STORAGE DEVELOPER CONFERENCE
SDC 21
BY Developers FOR Developers
Virtual Conference
September 28-29, 2021
A SNIA Event

PCIe® 6.0 Specification: A High-Performance Interconnect for Storage Networking Challenges

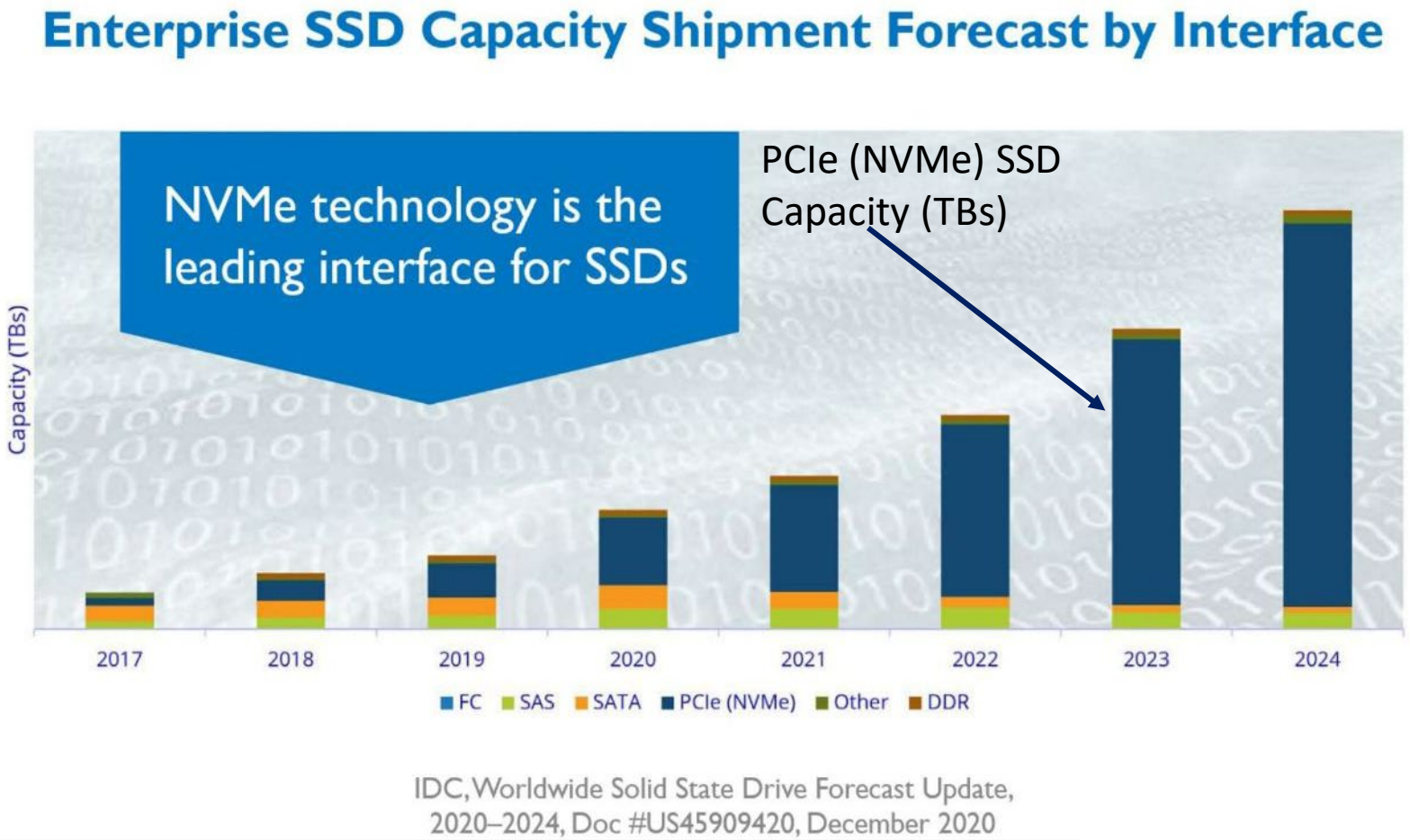
Dr. Mohiuddin Mazumder
Co-chair, PCI-SIG® Electrical Work Group
Senior Principal Engineer
Intel Corporation

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
- Ordered Sets handling with high error rate
- Low Power enhancements: L0p
- Area reduction: Shared Credits
- Key Metrics and Requirements for PCIe 6.0 Specification – Evaluation
- Conclusions and Call to Action

PCIe® Interconnect for Storage

- ~ 70-80% of data-center SSDs use PCIe as the interface because of its
  - High performance – data rate and scalable widths (x1, x2, x4, x8, x16)
  - Low latency – directly connects storage device to the host
  - Extended RAS (Reliability, Availability, and Serviceability) features
  - Standard form factors (e.g., m.2, u.2, Add-In-Card, and EDSFF)
  - Low power
  - Reduced Total Cost of Ownership
- Examples of usages that benefit most from PCIe (NVMe®) SSDs
  - Database, AI/ML, HPC, Virtualization, Edge Computing, Automotive, Gaming, 3D Graphics



IDC, Worldwide Solid State Drive Forecast Update, 2020–2024, Doc #US45909420, December 2020

PCIe® architecture delivers a high performance, low-latency interconnect between the storage SSDs and the host CPU/switch

