DATA IS POTENTIAL

NVMe/PCIe - HDD Ramifications

OCP Discussion of Options for PCIe/NVMe HDD Devices

Jon Trantham, Seagate Research

January 23rd, 2019

SEAGATE

Legal Disclaimers

Disclaimer

This document is provided as is with no warranties whatsoever, including any warranty of merchantability, non-infringement, fitness for any particular purpose, or any warranty otherwise arising out of this proposal. Seagate disclaims all liability for infringement of proprietary rights, relating to use of information in this proposal. This is only a proposal, and as such is subject to change. No license, express or implied, by estoppel or otherwise, to any intellectual property rights is granted herein. Opinions expressed herein are from Seagate and may not reflect those of other drive suppliers.

Trademarks

PCI, PCI Express, PCIe Express U.2, PCIe, ExpressModule, and PCI-SIG are trademarks or registered trademarks of PCI SIG. SNIA is a trademark of the Storage Networking Industry Association.

Other company and product names may be trademarks of the respective companies with which they are associated.

Definitions, as used in this presentation:

Definitions:

- **SFF-8639**: refers to the SFF Committee-defined SFF-8639 connector (not its wiring)
- **“U.2”**: generally refers to the “PCIe Express U.2™” implementation as defined in:
  - *PCI Express SFF-8639 Module Specification Rev. 3.0, V1.0* by the *PCI-SIG*
  - *and earlier in “Enterprise SSD Form Factor” in Versions 1.0A* by the *SSD Form Factor Work Group*

PCI Express U.2™ is a registered service mark of PCI-SIG.

- **“U.3”**: refers to the PCIe implementation as defined in:
  - *SFF-TA-1001, Rev 1.0A*

Note: SFF-TA-1001, as written, additionally requires U.2 backwards compatibility support

- *U.3* in this document refers only to the new SAS/PCIe common pinout defined in SFF-TA-1001

Executive Summary

- NVMe-interface HDDs will likely have value for OCP systems
- Including all current PCIe/NVMe options in HDDs will be costly
  - OCP systems will likely not use most of the options that add costs
- We recommend forming a task group to develop OCP NVMe HDD standards

NVMe Implementation Tradeoffs on HDDs

SYSTEM DESIGN IMPACT / CHANGES

LOWER COSTS

The main question:

Do we prefer no changes to current system designs, or to pursue lower-cost systems with some changes?

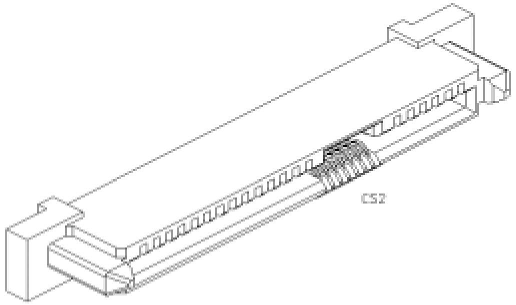
FEATURES IN QUESTION

NVMe Feature	SSD Path	HDD Pref	Purpose	Comments
Lane Count	4	2	Extra bandwidth	HDDs will only need up to 2 lanes of bandwidth for foreseeable future
Lane Config.	U.2, U.3, both	U.3	SSD Compatibility	Discussed later
Dual-Port Signal	1 x4 vs. 2 x2	1 x2 vs. 2 x1	Align with lane count	Should match lane count
Spread Control	Both Common Refclk & SRIS	TBD	PCIe spread spectrum	Don't expect systems will want clocks everywhere, but requires HBAs w/SRIS
Out-of-band I/O	SMBus	TBD	OOB mgmt. interface	Extra system signals & drive costs
Unpowered VPD	SMBus + EEPROM	TBD	System can query devices w/o main pwr.	Requires SMBus, EEPROM
Resets	H/W pins	TBD	Device / link reset	Extra system signals
Connector	SFF-8639	TBD	Needed for extra lanes	Discussed later
12V-only	12V-only	TBD	Costs/ less complexity	Most SSDs are 12V-only

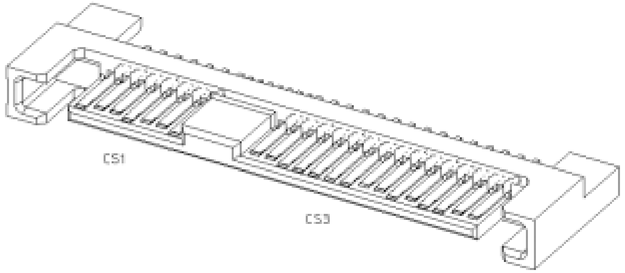
We would like to close / get to commonality on the TBDs

Background: Connectors

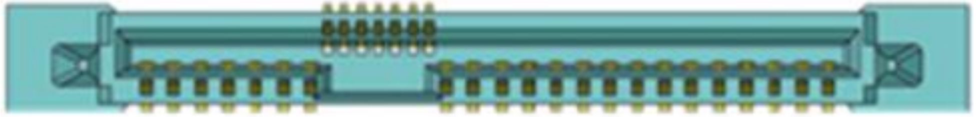
SFF-8482: Defines the connector used with SAS HDD/SSD drives

