STORAGE DEVELOPER CONFERENCE

SDC 23

BY Developers FOR Developers

PCIe® 7.0 Specification: 128 GT/s Bandwidth for Future Data-Intensive Markets

Presented by:
Dr. Debendra Das Sharma
Intel Senior Fellow and co-GM, Memory and I/O Technologies
Member, PCI-SIG® Board

Agenda

- ❑Introduction: Evolution of PCI Express® Technology
- ❑PCIe® 6.0 and PCIe 7.0 Specifications: A Deep Dive
- ❑PCI Express Technology and Storage
- ❑Form Factors
- ❑Optical-friendly PCIe Technology
- ❑Compliance
- ❑Conclusions

Evolution of PCI Express® Specification

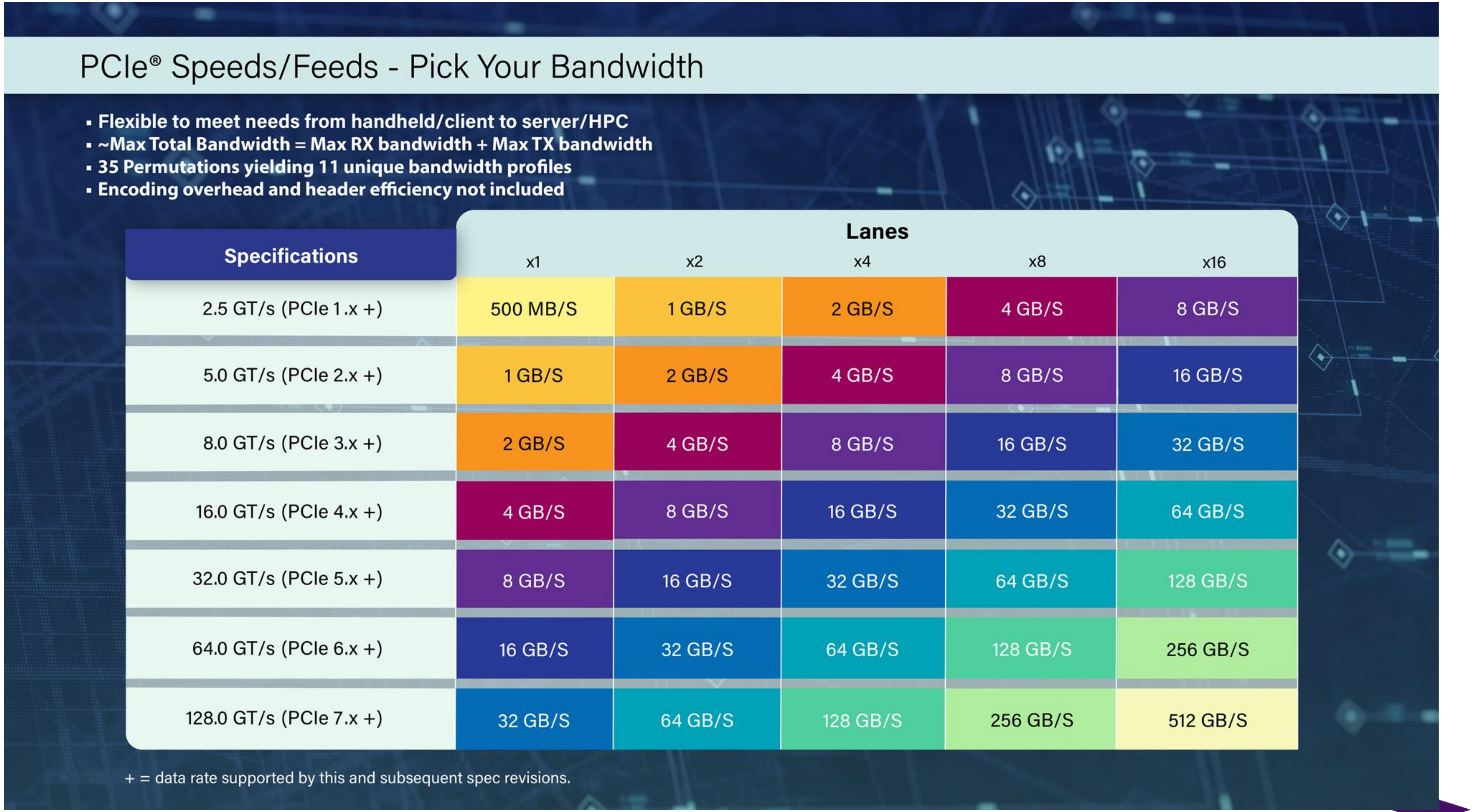
- PCIe® specification doubles the data rate every generation with full backwards compatibility every 3 years
- Ubiquitous I/O across the compute continuum: PC, Hand-held, Workstation, Server, Cloud, Enterprise, HPC, Embedded, IoT, Automotive, AI
- One stack / same silicon across all segments with different form-factors; a x16 PCIe 5.0 device interoperates with a x1 PCIe 1.0 device!
- PCIe 7.0 specification currently at Rev 0.3 level maturity – making good progress

Revision	Max Data Rate	Encoding	Signaling
PCIe 1.0 (2003)	2.5 GT/s	8b/10b	NRZ
PCIe 2.0 (2007)	5.0 GT/s	8b/10b	NRZ
PCIe 3.0 (2010)	8.0 GT/s	128b/130b	NRZ
PCIe 4.0 (2017)	16.0 GT/s	128b/130b	NRZ
PCIe 5.0 (2019)	32.0 GT/s	128b/130b	NRZ
PCIe 6.0 (2022)	64.0 GT/s	1b/1b (Flit Mode*)	PAM4
PCIe 7.0 (2025)	128.0 GT/s	1b/1b (Flit Mode*)	PAM4

(*Flit Mode also enabled in other Data Rate with their respective encoding)

PCIe architecture continues to deliver bandwidth doubling for 7 generations spanning 3 decades! An impressive run!

PCI Express® Specifications: Speeds and Feeds



PCIe® Speeds/Feeds - Pick Your Bandwidth

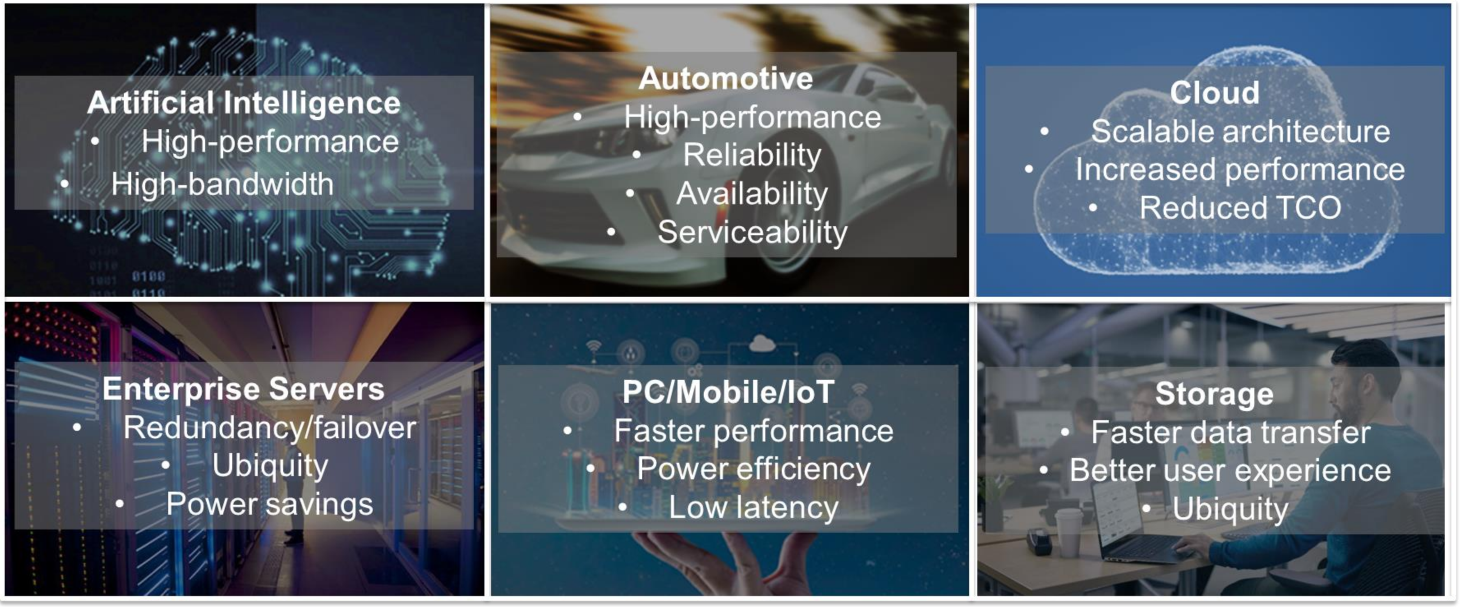
- **Flexible to meet needs from handheld/client to server/HPC**
- **~Max Total Bandwidth = Max RX bandwidth + Max TX bandwidth**
- **35 Permutations yielding 11 unique bandwidth profiles**
- **Encoding overhead and header efficiency not included**

Specifications	**Lanes**				
	x1	x2	x4	x8	x16
2.5 GT/s (PCIe 1.x +)	500 MB/S	1 GB/S	2 GB/S	4 GB/S	8 GB/S
5.0 GT/s (PCIe 2.x +)	1 GB/S	2 GB/S	4 GB/S	8 GB/S	16 GB/S
8.0 GT/s (PCIe 3.x +)	2 GB/S	4 GB/S	8 GB/S	16 GB/S	32 GB/S
16.0 GT/s (PCIe 4.x +)	4 GB/S	8 GB/S	16 GB/S	32 GB/S	64 GB/S
32.0 GT/s (PCIe 5.x +)	8 GB/S	16 GB/S	32 GB/S	64 GB/S	128 GB/S
64.0 GT/s (PCIe 6.x +)	16 GB/S	32 GB/S	64 GB/S	128 GB/S	256 GB/S
128.0 GT/s (PCIe 7.x +)	32 GB/S	64 GB/S	128 GB/S	256 GB/S	512 GB/S

+ = data rate supported by this and subsequent spec revisions.

Bandwidth Drivers for PCI Express® Specifications

- Device side: Networking (800Gb/s -> 1.6 Tb/s), Accelerators, FPGA/ ASICs, Memory (need more memory b/w)
- Alternate Protocols (CXL™, proprietary SMP cache coherency protocols for multi-socket servers) on PCIe® architecture
- As compute capability grows exponentially, so does I/O bandwidth
  - Platform already has hundreds of lanes for I/O => speed has to go up
- But ... we need to meet the cost, performance, power metrics as an ubiquitous I/O with hundreds of Lanes in a platform



Artificial Intelligence
- High-performance
- High-bandwidth

Automotive
- High-performance
- Reliability
- Availability
- Serviceability

Cloud
- Scalable architecture
- Increased performance
- Reduced TCO

Enterprise Servers
- Redundancy/failover
- Ubiquity
- Power savings

PC/Mobile/IoT
- Faster performance
- Power efficiency
- Low latency

Storage
- Faster data transfer
- Better user experience
- Ubiquity

(New Usage Models: Cloud, AI/ Analytics, Edge)

New usage models are driving bandwidth demand – doubling every three years

PCIe® 6.0 and PCIe 7.0 Specifications:

A Deep Dive

Key Metrics for PCIe® 6.0/ 7.0 Architecture: Requirements

Metrics	Expectations
Data Rate	PCIe 6.0 data rate @ 64 GT/s -> PCIe 7.0 data rate @ 128.0 GT/s, PAM4 signaling(double the bandwidth per pin every generation)
Latency	<10ns adder for Transmitter + Receiver over 32.0 GT/s (including FEC) (We can not afford the 100ns FEC latency as networking does with PAM4)
Bandwidth Inefficiency	<2 % adder over PCIe 5.0 specification across all payload sizes
Reliability	0 < FIT << 1 for a x16 (FIT – Failure in Time, number of failures in 10^9 hours)
Channel Reach	Similar to PCIe 5.0 specification under similar set up for Retimer(s) (maximum 2)
Power Efficiency	Better than PCIe 5.0 specification
Low Power	Similar entry/ exit latency for L1 low-power state Addition of a new power state (L0p) to support scalable power consumption with bandwidth usage without interrupting traffic
Plug and Play	Fully backwards compatible with PCIe 1.x through PCIe 5.0/6.0 specifications
Others	HVM-ready, cost-effective, scalable to hundreds of Lanes in a platform

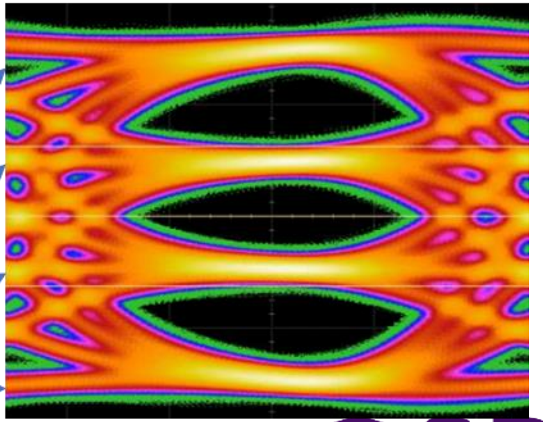
Right trade-offs to meet each of these metrics!