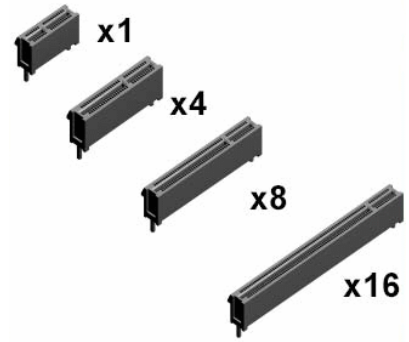
Examples of PCIe® Form Factors and Usages

Ethernet Network Adapter

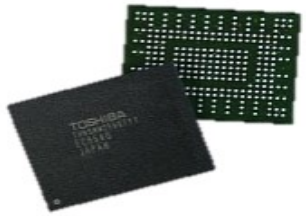
FPGA Accelerator

CEM Cards

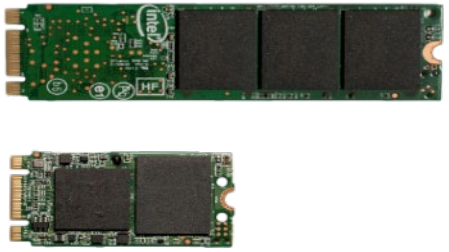
Storage



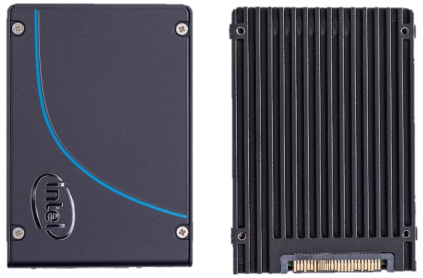
CEM Connectors



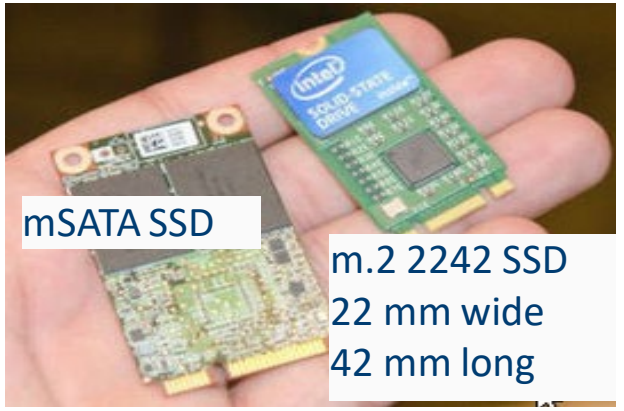
BGA



M.2

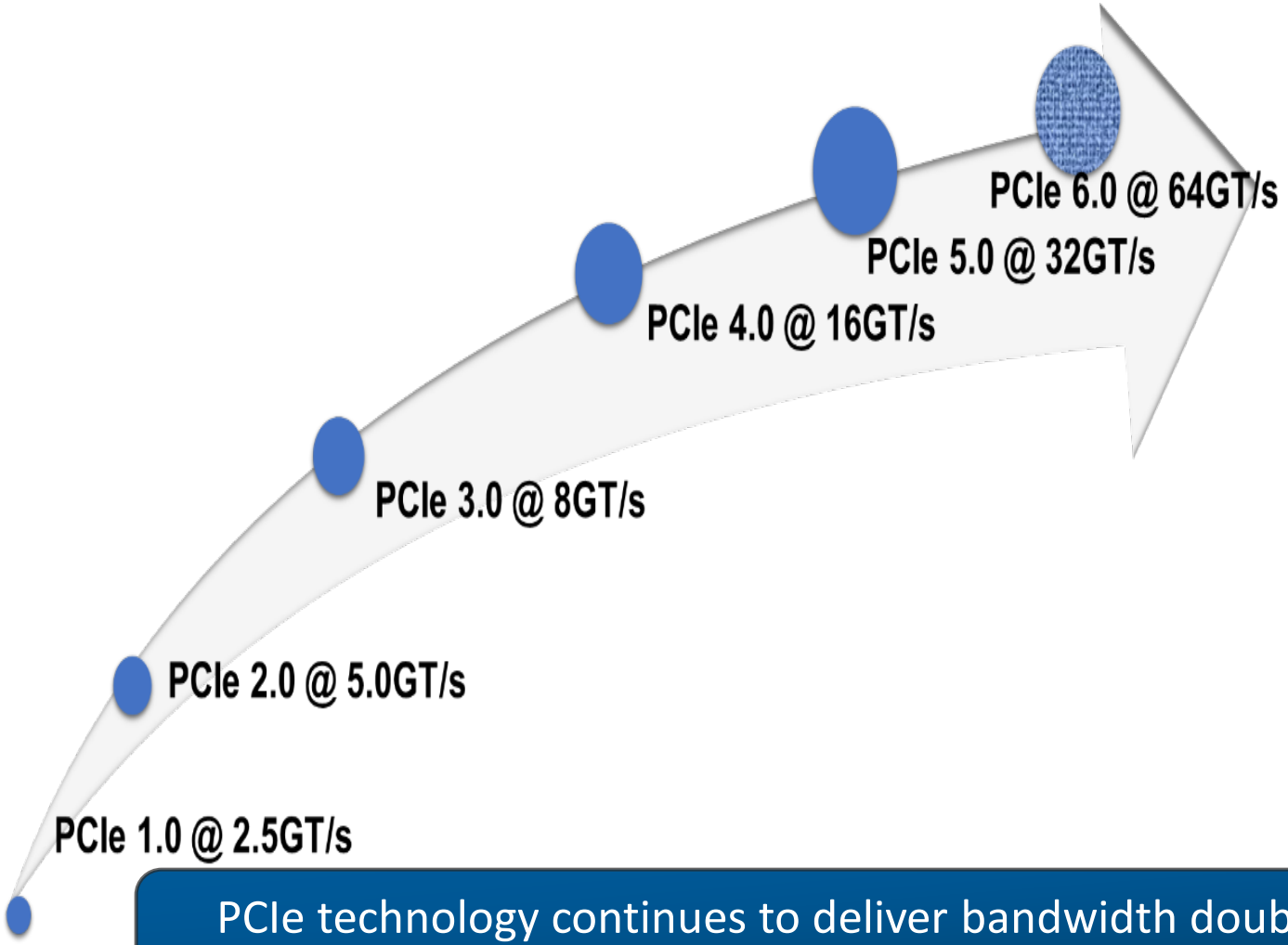


U.2 2.5in (SFF-8639)



Many form factors and applications drive the bandwidth scaling of PCIe technology

Progression of PCIe® Technology Bandwidth



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 (WIP)	64.0 (PAM-4, Flit)	2021*

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

PCI Express® Architecture Advantages

Data Center | Mobile | Embedded

- Single PHY standard
