PCI EXPRESS® 6.0: A LOW LATENCY, HIGH BANDWIDTH, HIGH RELIABILITY AND COST-EFFECTIVE INTERCONNECT WITH 64.0 GT/S PAM-4 SIGNALING

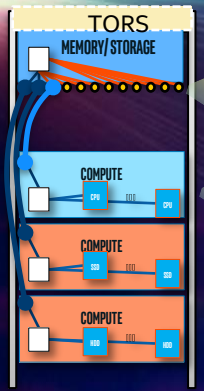
Dr. Debendra Das Sharma
Intel Fellow and Director of I/O Technologies and Standards
Data Center Group
Director, PCI-SIG Board and co-chair CXL Board Technical Task Force

Agenda

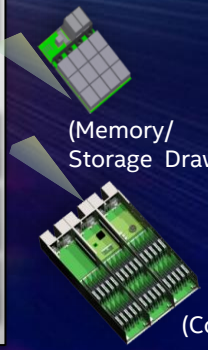
- Introduction to PCI Express® and Load-Store Interconnects
- Evolution of Data Rates in PCI Express
- Key Metrics and Requirements for PCIe 6.0
- PAM-4 and Error Assumptions/ Characteristics
- Error Correction and Detection: FEC, CRC, and Retry
- Flit Mode
- Key Metrics and Requirements for PCIe 6.0 – Evaluation
- Conclusions

CLOUD COMPUTING LANDSCAPE TODAY

(RACK OF SERVERS)



LD/ ST I/O INSIDE DRAWER AND RACK

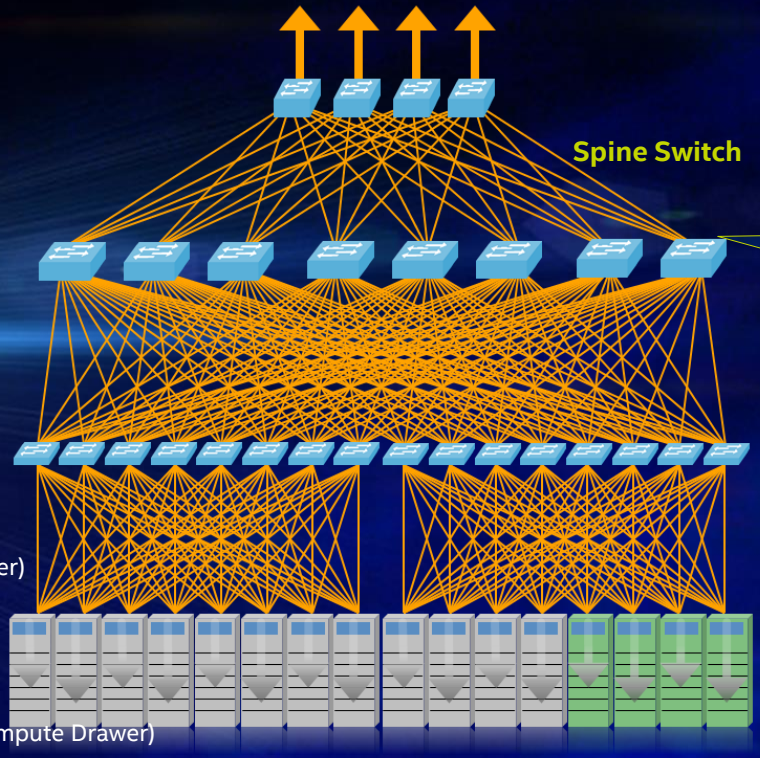


(Memory/ Storage Drawer)

(Compute Drawer)