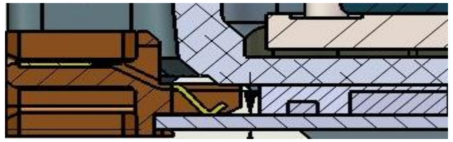


BACKPLANE FIXED (RECEPTACLE) CONNECTOR

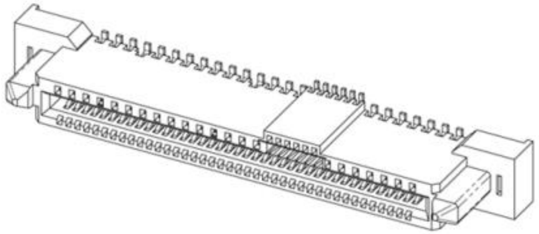


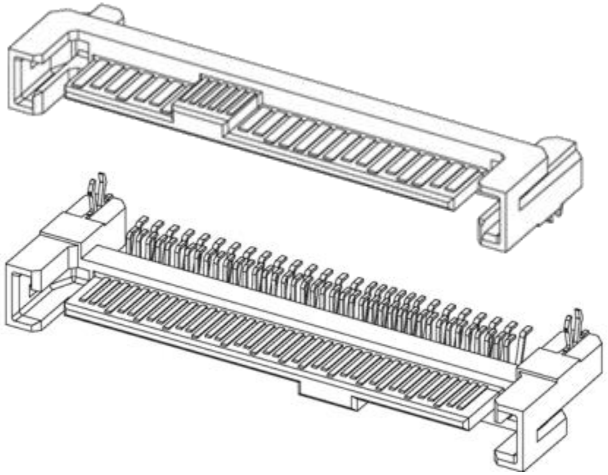
DEVICE FREE (PLUG) CONNECTOR

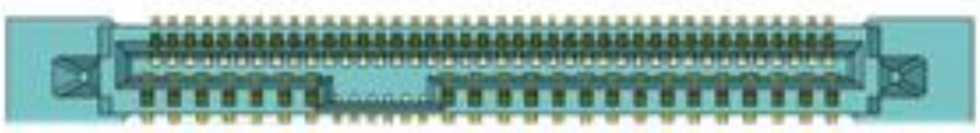


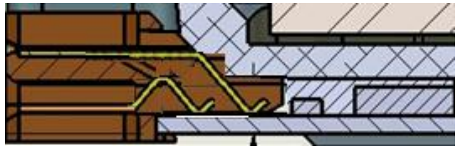


SFF-8639: Defines a connector commonly used on PCIe/NVMe SSDs









Background: U.2 vs. U.3 Lane Wiring

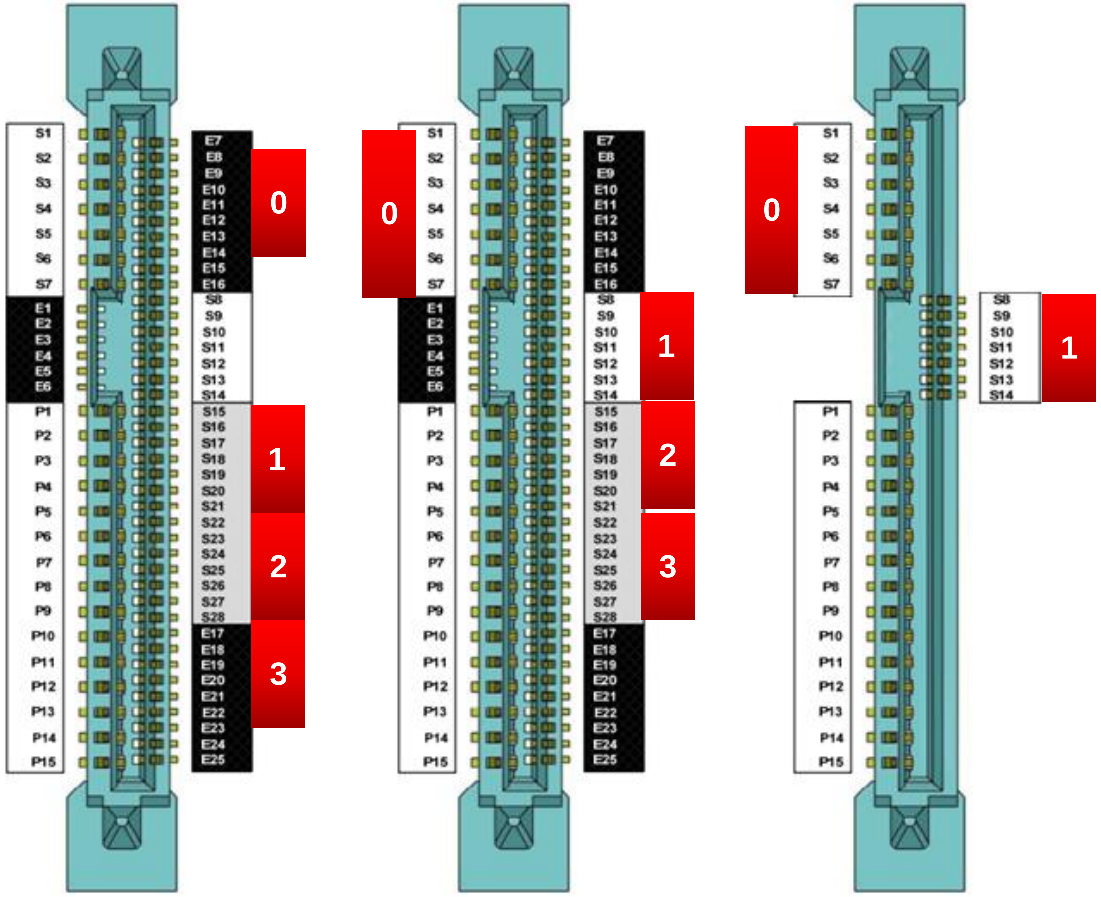
“U.2 wiring”:

- NVMe lanes are on **different pins** than SAS
- Popular wiring for many SSDs

“U.3 wiring”:

- NVMe lanes are on the **same** pins as SAS
- Works well w/tri-mode (NVMe, SAS, SATA) HBAs