- Low power and high performance
- Alternate protocol support
- Doubling the bandwidth for sixth generation with full backwards compatibility
- A variety of standard form factors
- A robust and mature compliance and interoperability program

Backwards compatibility enable broad ecosystem and makes PCIe architecture a low-cost I/O for diverse applications

Key Metrics for PCIe 6.0 Specification: Requirements

Metrics	Expectations
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) (PCIe usages cannot 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!

PAM4 Signaling at 64 GT/s

PAM4 signaling: Pulse Amplitude Modulation 4-level

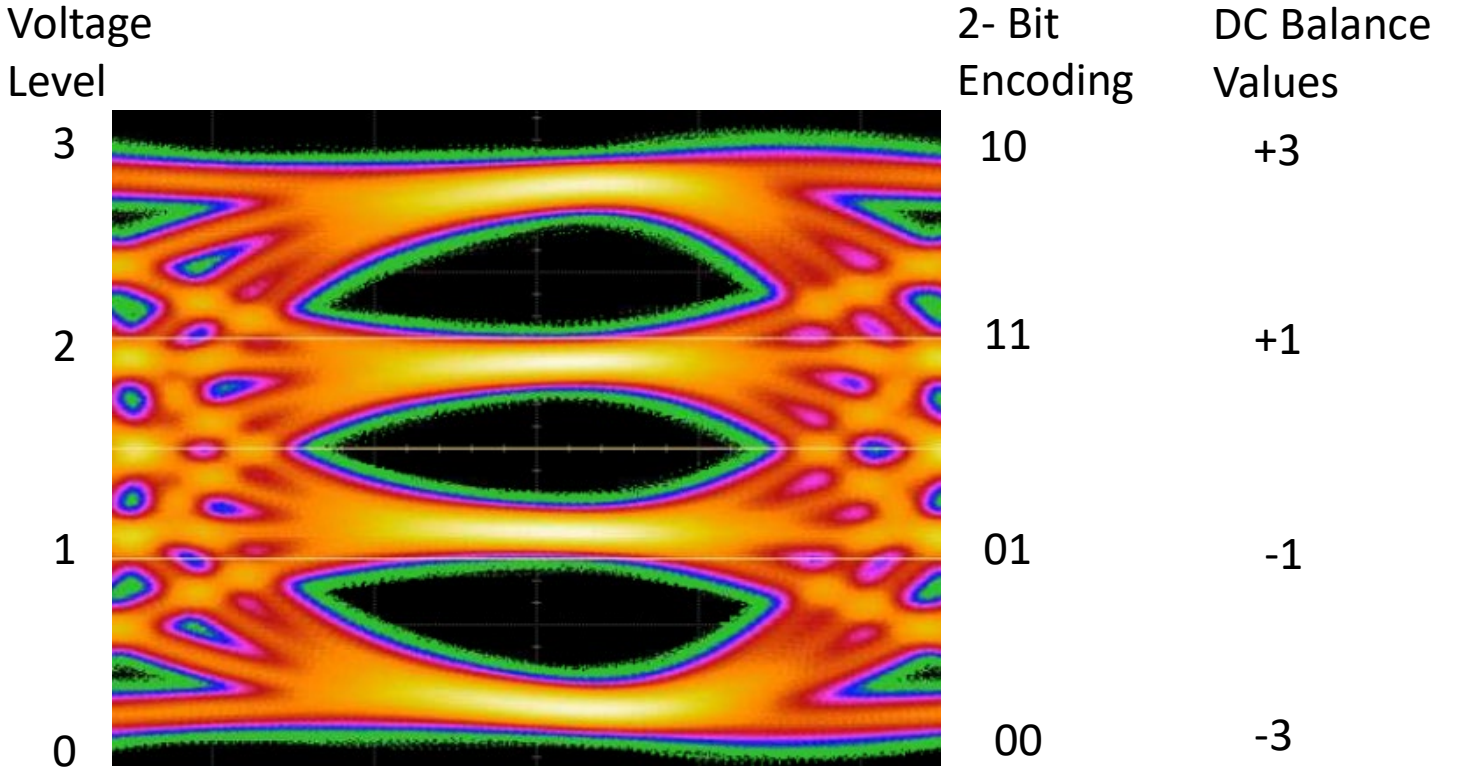
- 4 levels (2 bits) in same Unit Interval (UI); 3 eyes
- Helps channel loss (same Nyquist as 32.0 GT/s)