Core/ Edge Network & Inter-DC Network

Spine Switch



Hyperscale Data Center and Edge

Silicon Photonics

High-bandwidth connectivity at 100G and beyond

Programmable Switching

P4-programmable scale-out fabric with uncompromising performance

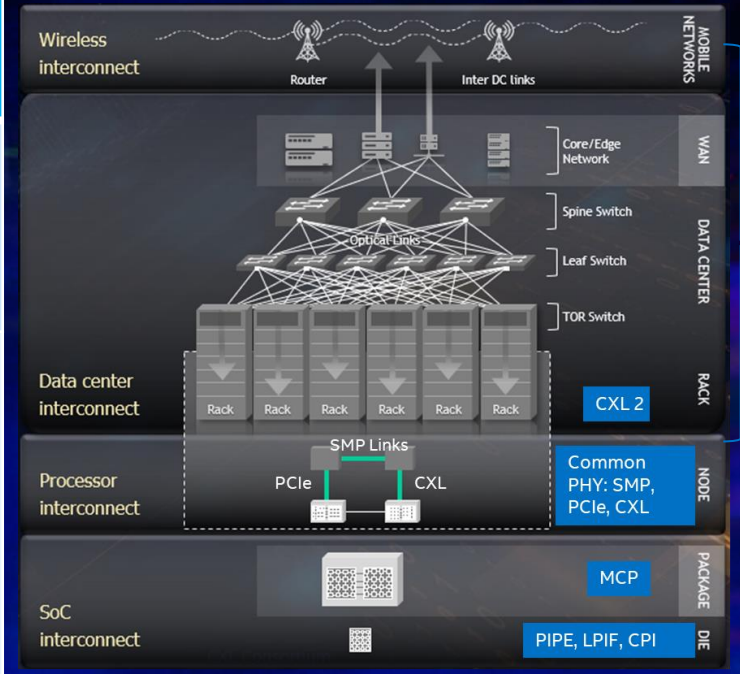
Rack of Servers

Programmable infrastructure acceleration for demanding data movement with Smart NIC

Data Center as a Computer –Interconnects are key to driving warehouse scale efficiency!

TAXONOMY, CHARACTERISTICS, AND TRENDS OF INTERCONNECTS

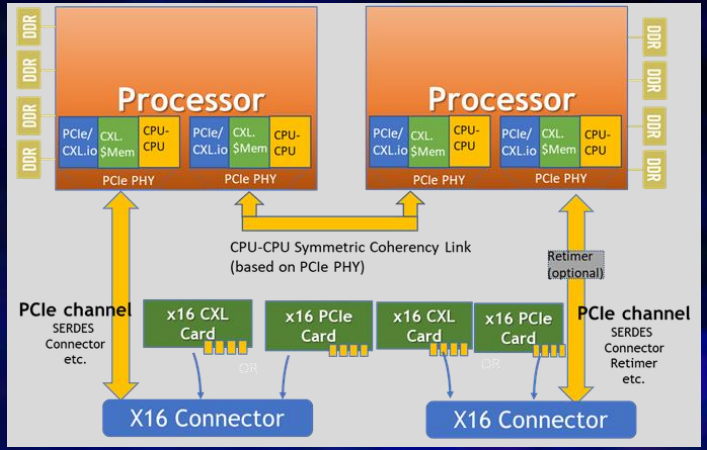
Category	Type and Scale	Data Rate/ Characteristics	PHY Latency (Tx + Rx)
Latency Tolerant (Narrow, fast)	Networking / Fabric Data Center Scale	56/ 112 GT/s-> 224 GT/s (PAM4) 4-8 Lanes, cables/ backplane	100+ ns w/ FEC (20ns+ w/o FEC)
Latency Sensitive (Wide, high speed)	Load-Store I/O Arch. Ordering (PCIe/ CXL / UPI) Node level -> sub-Rack	32 GT/s (NRZ) -> PCIe Gen6 64 GT/s (PAM4) Hundreds of Lanes Power, Cost, Si-Area, Backwards Compatible, Latency, On-board -> cables/ backplanes	<10ns (Tx+ Rx: PHY-PIPE) 0-1ns FEC overhead



Latency Sensitive Load-Store I/O moving to 64.0 GT/s using PAM-4: innovations on track to meet latency, area, and cost challenges

Load-Store I/O: 101

- Ability to directly access memory (CPU, I/O)
- Memory mapped into system memory space
  - Coherent or Non-coherent access
  - Accesses across PCIe non-coherent
  - Accesses across CXL can be either
- Some form of ordering or cache coherency
  - PCIe: Producer-Consumer Ordering Semantics
- Transactions are guaranteed to be delivered and completed in a reasonable time
  - no dropped packets, no software based retry
  - typically hardware based link level replay on error
- Timeout and Error reporting hierarchy
  - Hardware based error containment guarantees