Note: current U.3 specs also require U.2 support

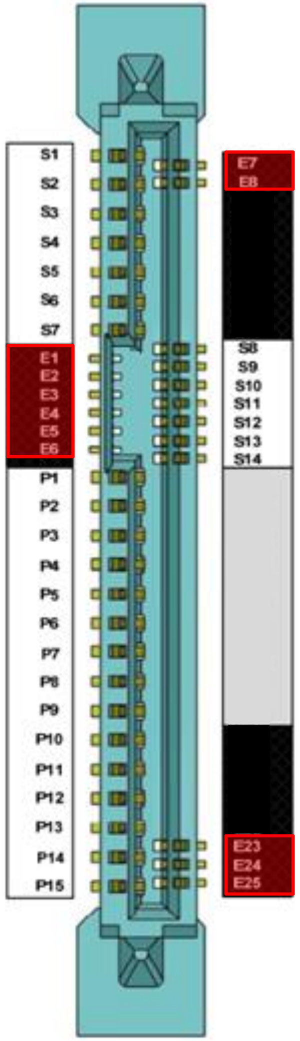


U.2

U.3

SAS

Background: U.3 SFF-8639 Pins Usable for 2-lane HDDs



Eleven new pins on the U.3 SFF-8639 connector would be needed for a U.3 HDD:

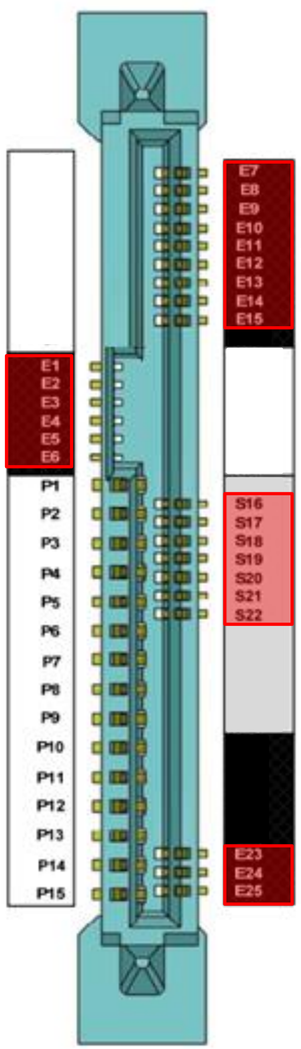
E1: REFCLKB+
E2: REFCLKB-
E3: +3.3Vaux (used to supply SMBus EEPROM)
E4: PERSTB#
E5: PERST#
E6: IFDet2#
E7: REFCLK+
E8: REFCLK-
E23: SMBCLK
E24: SMBDAT
E25: DualPortEn#

Twenty-Eight pins are not used / essential:

E9, E12, E15: Ground
E10, E11, E13, E14: Open
E16: HPT1 (don't care, should be grounded)
S15: HPT0 (don't care, should be grounded)
E17-E22: Open/Grounded (Unused U.2 Port) pins
S16-S22: 3rd NVMe Port
S23-S28: 4th NVMe Port

All Twenty-Nine existing SFF-8482 pins are also required for U.3

Background: U.2 SFF-8639 Pins Usable for 2-lane HDDs



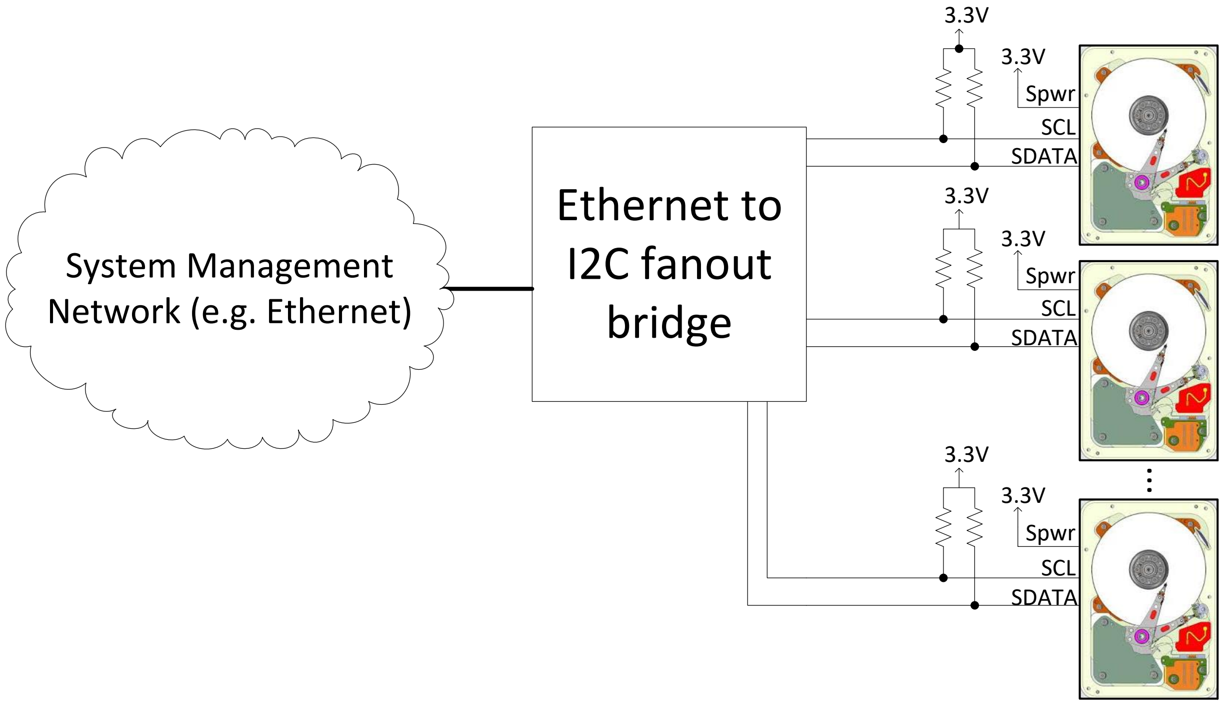
Twenty-five new pins on the U.2 SFF-8639 connector potentially would be needed for a U.2-HDD:
E1: REFCLKB+
E2: REFCLKB-
E3: +3.3Vaux (used to supply SMBus EEPROM)
E4: PERSTB#
E5: PERST#
E6: IFDet2#
E7: REFCLK+
E8: REFCLK-
E9-E15: PCIe Port 0
E23: SMBCLK
E24: SMBDAT
E25: DualPortEn#
S16-S22: PCIe Port 1