Reduced voltage levels (EH) and eye width increases susceptibility to errors

Gray Coding to help minimize errors in UI

Precoding to minimize errors in a burst

Voltage levels at Tx and Rx define encoding

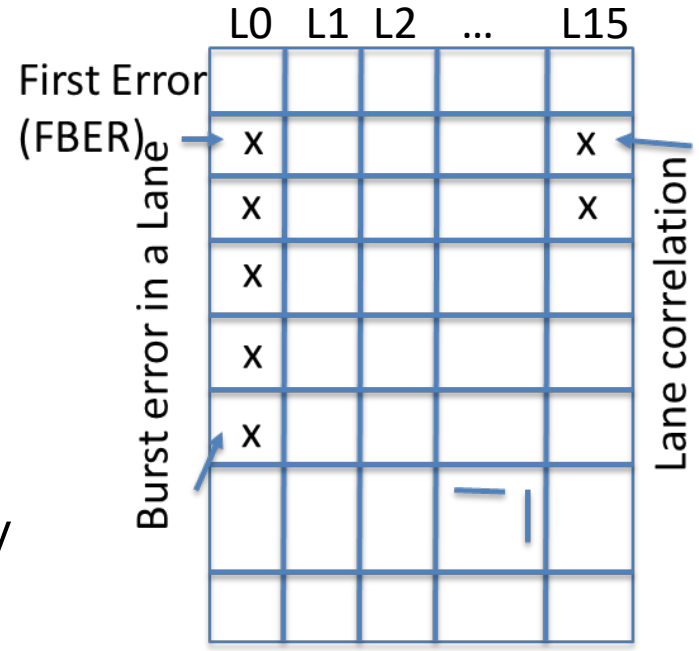


Sensitivity to noise (xtalk, reflection, and device-related) is a key challenge

Error Assumptions and Characteristics with 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



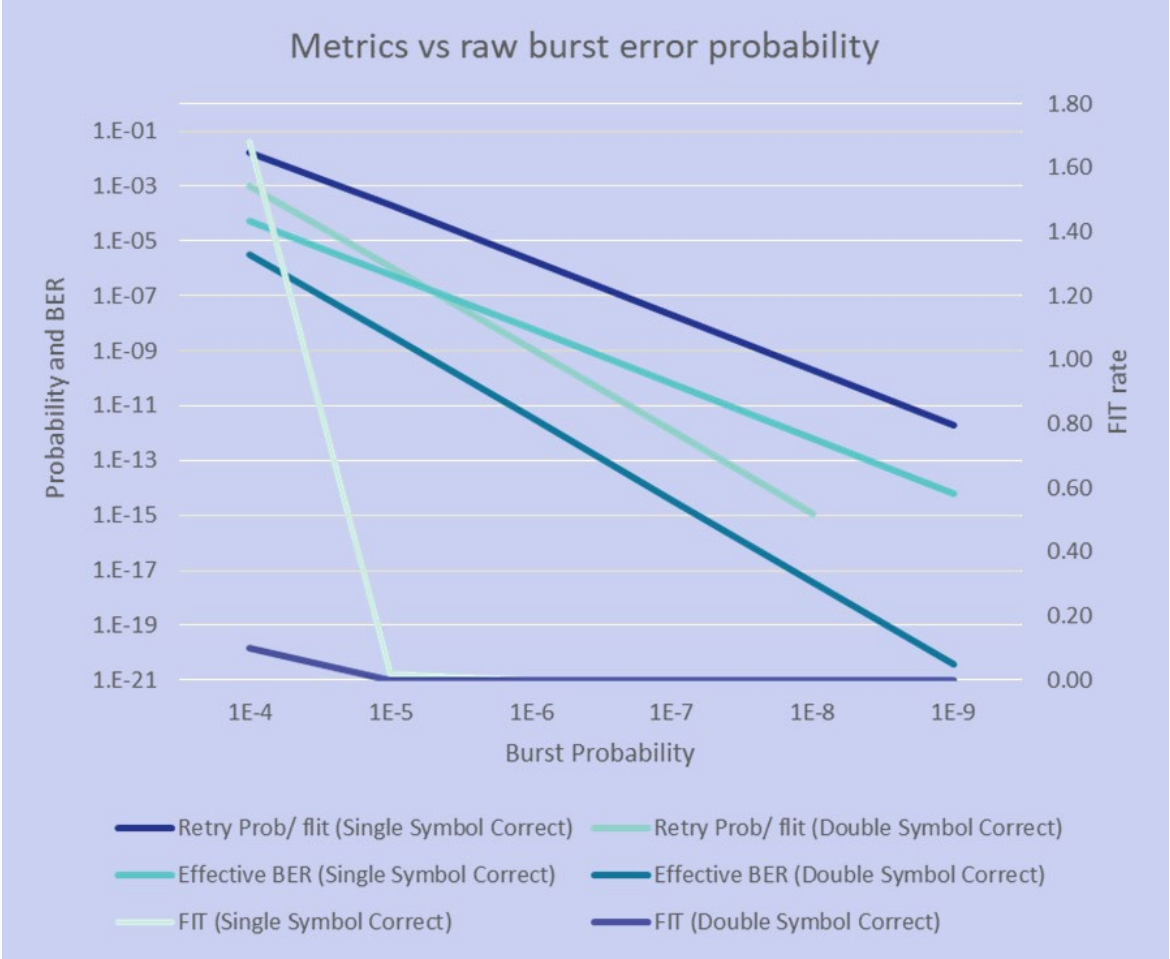
Handling Errors and Metrics Used for Evaluation