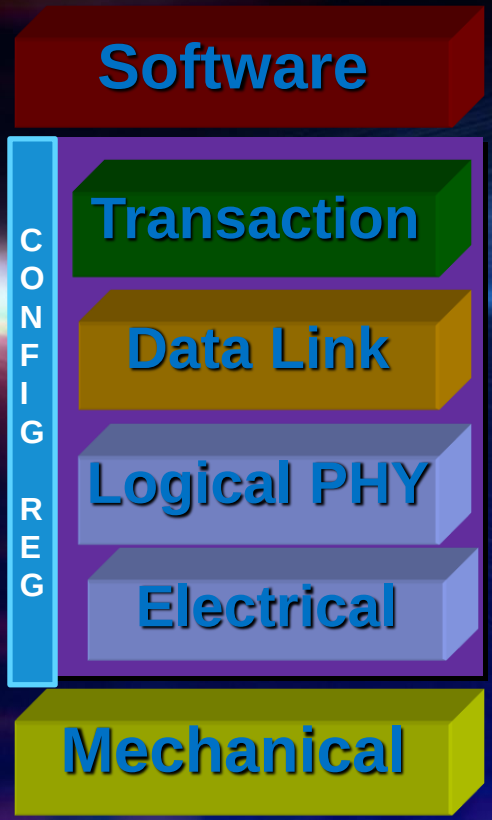


(Load-Store I/O (PCIe, CXL, SMP coherency) based on a common PCIe PHY => PCIe needs to stay Low-latency with 0-latency add generationally)

Device A	Device B
Write Data	Read Flag
Write Flag	Read Data

(Producer Consumer Ordering Model: Reading updated Flag guarantees reading updated Data)

PCI EXPRESS: LAYERED PROTOCOL



⇦ PCI compatibility, configuration, driver model
⇦ PCIe architecture enhanced configuration model

⇦ Split-transaction, packet-based protocol
⇦ Credit-based flow control, virtual channels

⇦ Logical connection between devices
⇦ Reliable data transport services (CRC, Retry, Ack/Nak)

⇦ Physical information exchange
⇦ Interface initialization and maintenance

⇦ Market segment specific form factors
⇦ Evolutionary and revolutionary

PCI-SIG®: An Open Industry Consortium

Founded in 1992

Organization that **defines the PCI Express® (PCIe®) I/O bus specifications and related form factors**

830+ member companies located worldwide

Creating specifications and mechanisms to **support compliance** and **interoperability**

Board of Directors 2020 – 2021

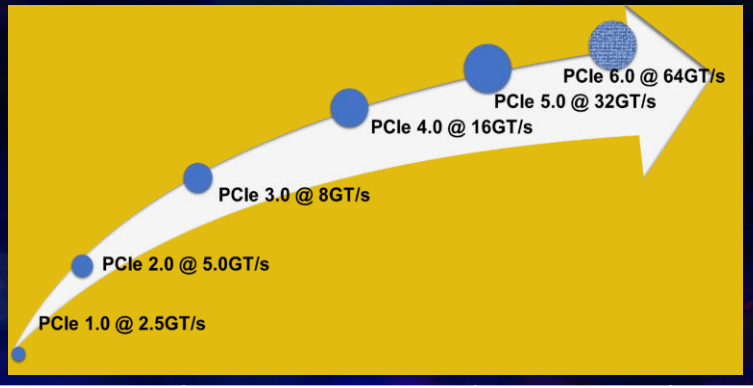
AMD
DELL EMC
IBM
intel
KEYSIGHT TECHNOLOGIES
NVIDIA
Qualcomm
SYNOPSYS
Silicon to Software

Agenda

- Introduction to PCI Express® and Load-Store Interconnects
- Evolution of Data Rates in PCI Express
- Key Metrics and Requirements for PCIe 6.0
- PAM-4 and Error Assumptions/ Characteristics
- Error Correction and Detection: FEC, CRC, and Retry
- Flit Mode
- Key Metrics and Requirements for PCIe 6.0 – Evaluation
- Conclusions

Evolution of PCI-Express

- Double data rate every gen in ~3 years
- Full backward compatibility
- Ubiquitous I/O: PC, Hand-held, Workstation, Server, Cloud, Enterprise, HPC, Embedded, IoT, Automotive, AI
- One stack / silicon, multiple form-factors
- Different widths (x1/ x2/ x4/ x8/ x16) and data rates fully inter-operable
  - a x16 Gen 5 interoperates with a x1 Gen 1!
- PCIe deployed in all computer systems since 2003 for all I/O needs