Twenty-Eight pins are not used / essential:
S1–S14: SAS Ports
S15: HPT0 (don’t care, should be open)
S23-S28: 3rd NVMe Port
E16: HPT1 (don’t care, should be open)
E17-E22: 4th NVMe Port

Fifteen SFF-8482 P1-P15 pins are required for U.2

Background: SMBus-OOB Drive Management Interface

- PCIe device management can be done in-band (over PCIe) or out-of-band (via SMBus)
- SMBus (or SMB) uses a simple I^{2}C interface as an out-of-band communication bus
- Some SMBus designs allow for power to be applied with the rest of the device de-energized
  - Host can partially energize and query devices VPD without the device being fully powered-up
- SMBus adds to system complexity - must fan out independent I^{2}C interfaces to all drives


System Management Network (e.g. Ethernet)
Ethernet to I2C fanout bridge
3.3V
3.3V
Spwr
SCL
SDATA
3.3V
3.3V
Spwr
SCL
SDATA
3.3V
3.3V
Spwr
SCL
SDATA

Potential “NVMe on HDD” Configurations

Drive Connector	Lane Config	Power Supplies	Discrete H/W Reset Support	Port Config Control (1x2 vs. 2x1)	SRIS	Feasibility (10=high)	Cost
“SAS” (SFF-8482)	U.3	5V & 12V	No	S/W-only	Required	10	$*
	U.3	12V-Only	Yes**	S/W or H/W**	Required	7	$$*
“NVMe” (SFF-8639)	U.3	12V-Only	Yes	S/W or H/W	Optional	5	$$$*
	U.2	12V-Only	Yes	S/W or H/W	Optional	4	$$$*
	U.2 & U.3	12V-Only	Yes	S/W or H/W	Optional	1	$$$$*

* SMBus support adds costs and may not be feasible on all HDD products. SMBus signals are wired to different pin locations on the SFF-8482 connector than on SSDs. Supporting SMBus vital product data while drive is de-energized (via I2C EEPROM & host-supplied power) adds further costs to the product.

** Reset & Port control signals are on different locations

Note: The NVMe Management Interface command set is available via NVMe interface regardless of whether SMBus support is included.

Potential "NVMe on HDD" Configurations – Likely Candidates

Drive Connector	Lane Config	Power Supplies	Discrete H/W Reset Support	Port Config Control (1x2 vs. 2x1)	SRIS	Feasibility (10=high)	Cost
"SAS" (SFF-8482)	U.3	5V & 12V	No	S/W-only	Required	10	$*
	U.3	12V-Only	Yes**	S/W or H/W**	Required	7	$$*
"NVMe" (SFF-8639)	U.3	12V-Only	Yes	S/W or H/W	Optional	5	$$$*
	U.2	12V-Only	Yes	S/W or H/W	Optional	4	$$$*
	U.2 & U.3	12V-Only	Yes	S/W or H/W	Optional	1	$$$$*

Potential “NVMe on HDD” Configuration – Our Preference

Drive Connector	Lane Config	Power Supplies	Discrete H/W Reset Support	Port Config Control (1x2 vs. 2x1)	SRIS	Feasibility (10=high)	Cost
“SAS” (SFF-8482)	U.3	5V & 12V	No	S/W-only	Required	10	$*
	U.3	12V-Only	Yes**	S/W or H/W**	Required	7	$$*
“NVMe” (SFF-8639)	U.3	12V-Only	Yes	S/W or H/W	Optional	5	$$$*
	U.2	12V-Only	Yes	S/W or H/W	Optional	4	$$$*
	U.2 & U.3	12V-Only	Yes	S/W or H/W	Optional	1	$$$$*