- 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 enable low FEC latency (<2ns) to meet the performance needs of performance critical Load/Store I/O

Our Approach: Light-Weight FEC and Retry

- Light-weight FEC, strong CRC, and keep the overall latency (including retry) low so that the Load/Store 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 with FBER of 1E-6 to meet the metrics (latency, area, power, bandwidth)

Flit Encoding PCIe® 6.0 Architecture: Low-latency with High Efficiency

- Flit (flow control unit) based: FEC needs fixed set of bytes
- Correction 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 T(DL)LP CRC
  - Improved bandwidth utilization due to overhead amortization
  - Flit Latency: 2ns x16, 4ns x8, 8 ns x4, 16 ns x2, 32 ns x1
  - Guaranteed Ack and credit exchange => low Latency, low storage
- Optimization: Retry error flit only with existing Go-Back-N retry

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

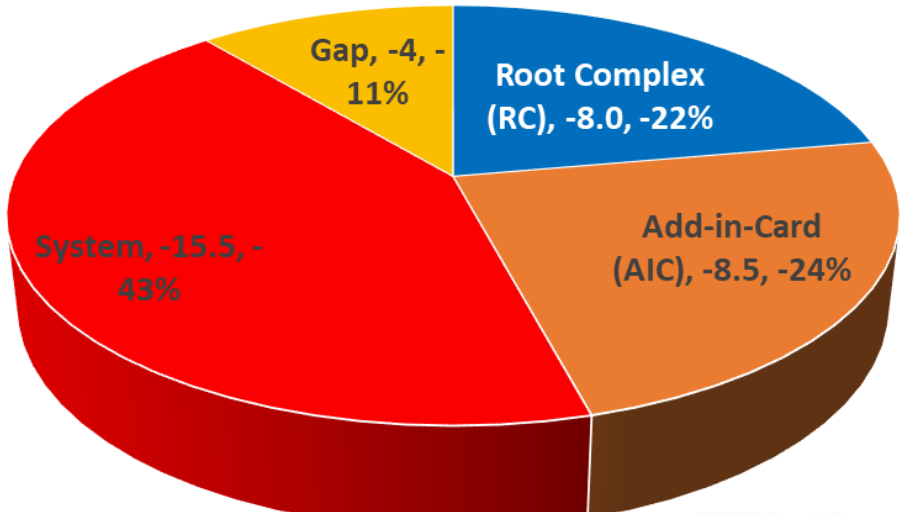
Low latency improves performance and reduces area

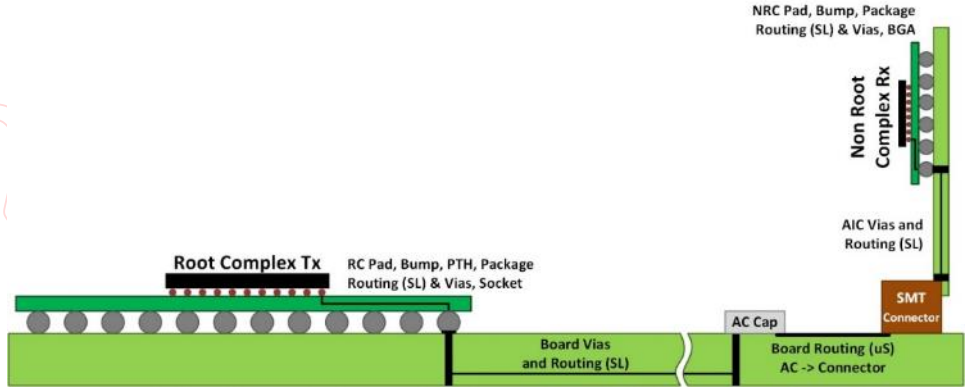
Electrical Improvements to Achieve Low-latency

- First Bit Error Rate (FBER) < 10-6
  - Pad-to-pad channel loss < 32 dB at 16 GHz
  - Significant crosstalk and reflection reduction
  - Reference clock and CDR improvement
  - Jitter reduction compared to PCIe® 5.0 technology ~ 2x
  - Improvement in Reference Equalization: Tx 2nd pre-cursor, improved CTLE peaking and bandwidth, and 16-tap DFE
- Minimize Burst Error Probability
  - PAM4 precoding
  - Gray coding
  - Limits on DFE taps

Pad-to-Pad Loss and System Routing Length

Loss Parameters	PCIe® 5.0 Specification Rev 1.0 (dB)	PCIe 6.0 Specification Rev 0.9 (dB)
Pad-to-Pad Loss at 16 GHz	-36	-32
Root Complex (RC)	-9.0	-8.0
Add-in-Card (AIC)	-9.5	-8.5
System	-17.5	-15.5





