SCSI Standards and Technology Update

Marty Czekalski

President, SCSI Trade Association

Interface and Emerging Architecture Program Manager - Seagate Technology

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

SCSI Standards and Technology Update

SCSI continues to be the backbone of enterprise storage deployments and has rapidly evolved by adding new features, capabilities, and performance enhancements.

This talk will include an up-to-the-minute recap of the latest additions to the SAS standard and roadmaps. It will focus on the status of 12Gb/s SAS staging, advanced connectivity solutions such as MultiLink SAS™ and cover SCSI Express, a new transport of SOP (SCSI over PCIe). Presenters will also provide updates on new SCSI feature such as atomic writes, remote copy, and initial work on 24Gb/s SAS.

SCSI Standards and Technology Update

SNIA Education

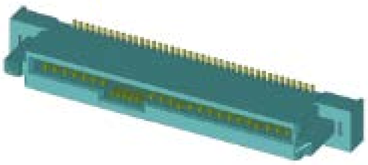
- Optimized solid state SCSI initiative
  - SCSI Express
- New SCSI features for FLASH and performance
- Express Bay
- Beyond 12Gb/s SAS
- Zoned Block Commands

Serial Attached SCSI

SCSI SF

Optics

SCSI EXPRESS

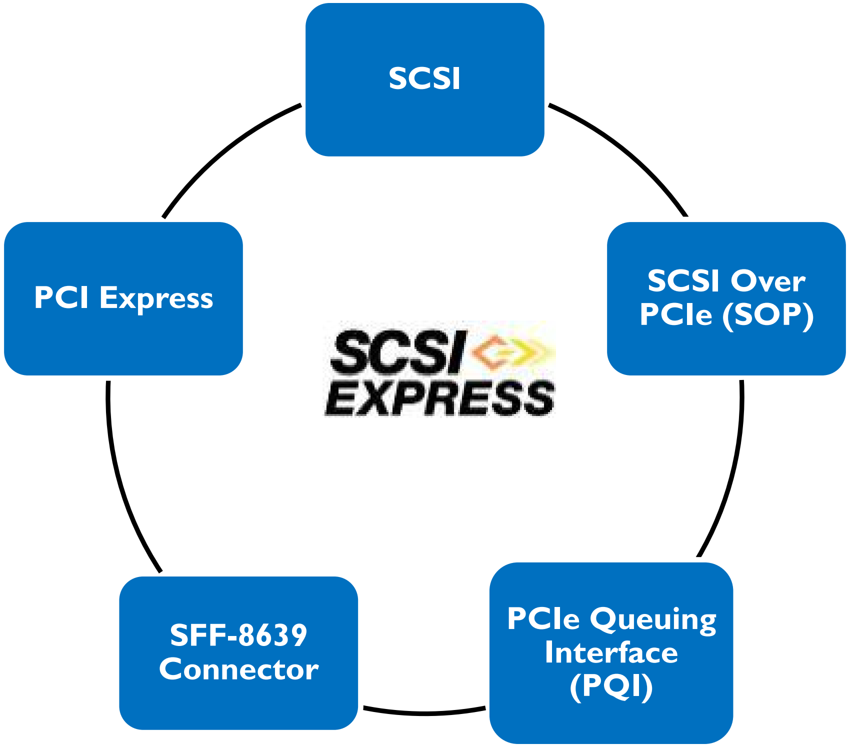


SCSI Express Overview

- 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?
  - Delivers proven enterprise storage for PCIe-based storage devices in a standardized ecosystem
  - Takes advantage of lower latency PCIe to improve performance
  - Offers 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

SOP/PQI Status

- Not just for SSD
  - Can be used for devices, HBAs (bridges), and RAID controllers
- 1.0 versions of the specifications completed in T10
  - Undergoing finalization/publication process in INCITS

New SCSI Features for FLASH and Performance

- Extended copy feature
- Atomic Writes
- Hinting
- SCSI-SF
- Power limit control

Extended Copy Using Tokens

- New feature allowing direct movement of data between storage devices on the same fabric
  - Leverages the ability of SCSI devices to act as both an initiator and a target
- Greatly improves performance
- Greatly reduces overhead
  - Eliminates multiple passes of data over PCIe
  - Eliminates use of system memory as a buffer

Atomic Operations

- Atomic Write – all or nothing is written
  - For single commands and across non-contiguous LBA ranges (Scatter)
- Atomic Read - data read is consistent at a point in time
  - No partial updates in process
  - Multiple extents (Gather)
- Benefits:
  - Simplifies resilient system designs
    - Database, file system, etc.
  - Improves system performance in these applications

Atomic Writes

- Single extent Atomic Writes
  - Proposal 14-043r4 accepted for inclusion in SBC-4
- Scatter/Gather, Writes/Reads with option for Atomic still in discussion
  - 12-086r5/12-087r5 latest versions
  - R6 version documents assigned, but not uploaded yet
  - Improved efficiencies for database and file systems
  - More discussion needed, acceptance timeframe TBD
- Granular Atomic Operations 14-034r2
  - Granularity – Command field specifying allowed granularity
  - Maximum Atomic transfer length and Atomic granularity size attributes – VPD page attributes
  - Nearly complete, expect acceptance at next T10 meeting cycle