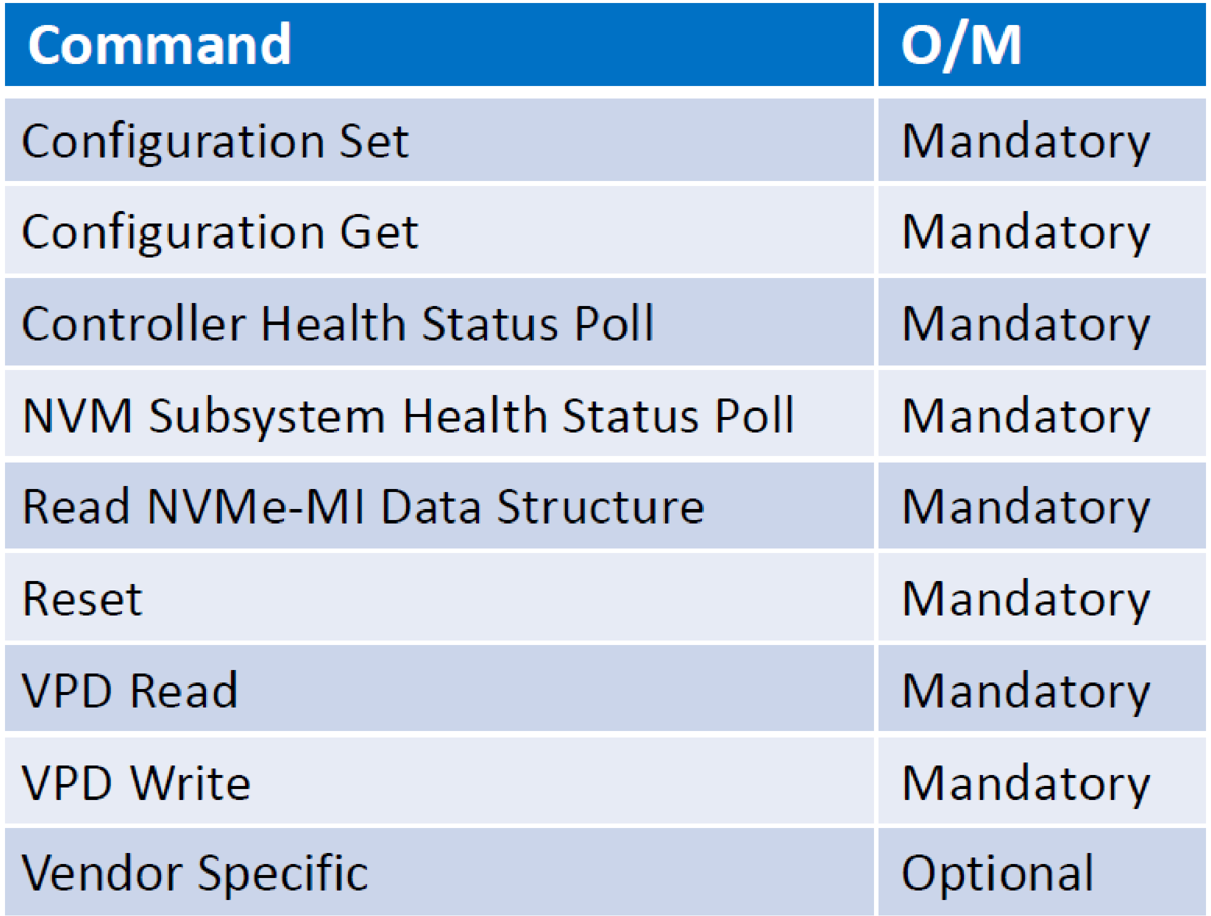
Summary

- NVMe HDD interface standards are under development
- NVMe HDDs will likely have value for OCP systems
- Full spec-compliant PCIe/NVMe feature support for HDD OCP systems is overkill
  - U.3 support requires U.2 backwards compatibility and was written for two-x2 and one-x4 lanes
  - Two-x1 or one-x2 PCIe lanes provide more bandwidth than will be needed for the foreseeable future
  - Maintaining U.2 backwards compatibility adds cost and will have no value to most customers
  - Other features, such as SMBus may not be desirable in high-density systems

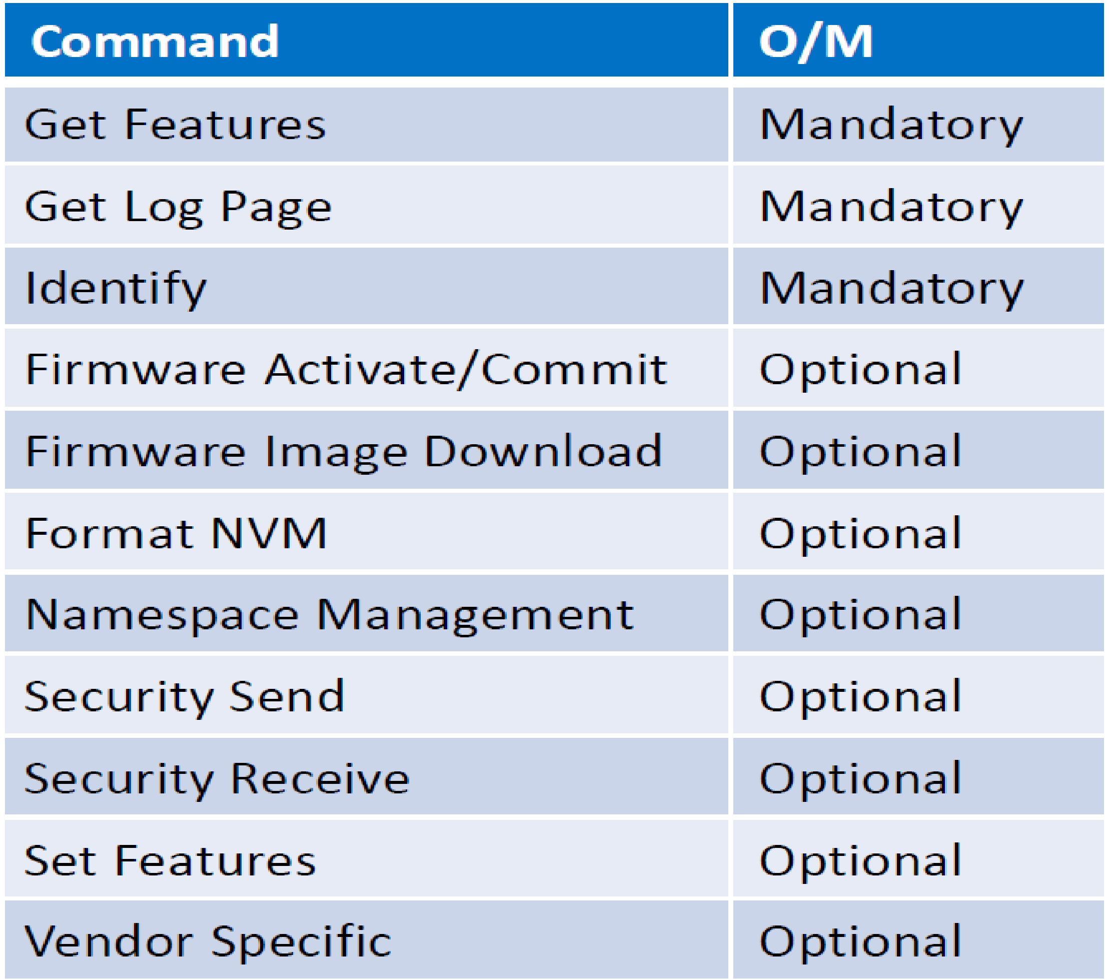
We recommend forming a task group to develop an OCP NVMe standard for HDDs

Backup

PCIe – SMBus MI Commands



Command	O/M
Configuration Set	Mandatory
Configuration Get	Mandatory
Controller Health Status Poll	Mandatory
NVM Subsystem Health Status Poll	Mandatory
Read NVMe-MI Data Structure	Mandatory
Reset	Mandatory
VPD Read	Mandatory
VPD Write	Mandatory
Vendor Specific	Optional