13” system routing requires -32 dB pad-to-pad loss support and PCB loss of 1.0 dB/in

Summary of Key Electrical Changes in PCIe 6.0 Technology

- 64 GT/s PAM4 requires FEC
- Raw BER (pre-FEC and before any DFE burst error): 1e-06
- Pad-to-pad loss: -32 dB at 16 GHz
- Reference Package Models: ~3-6 dB improvement in reflection and xtalk
- Ref CLK Rj RMS: 100 fs (clean), 150 fs in system simulations
- 4-tap Tx equalization with a 2nd Tx pre-cursor – New Preset Table for 64 GT/s
- Tx precoding and gray coding mandatory
- ~2x improvement in silicon jitter
- Reference Rx: Improved CTLE and 16-tap DFE
- Rx eye mask (Top eye: 0.10 UI and 6.0 mV)

A holistic approach to improve Electrical, Logical, and Protocol layers was key to achieve a low-latency PAM4 solution

Assessment of Channel Reach at 64 GT/s

TX

- Rev0.9 Jitter Specification
- Fixed best TxEq: Pre2, Pre1, Post1 -> 0.04,-0.2,0
- Tx SNDR: 34 dB
- Tx R_{TERM}: 45 Ohm (to account for DC loss)
- Voltage swing: 0.8V (1.0V for NEXT)
- 2 FEXT / 3 NEXT
- Rise/Fall Time: 0.2 UI

RX

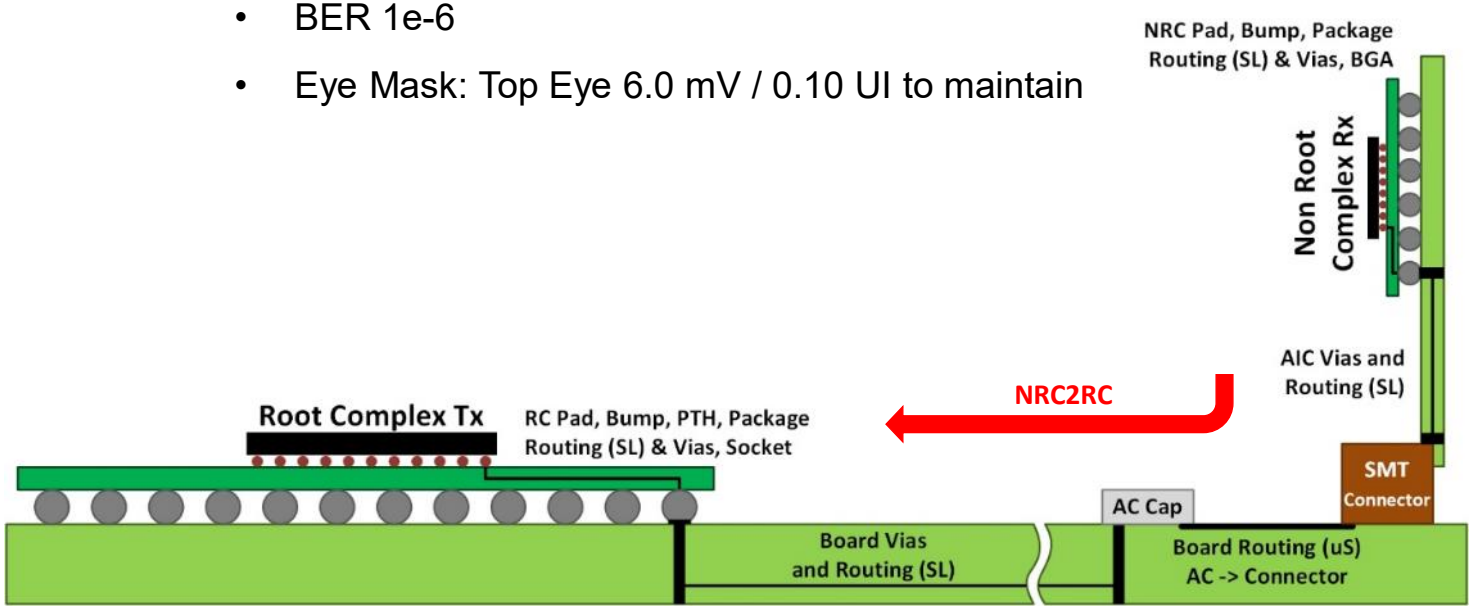
- CTLE: Rev 0.9 Spec
- DFE: 16-tap, h1/h0: 0.55, 10-bit Quantization (~1mV)
- Rx R_{TERM}: 50 Ohm