Atomic Operations

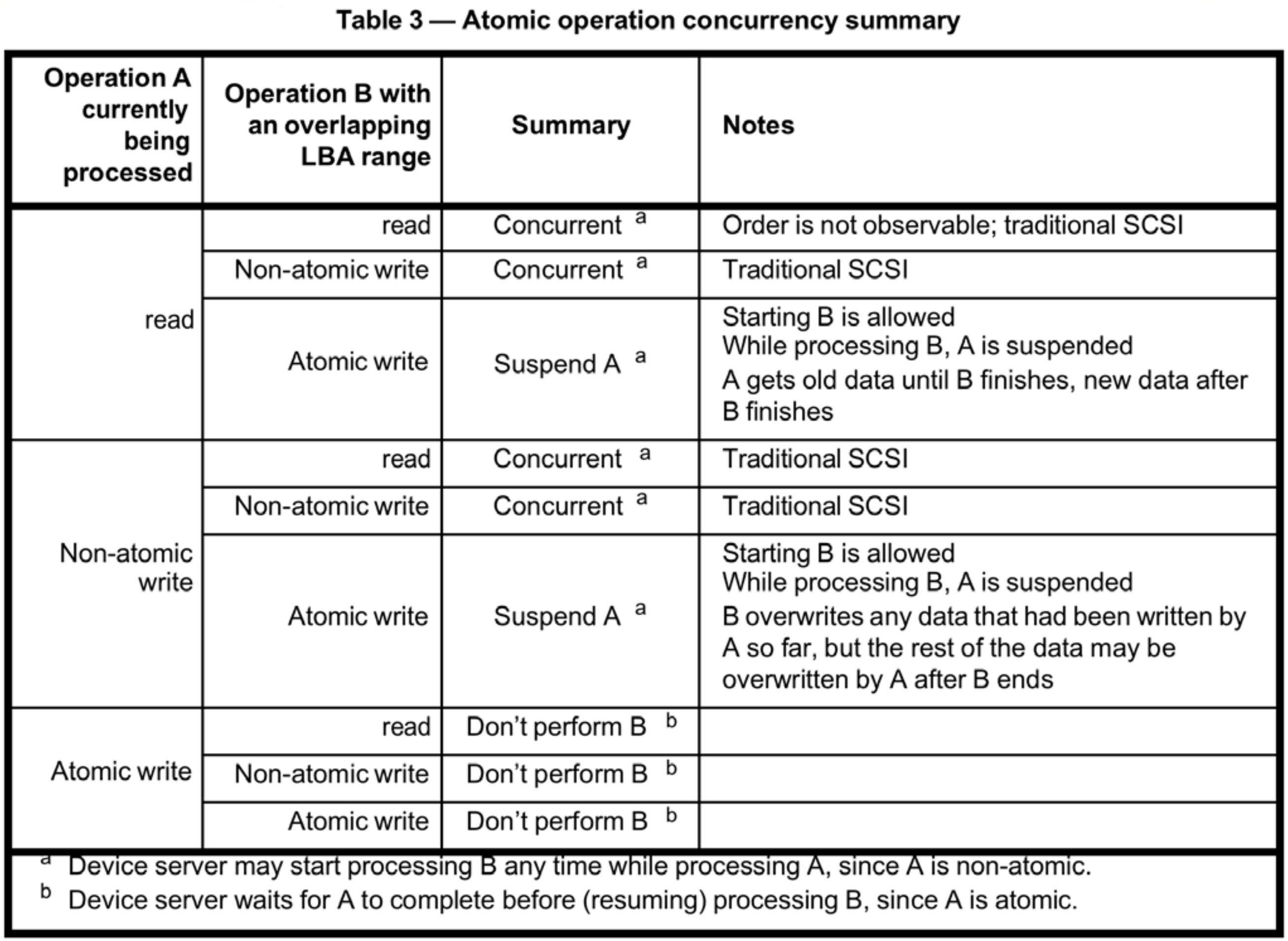


Table 3 — Atomic operation concurrency summary

Operation A currently being processed	Operation B with an overlapping LBA range	Summary	Notes
read	read	Concurrent [a]	Order is not observable; traditional SCSI
	Non-atomic write	Concurrent [a]	Traditional SCSI
	Atomic write	Suspend A [a]	Starting B is allowed While processing B, A is suspended A gets old data until B finishes, new data after B finishes
Non-atomic write	read	Concurrent [a]	Traditional SCSI
	Non-atomic write	Concurrent [a]	Traditional SCSI
	Atomic write	Suspend A [a]	Starting B is allowed While processing B, A is suspended B overwrites any data that had been written by A so far, but the rest of the data may be overwritten by A after B ends
Atomic write	read	Don't perform B [b]	
	Non-atomic write	Don't perform B [b]	
	Atomic write	Don't perform B [b]	

[a] Device server may start processing B any time while processing A, since A is non-atomic.

[b] Device server waits for A to complete before (resuming) processing B, since A is atomic.

From T10 proposal 14-043r4

Logical Block Markup Descriptor (Hinting)

- Hinting proposal reworked and is now called Logical Block Markup Descriptors (14-052r6)
- Intended to be a consistent interface across SBC-4, and ACS-4
  - Will likely be placed as an annex in SAT-4 and referenced by both standards
- Access Patterns
  - Overall Frequency, Read/Write Frequency, Write Sequentially, Read Sequentially, Subsequent I/O, OSI Proximity

Forced UNMAP (now Write Zeros)

- Originally started as Forced Unmap, but after much discussion the desired behavior could be satisfied with a Write Zeros command (14-071r0).
  - Combine the command interface format of the UNMAP and the logical block provisioning and write properties of the WRITE SAME command
- T13 will also incorporate an equivalent command

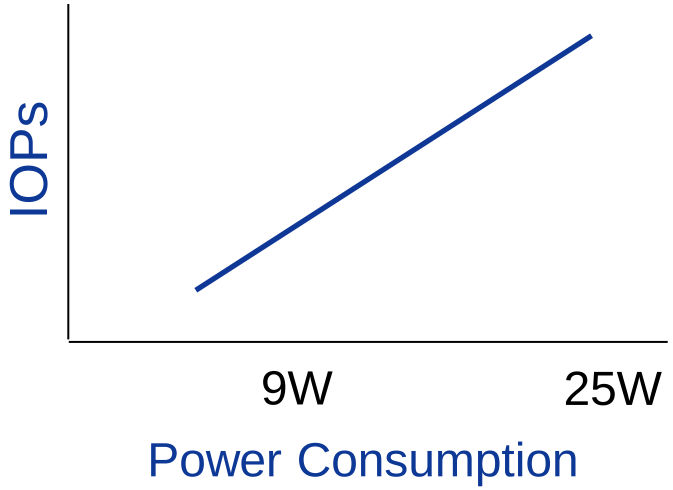
SCSI-SF (Simplified Features)

- SCSI contains a rich feature set with multiple methods and options
- SCSI-SF is targeted as a common subset of features for increased efficiency of implementations, qualifications and maintenance
- Little progress, completion TBD