PAM4 Signaling at 64.0 and 128.0 GT/s

- PAM4 signaling: Pulse Amplitude Modulation 4-level
  - 4 levels (2 bits) in same Unit Interval (UI); 3 eyes
  - Helps channel loss (same Nyquist as 32.0 GT/s)
- Reduced voltage levels (EH) and eye width increases susceptibility to errors
- Gray Coding to reduce errors in each UI
- Precoding to minimize errors in a burst
- Voltage levels at Tx and Rx define encoding

Voltage Level	Tx Voltage	Rx Voltage (V)
0	-Vtx	V <= Vth1
1	-Vtx/3	Vth1 < V <= Vth2
2	+Vtx/3	Vth2 < V <= Vth3
3	+Vtx	V > Vth3

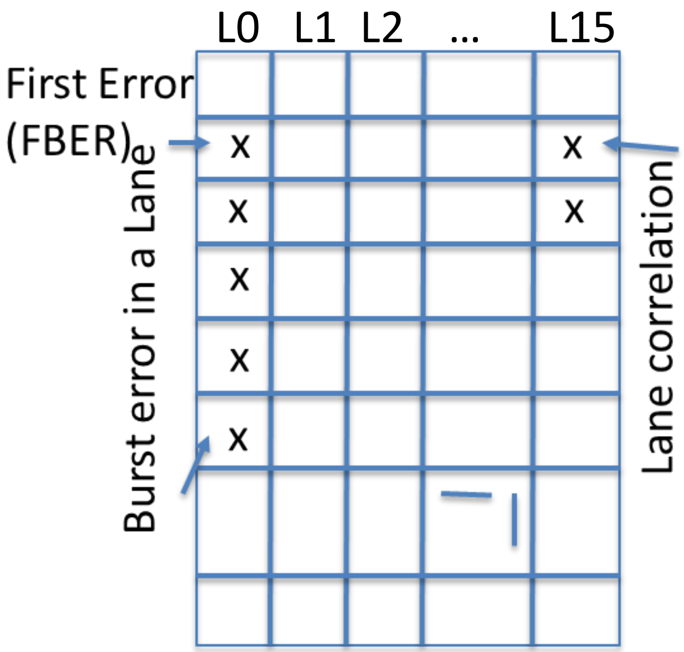
Scrambled 2-bit aligned value		Unscrambled 2-bit as well TS0 Ordered Sets	Voltage Level	DC-balance Values
Prior to Gray Coding	After Gray Coding			
10	11	11	3	+3
11	10	10	2	+1
01	01	01	1	-1
00	00	00	0	-3



Error Assumptions and Characteristics w/ PAM4

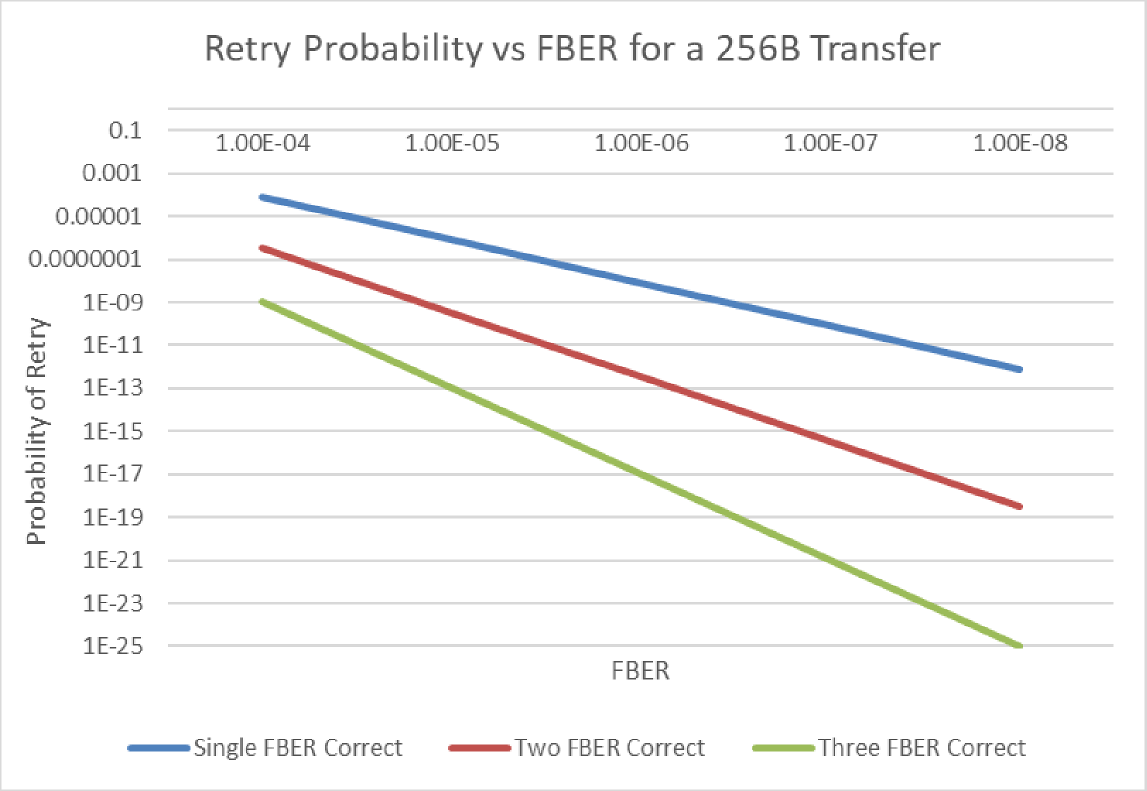
Parameters of interest: FBER and error correlation within Lane and across Lanes

- FBER – First bit error rate
  - Probability of the first bit error occurring at the Receiver
- Receiving Lane may see a burst propagated due to DFE
  - The number of errors from the burst can be minimized
    - Constrain DFE tap weights - balance TxEQ, CTLE and DFE equalization
- Correlation of errors across Lanes
  - Due to common source of errors (e.g., power supply noise)
  - Conditional probability that a first error in a Lane => errors in nearby Lanes
- BER depends on the FBER and the error correlation in a Lane and across Lanes



Our Approach: Light-weight FEC and Retry

- Light-weight FEC, strong CRC, and keep the overall latency (including retry) really low so that the Ld/St applications do not suffer latency penalty
- We are better off retrying a packet with 10^{-6} (or 10^{-5}) probability with a retry latency of 100ns vs having a FEC latency impact of 100ns with a much lower retry probability



Low latency mechanism w/ FBER of 1E-6 to meet the metrics (latency, area, power, bandwidth)

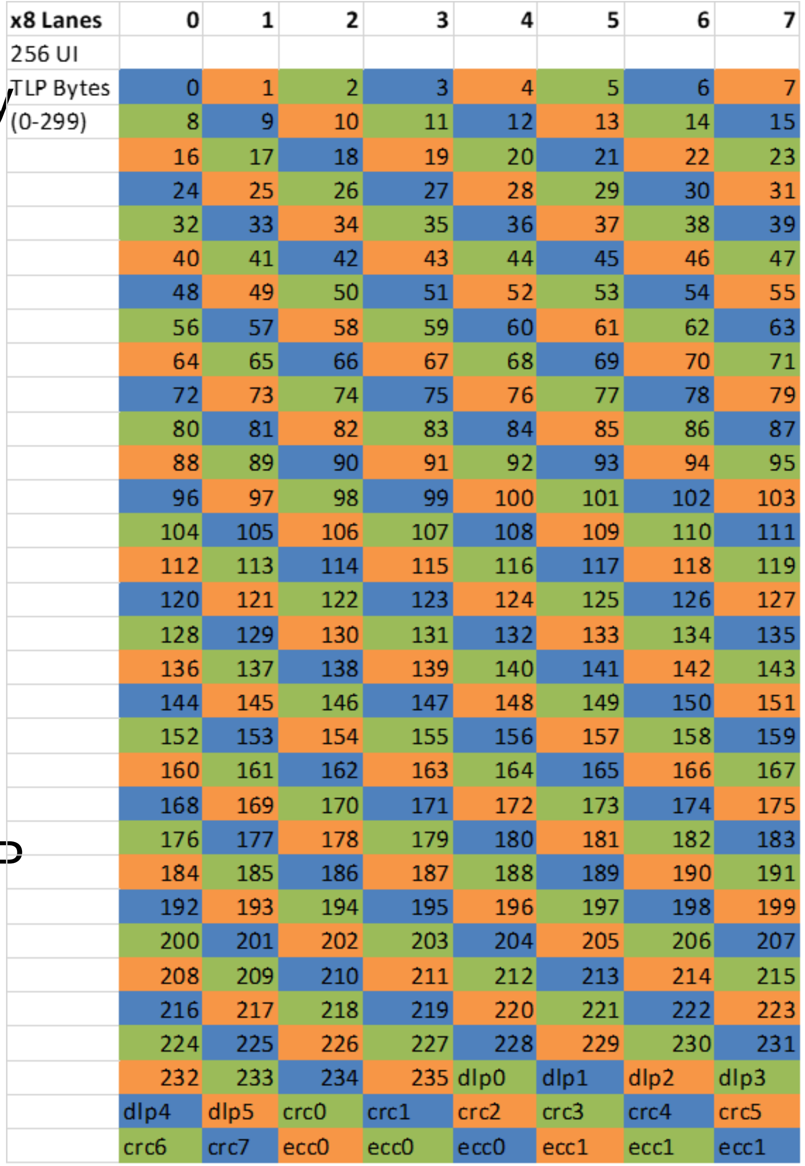
FLIT Encoding : Low-latency w/ High Efficiency

- FLIT (flow control unit) based: FEC needs fixed set of bytes
- Error Correction (FEC) in FLIT => CRC (detection) in FLITs => Retry at FLIT level
- Lower data rates will also use the same FLIT once enabled

- FLIT size: 256B
  - 236B TLP, 6B DLP, 8B CRC, 6B FEC
  - No Sync hdr, no Framing Token (TLP reformat), no TLP/DLLP CRC
  - Improved bandwidth utilization due to overhead amortization
  - FLIT Latency: 2ns x16, 4ns x8, 8ns x4, 16ns x2, 32ns x1
  - Guaranteed Ack and credit exchange => low Latency, low storage

- Optimization: Retry error FLIT only + existing Go-Back-N retry
- Other benefits of Flit Mode: scalability (future-proofing) with new TLP arrangement, making it easier to parse
- Once Flit mode is negotiated, it must be supported at all speeds