Command	O/M
Get Features	Mandatory
Get Log Page	Mandatory
Identify	Mandatory
Firmware Activate/Commit	Optional
Firmware Image Download	Optional
Format NVM	Optional
Namespace Management	Optional
Security Send	Optional
Security Receive	Optional
Set Features	Optional
Vendor Specific	Optional

Why supporting both U.2 and U.3 in HDDs is undesirable

HDDs will have likely have two PCIe lanes

- These lanes are in different locations on U.2 vs. U.3

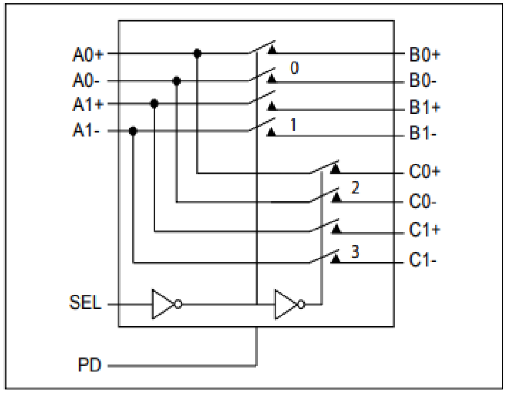
Supporting both U.2 & U.3 requires either:

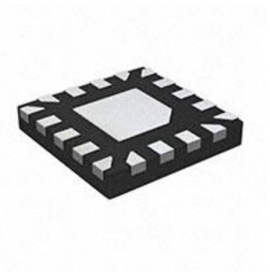
1. Additional PCIe lanes in each HDD controller
   - Increases controller ball count & cost
2. External bus switches (e.g. two of the parts on the right)
   - Increases drive components and costs by ~$2

- Supporting U.2 & U.3 requires SFF-8639

PI3DBS16212
3.3V, 1-20Gbps 1-Lane 2:1 Mux/De-Mux Switch

Block Diagram





The PI3DBS16212 is an 4 to 2 differential channel multiplexer/demultiplexer switch. This solution can switch multiple signal types up to data rate of 20Gbps.

https://www.diodes.com/assets/Databriefs/PI3DBS16212-Product-Brief.pdf

Conclusion: It is highly desirable to standardize around either U.2 or U.3 lane assignments

Using SFF-8482 connector for "PCIe-HDD" Benefits:

Using SFF-8482 would have these benefits:

1. Costs

- 2-lanes of PCIe have sufficient bandwidth for HDD PCIe/NVMe applications for the foreseeable future
- The SFF-8639 connector is estimated to add $0.50 - $1.00 per drive
  - The SFF-8639 connector adds 39 interconnects, of which only 11 will possibly be used on HDDs
- Using a SFF-8482 for PCIe reduces other costs that get passed to our customers:
  - inventory management: a single hardware SKU can fulfill customer SAS & PCIe requests
    - allows for common drive PCB hardware between SAS & PCIe
  - reduces in qualification expenses: drive vendors only have to qualify one PCBA for EMI/EMC

2. Form Factor Feasibility

- The SFF-8639 connector requires two rows of PCB interconnects
  - These additional I/Os may mechanically interfere with high-disk-count drive enclosure mechanics

Using SFF-8482 connector for “PCIe-HDD” Benefits

3. Reliability

- The current standards rely on the proper operation of 11 additional interconnects between system and drive
  - Device behavior in the current standard relies on discrete signals for mode configuration
  - These signals must have ESD, over-voltage, & fail-safe protection
  - These extra interconnections and dependencies are not essential for PCIe and can be eliminated
    - which should improve the overall system reliability

4. Maximizes industry enclosure design options

- Systems built with SFF-8639 receptacles can flexibly use both SFF-8482 HDDs and SFF-8639 SSD drives
- HDD-only PCIe systems can use SFF-8482 receptacles to reduce costs

Using SFF-8482 - Key Differences (from SFF-8639 U.3)

Using *SFF-8482* might require the following system accommodations (versus U.3):

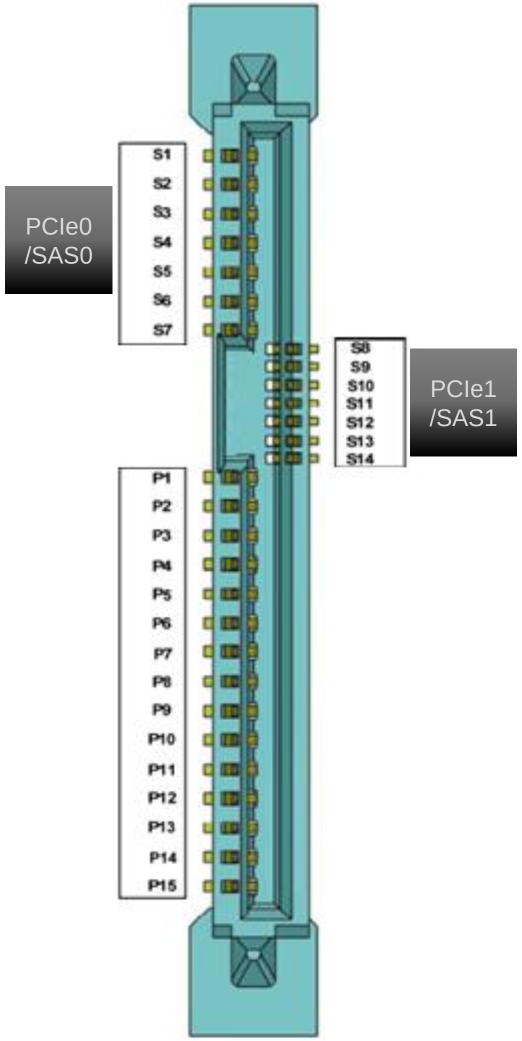
1. PCIe/SAS on U.3 lane locations only (with no U.2 backwards compatibility)
2. SRIS (independent REFCLK) support is mandatory on all hosts (switches) and devices
3. Discrete RESET (PERST/PERSTB) signals are eliminated on 5V/12V drives
4. Port configuration (Dual-port (2x1) vs. single-port (1x2)) control is done via the interface
5. SMBus (if used) pin locations are moved to different pins for HDDs
6. An additional configuration strap definition for device detection