Power Limit Control

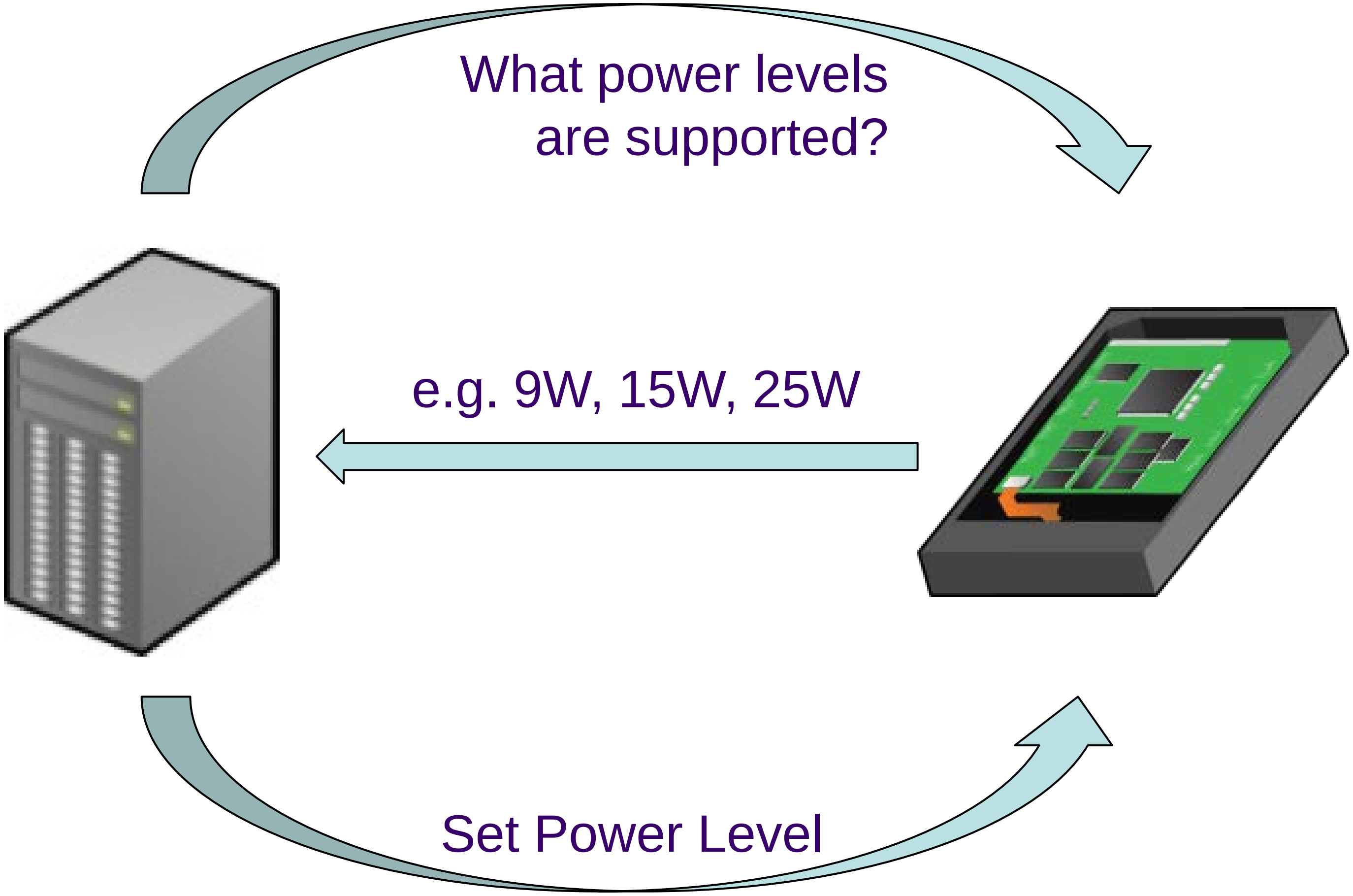
- $P \cong P_{base} + P_{I/O} * IOPs$

- Power Limit Control
  - Allows system and device to negotiate allowable power usage
  - Both SAS and PCIe have this capability



- For more bandwidth, additional links are needed

Power Limit Control



What power levels are supported?