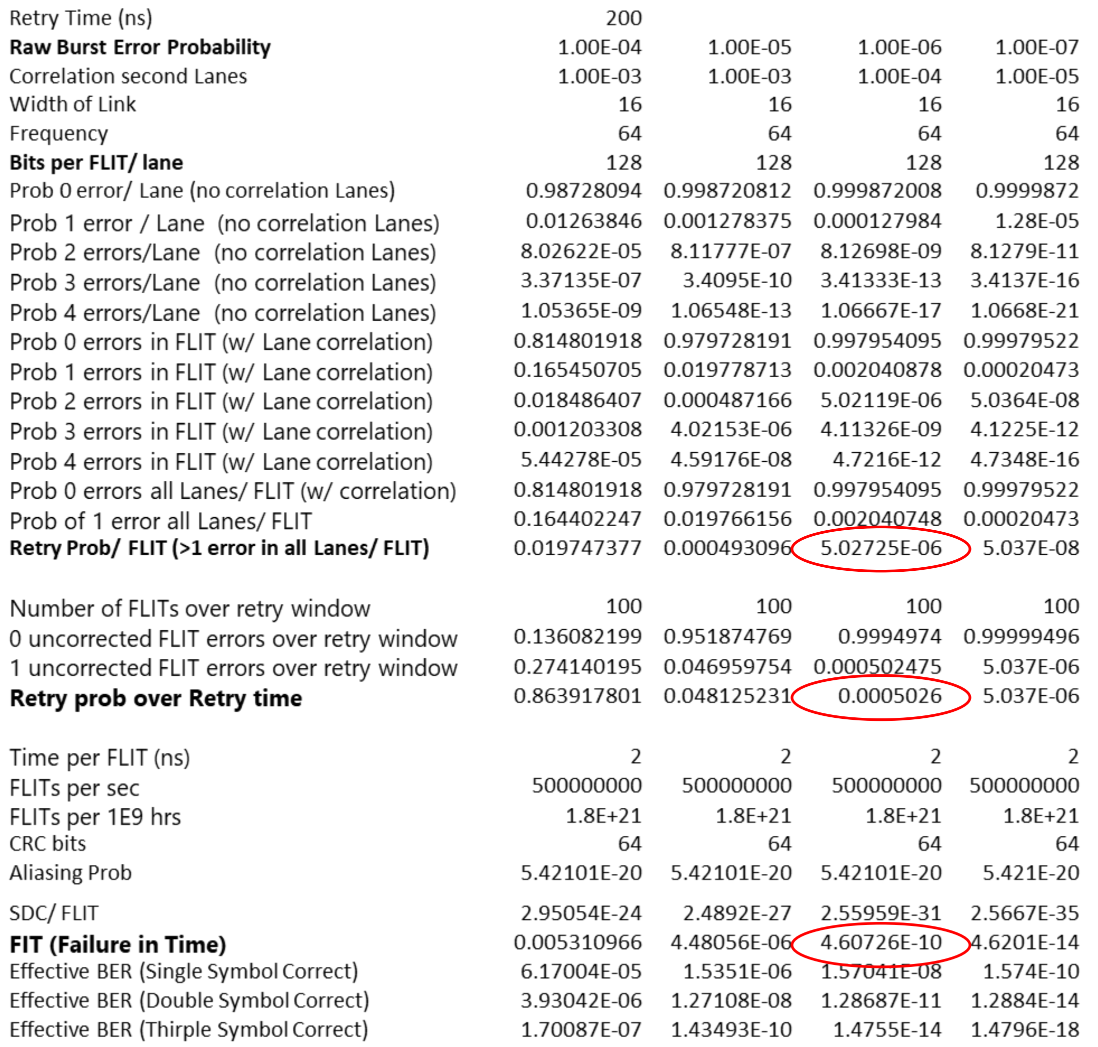
Low latency improves performance and reduces area



x8 Lanes	0	1	2	3	4	5	6	7
256 UI								
TLP Bytes	0	1	2	3	4	5	6	7
(0-299)	8	9	10	11	12	13	14	15
	16	17	18	19	20	21	22	23
	24	25	26	27	28	29	30	31
	32	33	34	35	36	37	38	39
	40	41	42	43	44	45	46	47
	48	49	50	51	52	53	54	55
	56	57	58	59	60	61	62	63
	64	65	66	67	68	69	70	71
	72	73	74	75	76	77	78	79
	80	81	82	83	84	85	86	87
	88	89	90	91	92	93	94	95
	96	97	98	99	100	101	102	103
	104	105	106	107	108	109	110	111
	112	113	114	115	116	117	118	119
	120	121	122	123	124	125	126	127
	128	129	130	131	132	133	134	135
	136	137	138	139	140	141	142	143
	144	145	146	147	148	149	150	151
	152	153	154	155	156	157	158	159
	160	161	162	163	164	165	166	167
	168	169	170	171	172	173	174	175
	176	177	178	179	180	181	182	183
	184	185	186	187	188	189	190	191
	192	193	194	195	196	197	198	199
	200	201	202	203	204	205	206	207
	208	209	210	211	212	213	214	215
	216	217	218	219	220	221	222	223
	224	225	226	227	228	229	230	231
	232	233	234	235	dlp0	dlp1	dlp2	dlp3
	dlp4	dlp5	crc0	crc1	crc2	crc3	crc4	crc5
	crc6	crc7	ecc0	ecc0	ecc0	ecc1	ecc1	ecc1

Retry Probability and FIT vs FBER Correlation

- Single Symbol Correct interleaved FEC plus 64-b CRC works well for raw FBER of 1E-6 even with high Lane correlation
- Retry probability per FLIT is 5 x 10^{-6}
- B/W loss is 0.05% even with go-back-n
- FIT is almost 0
- Can mitigate the bandwidth loss significantly by adopting retry only the non-NOP TLP FLIT



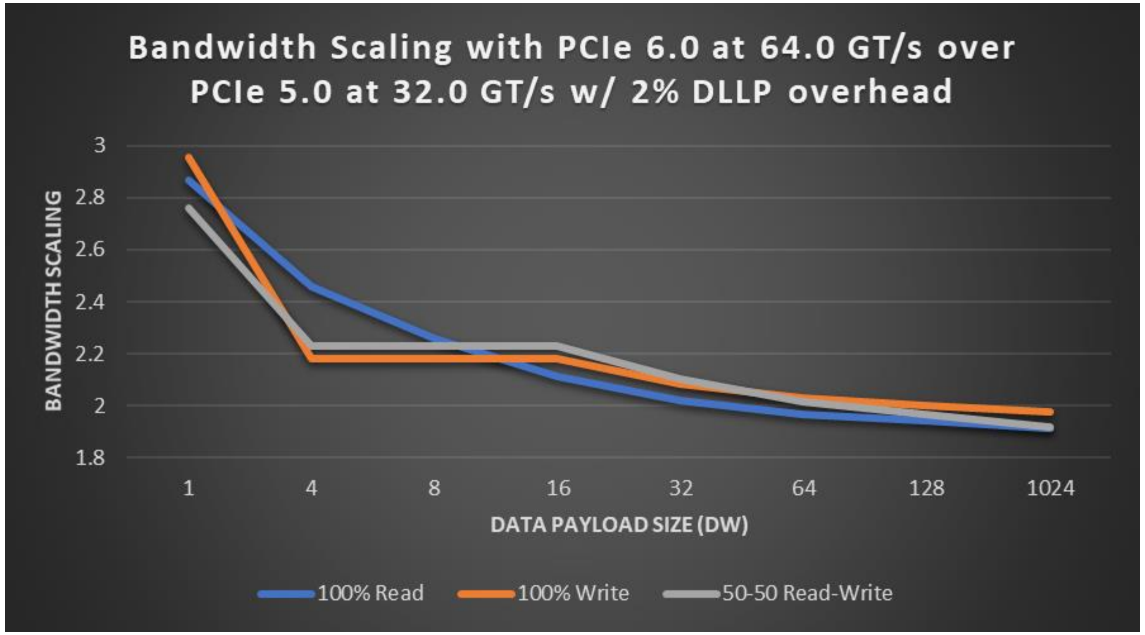
Retry Time (ns)	200			
Raw Burst Error Probability	1.00E-04	1.00E-05	1.00E-06	1.00E-07
Correlation second Lanes	1.00E-03	1.00E-03	1.00E-04	1.00E-05
Width of Link	16	16	16	16
Frequency	64	64	64	64
Bits per FLIT/ lane	128	128	128	128
Prob 0 error/ Lane (no correlation Lanes)	0.98728094	0.998720812	0.999872008	0.9999872
Prob 1 error / Lane (no correlation Lanes)	0.01263846	0.001278375	0.000127984	1.28E-05
Prob 2 errors/Lane (no correlation Lanes)	8.02622E-05	8.11777E-07	8.12698E-09	8.1279E-11
Prob 3 errors/Lane (no correlation Lanes)	3.37135E-07	3.4095E-10	3.41333E-13	3.4137E-16
Prob 4 errors/Lane (no correlation Lanes)	1.05365E-09	1.06548E-13	1.06667E-17	1.0668E-21
Prob 0 errors in FLIT (w/ Lane correlation)	0.814801918	0.979728191	0.997954095	0.99979522
Prob 1 errors in FLIT (w/ Lane correlation)	0.165450705	0.019778713	0.002040878	0.00020473
Prob 2 errors in FLIT (w/ Lane correlation)	0.018486407	0.000487166	5.02119E-06	5.0364E-08
Prob 3 errors in FLIT (w/ Lane correlation)	0.001203308	4.02153E-06	4.11326E-09	4.1225E-12
Prob 4 errors in FLIT (w/ Lane correlation)	5.44278E-05	4.59176E-08	4.7216E-12	4.7348E-16
Prob 0 errors all Lanes/ FLIT (w/ correlation)	0.814801918	0.979728191	0.997954095	0.99979522
Prob of 1 error all Lanes/ FLIT	0.164402247	0.019766156	0.002040748	0.00020473
Retry Prob/ FLIT (>1 error in all Lanes/ FLIT)	0.019747377	0.000493096	5.02725E-06	5.037E-08
Number of FLITs over retry window	100	100	100	100
0 uncorrected FLIT errors over retry window	0.136082199	0.951874769	0.9994974	0.99999496
1 uncorrected FLIT errors over retry window	0.274140195	0.046959754	0.000502475	5.037E-06
Retry prob over Retry time	0.863917801	0.048125231	0.0005026	5.037E-06
Time per FLIT (ns)	2	2	2	2
FLITs per sec	500000000	500000000	500000000	500000000
FLITs per 1E9 hrs	1.8E+21	1.8E+21	1.8E+21	1.8E+21
CRC bits	64	64	64	64
Aliasing Prob	5.42101E-20	5.42101E-20	5.42101E-20	5.421E-20
SDC/ FLIT	2.95054E-24	2.4892E-27	2.55959E-31	2.5667E-35
FIT (Failure in Time)	0.005310966	4.48056E-06	4.60726E-10	4.6201E-14
Effective BER (Single Symbol Correct)	6.17004E-05	1.5351E-06	1.57041E-08	1.574E-10
Effective BER (Double Symbol Correct)	3.93042E-06	1.27108E-08	1.28687E-11	1.2884E-14
Effective BER (Thirple Symbol Correct)	1.70087E-07	1.43493E-10	1.4755E-14	1.4796E-18

(Numbers get worse by 2x at 128.0 GT/s - still well within expectations)

Spec Requirement: FBER of 1E-6 with a burst of <=16
to meet the performance goals with a light-weight FEC

