12th ANNUAL WORKSHOP 2016

EVOLUTION OF PCI EXPRESS® AS THE UBIQUITOUS I/O INTERCONNECT TECHNOLOGY

Debendra Das Sharma, PhD
Senior Principal Engineer and Director I/O Technology and Standards
Data Center Group, Intel Corporation
Chair, PHY Logical Group, PCI-SIG®

April 7, 2016

AGENDA

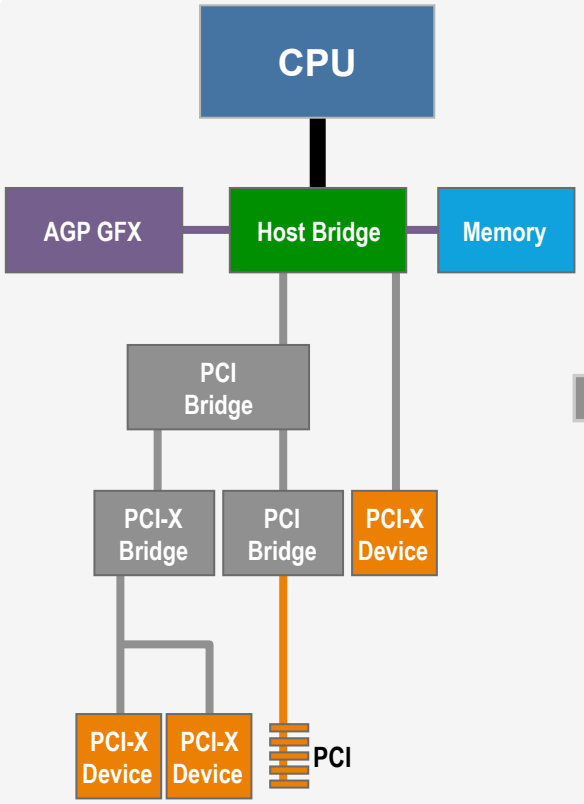
- Introduction
- PCI Express® (PCIe®) 4.0 Specification
- Retimers for Extending Channel Reach
- Form Factors
- Compliance
- Conclusions

EVOLUTION OF PCI/ PCI EXPRESS® TECHNOLOGY

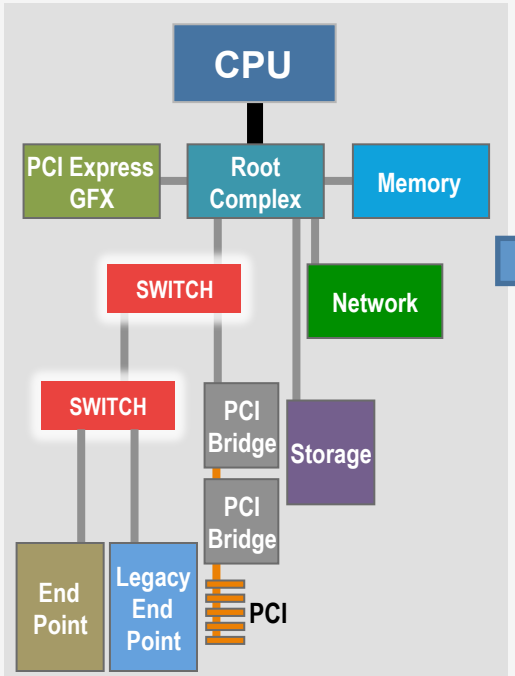
- **Peripheral Component Interconnect (PCI) started as bus-based PC interconnect in 1992**
  - 32 bit @ 33 MHz
  - Evolved to 64 bits @ 33/ 66/ 132 MHz
- **Moved to link-based serial interconnect with full-duplex differential signaling with PCI Express® (PCIe®) with backwards compatibility for software**
  - Doubling data rate every generation
- **Evolution from PC to HPC, servers, clients, hand-held, and Internet-Of-Things usage over three decades**