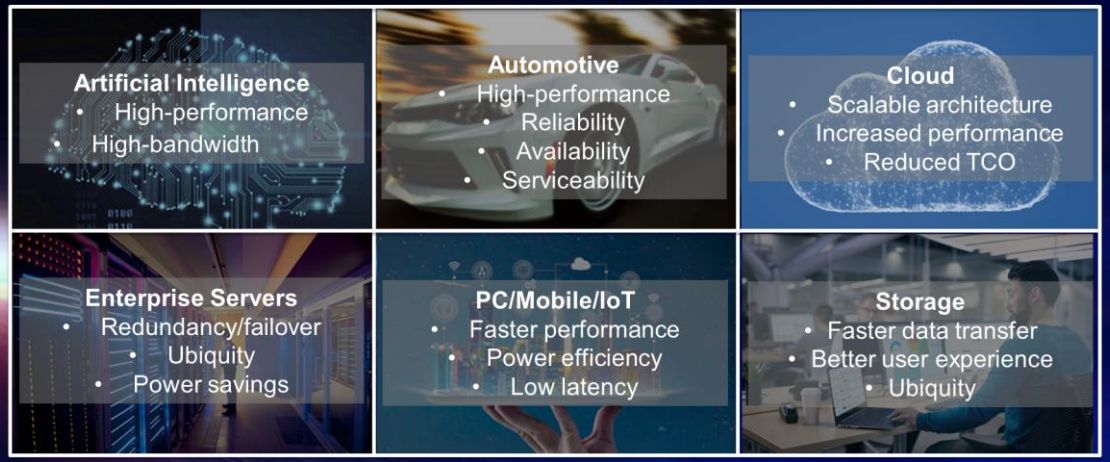


PCIe Specification	Data Rate(Gb/s) (Encoding)	x16 B/W per dirn**	Year
1.0	2.5 (8b/10b)	32 Gb/s	2003
2.0	5.0 (8b/10b)	64 Gb/s	2007
3.0	8.0 (128b/130b)	126 Gb/s	2010
4.0	16.0 (128b/130b)	252 Gb/s	2017
5.0	32.0 (128b/130b)	504 Gb/s	2019
6.0 (WIP)	64.0 (PAM-4, Flit)	1024 Gb/s (~1Tb/s)	2021*

* - Projected ** - bandwidth after encoding overhead

PCIe continues its impressive run of doubling bandwidth for six generations spanning 2 decades!

Bandwidth Drivers for PCIe 6.0



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

- Device side: Networking (800G in early 2020s), Accelerators, FPGA/ ASICs, Memory
- Alternate Protocols on PCIe
- As the per socket compute capability grows at an exponential pace, so does I/O needs – we have already added a lot of Lanes per socket (currently 128 Lanes) => 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

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

Key Metrics for PCIe 6.0: Requirements

Metrics	Requirements
Data Rate	64 GT/s, PAM4 (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 PAM-4)
Bandwidth Inefficiency	<2 % adder over PCIe 5.0 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 under similar set up for Retimer(s) (maximum 2)
Power Efficiency	Better than PCIe 5.0
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
Others	HVM-ready, cost-effective, scalable to hundreds of Lanes in a platform

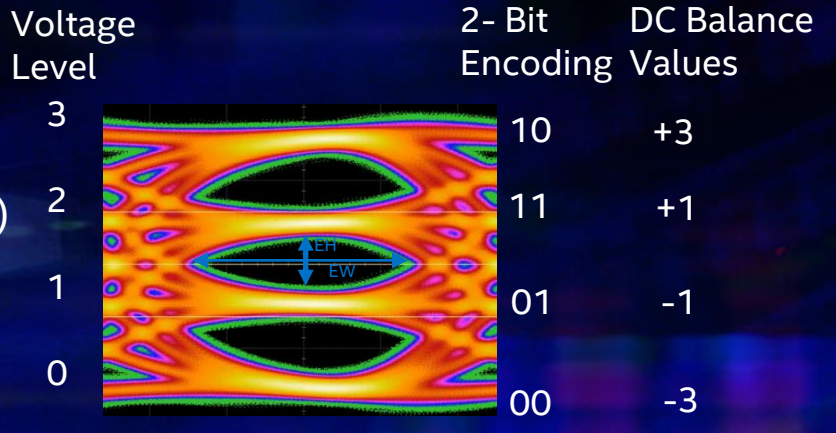
Need to make the right trade-offs to meet each of these metrics!