Downsides of a SFF-8482 approach

1. **Systems changes are needed:**
   - 5V/12V Systems may no longer rely on discrete reset pins. Resets must be done in-band.
   - An additional configuration strap is used for device detection
   - PCIe Switches/HBAs must support SRIS
   - SMBus (if used) must be switched / muxed between two locations
   - Server BIOS / array software will likely have collateral changes
2. **Care must be exercised with port configuration**
   - Port configuration (2x1 vs. 1x2) is stored in the drive’s non-volatile memory
   - Default port configuration is set during manufacturing according to customer preferences
   - Mismatched drives & systems may not link properly
     - Re-configuring drives in the field may require forcing a x1 (single-lane) connection in software or hardware

The following slides show the impact of one possible SFF-8482 approach in greater detail

SFF-8482 Impact – Change #1



“U.3” PCIe / SAS lane locations

- PCIe-HDDs will follow the U.3-defined lane pin assignments
  - HDD PCIe Lane 0 is on pins S1-S7, Lane 1 is on pins S8 – S14
- Backwards compatibility (of U.2 PCIe lane locations) is not supported

SFF-8482 Impact – Change #2

SRIS is required

- SRIS – (Separate RefClk with Independent Spread-spectrum Control) support is required

Background Notes:

- SRIS is already supported in many newer PCIe switches / devices
- Running REFCLK signals to every slot in an array is an EMI source in systems
- REFCLK was originally used in PCs to reduce card costs (avoids local oscillators / crystals on PCIe cards)
- REFCLK remained as a way to synchronize down-spread spectrum in a system
- U.2 & U.3 define **SRIS** – Separate RefClk with Independent SSC and **SRNS** – Separate Refclk with no SSC
- HDDs have internal reference clock oscillators and do not need it

SFF-8482 Impact – Change #3

Device support of discrete reset signals (PERST / PERSTB) is not required OR 12V-only is required

- SFF-8482 PCIe 12V/5V HDDs will not contain discrete reset signal connections
- SFF-8482 PCIe 12V-only HDDs may support PERST on P7 and PERSTB on P8 (signals must be 5V-FailSafe)

Notes:

- Reset was originally used in PCs to reduce card costs (avoids power-good monitors on cards)
- HDDs have traditionally monitored their power and generated cold resets autonomously
- Hot Resets can be generated in-band using TS1 ordered sets
- Cold Resets can be forced using P3 PWRDIS power control pin

Additional Power-On Reset Timing Requirements Change:

- Autonomous (cold power-on / PWRDIS) reset timing response is relaxed to 100ms after power-good

Notes:

- Existing tight timing requirements were driving unnecessary device costs and complexity
- Relaxed timing should help accommodate for variability of system power in large arrays

SFF-8482 Impact – Change #4

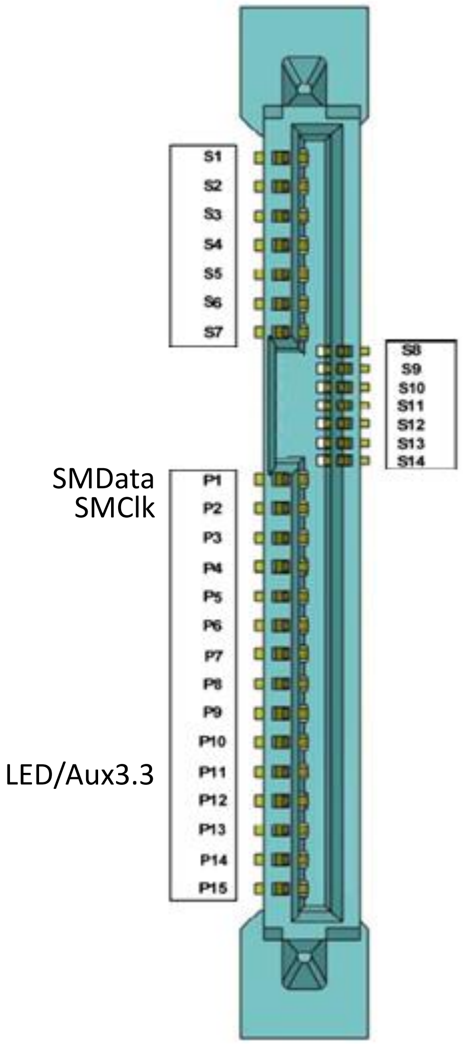
Dual Port Control

- Dual port (two independent x1 lanes) versus single port (one x2 lane) control is done via SetFeatures

Notes:

- HDD-PCIe devices may support two independent PCIe x1 links or a single PCIe link with up to two lanes
- HDDs lane default configuration will be set during device manufacturing per customer requirements
- Customers can change from the default configuration via SetFeatures control bit
- This is akin to SSDs mode selection, which is done via a *DualPortEn#* discrete signal

SFF-8482 Impact – Change #5



SMBus Changes

- SMBus is an optional 2-wire feature used for side-channel device detection and configuration
- On SSDs w/SMBus, the SMBus can be queried with the device depowered by supplying 3.3Vaux

For PCI-HDD devices with SMBus support:

- SMData is located on pin P1
- SMClk is located on pin P2
- Unpowered device SMBus power is supplied by applying 3.3V to the ACTIVITYLED# pin
  - Current on this pin must be limited by host to 15mA (e.g. via active circuit or 220ohm resistor)
- Host usage of (optional) Wake# signal on pin P1 must not interfere with SMBus traffic

  Note: HDD Support of discrete Wake# signal is not planned

SFF-8482 Impact – Change #6

New Device Type Definition

- PCIe-HDD devices are identified via new device type encoding as shown below
  - On SAS/SATA drives, pin P4 must be pulled to ground within 100ms of power application