EVOLUTION OF PCI TECHNOLOGY

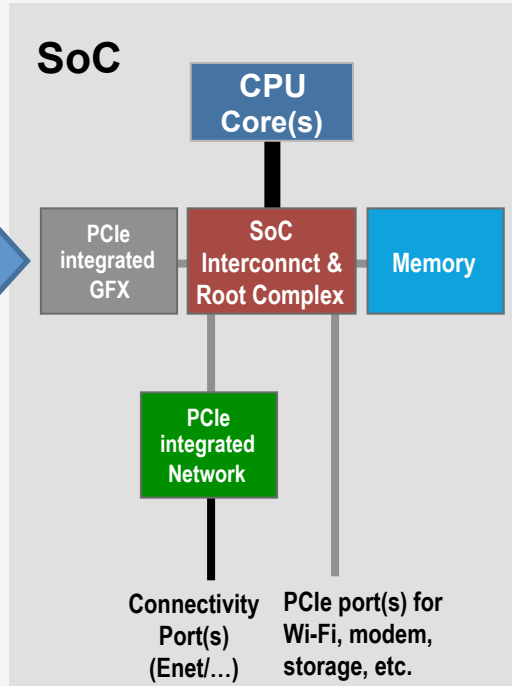
PCI/PCI-X SYSTEM



PCI EXPRESS® BASED SYSTEMS

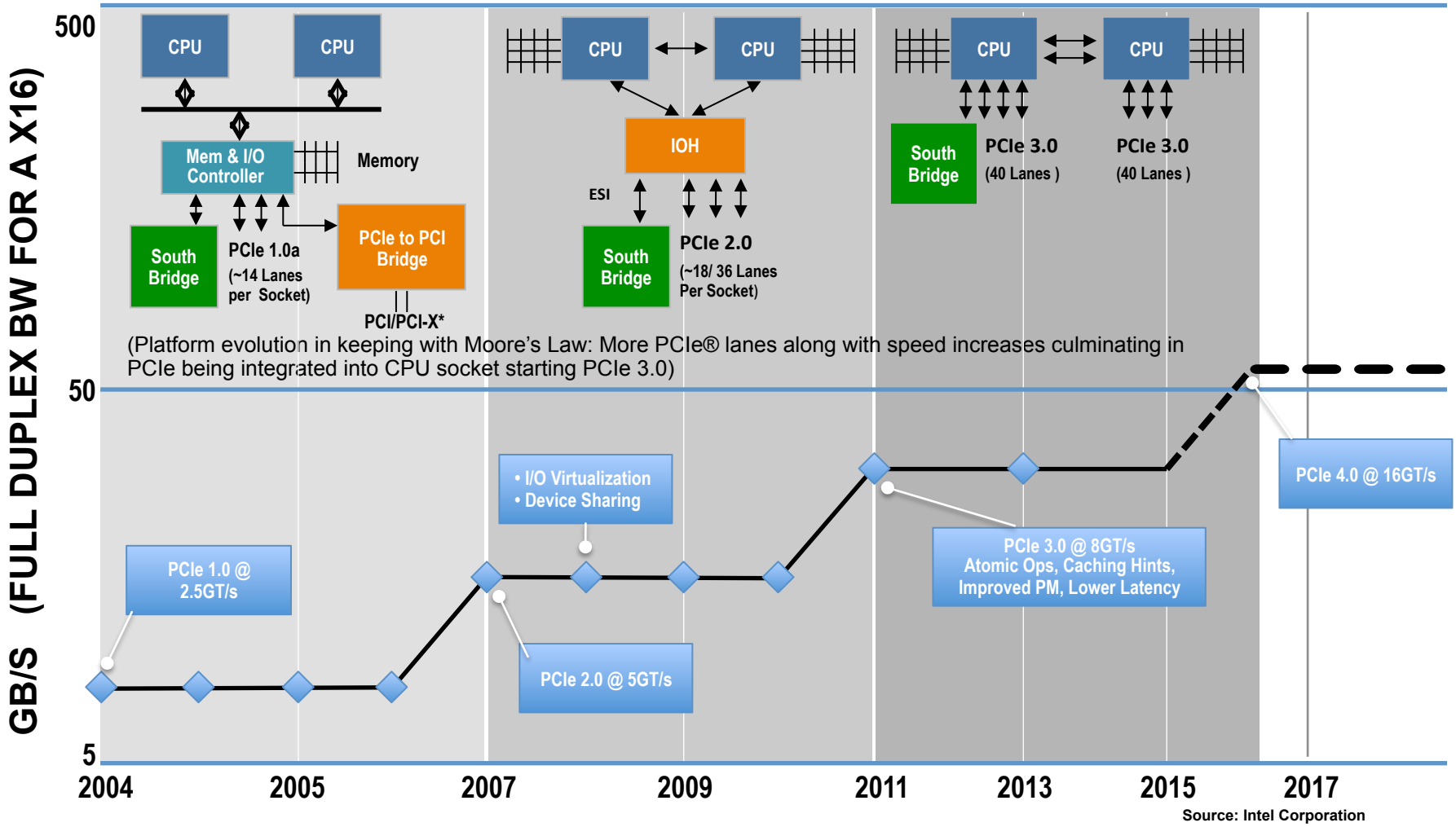


Large system with lots of open PCIe® slots – e.g., servers, workstations



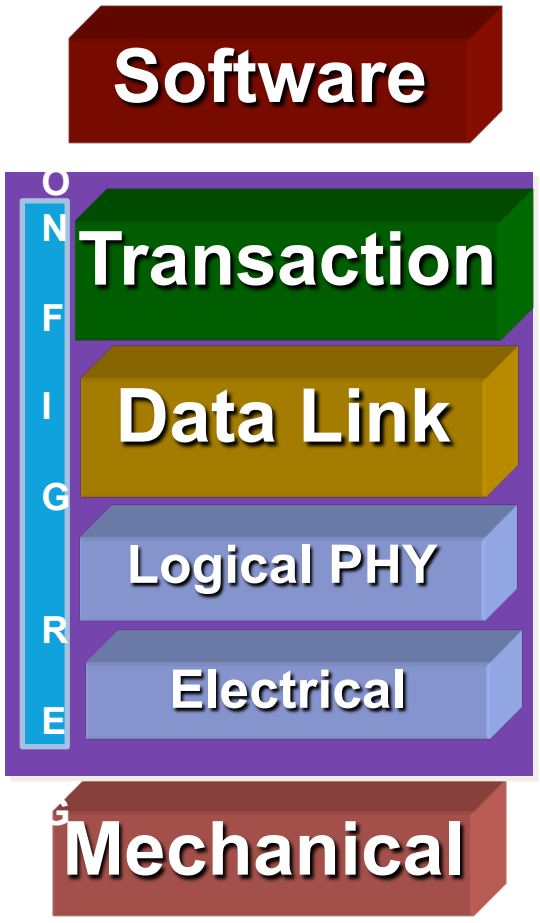
Highly integrated systems including hand-held devices

PCI EXPRESS® ROADMAP AND PLATFORM EVOLUTION



Source: Intel Corporation

PCI EXPRESS®: A LAYERED ARCHITECTURE



⇦ PCI compatibility, configuration, driver model
⇦ PCI Express 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 Express Layering - Enabler for Modularity and Reuse

PCI EXPRESS®: A LOW-POWER INTERCONNECT

Item	PCIe® 3.0	PCIe® 2.0	M-PHY Gear 3
Line Speed [Gbps]	8	5	5.83
PHY Overhead	128/130, 1[GB/s]	8/10, 500[MB/s]	8/10, 583[MB/s]
Active Power [mW]	60 (L0)	46 (L0)	58 (HS)
Standby Power [mW]	0.11 (L1.2)	0.11 (L1.2)	0.2 (Hibern8)
MB/mJ (higher = better)	14-18	8-12	8-12