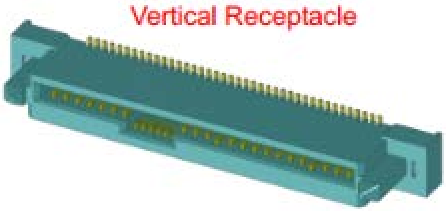
e.g. 9W, 15W, 25W

Set Power Level

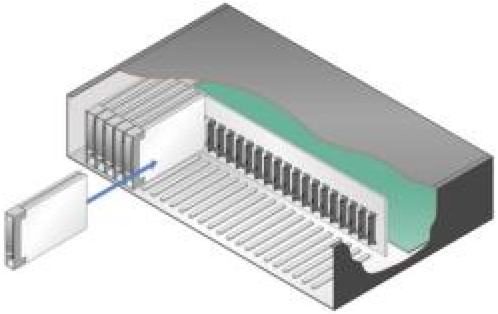
Express Bay Components

25W Power

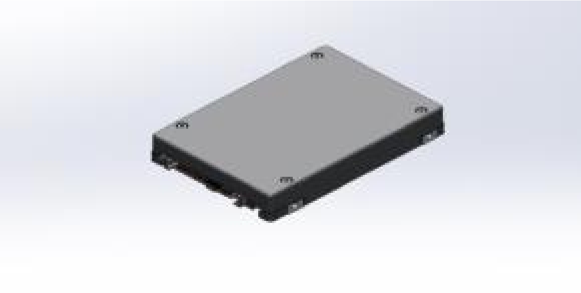
Cooling



Multifunction Connector

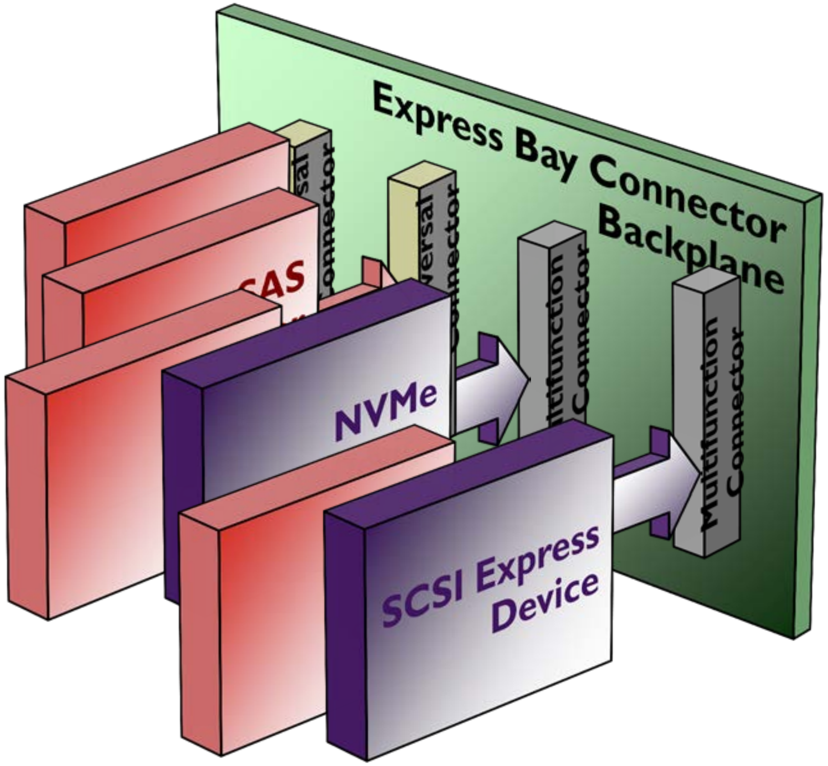


Accessibility / Serviceability



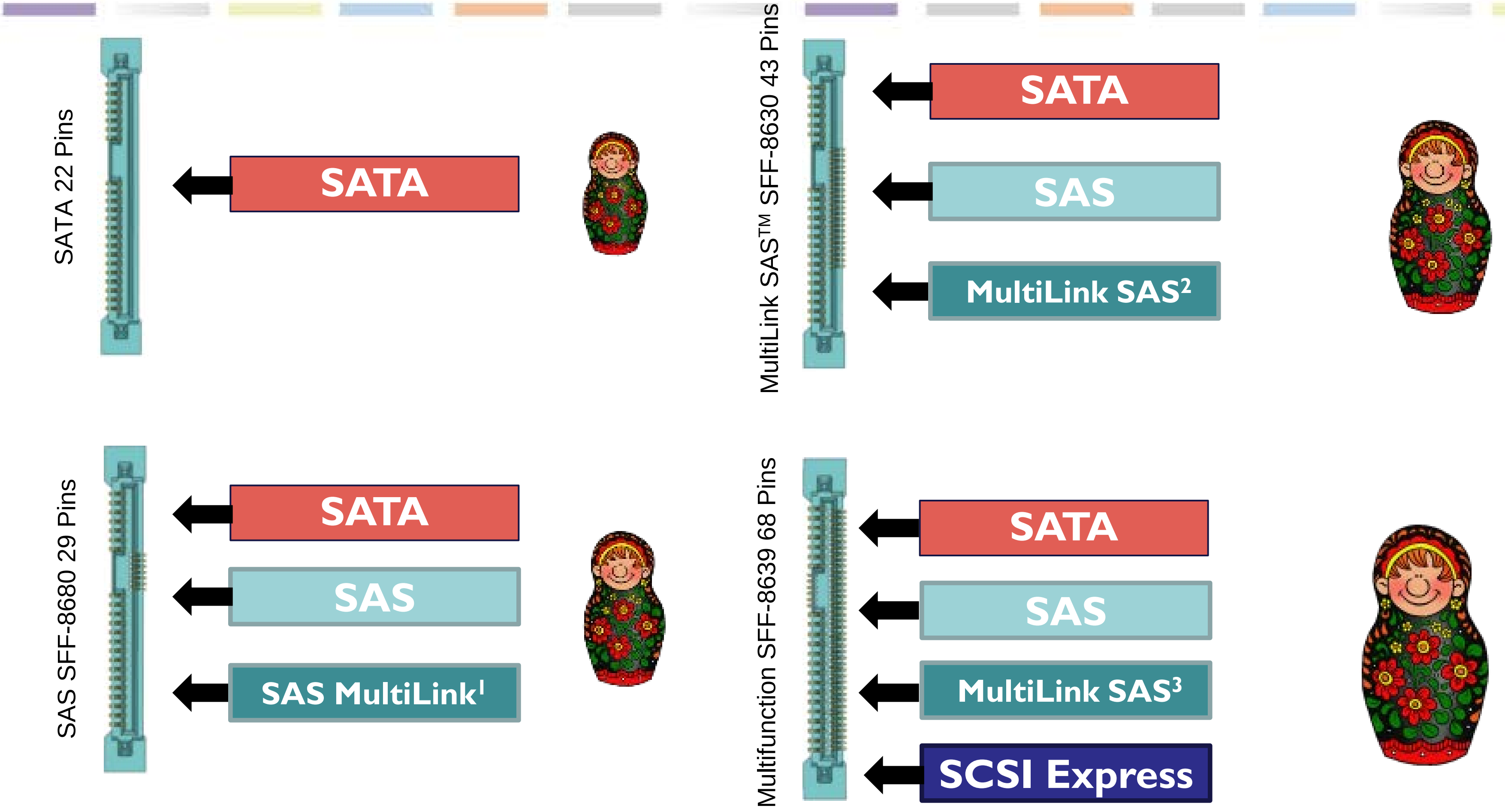
Traditional Drive Form Factor

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



SATA 22 Pins

SATA

SAS SFF-8680 29 Pins

SATA

SAS

SAS MultiLink[1]

MultiLink SAS™ SFF-8630 43 Pins

SATA

SAS

MultiLink SAS[2]

Multifunction SFF-8639 68 Pins

SATA

SAS

MultiLink SAS[3]

SCSI Express

[1] Max two links operational

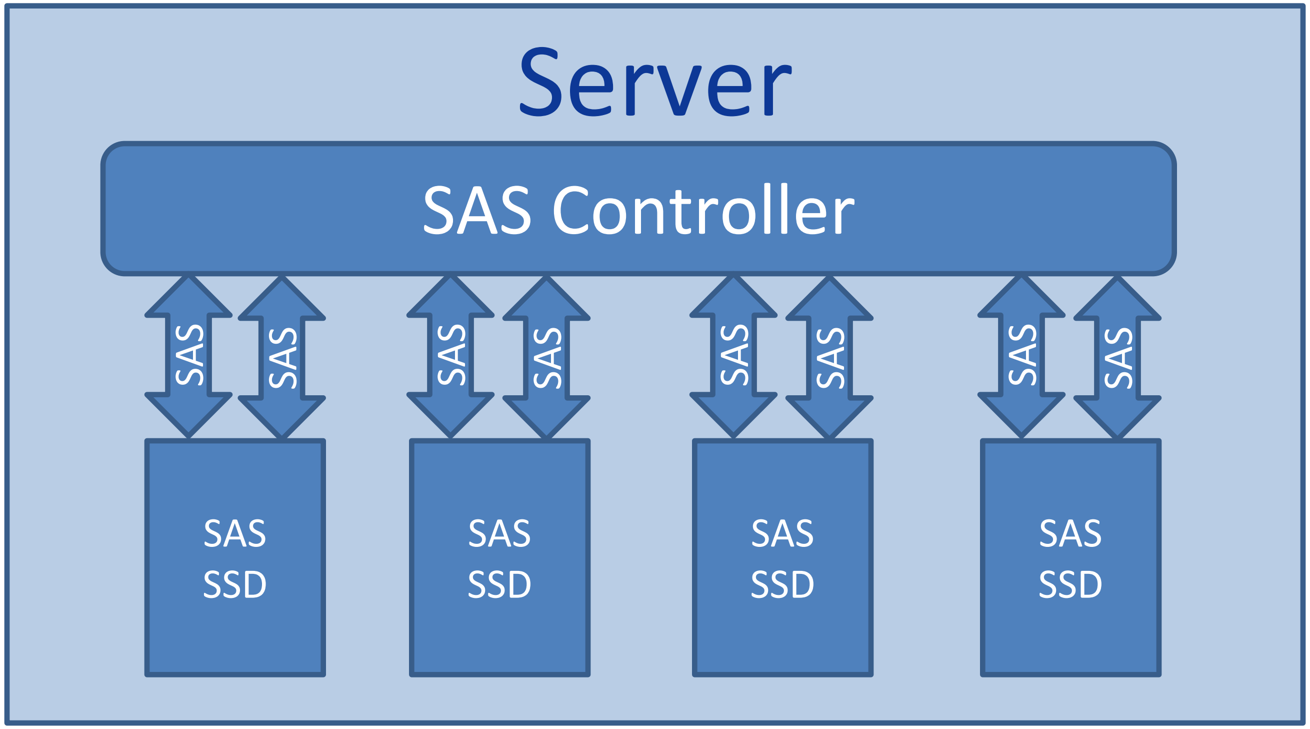
[2] Four links operational

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

Bandwidth per Device Connected

	SATA	SAS	Wide-port SAS	PCIe
No of Ports / Lanes	1	1	2	4
Transfer Rate per Port/Lane	Half-duplex 6 Gb/s	Full-duplex 12 Gb/s	Full-duplex 12 Gb/s	Full-duplex 8 Gb/s
Max Bandwidth	0.6 GB/s	2.4 GB/s	4.8 GB/s	8 GB/s
Interface 4KB Random I/O Capability	100K	450K	900K	1500K