PCIe® 6.0 Specification FLIT Mode Bandwidth at 64.0 GT/s

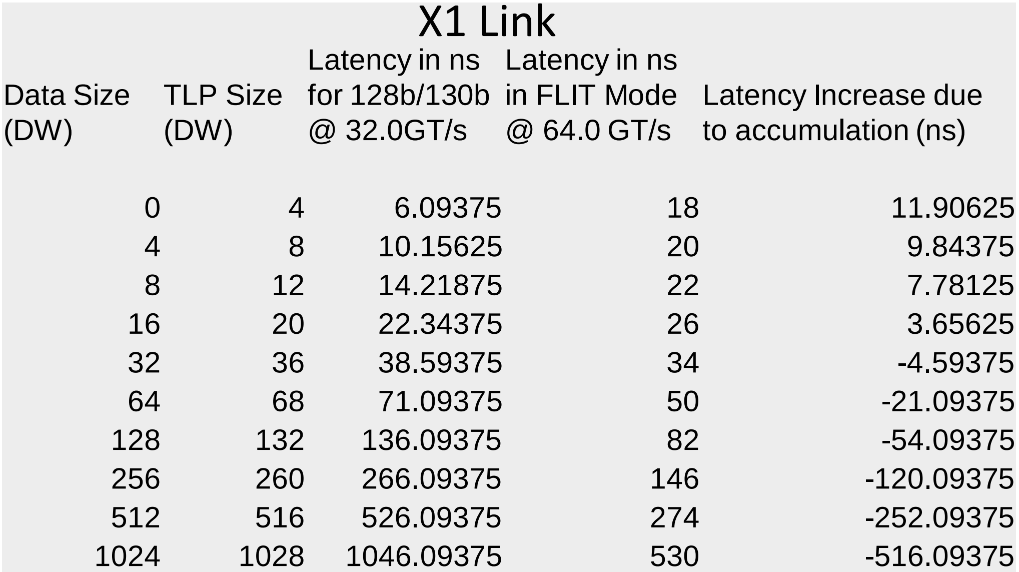
- Bandwidth increase = 2X (BW efficiency of FLIT mode) / (BW efficiency in non-FLIT mode)
- Overall we see a >2X improvement in bandwidth (benefits most systems)
  - Efficiency gain reduces as TLP size increases
  - Beyond 512 B (128 DW) payload goes below 1
- Bandwidth efficiency improvement in FLIT mode due to the amortization of CRC, DLP, and ECC over a FLIT (8% overhead) – works out better than sync hdr, DLLP, Framing Token per TLP, and 4B CRC per TLP overheads in PCIe 5.0 specification
- Expect 2X increase in bandwidth at 128.0 GT/s with PCIe 7.0 specification



Bandwidth Efficiency improvement causes > 2X bandwidth gain for up to 512B Payload in 64.0 GT/s FLIT mode

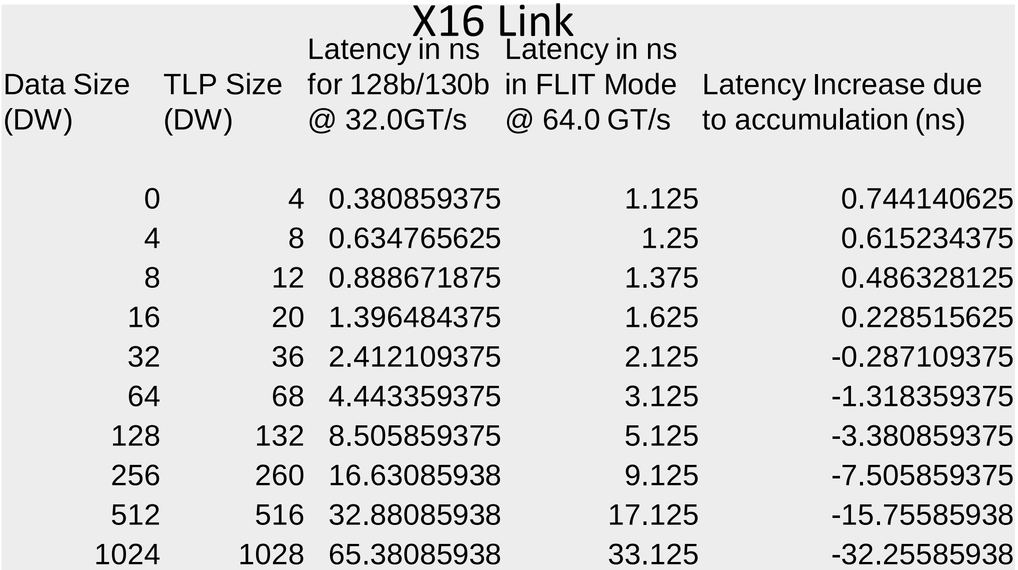
Latency Impact of FLIT Mode

- FLIT accumulation in Rx only (Tx pipeline)
- FEC + CRC delay expected to be ~ 1-2 ns
- Expected Latency savings due to removal of sync hdr, fixed FLIT sizes (no framing logic, no variable sized TLP/ CRC processing) is not considered in Tables here
- With twice the data rate and the above optimizations, realistically expect to see lower latency except for x2 and x1 for smaller payload TLPs –worst case ~10ns adder
- Latency expected to improve at 128.0 GT/s as the accumulation time halves from 64.0 GT/s



X1 Link

Data Size (DW)	TLP Size (DW)	Latency in ns for 128b/130b @ 32.0GT/s	Latency in ns in FLIT Mode @ 64.0 GT/s	Latency Increase due to accumulation (ns)
0	4	6.09375	18	11.90625
4	8	10.15625	20	9.84375
8	12	14.21875	22	7.78125
16	20	22.34375	26	3.65625
32	36	38.59375	34	-4.59375
64	68	71.09375	50	-21.09375
128	132	136.09375	82	-54.09375
256	260	266.09375	146	-120.09375
512	516	526.09375	274	-252.09375
1024	1028	1046.09375	530	-516.09375



X16 Link

Data Size (DW)	TLP Size (DW)	Latency in ns for 128b/130b @ 32.0GT/s	Latency in ns in FLIT Mode @ 64.0 GT/s	Latency Increase due to accumulation (ns)
0	4	0.380859375	1.125	0.744140625
4	8	0.634765625	1.25	0.615234375
8	12	0.888671875	1.375	0.486328125
16	20	1.396484375	1.625	0.228515625
32	36	2.412109375	2.125	-0.287109375
64	68	4.443359375	3.125	-1.318359375
128	132	8.505859375	5.125	-3.380859375
256	260	16.63085938	9.125	-7.505859375
512	516	32.88085938	17.125	-15.75585938
1024	1028	65.38085938	33.125	-32.25585938

Meets or exceeds the latency expectations

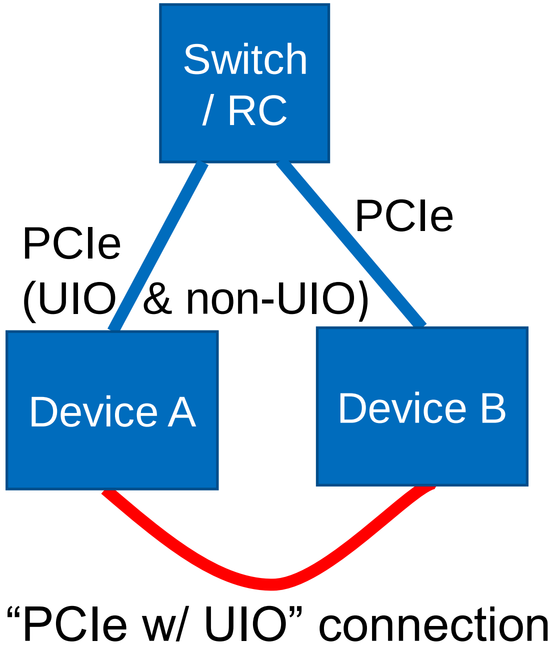
Key Metrics for PCIe® 6.0 Specification: Evaluation

Metrics	Expectations	Evaluation
Data Rate	64 GT/s, PAM4 (double the bandwidth per pin every generation)	Meets
Latency	<10ns adder for Transmitter + Receiver over 32.0 GT/s (including FEC) (We can not afford the 100ns FEC latency as n/w does with PAM-4)	Exceeds (Savings in latency with <10ns for x1/ x2 cases)
Bandwidth Inefficiency	<2 % adder over PCIe 5.0 specification across all payload sizes	Exceeds (getting >2X bandwidth in most cases)
Reliability	0 < FIT << 1 for a x16 (FIT – Failure in Time, failures in 10^9 hours)	Meets
Channel Reach	Similar to PCIe 5.0 specification under similar set up for Retimer(s) (maximum 2)	Meets
Power Efficiency	Better than PCIe 5.0 specification	Design dependent – expected to meet
Low Power	Similar entry/ exit latency for L1 low-power state Addition of a new power state (L0p) to support scalable power consumption with bandwidth usage without interrupting traffic	Design dependent – expected to meet; L0p looks promising
Plug and Play	Fully backwards compatible with PCIe 1.x through PCIe 5.0 specification	Meets
Others	HVM-ready, cost-effective, scalable to hundreds of Lanes in a platform	Expected to Meet

Meets or exceeds requirements on all key metrics. Expect same results for 128.0 GT/s

Unordered I/O (UIO): QoS and path to Multi-Pathing

- PCI/PCIe enforce Producer/Consumer via fabric-enforced ordering rules
  - Problem: Limits performance
  - Problem: PCIe Posted Writes don’t match other SoC fabric semantics; Requester doesn’t (directly) know if/when the write has actually completed
  - Problem: Mismatched write performance to multiple destinations cause ~global stalls
- Relaxed Ordering (RO), and ID-Ordering (IDO) not commonly used
  - Problem: Still need “flag” operations to use PCI baseline ordering
  - Problem: RO/IDO not intended to support cases with multiple paths (see example at right)
- Goals:
  - Enable higher performance, esp. multiple-paths, via source-ordering
  - Fully backwards compatible with existing producer-consumer model
  - Simplest possible discovery/configuration



Example illustrating multiple paths between communicating devices future

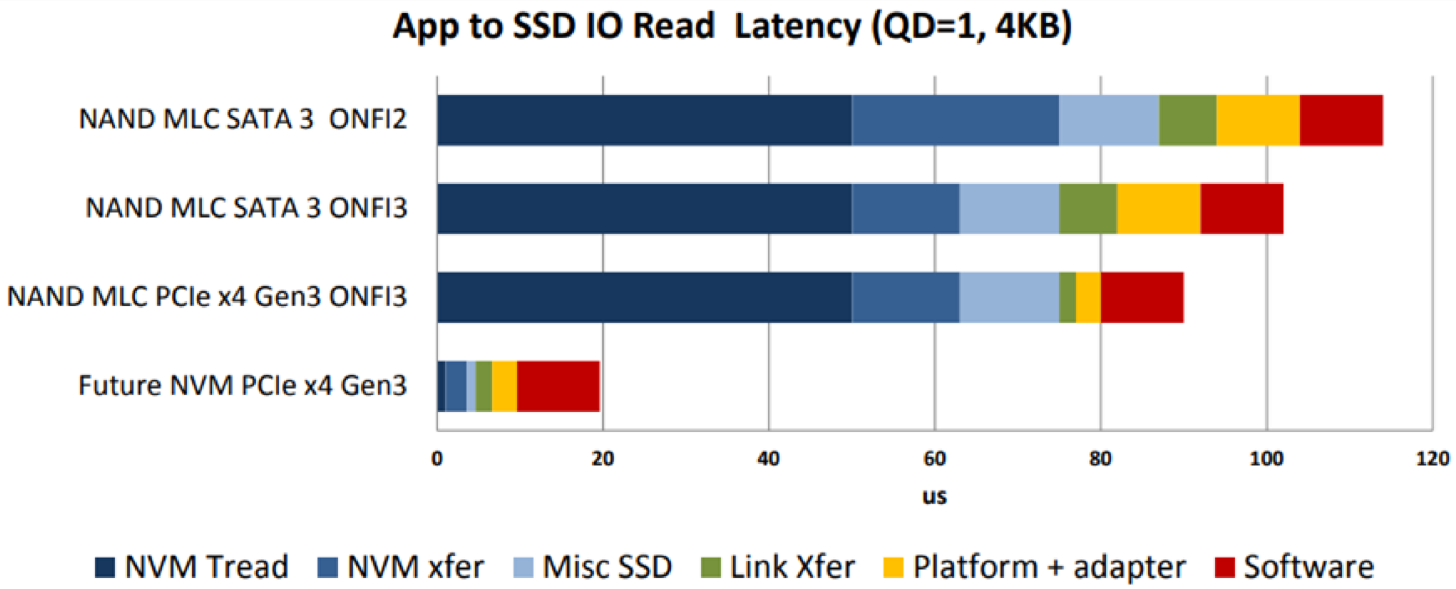
UIO Details

- UIO on one or more dedicated non-VC0 channel
  - VC0 always for traditional ordering model
  - Reuse existing FC mechanisms: 5 transactions total – all new TTYPEs
    - “Posted”: UIO Memory writes (gets completion)
    - Non-Posted: UIO Memory Read
    - Completion: UIO Memory Read Completion with data, UIO Mem Rd Completion without data; UIO Memory Write completion (no data)
  - Must be enabled end-to-end
  - UIO and non-UIO don’t mix – different VCs also ensures QoS
    - E.g., Persistent Memory access vs regular memory access w/ UIO is two different VCs (non-0)
    - VCs are now easier/ cheaper to implement with shared credits
  - No ordering – all ordering enforced at source (i.e., don’t do the flag write till all the data that is covered by it is completed)

