2022 OFA Virtual Workshop

PCIE® 6.0 SPECIFICATION: A HIGH-PERFORMANCE I/O INTERCONNECT FOR ADVANCED NETWORKING APPLICATIONS

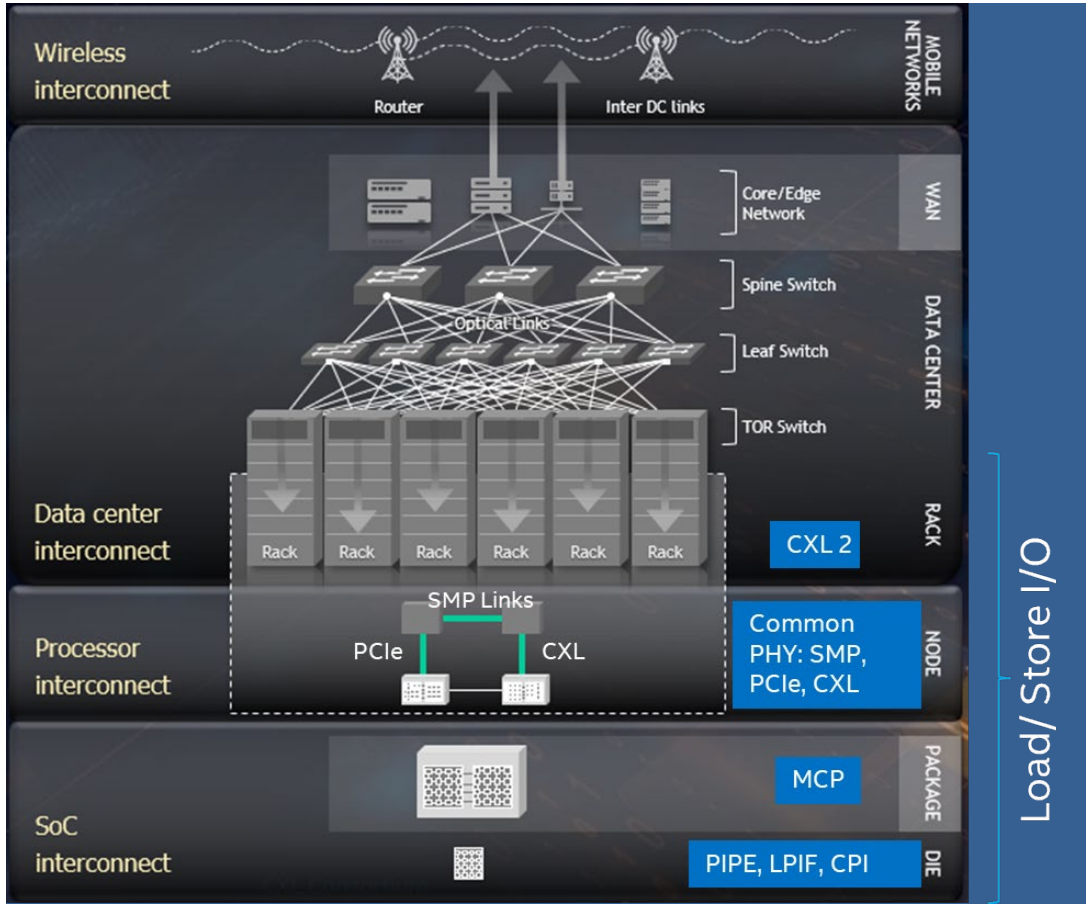
Dr. Debendra Das Sharma

Intel Senior Fellow, Intel Corporation
Director, PCI-SIG® Board

AGENDA

- Background
- Key Metrics and Requirements for PCIe® 6.0 Specification
- PAM4 and Error Assumptions/ Characteristics
- Error Correction and Detection: FEC, CRC, and Retry
- Flit Mode
- Low Power enhancements: L0p
- Key Metrics and Requirements for PCIe 6.0 Specification – Evaluation
- Conclusions and Call to Action