Source: IDF Sept'15

Synopsys* Published Power Data

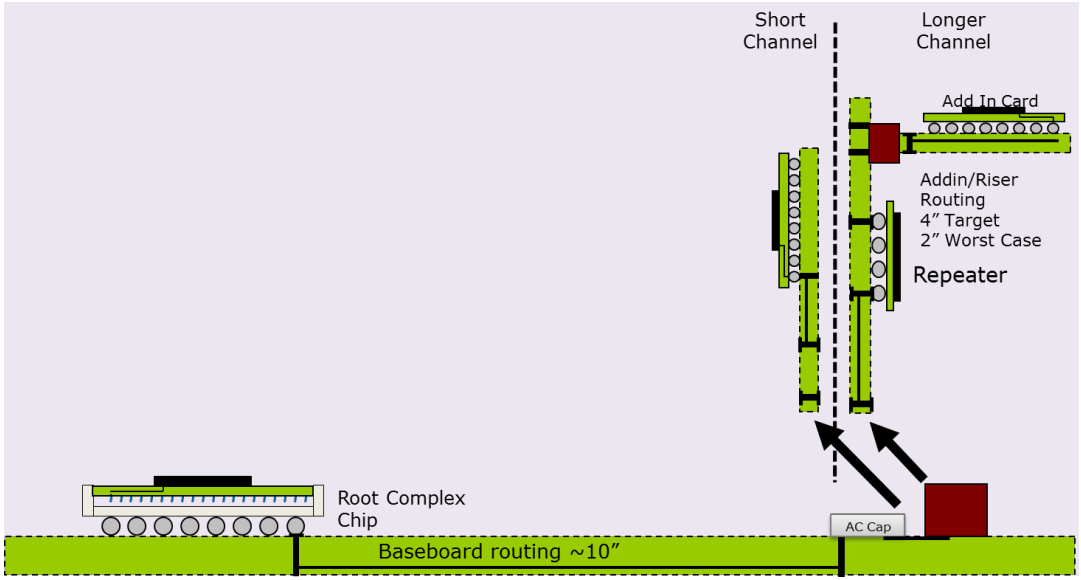
- 5 mW/Gb/Lane – Active
- 10 uW/Lane – Standby
- Source: http://news.synopsys.com/2015-05-21-Synopsys-Announces-Industrys-Lowest-Power-PCI-Express-3-1-IP-Solution-for-Mobile-SoCs

AGENDA

- Introduction
- PCI Express® (PCIe®) 4.0 Specification
- Retimers for Extending Channel Reach
- Form Factors
- Compliance
- Conclusions

PCI EXPRESS® 4.0 SPEED AND CHANNEL

- PCIe® 4.0 data rate: 16.0 GT/s
- Fully backwards compatible with PCIe 3.x (8.0 GT/s), PCIe 2.x (5.0 GT/s) and PCIe 1.x (2.5 GT/s); Preserves decades of ecosystem investment and innovation
- Low cost, high performance, low power I/O technology
- Connector improvements to reduce cross-talk and improve insertion loss at 8G Nyquist
- 2 connector 20” server PCIe topology needs either retimer or ultra low-loss PCB to operate at 16.0 GT/s