Other

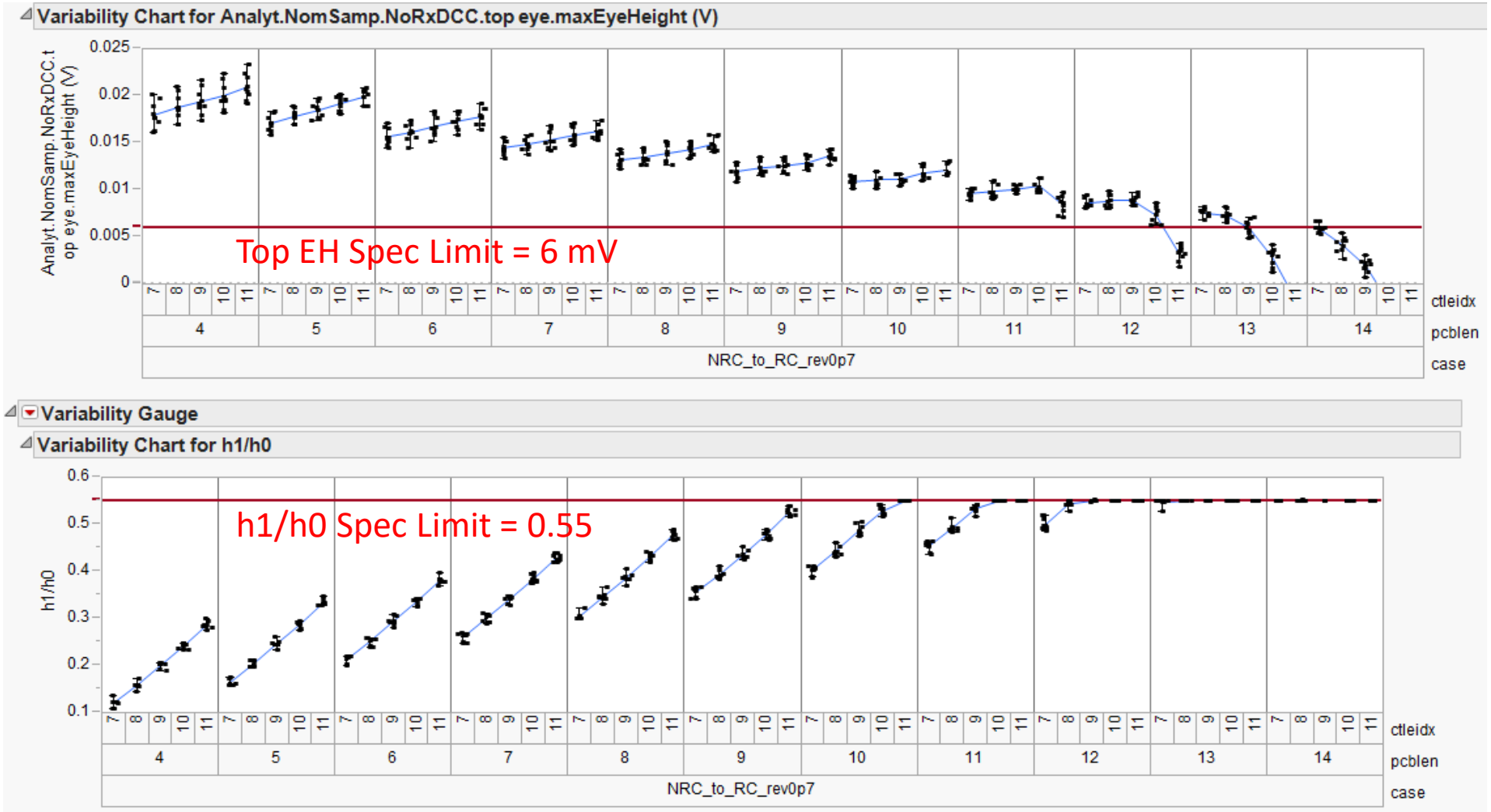
- BER 1e-6
- Eye Mask: Top Eye 6.0 mV / 0.10 UI to maintain

Channel

- Inductive coil based Rev0.7 Package models
- PCB length: 4” - 14”, AIC length: 4”
- Best available CEM connector
- BB and AIC impedance variation: low, nom, high -> 9 cases
- Directions: NRC to RC