Wide Port SAS for Increased Throughput



2.4 GB/s full-duplex per SSD

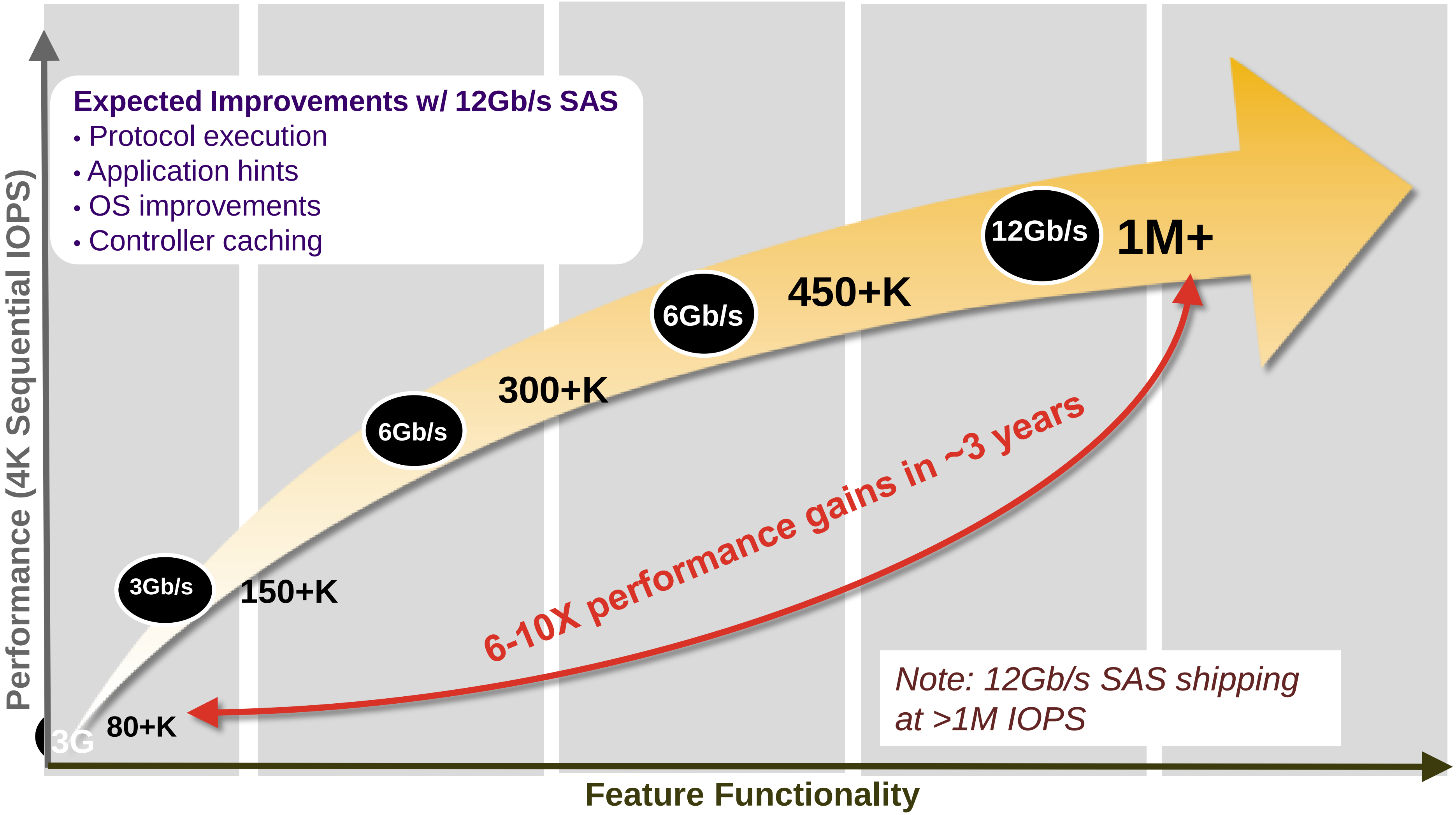
Express Bay Summary

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

Beyond 12Gb/s – SAS Continues to Evolve

- SAS roadmap moving forward
  - 24Gb/s SAS specification in development
  - Preserve 6Gb/s SAS, 12Gb/s SAS, and 6 Gb/s SATA usage models and compatibility
  - 12Gb/s SAS controllers now shipping at >1 million IOPs
- Proven reliability, high availability, and serviceability
- SAS ecosystem in place
  - Test & measurement equipment
  - Internal & external connectors and cabling
  - HDDs and SSDs shipping today