Row Pass Col?	UIO Write	UIO Read	UIO Completion
UIO Write	Permitted	Permitted	Permitted
UIO Read	Permitted	Permitted	Permitted
UIO Completion	Yes	Yes	Permitted

PCI Express® Technology and Storage

PCIe® SSDs for Storage

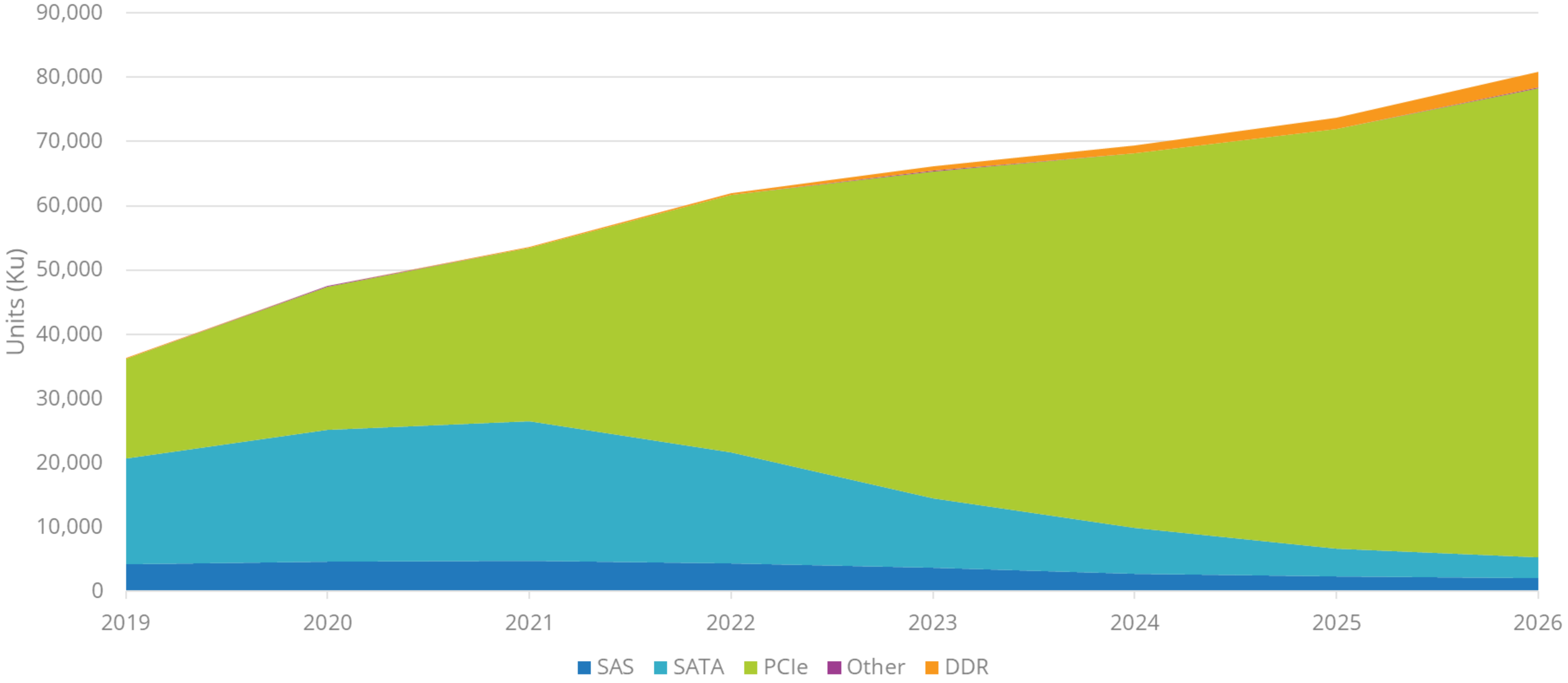


Source: FMS 2013 "[NVMe Express Overview & Ecosystem Update]"

- **PCI Express® architecture is a great interface for SSDs**
  - Stunning performance 8 GB/s per lane/ direction (PCIe 6.0 specification x1 @ 64.0 GT/s)
  - Lane scalability 32/ 16 GB/s per device (x4/ x2)
  - Lower latency Platform + Adapter: 10 µsec down to 1 µsec
  - Lower power No external SAS IOC saves 7-10 W
  - Lower cost No external SAS IOC saves $
  - CPU-integrated PCIe lanes Up to 128 PCIe 3.0 specification

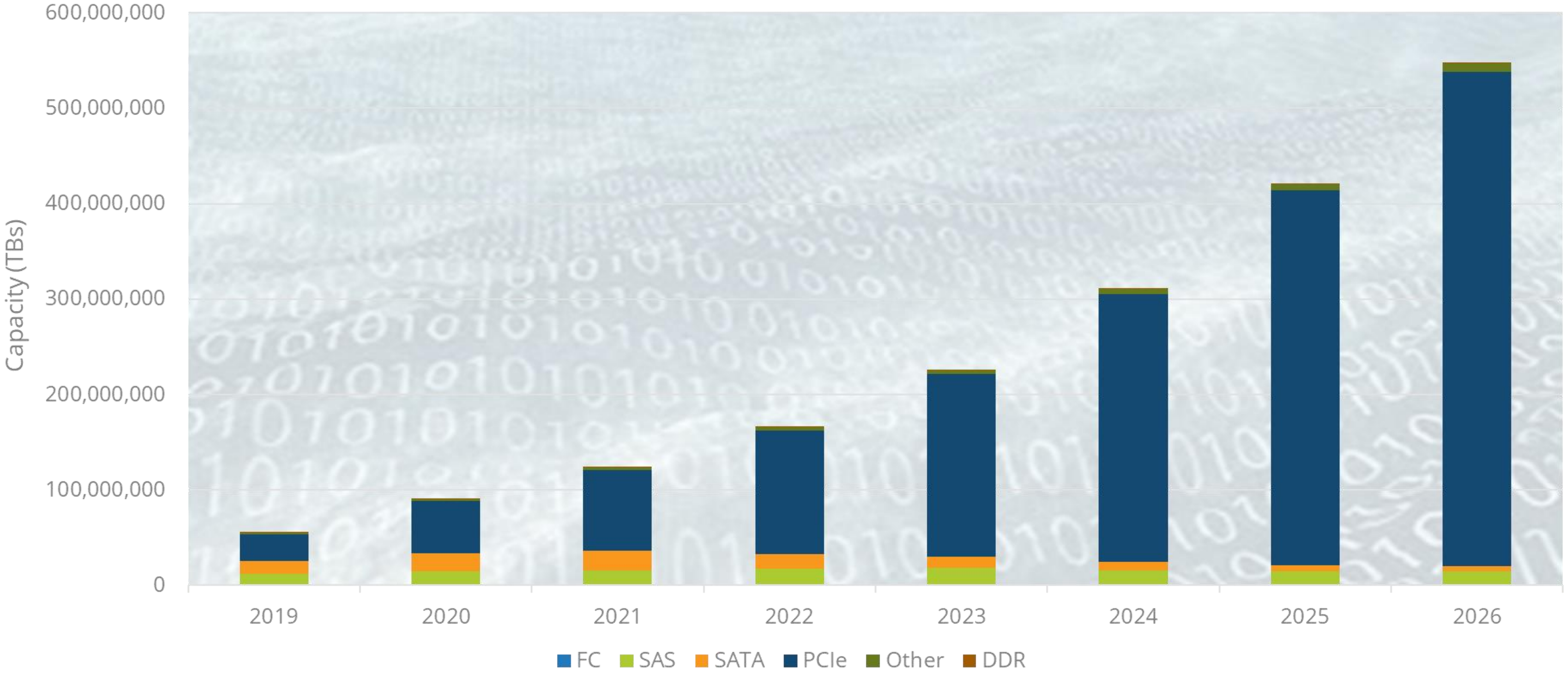
- **With NVM Express® and PCIe technology evolution, storage is no longer the bottleneck**

Enterprise SSD Unit Shipment Forecast by Interface



Source: Worldwide Solid State Storage Forecast, 2022–2026 (May 2022) IDC #US47831722

Enterprise SSD Capacity Shipment Forecast by Interface



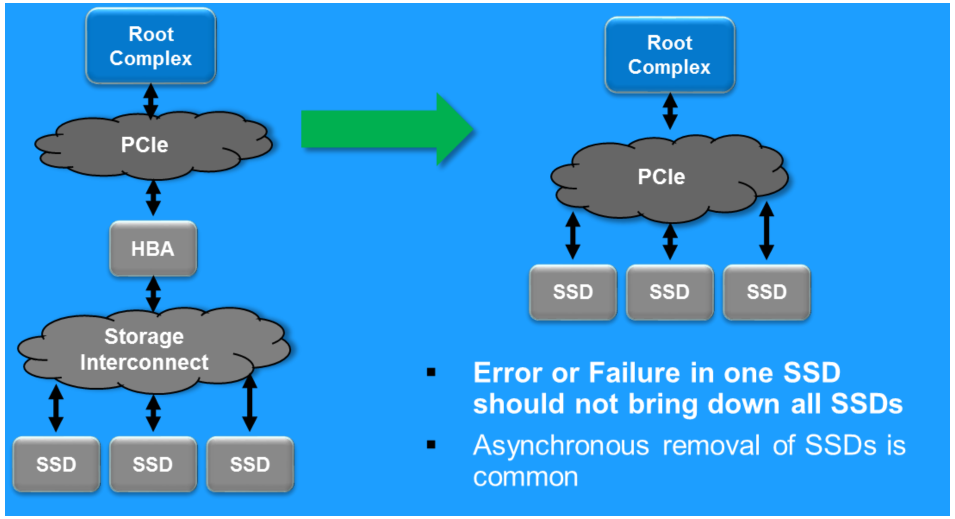
Source: Worldwide Solid State Storage Forecast, 2022–2026 (May 2022) IDC #US47831722

RAS Features

- PCIe® architecture supports a very high-level set of Reliability, Availability, Serviceability (RAS) features
- All transactions protected by CRC-32 for non-Flit Mode and 6B FEC + 8B CRC for Flit Mode and Link level Retry, covering even dropped packets
- Error injection mechanism along with elaborate error logging in Flit Mode
- Transaction level time-out support (hierarchical)
- Well defined algorithm for different error scenarios
- Advanced Error Reporting mechanism
- Support for degraded link width / lower speed
- Support for hot-plug (planned and surprise)