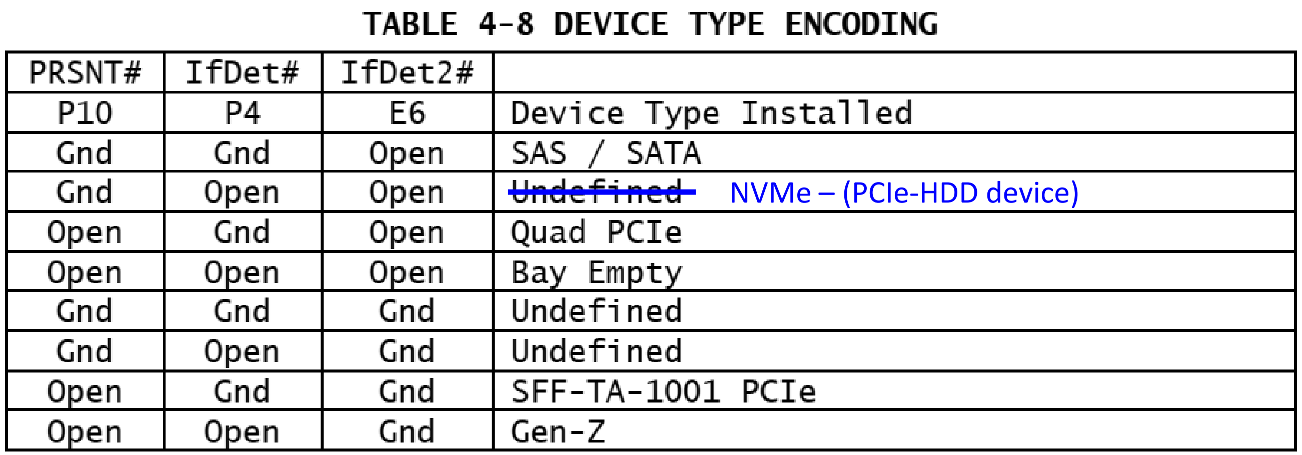


TABLE 4-8 DEVICE TYPE ENCODING

PRSNT#	IfDet#	IfDet2#	
P10	P4	E6	Device Type Installed
Gnd	Gnd	Open	SAS / SATA
Gnd	Open	Open	~~Undefined~~ NVMe – (PCIe-HDD device)
Open	Gnd	Open	Quad PCIe
Open	Open	Open	Bay Empty
Gnd	Gnd	Gnd	Undefined
Gnd	Open	Gnd	Undefined
Open	Gnd	Gnd	SFF-TA-1001 PCIe
Open	Open	Gnd	Gen-Z

SFF-TA-1001 vs. SFF-8482 Approach Pinout Comparison

	SFF-TA-1001		PCIe-HDD	
PIN	HOST	DEVICE PLUG	HOST	DEVICE PLUG
P1	WAKE#	WAKE#	**SMBDAT/WAKE#**	**SMBDAT/WAKE#**
P2	Reserved		**SMBCLK**	**SMBCLK**
P3	PWRDIS	PWRDIS	PWRDIS	PWRDIS
P4	IfDet#	IfDet#	IfDet#	IfDet#
P5	Ground	Ground	Ground	Ground
P6	Ground	Ground	Ground	Ground
P7	+5V		+5V/PRESETA*	+5V/PRESETA*
P8	+5V		+5V/PRESETB*	+5V/PRESETB*
P9	+5V		+5V/DUALPT*	+5V/DUALPT*
P10	PRSNT#	PRSNT#	PRSNT#	PRSNT#
P11	ACTIVITY LED	ACTIVITY#	**ACTIVITY LED 3.3V Aux**	**ACTIVITY# 3.3V Aux**
P12	Ground	Ground	Ground	Ground
P13	+12V	+12 V Precharge	+12V	+12 V Precharge
P14	+12V	+12 V	+12V	+12 V
P15	+12V	+12 V	+12V	+12 V

*** - 12V Only**

	SFF-TA-1001		PCIe-HDD
PIN	HOST	DEVICE PLUG	DEVICE PLUG
S1	Ground	Ground	Ground
S2	TX p0	PETp0	PETp0
S3	TX n0	PETn0	PETn0
S4	Ground	Ground	Ground
S5	RX n0	PERn0	PERn0
S6	RX p0	PERp0	PERp0
S7	Ground	Ground	Ground
S8	Ground	Ground	Ground
S9	TX p1	PETp1	PETp1
S10	TX n1	PETn1	PETn1
S11	Ground	Ground	Ground
S12	RX n1	PERn1	PERn1
S13	RX p1	PERp1	PERp1
S14	Ground	Ground	Ground
S15	Ground	HPT0	
S16	Ground	Ground	
S17	TX p2	PETp2	
S18	TX n2	PETn2	
S19	Ground	Ground	
S20	RX n2	PERn2	
S21	RX p2	PERp2	
S22	Ground	Ground	
S23	TX p3	PETp3	
S24	TX n3	PETn3	
S25	Ground	Ground	
S26	RX n3	PERn3	
S27	RX p3	PERp3	
S28	Ground	Ground	

	SFF-TA-1001		PCIe-HDD
Pin	Host	DEVICE PLUG	DEVICE PLUG
E1	REFCLKB+	REFCLKB+	
E2	REFCLKB-	REFCLKB	
E3	+3.3V Aux	+3.3 Vaux	
E4	PERSTB#	PERSTB#	
E5	PERST#	PERST#	
E6	IfDet2#	Ground	
E7	REFCLK+	REFCLK+	
E8	REFCLK-	REFCLK	
E9	Ground	Ground	
E10			
E11			
E12	Ground	Ground	
E13			
E14			
E15	Ground	Ground	
E16	HPT1	HPT1	
E17			
E18			
E19	Ground	Ground	
E20			
E21			
E22	Ground	Ground	
E23	SMBCLK	SMBCLK	
E24	SMBDAT	SMBDAT	
E25	DualPortEn#	DualPortEn#	