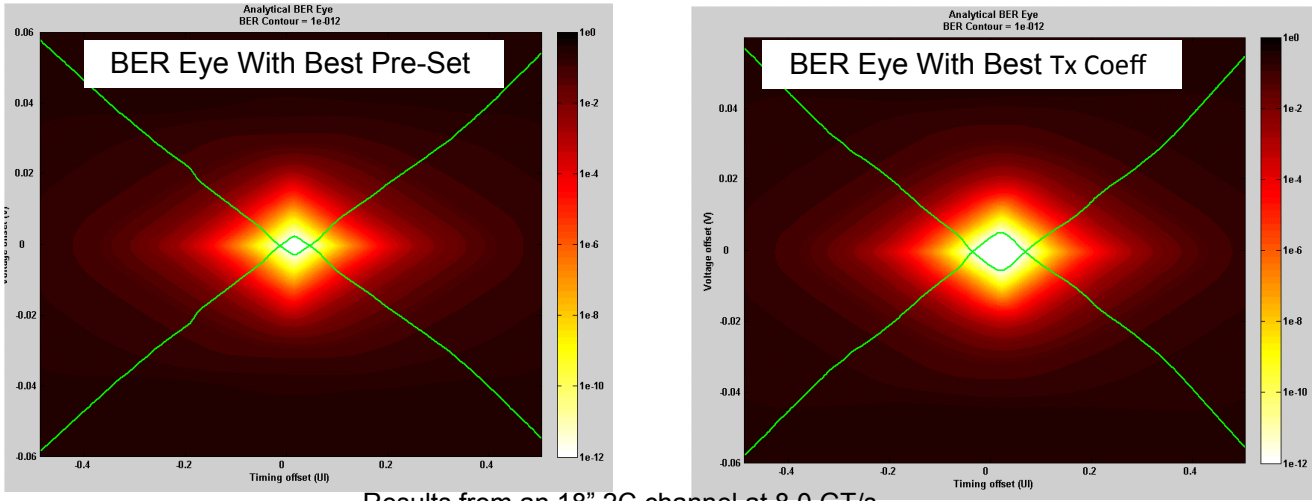


Source: Intel Corporation

TRANSMITTER EQUALIZATION

- **2.5 GT/s and 5.0 GT/s: Fixed de-emphasis for Link**
- **8.0 GT/s and 16.0 GT/s: Analysis demonstrates need for per Tx-Rx EQ**
  - Variations in receiver design, channel, PVT
  - Adjust each Tx by its Rx individually
  - Start with a preset value and then adjust dynamically
- **4 Stages:**
  - Stage 0: Preset values communicated at a lower data rate to downstream component
  - Stage 1: Link tries to stabilize at the preset at 1E-4 BER
  - Stages 2 and 3: Each receiver asks its transmitter to adjust till it achieves 1E-12 or better BER

Co-efficient based Tx EQ provides better margin



Results from an 18” 2C channel at 8.0 GT/s

Source: Intel Corporation

LINK EQUALIZATION: STAGES 2 AND 3

Stage 2: Intended for Upstream Port to achieve BER <= 10^{-12}. Starts at the preset. Coefficients/presets are exchanged in sub-loops until this is accomplished within 24 ms. A port may decide not to make any new requests. Corresponds to Phase 2.

Example: start from preset 7 (coef=4/6)

1st sub-loop

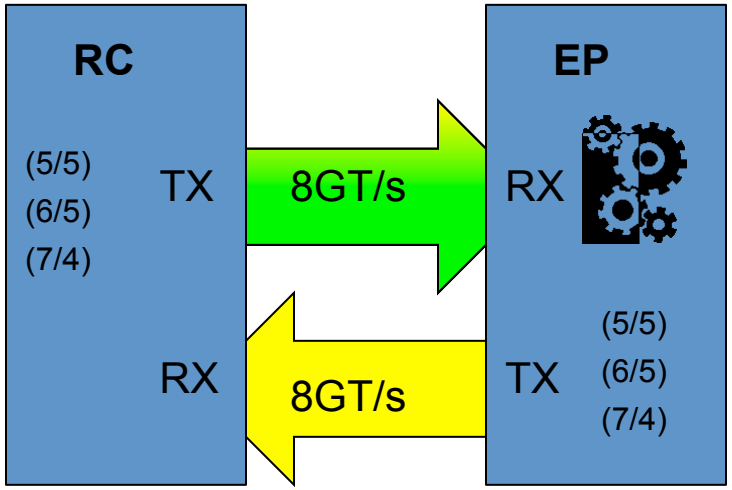
a. EP Rx eval reveals need for less post, more pre
b. EP sends (5/5) to RC
c. RC applies (5/5) to TX
d. RC echo's (5/5) to EP