DPC/ eDPC for RAS

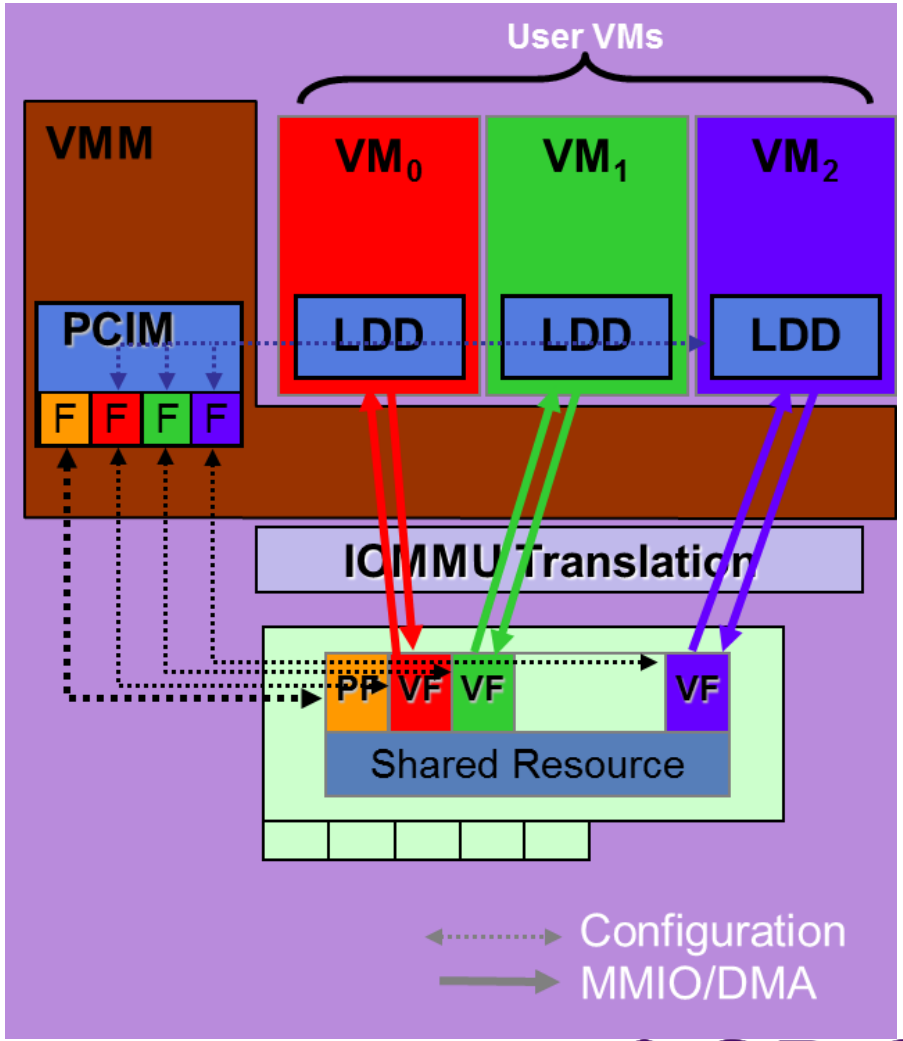
- (enhanced) Downstream Port Containment (DPC and eDPC) for emerging usages
- Emerging PCIe® technology usage models are creating a need for improved error containment/recovery and support for asynchronous removal (a.k.a. hot-swap)
- Defines an error containment mechanism, automatically disabling a Link when an uncorrectable error is detected, preventing potential spread of corrupted data
- Reporting mechanism with Software capability to bring up the link after clean up
- Transaction details on a timeout recorded (side-effect of asynchronous removal)
- eDPC: Root-port specific programmable response to gracefully handle DPC downstream



- **Error or Failure in one SSD should not bring down all SSDs**
- Asynchronous removal of SSDs is common

I/O Virtualization

- Reduces System Cost and power
- Single Root I/O Virtualization Specification
  - Released September 2007
  - Allows for multiple Virtual Machines (VM) in a single Root Complex to share a PCI Express® (PCIe®) adapter
- An SR-IOV endpoint presents multiple Virtual Functions (VF) to a Virtual Machine Monitor (VMM)
  - VF allocated to VM => direct assignment
- Address Translation Services (ATS) supports:
  - Performance optimization for direct assignment of a Function to a Guest OS running on a Virtual Intermediary (Hypervisor)
- Page Request Interface (PRI) supports:
  - Functions that can raise a Page Fault
- Process Address Space ID enhancement to support Direct assignment of I/O to user space

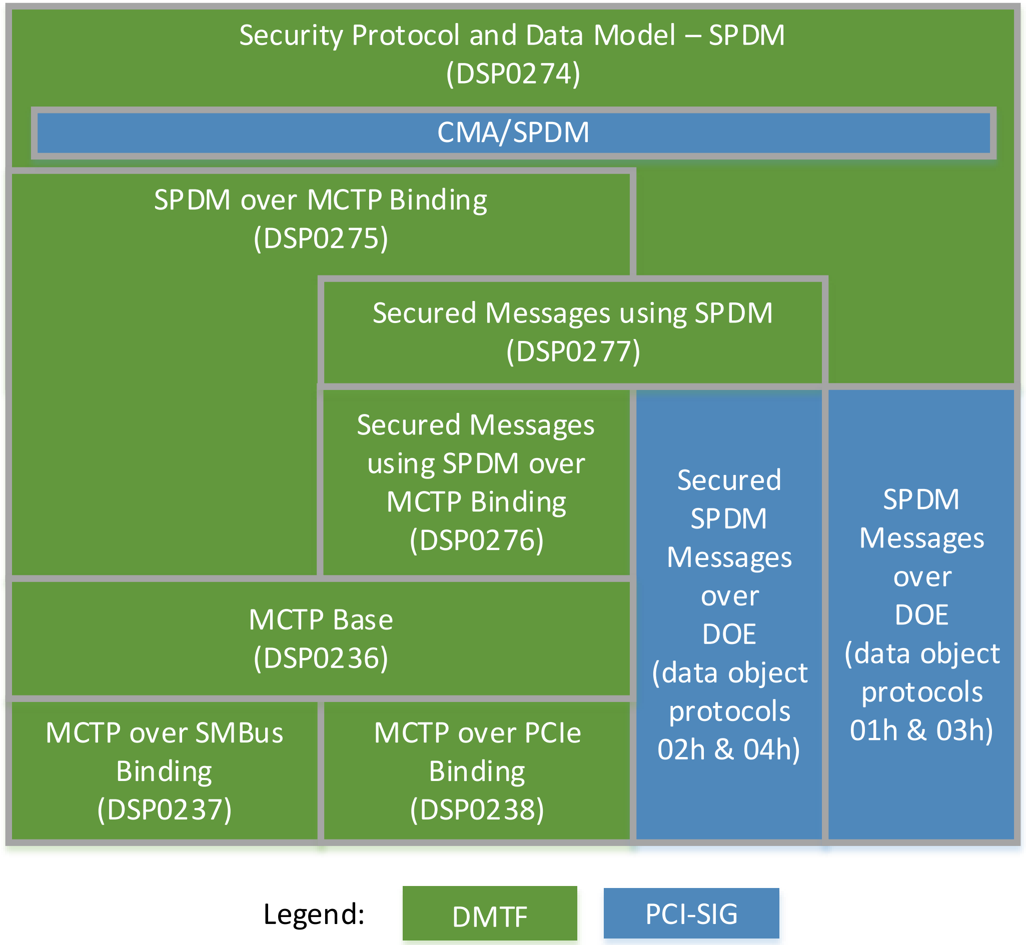


PCIe® Specification Security Capabilities

- Rationale: Key assets warrant improved security
  - Consumers: data integrity, confidentiality
  - Businesses & suppliers: reputation, revenue-stream, intellectual property, business continuity
  - Governments: national security, defense, elections, infrastructure
- Goal: Define foundational security capabilities for a wide spectrum of systems / devices / components
  - PCIe technology has a broad reach: Smart phones, tablets, PCs, servers, switches / routers, processors, memory/storage/IO modules, IoT devices, vehicles and more
  - Build on industry developments to provide consistency across multiple technologies – PCIe, CXL, USB, etc.
    - Including DMTF's (Distributed Management Task Force) Security Protocol and Data Model (SPDM) and Management Component Transport Protocol (MCTP) specifications
  - Build upon existing security standards (ISO, NIST, IEEE…) that are interconnect agnostic
  - Protect against multiple attacks: supply chain, physical, persistent, malicious components, etc.

Ack: Dave Harriman, Joe Cowan

PCI-SIG® & DMTF Specifications for Security



Security Protocol and Data Model – SPDM (DSP0274)
CMA/SPDM
SPDM over MCTP Binding (DSP0275)
Secured Messages using SPDM (DSP0277)
Secured Messages using SPDM over MCTP Binding (DSP0276)
Secured SPDM Messages over DOE (data object protocols 02h & 04h)
SPDM Messages over DOE (data object protocols 01h & 03h)
MCTP Base (DSP0236)
MCTP over SMBus Binding (DSP0237)
MCTP over PCIe Binding (DSP0238)
Legend:
DMTF
PCI-SIG

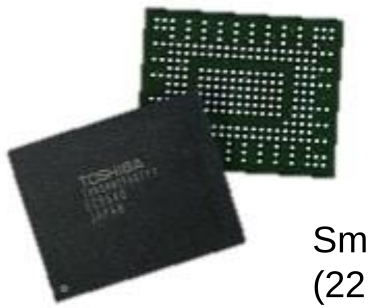
- SPDM defines a “toolkit” for authentication, measurement, and other security capabilities
- CMA/SPDM defines how SPDM is applied to PCIe® devices/systems
- DOE supports Data Object transport between host CPUs & PCIe components over PCIe technology
- Various MCTP bindings support Data Object transport over different interconnects
- IDE provides confidentiality, integrity, and replay protection for PCIe Transaction Layer Packets (TLPs)
- TDISP defines the security architecture and protocol device interface assignment to TEEs

Ack: Dave Harriman

Form Factors

PCIe® Architecture: One Base Specification - Multiple Form Factors

BGA



16x20 mm small and thin platforms

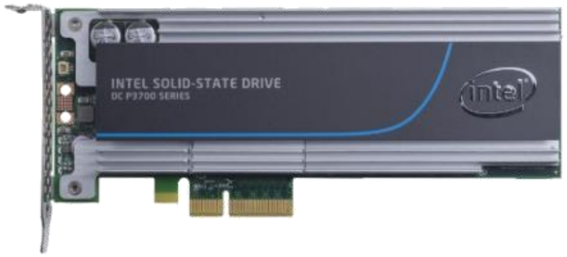
M.2

Smallest footprint (22mm x 30 to 110 mm): SSDs in boot slots, data center storage, WWAN

U.2 2.5in (aka SFF-8639)

SSDs x4 or 2 x2 w/ hot-plug

CEM Add-in-card



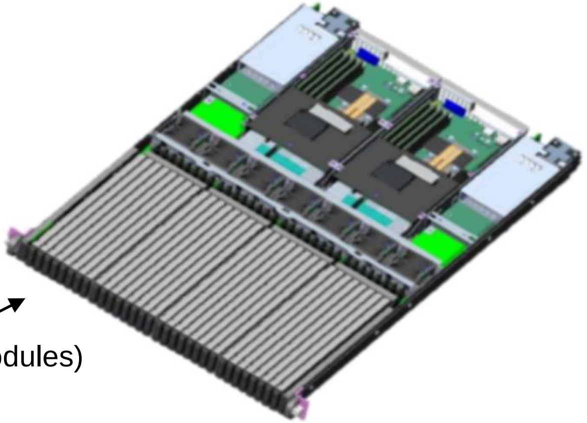
Widely used in systems w/ 4 HL options. Higher Power. Robust compliance program