INTERCONNECT: AN IMPORTANT PILLAR OF COMPUTE



Datacenter Interconnect drives performance @ scale

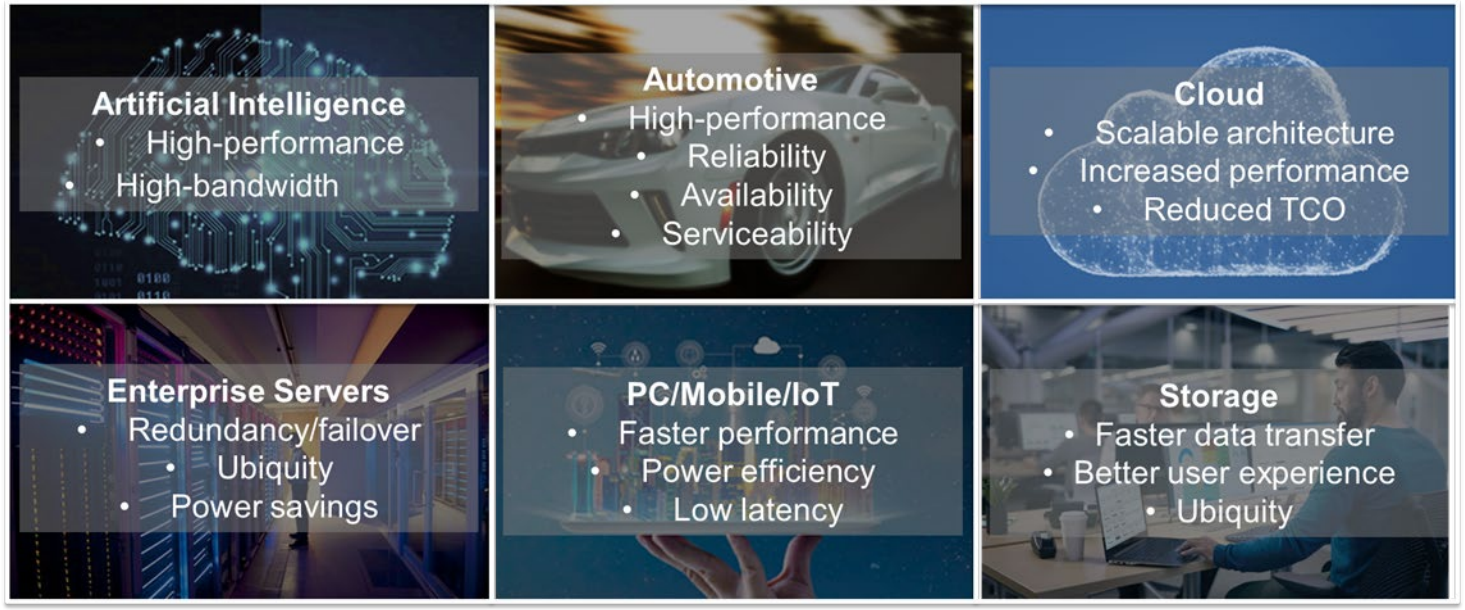
Warehouse scale computing:
compute performance, efficiency

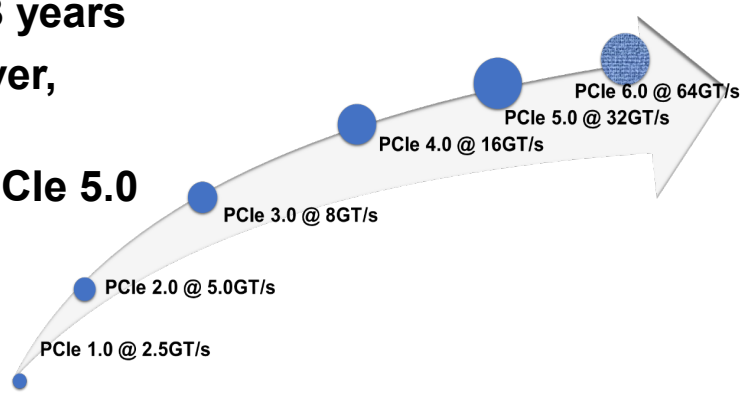
Processor interconnects (Load-Store I/O) drive Node/Rach level scaling a similar system for PC/ WS/ HH

SMP interconnects enable multi-socket scaling
PCIe® technology / CXL enables compute, memory and storage
Attach and dis-aggregation at Rack Level
Typically based on a common PCIe PHY
Networking at Rack level and beyond

PCIE® TECHNOLOGY: THE UBIQUITOUS LOAD-STORE I/O INTERCONNECT

- **PCIe® technology doubles the data rate 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**
- **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 Specification	Data Rate(GT/s) (Encoding)	Year
1.0	2.5 (8b/10b)	2003
2.0	5.0 (8b/10b)	2007
3.0	8.0 (128b/130b)	2010
4.0	16.0 (128b/130b)	2017
5.0	32.0 (128b/130b)	2019
6.0	64.0 (PAM4, Flit)	2022

(Bandwidth drivers for PCIe technology: Usages driving insatiable demand for compute, memory, storage, and networking bandwidth: PCIe technology is interconnect across these)

PCIe technology continues to deliver bandwidth doubling for six generations spanning 2 decades!