SAS Continues to Evolve – Performance Gains without Protocol Changes

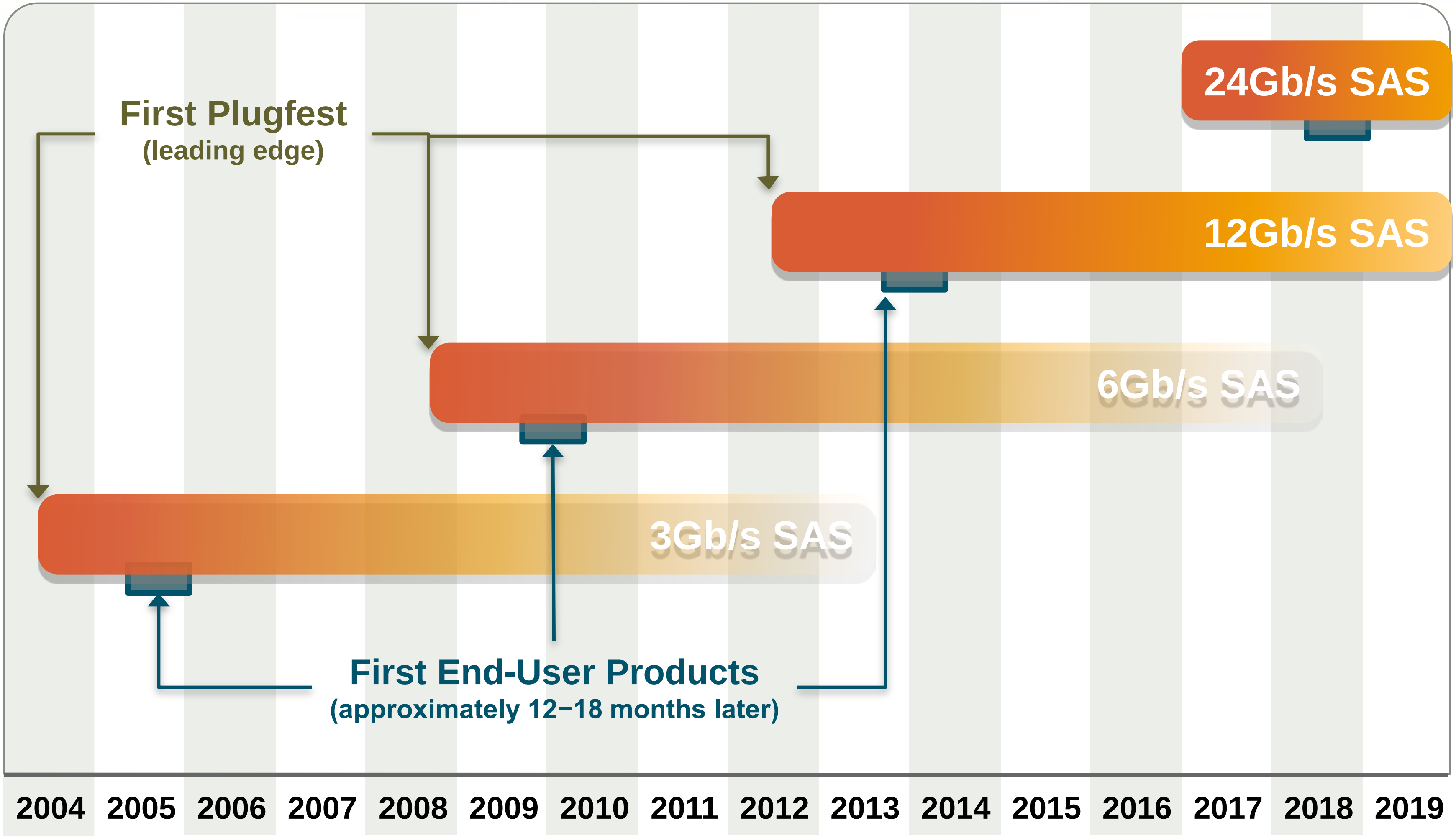


Expected Improvements w/ 12Gb/s SAS

- Protocol execution
- Application hints
- OS improvements
- Controller caching

Note: 12Gb/s SAS shipping at >1M IOPS

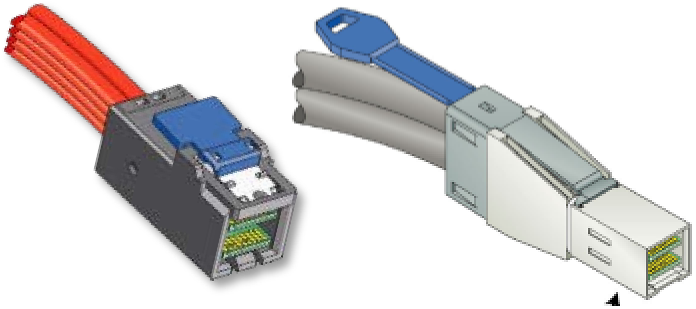
SAS Roadmap



** SAS Roadmap –SCSI Trade Association –March 2014*

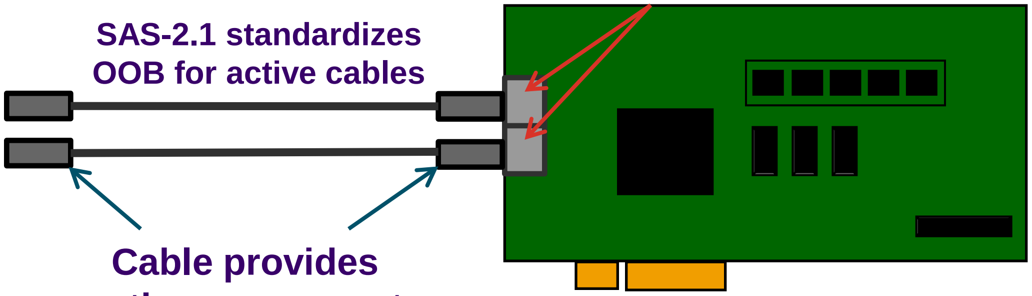
12Gb/s SAS External Interconnect

- Drive market consistency
- Simplified cable & connector options
- 2X density improvement
- Passive copper to 7m
- Active copper solution to 20m
- Active Optical (AOC) solution to 100m
- Managed connectivity standards



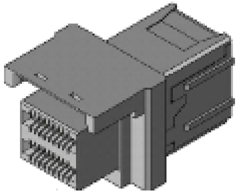
Internal similar to External

Supply Power for active cabling



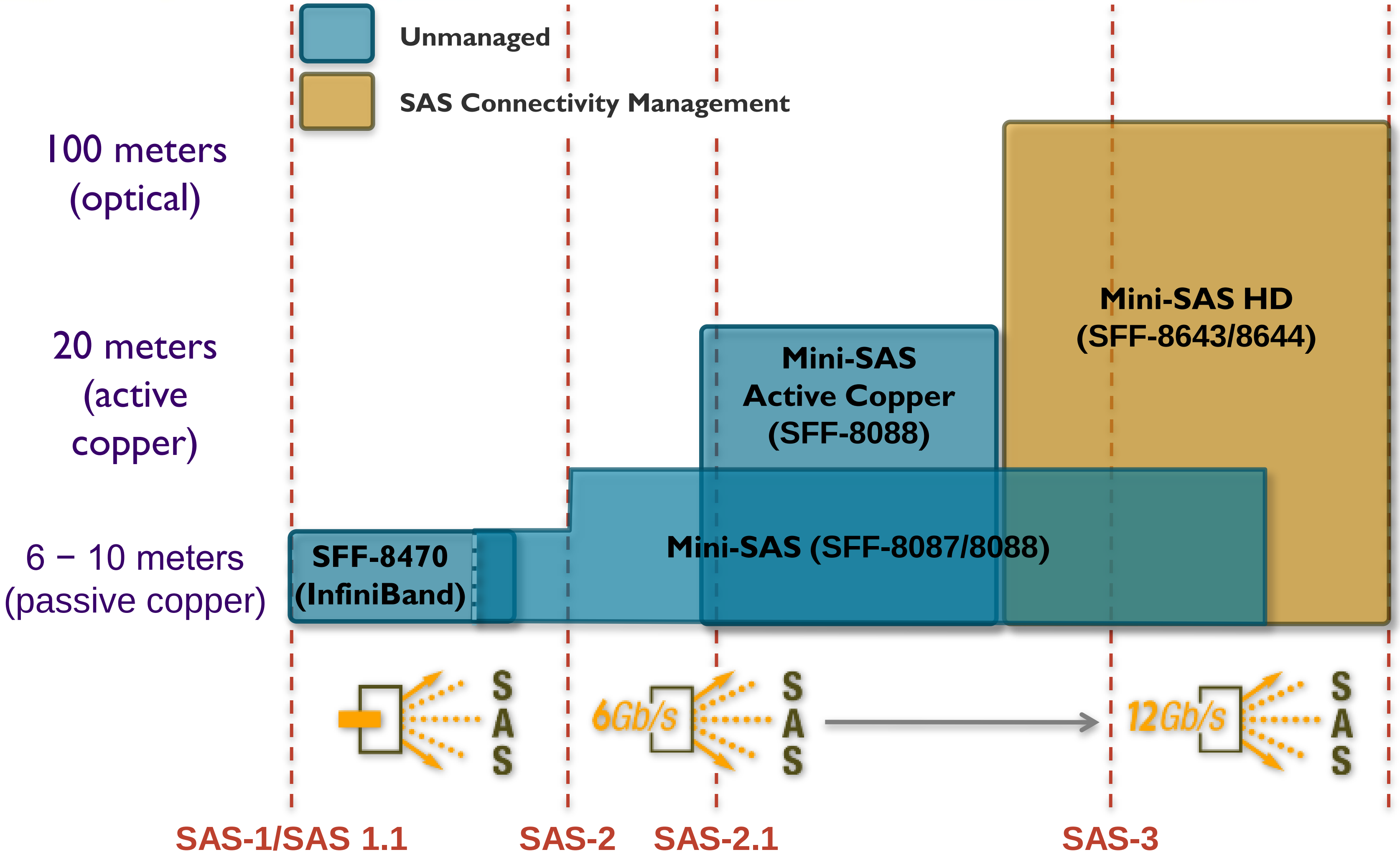
SAS-2.1 standardizes OOB for active cables

Cable provides active component for optical or copper



Passive, Active Copper, or Optical use same connector

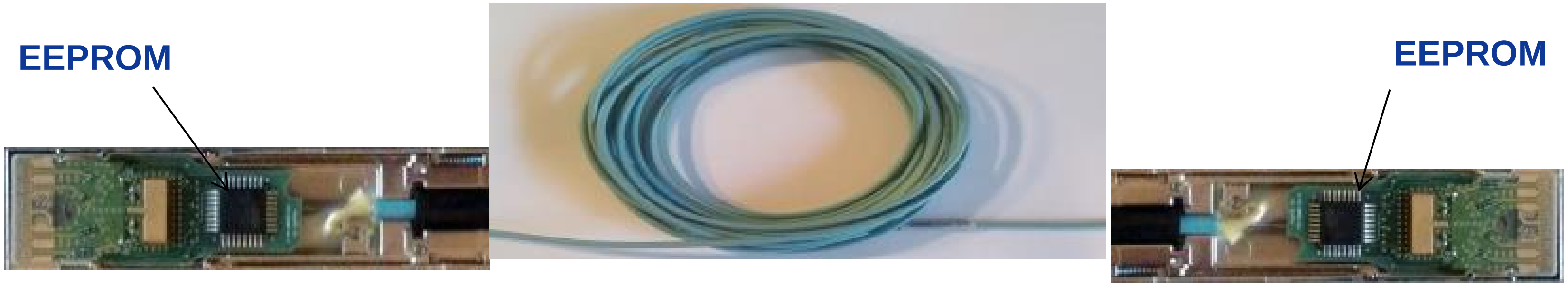
SAS Advanced Connectivity Roadmap


Unmanaged
SAS Connectivity Management
100 meters (optical)
20 meters (active copper)
6 – 10 meters (passive copper)
SFF-8470 (InfiniBand)
Mini-SAS (SFF-8087/8088)
Mini-SAS Active Copper (SFF-8088)
Mini-SAS HD (SFF-8643/8644)
SAS
6Gb/s
SAS
12Gb/s
SAS
SAS-1/SAS 1.1
SAS-2
SAS-2.1
SAS-3