High B/W: hand-held, IoT, automotive

High-end still and motion cameras

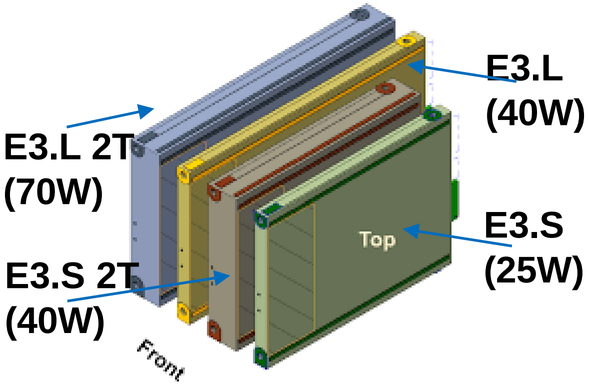
(Up to 36 Modules)

E1.S (SFF-TA-1006)



(Up to 32 Modules)

E1.L (SFF-TA-1007)

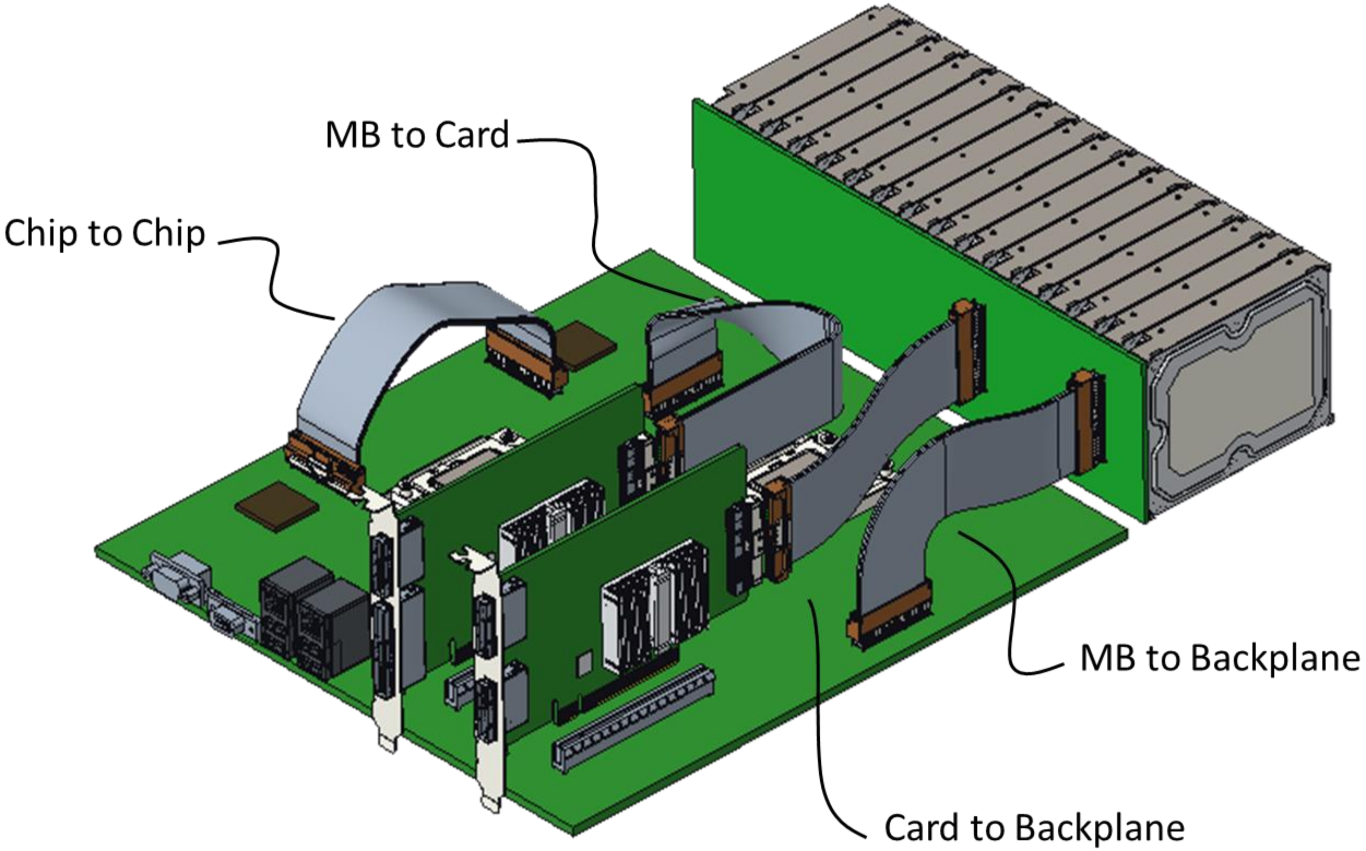


E3 Form-factors

Various Proprietary FFs for HPC Applications Multi-KW cards

Multiple Form-factors from the same silicon to meet the needs of different segments

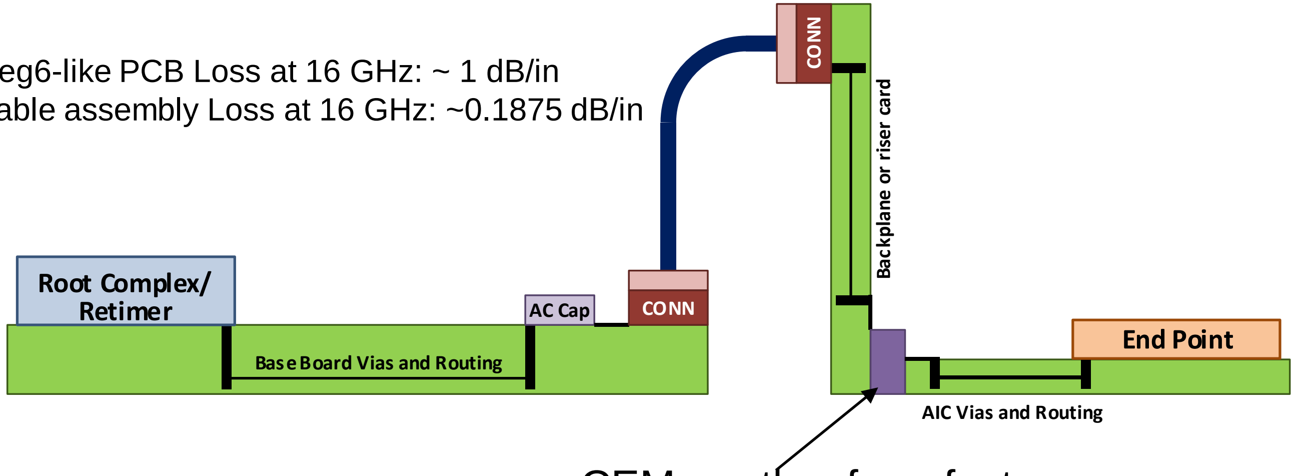
Cable Topology Support at Higher Speeds



Cable mitigates PCB loss limitations at 32+ GT/s and enables architectural flexibility

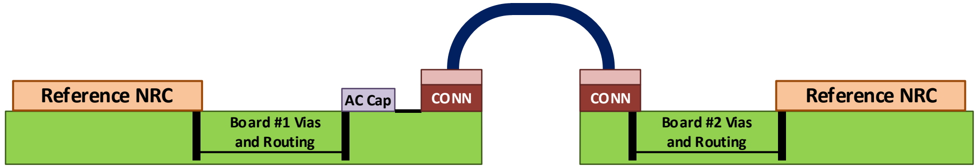
Internal and External Cable Topologies

Meg6-like PCB Loss at 16 GHz: ~ 1 dB/in
Cable assembly Loss at 16 GHz: ~0.1875 dB/in

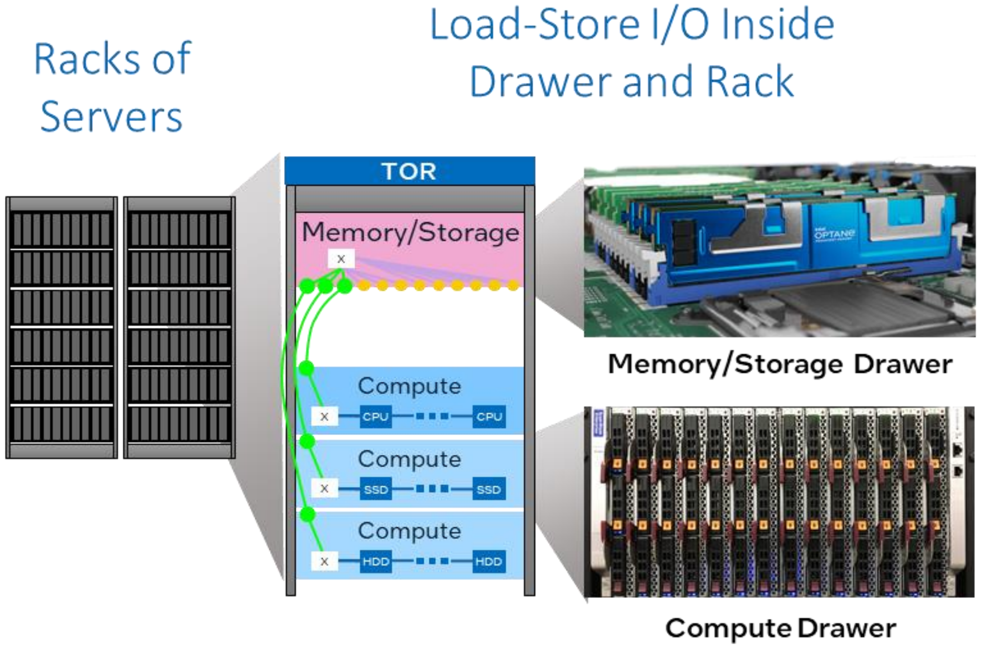


CEM or other form factor

A Typical Internal Cable Topology
(e.g., connecting a Riser/Backplane to the system board)



A Typical External Cable Topology
(e.g., connecting two boards within a rack)



Ack: Debendra's Keynote at FMS 2022

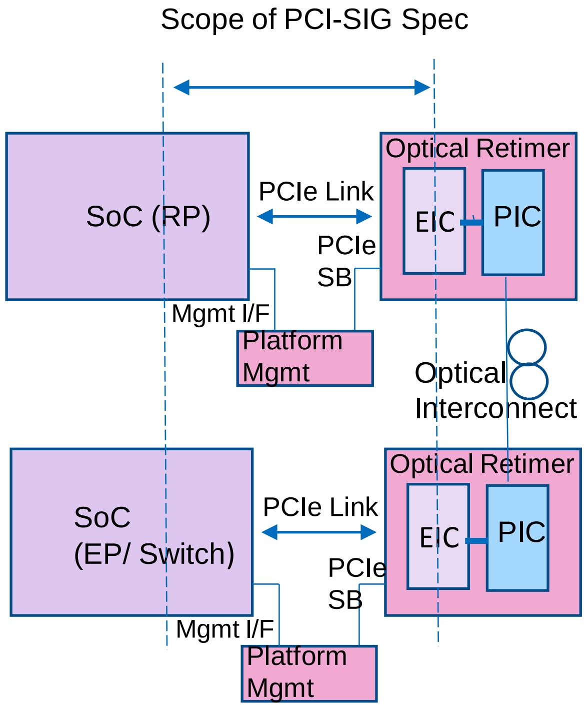
(Rack level dis-aggregation with CXL/ PCIe® technology enabled by PCIe cables – Electrical and Optical)

Development of internal and external cable specs for 32 and 64 GT/s are work-in-progress