KEY METRICS FOR PCIE® 6.0 SPECIFICATION: REQUIREMENTS

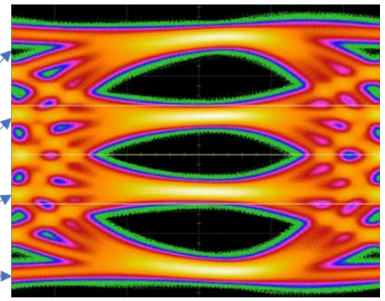
Metrics	Expectations
Data Rate	64GT/s, PAM4 (double the bandwidth per pin every generation)
Latency	<10ns adder for Transmitter + Receiver over 32.0 GT/s (including FEC) (We can not afford the 100ns FEC latency as networking does with PAM4)
Bandwidth Inefficiency	<2 % adder over PCIe 5.0 specification 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 specification through PCIe 5.0 specification
Others	HVM-ready, cost-effective, scalable to hundreds of Lanes in a platform

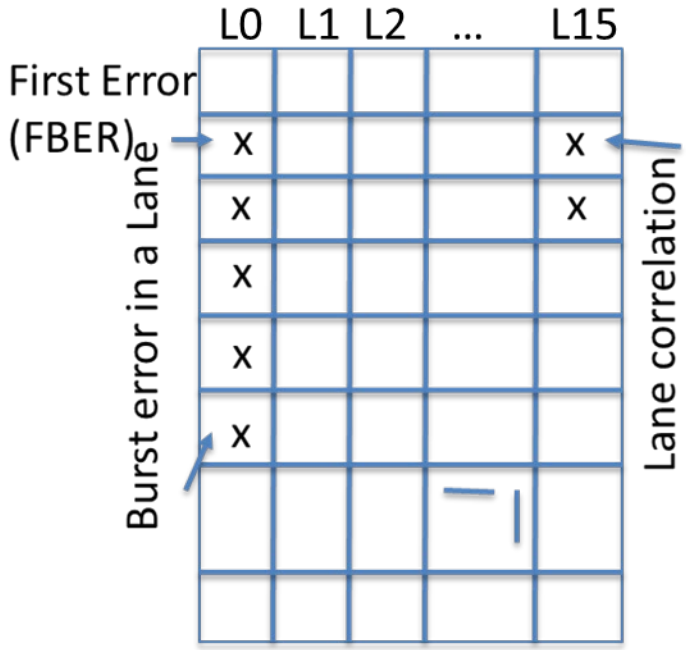
Need to make the right trade-offs to meet each of these metrics!

PAM4 SIGNALING AND ERROR ASSUMPTIONS FOR PCIE® 6.0 TECHNOLOGY

- **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**
  - Correlation of errors across Lanes (common source of errors such as power supply noise)
  - Correlation of errors on a Lane due to DFE (burst)
  - FBER: First bit error rate
- **Mitigation:**
  - Gray Coding to reduce errors in each UI
  - Precoding to minimize errors in a burst
  - Forward-error Correct (FEC) + Replay on CRC error