SCSI Standards and Technology Update

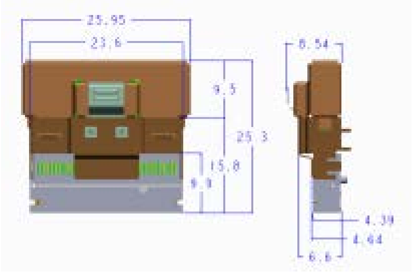
Managed Cable System

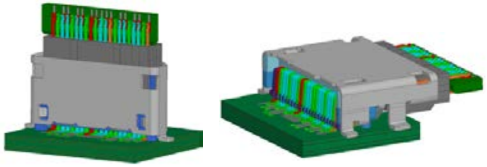
- New to SAS
- Managed cables simplify configuration and ease of use
- OoB (Out of Band) method of controlling the interface
- Every pluggable device has an EEPROM or microprocessor that communicates with the system via a low speed, two wire interface.
- Allows each port to support short passive copper cables to 100m active optical cables



New Mini Internal SAS Connector

- Servers are becoming much more constrained internally for space
  - Use of flat cable assemblies highly desirable
  - Small footprint and connector height constraints
- Two connectors currently under discussion
  - T10 doc # 13-236r1, SFF-8654
    - Eight links (x8), approx. 4.5 x 24 mm size
    - Four links (x4), approx. 4.5 x 13 mm size
  - T10 doc # 14-202r0
    - Four links (x4), approx. 3 x 13 mm size
  - Additional sideband pins
  - 24Gb/s capable target
- Target decision at next T10 meeting cycle
  - Expect use in servers currently under design at 12Gb/s SAS



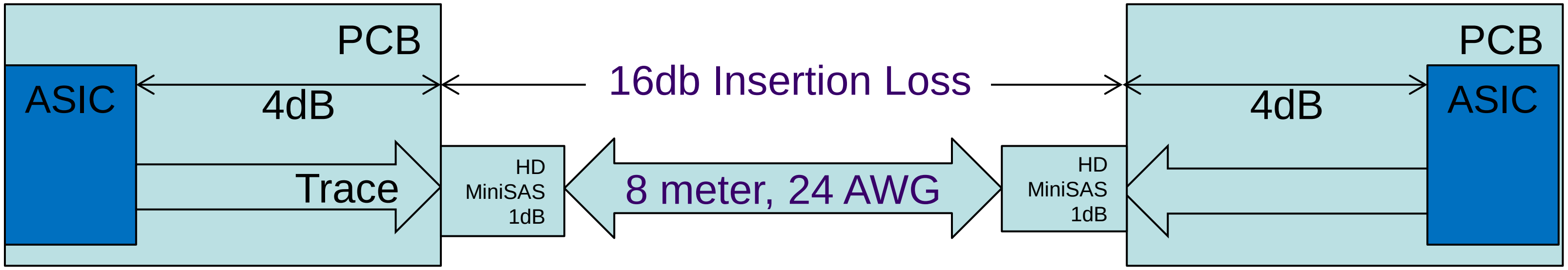


STA 24Gb/s SAS MRD

- Preserve existing SAS architecture
- Continue 6Gb/s SATA compatibility
- Maintain and support SAS backward compatibility
  - Must be backward compatible two generations: 6Gb/s SAS and 12Gb/s SAS
- Maximize link utilization when using devices operating at less than 24Gb/s
- Encourage improved storage system RAS attributes
- Double the transfer rate

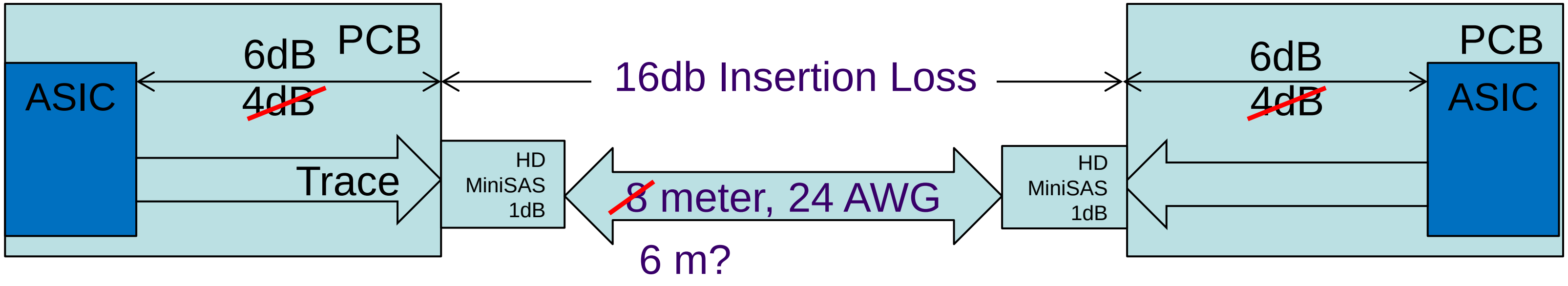
Basic Link Budgets

SAS 3.0 Total Insertion Loss = 24dB



Higher Frequency = More Insertion Loss

SAS 4.0 Total Insertion Loss = 28dB



Encoding vs. Line Rate

Encoding	Line Rate
8b10b	24.0 Gbps
64b66b	19.8 Gbps
128b130b	19.5 Gbps
256b257b	19.28 Gbps
512b513b	19.24 Gbps
1024b1025b	19.22 Gbps

- Longer encoding lengths offer similar bandwidth efficiencies and yield minimal reduction in line rate
- Longer encoding lengths increase buffering requirements and increase protocol handshake latency
- SAS-4 line rate range should be 19.5Gbps to 24Gbps

Forward Error Correction Performance (Last year's investigation)

	Encoding	FEC Bits	Overall coding length (bits)	Line Rate[1]	SI Gain[2] @ BER of 1e-15[3]	FEC Latency Adder[4]
	8b10b	0	8b10b	24.0 Gbps	0	0
	64b66b	0	64b66b	19.8 Gbps	0	0
	128b130b	0	128b130b	19.5 Gbps	0	0
✗	64b66b	14[5]	64b80b	24.0 Gbps	5.8 dB	~2.7 ns
✗	128b130b	16[5]	128b146b	21.9 Gbps	5.8 dB	~5.3 ns
✗	256b257b	18[5]	256b275b	20.63 Gbps	5.6 dB	~10.6 ns
✗	512b513b	20[5]	512b533b	19.99 Gbps	5.6 dB	~21.2 ns
✗	1024b1025b	88[6]	1024b1113b	20.87 Gbps	7.4 dB	~53.2 ns

[1] Raw data throughput of 19.2Gb/s.

[2] SI gain is addition IL that the system can tolerate (~2x the FEC gain at the slicer)

[3] Assumes 1e-15 as a target BER.

[4] Additional latency imposed by use of FEC

[5] Differential encoding and BCH algorithm for FEC.

[6] Reed-Solomon algorithm with T=4 for FEC.