Agenda

- Introduction to PCI Express® and Load-Store Interconnects
- Evolution of Data Rates in PCI Express
- Key Metrics and Requirements for PCIe 6.0
- PAM-4 and Error Assumptions/ Characteristics
- Error Correction and Detection: FEC, CRC, and Retry
- Flit Mode
- Key Metrics and Requirements for PCIe 6.0 – Evaluation
- Conclusions

PAM-4 Signaling at 64.0 GT/s

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



Encoding per UI (2bit)	Tx Voltage	Rx Voltage (V)
00	-Vtx	V <= Vth1
01	-Vtx/3	Vth1 < V <= Vth2
11	+Vtx/3	Vth2 < V <= Vth3
10	+Vtx	V > Vth3

Error Assumptions and Characteristics w/ PAM-4

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 FBER and the error correlation in a Lane and across Lanes

Agenda

- Introduction to PCI Express® and Load-Store Interconnects
- Evolution of Data Rates in PCI Express
- Key Metrics and Requirements for PCIe 6.0
- PAM-4 and Error Assumptions/ Characteristics
- Error Correction and Detection: FEC, CRC, and Retry
- Flit Mode
- Key Metrics and Requirements for PCIe 6.0 – Evaluation
- Conclusions

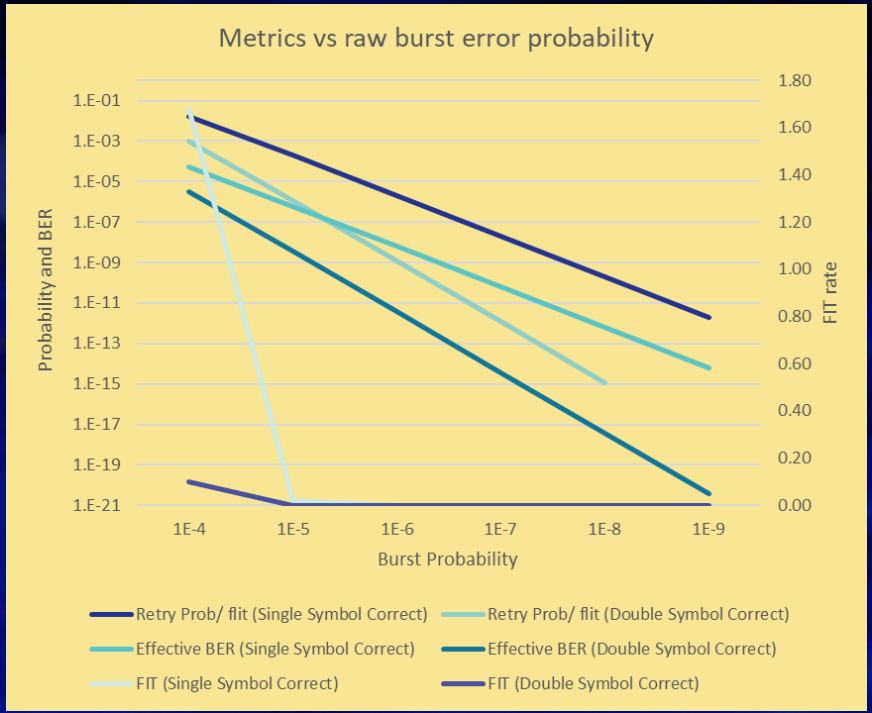
Handling Errors and Evaluation Metrics

- Two mechanisms to correct errors
  - Correction through FEC (Forward Error Correction)
    - Latency and complexity increases exponentially with the number of Symbols corrected
  - Detection of errors by CRC => Link Level Retry (a strength of PCIe)
    - Detection is linear: latency, complexity and bandwidth overheads
    - Need a robust CRC to keep FIT << 1 (FIT: Failure in Time – No of failures in 10^9 hours)
- Metrics: Prob of Retry (or b/w loss due to retry) and FIT
- Need to use both means of correction to achieve:
  - Low latency and complexity
  - Retry probability at acceptable level (no noticeable performance impact)
  - Low Bandwidth overhead due to FEC, CRC, and retry

Need to keep FEC correction latency low (2ns) to meet the performance needs of Load/Store I/O