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



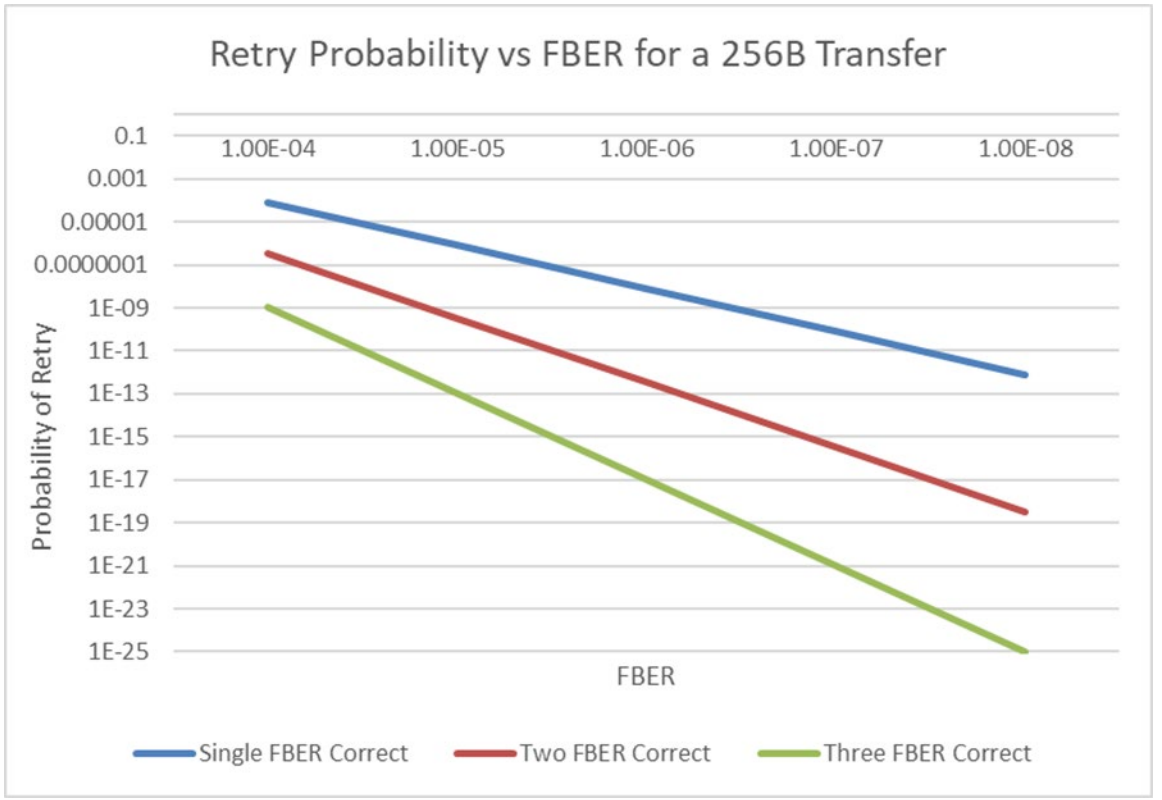


AGENDA

- Background
- Key Metrics and Requirements for PCIe® 6.0 Specification
- PAM4 and Error Assumptions/ Characteristics
- **Error Correction and Detection: FEC, CRC, and Retry**
- Flit Mode
- Low Power enhancements: L0p
- Key Metrics and Requirements for PCIe 6.0 Specification – Evaluation
- Conclusions and Call to Action

HANDLING ERRORS AND METRICS USED FOR EVALUATION

- **Two mechanisms to <u>correct</u> errors**
  - FEC (Forward Error Correction)
    - Latency and complexity increase exponentially with the number of Symbols corrected
  - Detection of errors by CRC => Link Level Retry (a strength of PCIe® technology)
    - 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)
- **<u>Metrics:</u> Prob of Retry (or b/w loss due to retry) and FIT**



(PCIe® 6.0 Specification Approach: Light-weight FEC with strong CRC and low-latency replay at link level – need FBER < 10^{-6})

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

AGENDA

- Background
- Key Metrics and Requirements for PCIe® 6.0 Specification
- PAM4 and Error Assumptions/ Characteristics
- Error Correction and Detection: FEC, CRC, and Retry
- **Flit Mode**
- Low Power enhancements: L0p
- Key Metrics and Requirements for PCIe 6.0 Specification – Evaluation
- Conclusions and Call to Action

