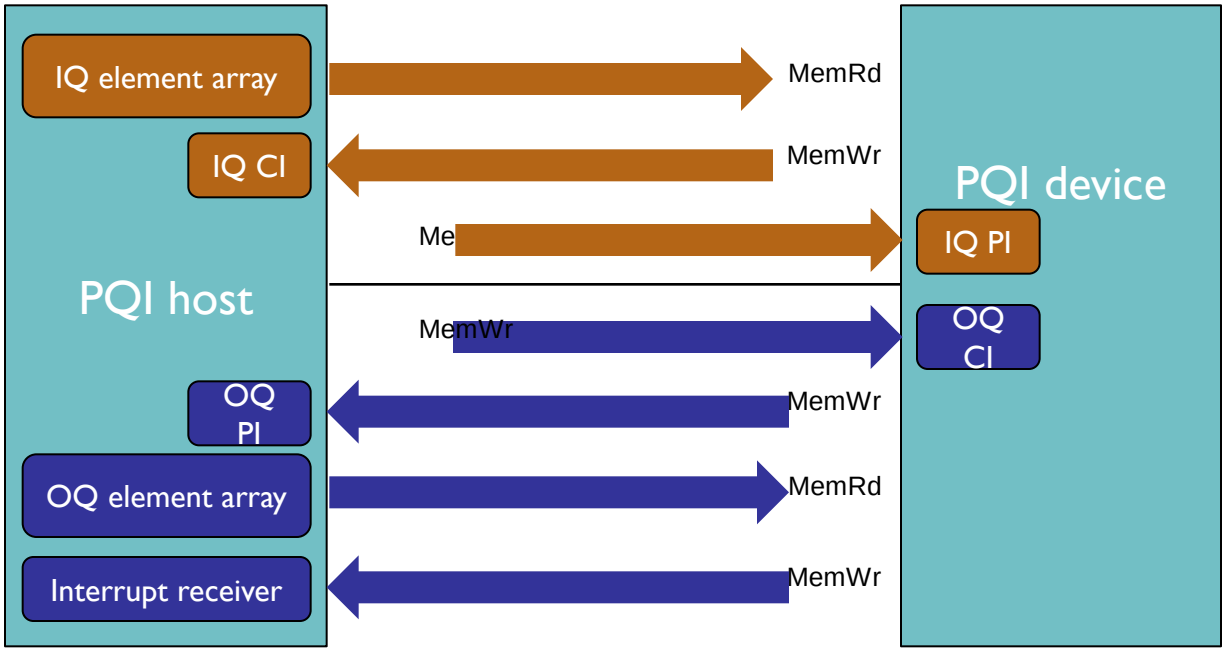
IQ and OQ Object Locations

SNIA Education

- Avoids PCI Express memory reads of PIs and CIs

- **IQ object locations**
  - IQ element array in host memory (typical)
  - IQ PI in PQI device memory space (always)
  - IQ CI in host memory (typical)

- **OQ object locations**
  - OQ element array in host memory (typical)
  - OQ PI in host memory (typical)
  - OQ CI in PQI device memory space (always)



Queue Types

- Administrator queues
  - Created via PQI device registers
    - Located in PQI device memory space
  - Single administrator IQ and administrator OQ
    - i.e., one administrator queue pair
  - IUs defined by PQI
- Operational queues
  - Created via PQI administrator functions
    - Delivered over the administrator queues
  - Any number of operational IQs and operational OQs
    - Not in pairs
    - Can be specific to different cores
  - IUs defined by the information unit layer standard
    - e.g., SCSI over PCI Express (SOP)