Current Direction on FEC vs. Materials

- Determined that DFE has significant impact on FEC
  - DFE tends to create burst errors that are not well matched to BCH codes; RS codes are better
- A well-designed 30 dB channel does not require FEC to meet BER 10^{-15}
  - OK for cables; backplanes will be more challenging
- Current choices (debate still underway)
  - No FEC, 128/130b encoding
    - Use better dielectric materials where needed
  - RS over multiple 128/130b frames (adds latency)
    - 2 frames (RS 300, 260) buys between 1-3 dB
    - 3 frames (RS 450, 390) buys between 3-5 dB
    - 4 frames (RS 586, 520) buys between 5-7 dB

Connectors for 24Gb/s SAS

- Mini SAS HD
  - Looks good, multiple vendors see a path to solutions
- Mini SAS
  - Not likely acceptable for 24Gb/s SAS
- QSFP+
  - Already 24Gb/s SAS capable
- New Internal mini connector
  - 24Gb/s SAS is a requirement for choosing the connector
- Drive connector
  - Looks encouraging, but margins tight

Primitive Encoding

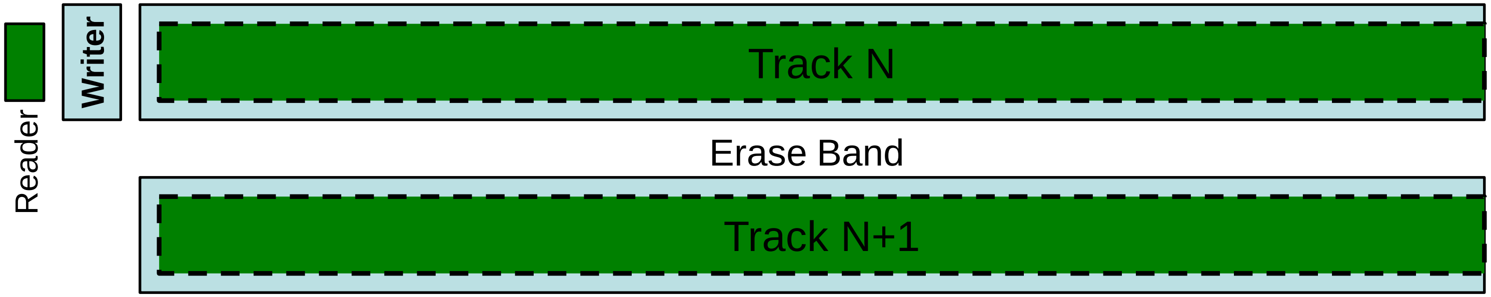
- 8b/10 Primitive Dwords (40 bits), become Primitive Packets (130 bits)
  - Primitive packet includes four “truncated’ primitives
- No differential encoding
- No scrambling of primitives
  - Use scrambled IDE segments for: Rate Match, Physical Link Rate Tolerance Management, idle periods
- INIT_SCRAM and INIT_SCRAM2
  - Synchronize scrambling phase for IDLE and SPL segments
  - Provide alignment of SAS-4 Packets
  - Provide for rapid resynchronization

24Gb/s SAS Summary

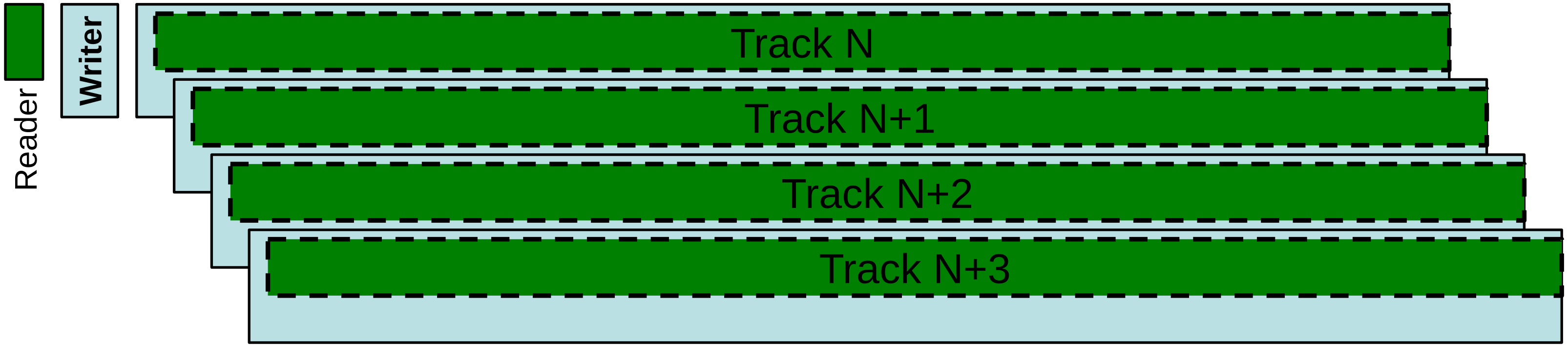
- 24Gb/s SAS is definitely possible
- 128/130 encoding
  - Primitive encoding scheme decided and being documented
- FEC vs. better board materials still under investigation
- Connector SI studies ongoing and nearing completion

Conventional versus SMR Writing

Conventional Writes



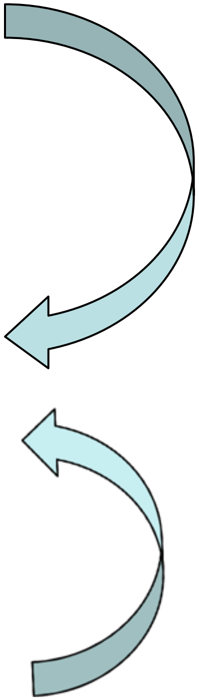
SMR Writes



SMR allows much higher track density and higher areal density growth rates

Overview of SMR Drive Types

- Motivated by Shingled Magnetic Recording, but interest expressed from SSD community
  - Current draft ZBC-r01a
- Drive Managed
  - Drive autonomously hides all SMR issues
  - Workloads can affect performance
- Host Aware
  - Superset of Drive Managed and Host Managed
  - Backward compatible
  - Extensions to ATA and SCSI command sets
- Host Managed
  - New device type
  - Extensions to ATA and SCSI command sets
  - Error conditions for some reads and writes
  - Not backward compatible



Comparison of ZBC Device Types

Style	SCSI Peripheral Device Type	ATA Device Signature	Zone Types	New Commands	New Rules
Drive Managed	00h: Direct Access Device	ATA	None	None	None
Host Aware	00h: Direct Access Device (with Host Aware flag)	ATA	Sequential Write *Preferred* and Conventional*	• Report Zones • Reset Write Pointer	None
Host Managed	14h: Host Managed Zoned Block Device	Host Managed Zoned	Sequential Write *Required* and Conventional*	• Report Zones • Reset Write Pointer	• No random writes to WP zones • No reads of unwritten data • Etc.

For more in-depth information see SNIA tutorial:
Shingled Magnetic Recording Models, Standardization, and Applications
by Tim Feldman and Mary Dunn

*optional

Summary

- SCSI Standards continue to evolve and adapt
- New features for performance and efficiency being added
- Proven stable protocol
- Don’t wait for final specification releases to implement features
- Follow T10 activities to ensure products meet current standards and take advantage of new features

Attribution & Feedback

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

Authorship History

Marty Czekalski

Updates:

Additional Contributors

Joe Foster
Rob Elliot
Mike James
Dave Allen
Greg McSorley
Tim Symons
Tim Feldman
Mary Dunn
STA members

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