Our Approach: Light-weight FEC and Retry

- Light-weight FEC & strong CRC
- FEC gets to a reasonable retry rate
- Keep latency (including retry) low
- We are better off retrying a packet with 10^{-6} (or 10^{-5}) probability with a retry latency of 100ns
  - better than always paying a FEC latency impact of 150ns+ in networking



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

Agenda

- Introduction to PCI Express® and Load-Store Interconnects
- Evolution of Data Rates in PCI Express
- Key Metrics and Requirements for PCIe 6.0
- PAM-4 and Error Assumptions/ Characteristics
- Error Correction and Detection: FEC, CRC, and Retry
- Flit Mode
- Key Metrics and Requirements for PCIe 6.0 – Evaluation
- Conclusions

Flit Mode with PCIe 6.0

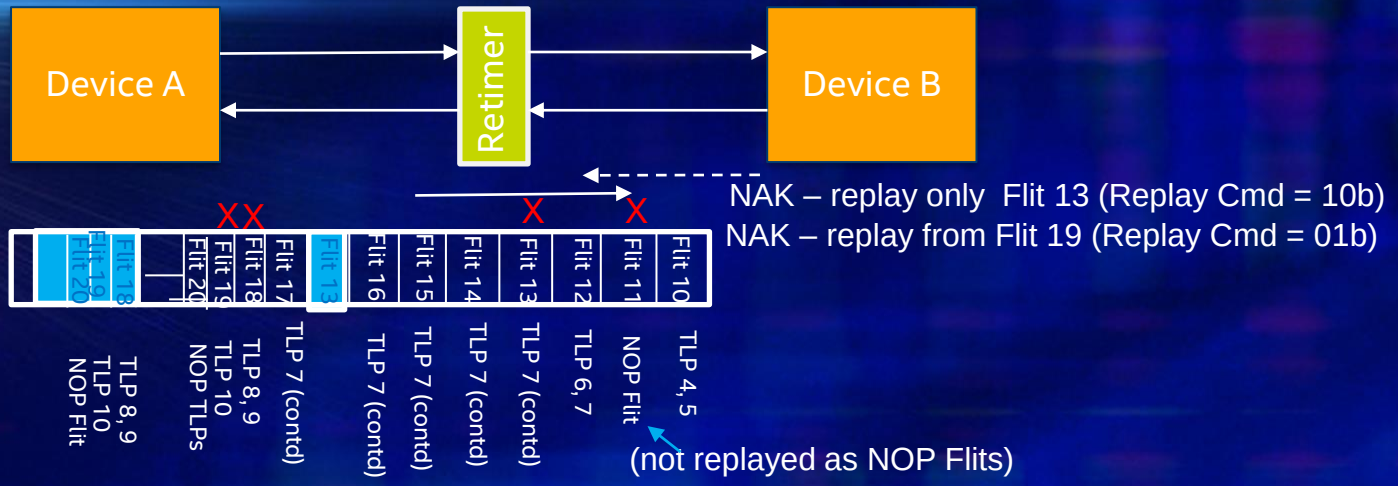
- Flit (flow control unit) based
- FEC needs fixed set of bytes
- Correction in flit => detection (CRC) in flits => Retry at flit level
- Lower data rates also in Flit Mode if enabled
- Flit size: 256B
  - 236B TLP, 6B DLP, 8B CRC, 6B FEC
  - No Sync hdr, Framing Token or per packet CRC
  - Improved b/w utilization w/ overhead amortization
  - Flit accumulation Latency:
    - 2ns x16, 4ns x8, 8 ns x4, 16 ns x2, 32 ns x1
  - Ack/ credit slot => low Latency/ low storage

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

Replay in Flit Mode

- Flit with NOP-only TLPs not replayed unless the subsequent flit also had an uncorrectable error
- On a replay, the Transmitter can choose to skip over the NOP Flits
- All replayed flits have the Replay Cmd = 11b (w/ Tx sequence number sent)