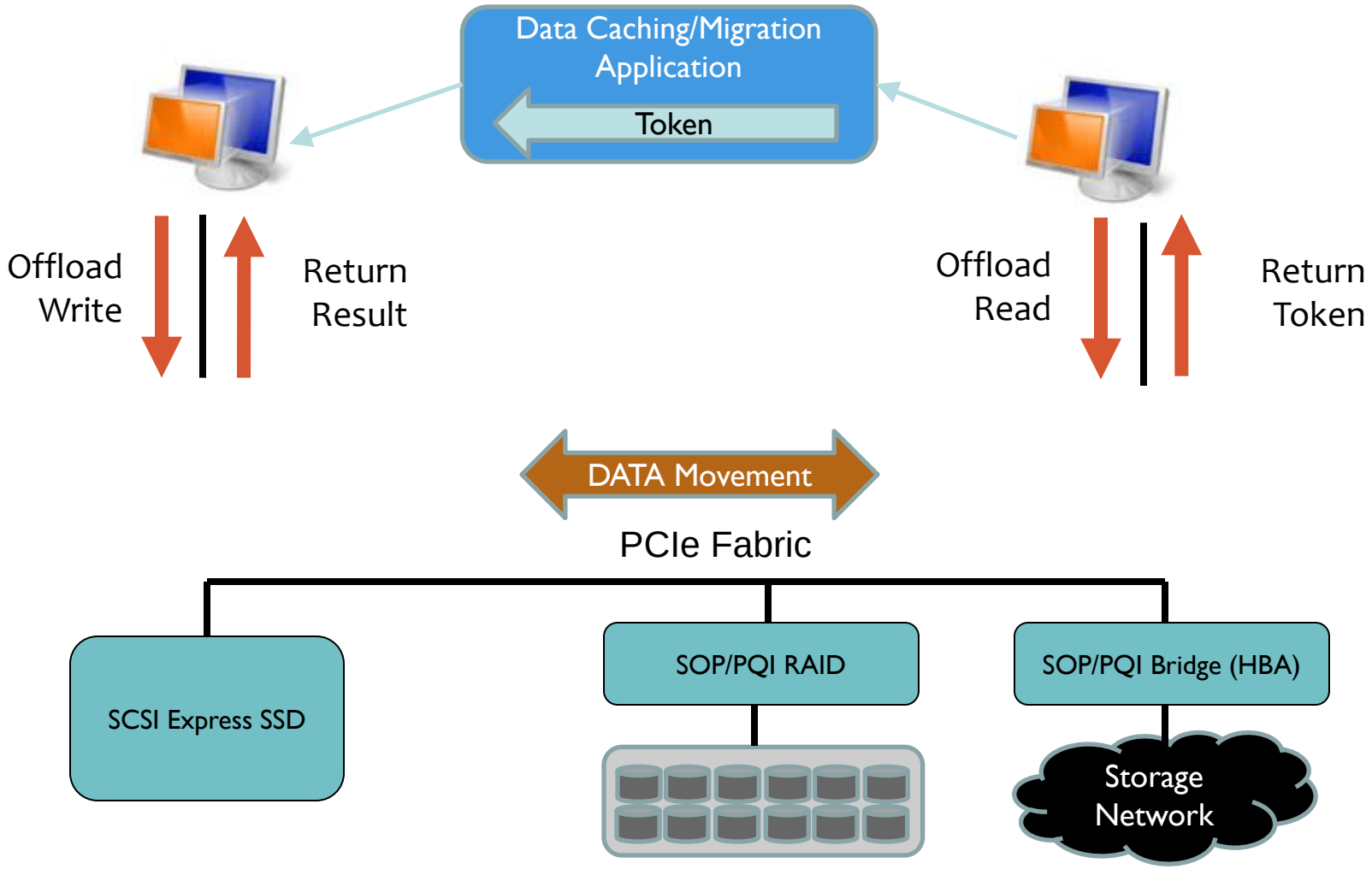
Key PQI Features

- Interrupt Coalescing
  - Single interrupt for multiple queue entries
  - Tuning via; count, min and max times
- Scatter Gather Lists (SGL)
  - Describes a data buffer
    - How it is distributed across non contiguous chunks of memory
  - Can be embedded in IUs or a separate list
  - Widely supported method across multiple OSs

SCSI – Looking to the Future

- SCSI Express
- 12Gb/s SAS, Multilink, 24Gb/s SAS
  - Performance and scalability
- Power Limit Control - up to 25W devices
  - Both SAS and SCSI Express
- Extended Copy Feature
- Atomic Writes
- Hinting & other NVM features

Extended Copy - Connecting the Tiers



Atomic Writes

- All or nothing written capability across multiple commands
  - For single commands and across non contiguous LBA ranges
- Benefits:
  - Simplifies resilient system designs
    - Database, file system, etc.
  - Improves system performance in these applications

Hinting & Other NVM Features

- Pass "hints" to devices to make operations more efficient and increase performance
  - Targeted at SSDs and hybrid drives, but also useful for HDDs
  - Device modifies how data is stored based on type
- Direct attached devices don't need to continually OPEN and CLOSE connections
  - Can be implemented within the existing standard
  - Reduces latency on both SSDs and HDDs
- NVM features and programming interfaces
  - Leverages ongoing work in SNIA and T10

SCSI Express Summary

- Proven SCSI protocol combined with PCIe creating an industry standard path to PCIe-based storage

- Enterprise storage for PCIe based storage devices
- Increased performance through lower latency
- Coexistence with SAS via Express Bay and common command set
- Unified management and programming interface

Attribution & Feedback

The SNIA Education Committee thanks the following individuals for their contributions to this Tutorial.

Authorship History

8/28/2012 Marty Czekalski
8/24/2012 Mike James

Updates:
9/3/2013 Marty Czekalski
3/20/2014 Marty Czkalski

Additional Contributors

David Allen
Rob Elliott
Ie-Wei Njoo
Bret Gibbs
Mike James
Joe Foster
Greg McSorley
STA Members

Please send any questions or comments regarding this SNIA Tutorial to ***tracktutorials@snia.org***