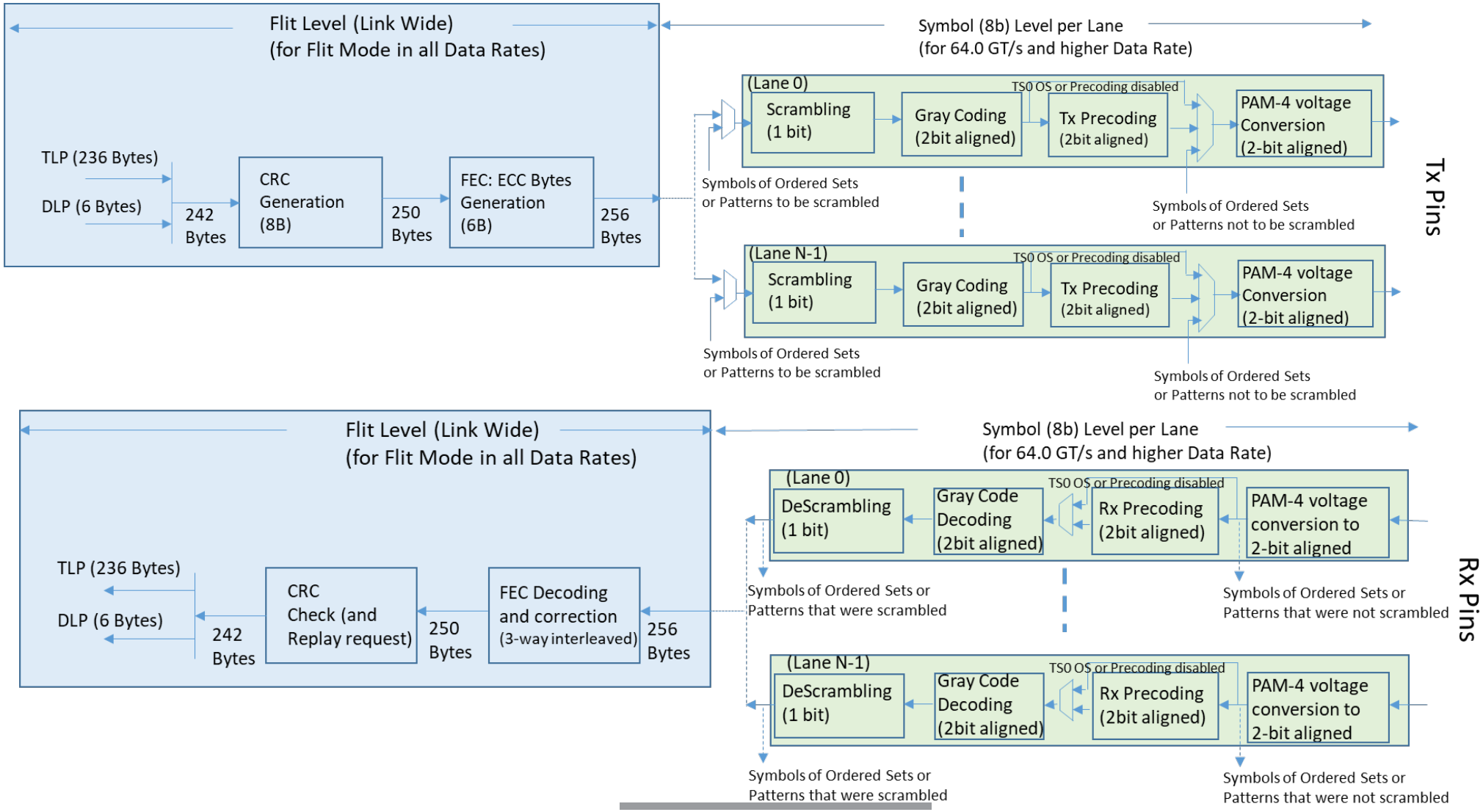
FLIT MODE: LOW LATENCY AND HIGH LINK EFFICIENCY

- **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
  - FEC: 3-way interleaved, single symbol correct – corrects a burst within 3 bytes
  - CRC: RS based – up to 8 byte errors guaranteed detect beyond that 2^{-64} aliasing
  - Low latency: <2ns FEC correct and CRC check
  - 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 with existing Go-Back-N retry**

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

TRANSMIT AND RECEIVE PATHS



REPLAY IN FLIT MODE

- **Once in Flit mode, we are always in Flit mode at all speeds (2.5 GT/s through 64.0 GT/s)**
- **NOP-Flits do not consume sequence number, not added to Tx Replay buffer, and not replayed**
- **Replayed Flits start with the Replay Cmd = 00b (w/ Tx sequence number sent) (alternate between 3 Flits with Replay Cmd = 00b followed by x Flits with Ack/ Nak; x = 3 in Normal Flit Exchange, x = 1 in Seq Num handshake)**
- **Ability to switch from selective Nak to Standard Nak**