Retry Probability and FIT vs FBER/ Correlation

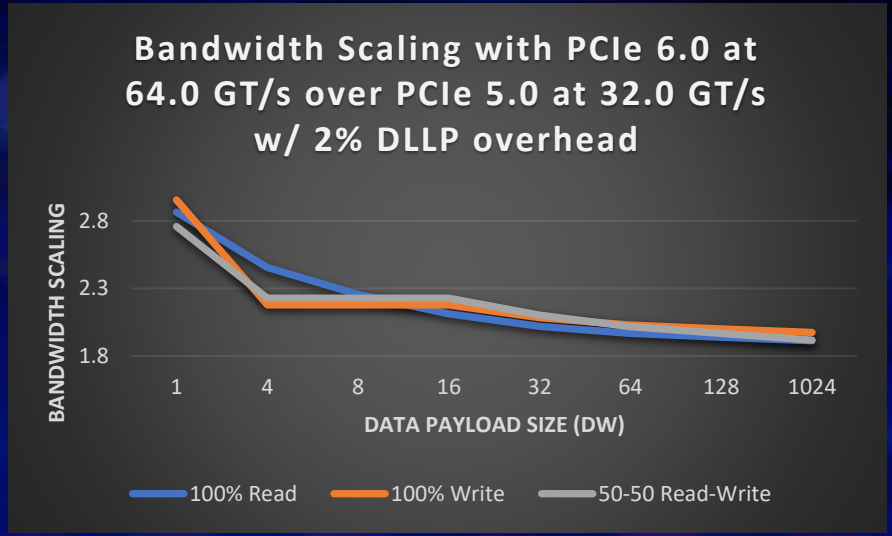
- Single Symbol Correct interleaved FEC plus 64-b CRC works really well for raw FBER of 1E-6 even with high Lane correlation
  - –Retry probability per flit is 5 x 10^{-6}
  - –B/W loss is 0.05% even with go-back-n
  - –FIT is almost 0
  - –Can mitigate the bandwidth loss significantly by adopting retry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

FBER 1E-6 meets the performance goals with a light-weight FEC

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

PCIe 6.0 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 2
- 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



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

(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

Agenda

- Introduction to PCI Express® and Load-Store Interconnects
- Evolution of Data Rates in PCI Express
- Key Metrics and Requirements for PCIe 6.0
- PAM-4 and Error Assumptions/ Characteristics
- Error Correction and Detection: FEC, CRC, and Retry
- Flit Mode
- Key Metrics and Requirements for PCIe 6.0 – Evaluation
- Conclusions

Key Metrics for PCIe 6.0: Evaluation

Metrics	Expectations	Evaluation (Trend)
Data Rate	64 GT/s, PAM4 (double bandwidth per pin every generation)	Meets (must do)
Latency	<10ns adder for Transmitter + Receiver over 32.0 GT/s (including FEC)	Exceeds (Savings in latency with <10ns for x1/ x2 cases)
Bandwidth Inefficiency	<2 % adder over PCIe 5.0 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 hrs)	Meets
Channel Reach	Similar to PCIe 5.0 under similar set up for Retimer(s) (maximum 2)	Meets
Power Efficiency	Better than PCIe 5.0	Design dependent – expected to meet
Low Power	Similar entry/ exit latency for L1 low-power state Addition of a new power state (L0p) for scalable power consumption with bandwidth usage	Design dependent – expected to meet; L0p looks promising
Plug and Play	Fully backwards compatible with PCIe 1.x through PCIe 5.0	Meets
Others	HVM, cost-effective, scales to hundreds of Lanes in platform	Expected to Meet

On track to meet or exceed requirements on all key metrics

Conclusions

- PCIe 6.0 is at Rev 0.7 level; Rev 0.9 imminent
- Very challenging in multiple fronts
  - New signaling with PAM-4: tradeoff around errors/ correlation, channels, performance/ area, and circuit complexity to double the bandwidth
  - Metrics (latency, bandwidth efficiency, area, cost, power) which are significantly more challenging than what other standards have done with PAM-4 at lower speeds
    - e.g., 100+ ns FEC latency on other standards vs our single digit ns latency targets; 12+% bandwidth inefficiency in other standards vs <2% inefficiency targets for us)
  - We are on track to exceed or meet the requirements
  - Need to continue to do due diligence though analysis, simulations, and test silicon characterization to ensure we have a robust specification
  - We have the combined innovation capability of 800+ members with a track record of delivering flawlessly against challenges for more than two decades – we will deliver this time also!!
- The journey continues …

THANK YOU

FOR YOUR CONTINUED SUPPORT
TOGETHER WE WILL CONTINUE TO
BUILD GREAT PRODUCTS!!