2nd sub-loop

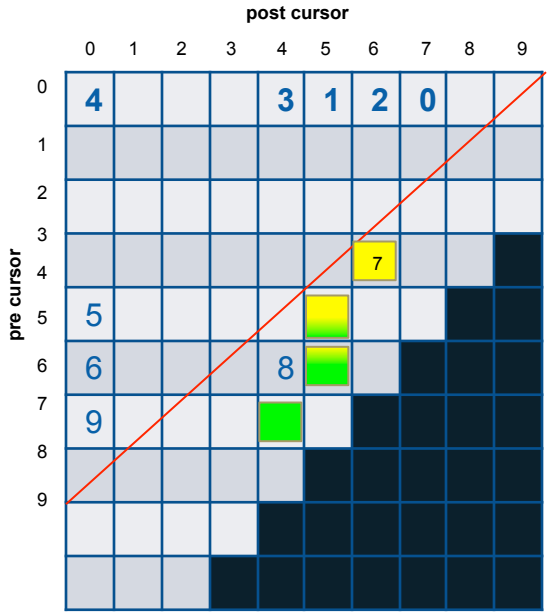
a. EP Rx eval needs more pre, post ok
b. → d. repeat with (6/5

3rd sub-loop finds good result with (7/4) so moves to phase 3

- Receiver full swing (FS) defines granularity of coeff
  - ✓ Table at bottom-right is for illustrative purposes
  - ✓ X-axis is pre-cursor, y-axis post-cursor, diagonal defines the boostline
  - ✓ Each tile represents a coeff (e.g. p7=4/6, p8=5/5, etc)
  - ✓ Numbers in tiles represent presets; black tiles are illegal coeff space



Source: Intel Corporation



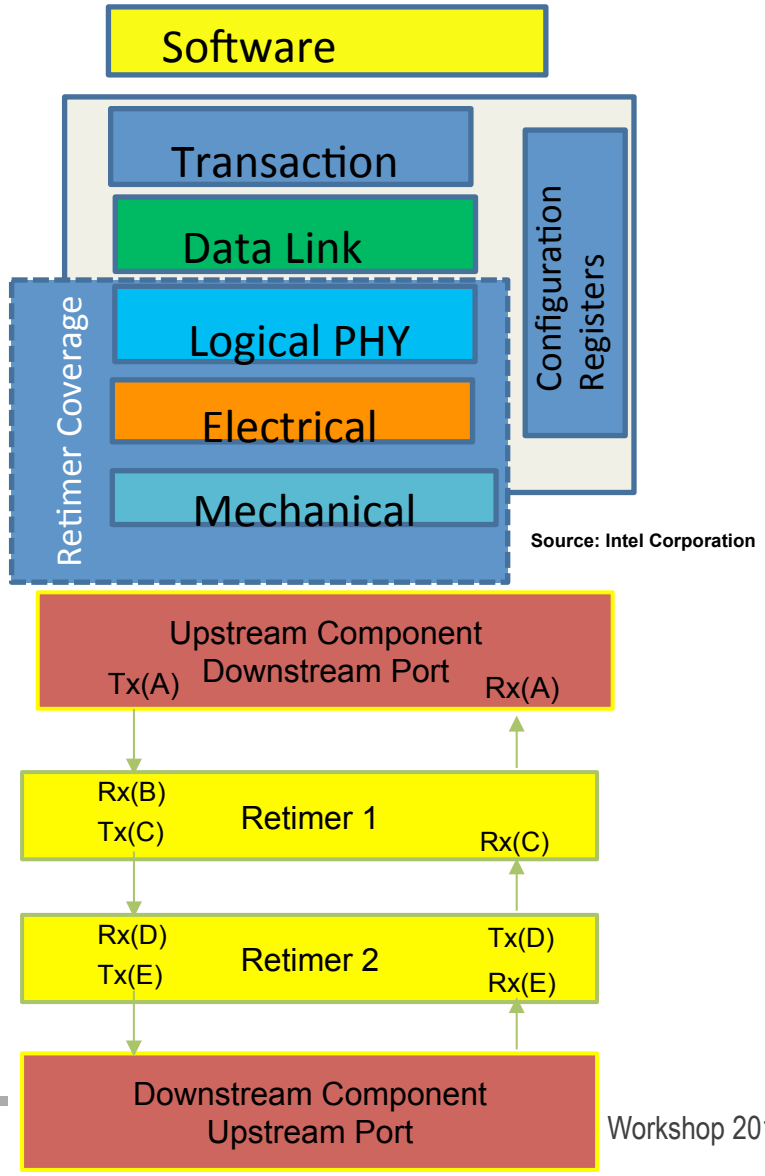
Stage 3/Phase 3 is same as phase 2 in opposite direction
Downstream Port may skip Phase 2/ 3 if presets are good enough for the Link