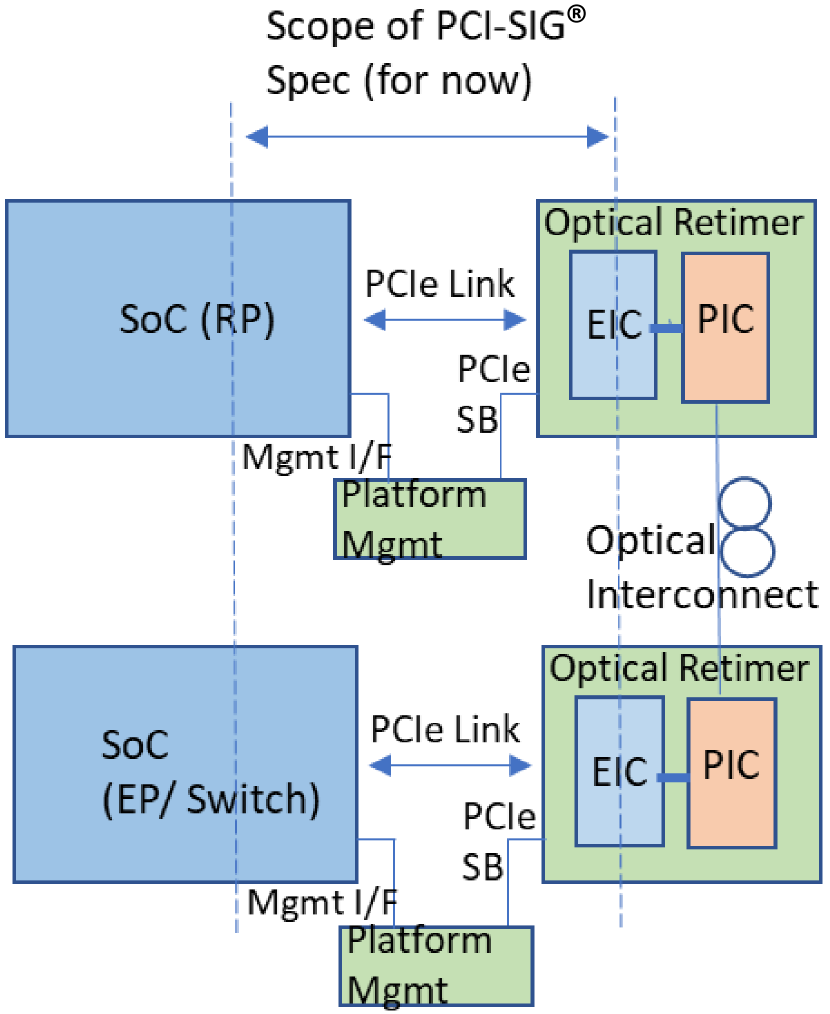
Optical-friendly PCI Express® Technology

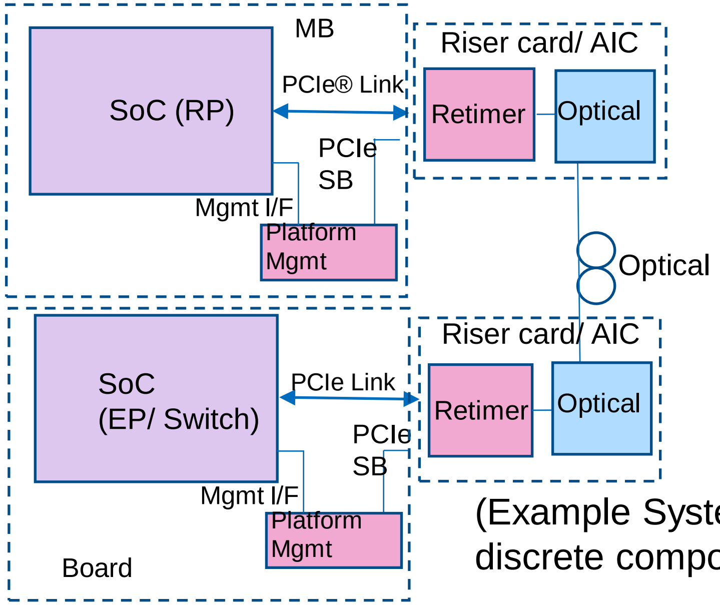
Problem Statement

- Optical has the promise of high bandwidth density and reach across a Rack/ Pod
  - Use case: Resource Pooling/ Sharing; Using PCIe® technology for developing composable systems with fabric topologies
  - Pros: Small (unlike copper cables which occupy more space) and reach (order of tens of meters vs Cu 1m/2m)
  - Cons: multiple technologies, cost vs copper – let this play out
- PCI-SIG® has launched an optical cable WG:
  - PHY Logical enhancements (comprehend sideband, mapping of mainband bit stream), including a Retimer-based approach
    - Enhancements to the Port / Pseudo-Port depending on the chosen implementation
  - Form-factor, if needed
  - Define in a way to enable on-package optics
- <u>Assumptions:</u>
  - Same type of Retimer is in both ends (optical technology neutrality)
  - EIC-PIC interface does not need to be defined
  - Complementary to copper cable work (different reach optimization)



Some Possible Implementations

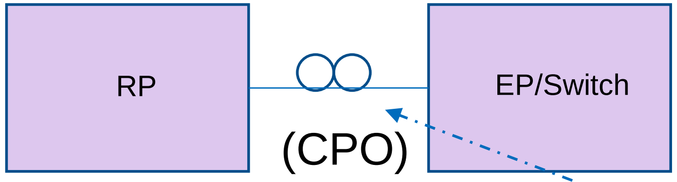




(Example System with discrete components)

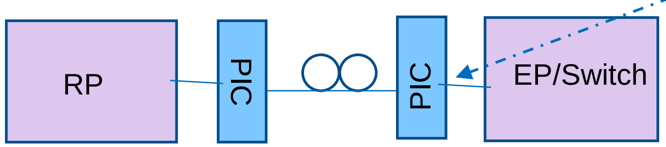
Can also have a copper cable connect the two Retimers (i.e., no optical)
(Want no separate channel/ wire/ fiber for side-band after Retimer)

(a: Retimer/ EIC/ PIC on riser/AIC)



(b: Retimer/ EIC/ PIC on package)

The optical or EIC-PIC signaling will not be spec'ed but the same Flit/ Ordered Sets defined can be used



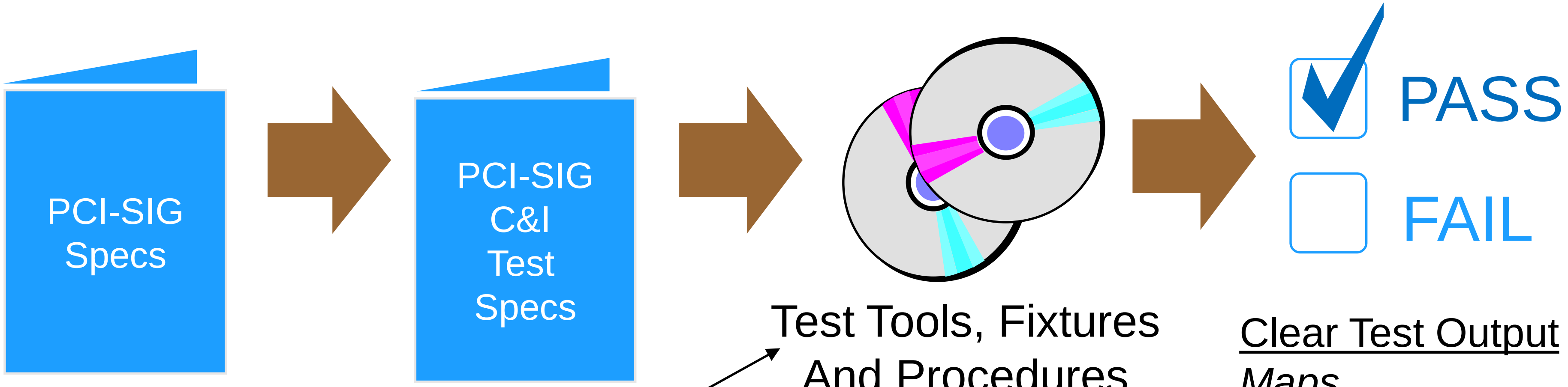
(c: Integrated EIC on package, Discrete PIC)

Same Base Spec ECN(s). Form-Factor can be looked at separately for various implementations

(Multiple permutations for on/off-package possible on either end: Port, Retimer/ EIC and PIC on 3, 2, 1 packages. Each side can have different levels of integration)

Compliance

PCI-SIG®: From Spec to Compliance



<u>PCI-SIG Specs</u>
Describes
Requirements

Base specification and Form-Factor specs

<u>C&I (Compliance and Interop) Test Specs</u>
Describes
Test criteria based on spec requirements

- Test Definitions
- Pass/Fail Criteria

Test Tools, Fixtures And Procedures

<u>Test H/W & S/W</u>
Validates
Test criteria

- Compliance
- Interoperability

<u>Clear Test Output</u>
Maps

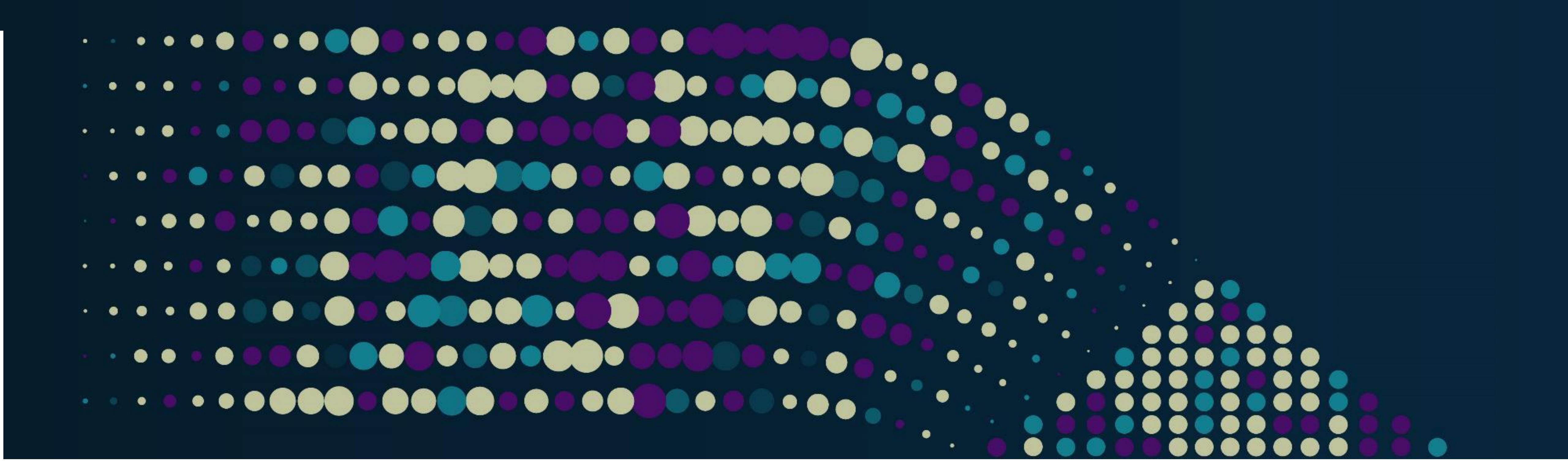
- Directly to Test Spec

Predictable path to design compliance

Conclusions

Source: Intel Corporation

- Single standard covering the entire compute continuum
- Predominant direct I/O interconnect from CPU with high bandwidth and used for alternate protocols with coherency and memory semantics
    - Low-power, High-performance
- Currently working on 7th generation: 128 GT/s, PAM4, Same FEC/ CRC/ Retry mechanism as PCIe® 6.0 specification with full backward compatibility
    - Expecting flat latency, high reliability, and improved power efficiency
- A robust and mature compliance and interoperability program



Please take a moment to rate this session.

Your feedback is important to us.