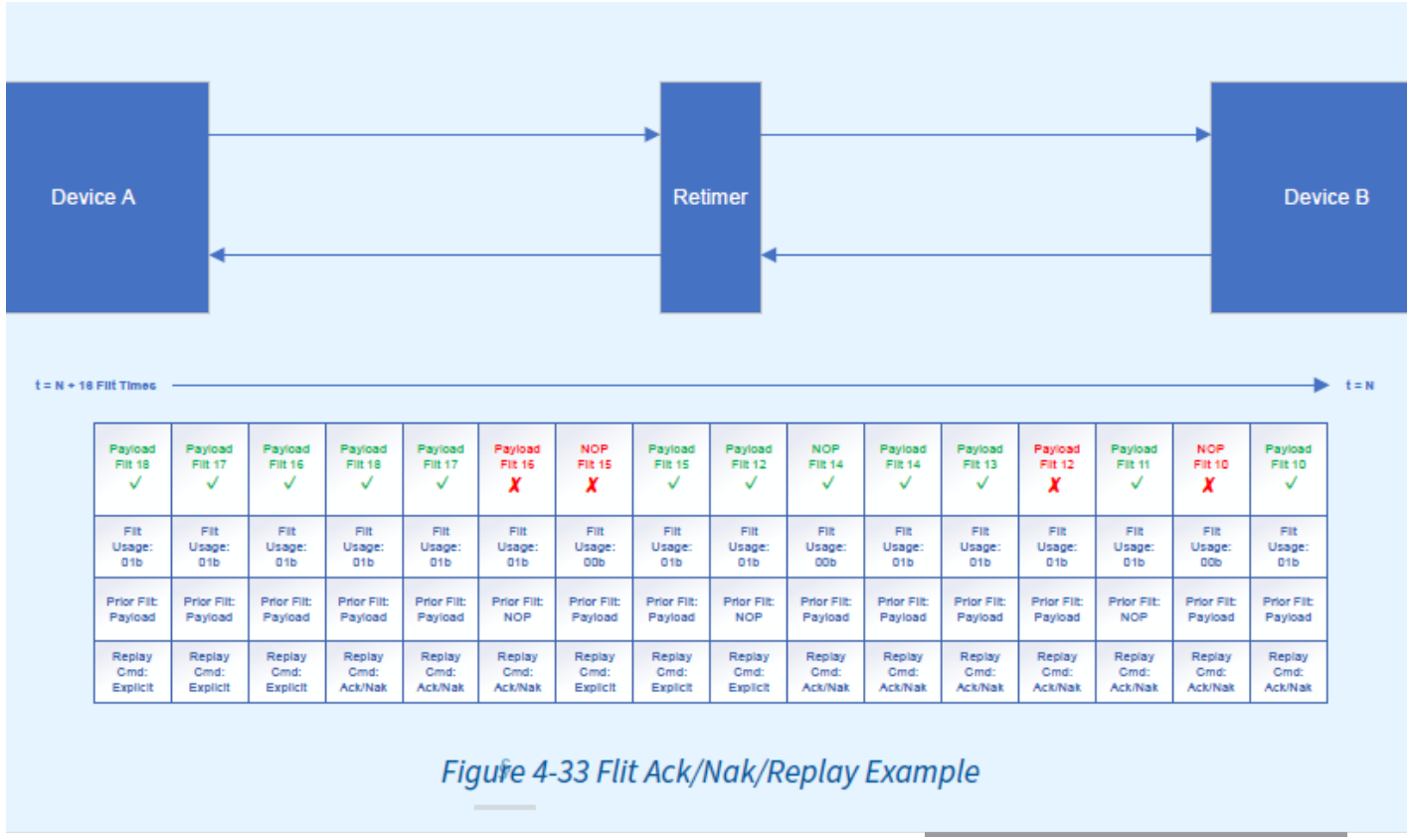


Figure 4-33 Flit Ack/Nak/Replay Example

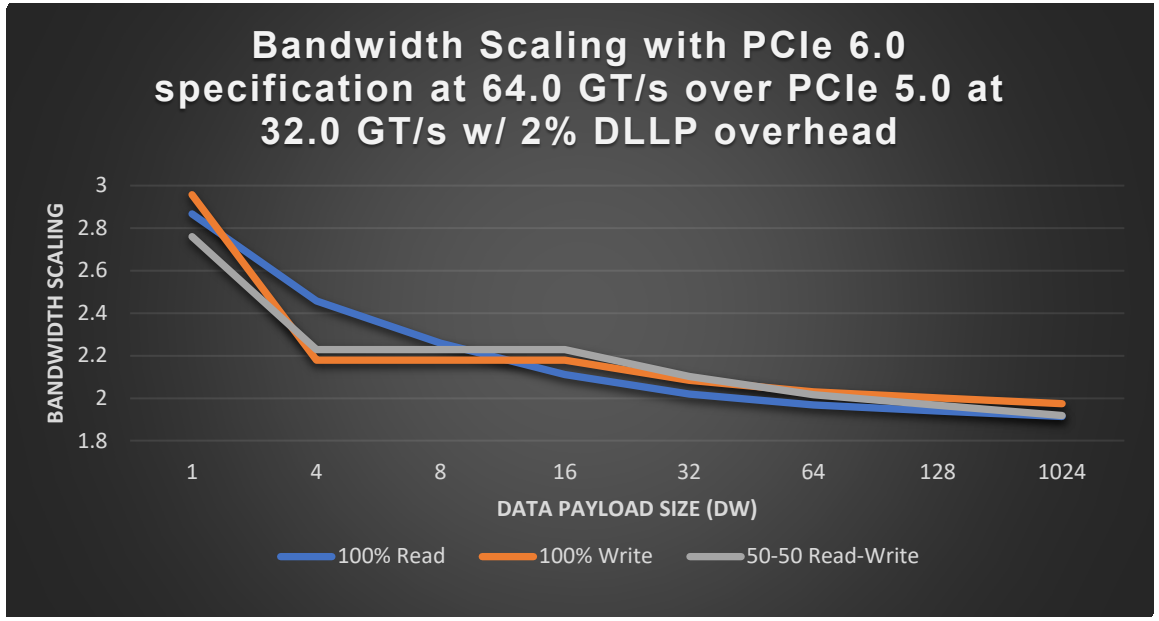
Example (Normal Flit Exchange Phase):

- B has Ack'ed till Flit 10 back to A
- Corrupted Flit 10 but Flit 11 indicates it was NOP – so no replay request
- Corrupted Flit 12; Fli 13 indicates it was payload => B sends 3 consecutive Flits with `selective Nak 11' (another invalid Flit prior to receiving Flit 12 would cause standard Nak)
- A sends Flit 12 with "Explicit Sequence Numbers" in 3 consecutive Flits (12, 15, 15)
- B sees two Flits in error (NOP Flit 15, Payload Flit 16) => B sends "standard Nak 15" in 3 consecutive Flits to A
- A removes Flits 12-15 from its Tx Replay buffer and replays Flits 16, 17, 18 from its Tx Replay buffer with explicit sequence number

KEY PERFORMANCE METRICS WITH PCIE® 6.0 SPECIFICATION: RETRY PROBABILITY, FIT, BANDWIDTH EFFICIENCY, LATENCY

FBER/ Retry Time	10^{-6}/ 100ns	10^{-6}/ 200ns	10^{-6}/ 300ns	10^{-5}/200ns
Retry probability per flit	$5x10^{-6}$	$5x10^{-6}$	$5x10^{-6}$	0.048
B/W loss with go-back-n (%)	0.025	0.05	0.075	4.8
FIT	4×10^{-7}	4×10^{-7}	4×10^{-7}	4×10^{-4}

Reasonable probability of retry and b/w loss at 10^{-6} FBER
FIT is close to 0 due to strong CRC



Better B/W efficiency for small packets (important for networking) in Flit Mode due to CRC amortization and removal of PHY overheads

x1

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

Overall latency improvement due to faster rate

x16

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

AGENDA

- Background
- Key Metrics and Requirements for PCIe® 6.0 Specification
- PAM4 and Error Assumptions/ Characteristics
- Error Correction and Detection: FEC, CRC, and Retry
- Flit Mode
- **Low Power enhancements: L0p**
- Key Metrics and Requirements for PCIe 6.0 Specification – Evaluation
- Conclusions and Call to Action