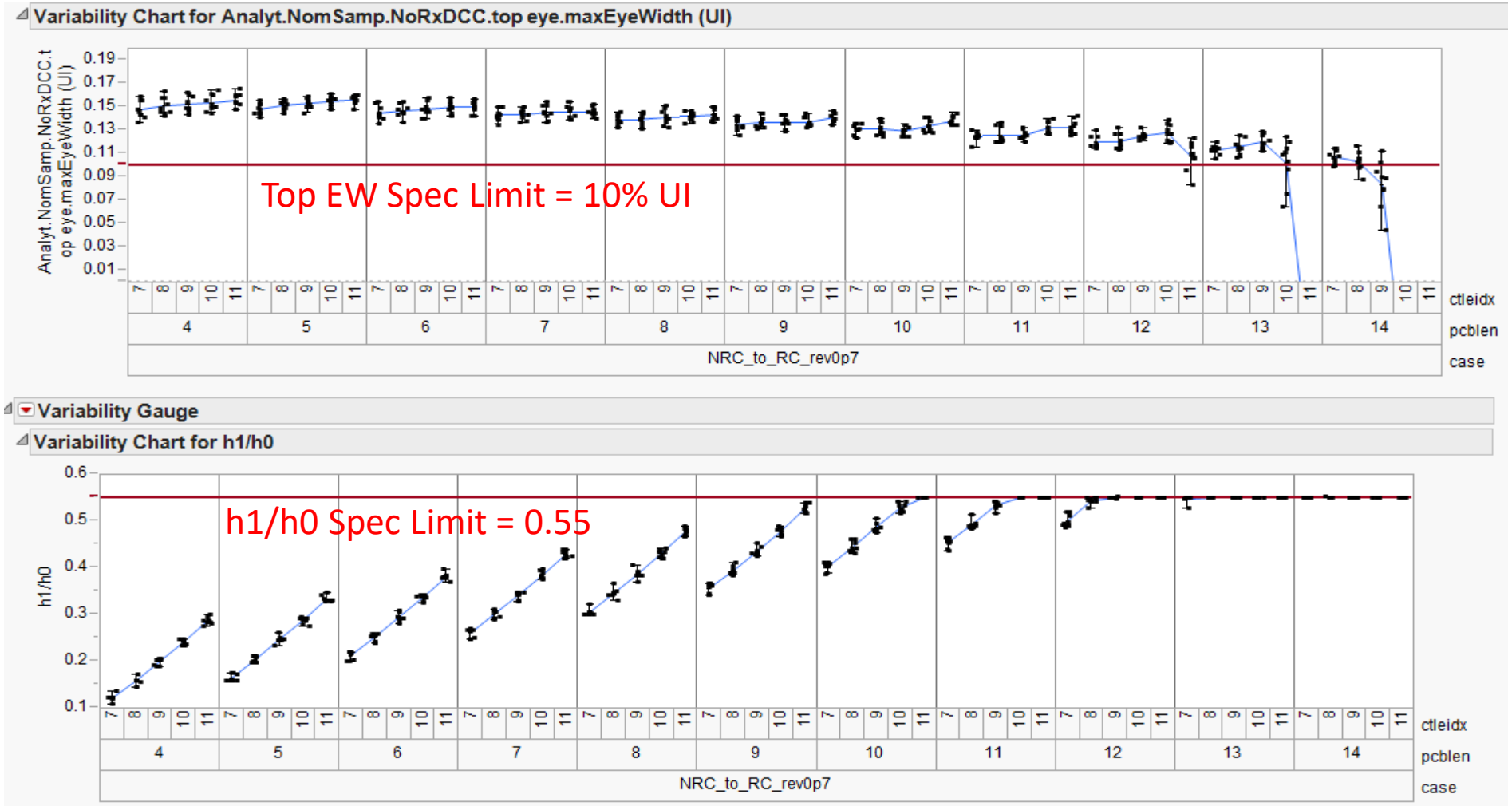
Eye Height vs. System Routing Length



CTLE Index 7: DC Gain = - 9 dB
CTLE Index 11: DC Gain = - 5 dB

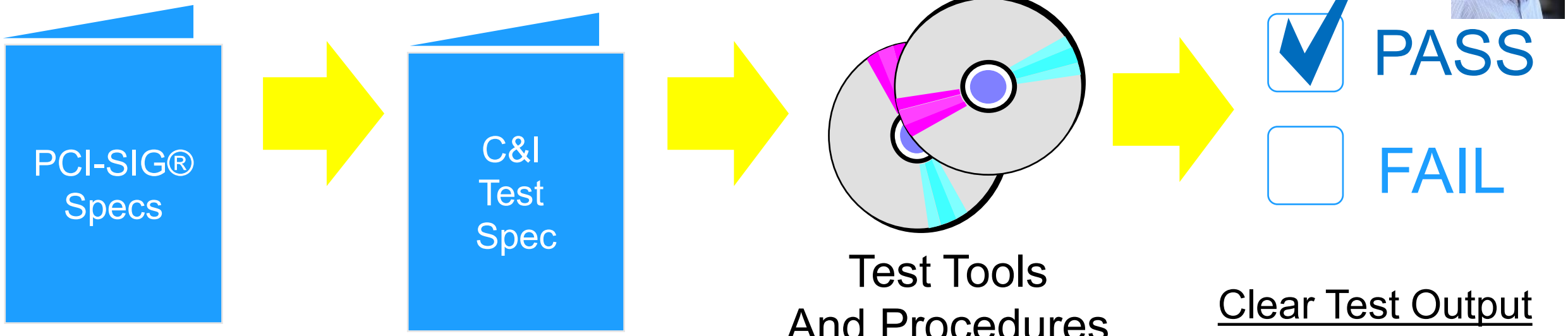
Limits on DFE tap coefficients imposed to minimize DFE error burst probability

Eye Width vs. System Routing Length



Compliant channel solutions are feasible for <=13” system and <=4” AIC routing

PCIe® Technology Compliance Process



PCI-SIG® Specs
Describes

Device requirements
- Base and CEM specs

C&I Test Specs
Define

Test criteria based on spec requirements
- Test Definitions
- Pass/Fail Criteria

Test Tools And Procedures
Test H/W & S/W
Validates

Test criteria
- Compliance
- Interoperability

Clear Test Output
Maps
- Directly to Test Spec

Predictable path to design compliance

Test and various form-factor Specs get developed within 1-2 years of Base Spec completion

PCIe 6.0 Specification: Key Messages

- PCIe 6.0 architecture can meet the needs of storage interconnect solution in the foreseeable future
- 64 GT/s PAM4 - doubles the bandwidth with backwards compatibility
- 64-bit CRC, a light FEC, FBER < 1e-06, and link level retry with low (~ 1e-05) retry probability enable low-latency (< 2 ns for x16) and high reliability (FIT << 1)
- PCIe 5.0 architecture-like channel reach is feasible with improvements in circuits and channels

Please take a moment to rate this session.

Your feedback is important to us.