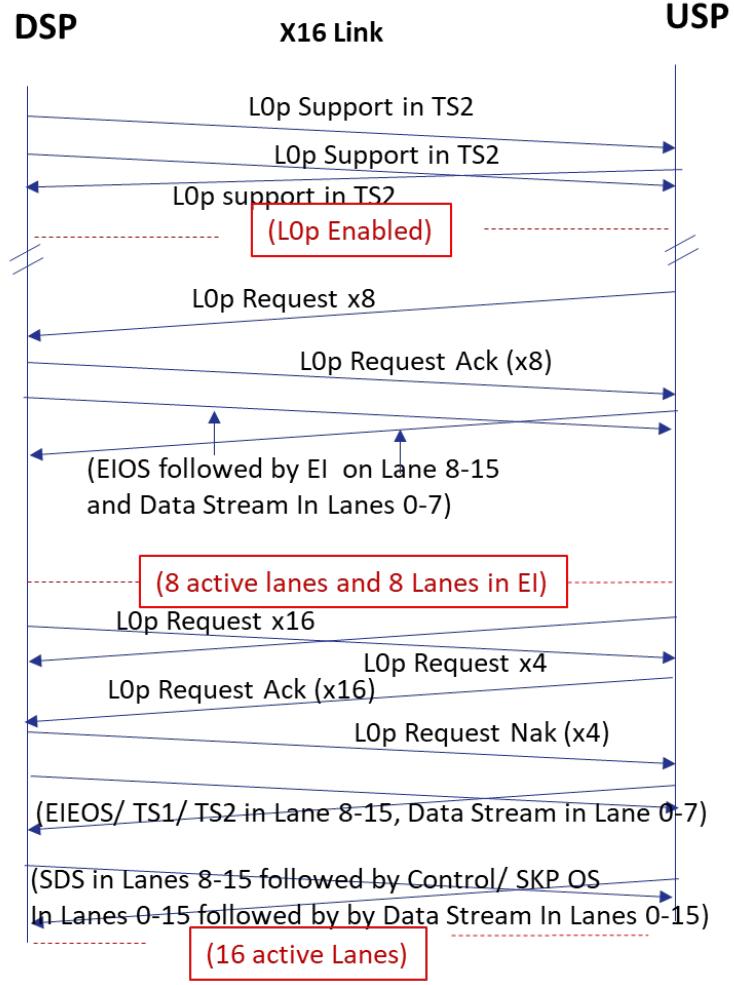
MOTIVATION FOR A NEW LOW-POWER STATE IN PCIE® 6.0 SPECIFICATION

- **Existing Methods: L0s, L1, Dynamic Link Width (DLW), Speed Change**
  - L0s power savings meagre – not supported if Flit mode is negotiated
  - L1 offers good power savings – but high entry and exit times
  - DLW - width can be modulated with bandwidth demand
    - Cons: High exit latency due to entire Link being in Recovery and Configuration
    - Ex: on a higher bandwidth demand, link retrains for tens of µsecs prior to width increase => traffic stalls for that time
  - Speed Change saves bandwidth
    - Cons: Less power savings than DLW and msecs of delay through speed transition

Need a new low-power state (L0p): power consumption proportionate to bandwidth usage, without impacting traffic flow

L0P TRANSITION WITH EXAMPLES

- **L0p (w/ Flit Mode) negotiated in Configuration.Complete (TS2)**
- **L0p width transition can be initiated by either side**
  - Link Partner either Acks or Naks in 1 usec
  - 2usec without a response: requestor abandons or re-requests
  - On a simultaneous request the one with the higher width wins and “Nak”s its Link Partner’s request if L0p.priority is not set
  - If L0p.priority is set, lower width wins
- **Reducing Link width: (sequence)**
  - Control SKP OS on all active Lanes
  - EIOS/ EI on Lanes to be turned off
  - Data Stream on reduced width
- **Increasing Link width: (sequence)**
  - EIEOS/ TS1/ TS2 training on Lanes to be activated while Data Stream continues on active Lanes
  - SDS sequence (>=8G) on Lanes to be activated (Data Stream continues on active Lanes)
  - (Control) SKP OS on all Lanes with new width
  - Data Stream on new width



KEY METRICS FOR PCIE® 6.0 SPECIFICATION: EVALUATION

Metrics	Expectations	Evaluation
Data Rate	64GT/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 PAM4)	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 specification 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

CONCLUSIONS AND CALL TO ACTION

- **PCIe® 6.0 specification completed**
- **We met the challenges on multiple fronts**
  - New signaling with PAM4: 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 PAM4 at lower speeds
  - We have exceeded or met the requirements
  - PCIe 6.0 specification is the reflection of the combined innovation capability of 900+ members with a track record of delivering flawlessly against challenges for more than two decades!!
- **Plan for products … Expect Networking to be an early adopter**

2022 OFA Virtual Workshop

THANK YOU

Dr. Debendra Das Sharma

Intel Senior Fellow, Intel Corporation

Director, PCI-SIG® Board