AGENDA

- Introduction
- PCI Express® (PCIe®) 4.0 Specification
- Retimers for Extending Channel Reach
- Form Factors
- Compliance
- Conclusions

RETIMER TO EXTEND CHANNEL REACH

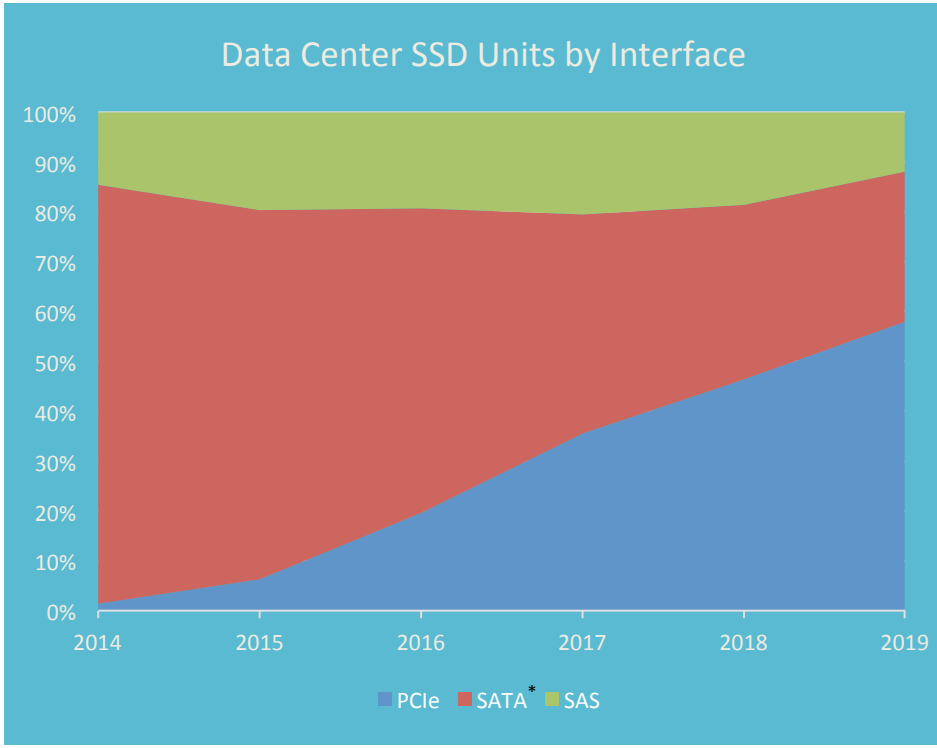
- Channel dxtension devices
- Part of PCIe® base spec (3.1+)
- Up to two retimers
- Critical for longer server channels in PCIe 4.0 architecture
- Has the electrical and PHY Logical – no link/transaction layer, no config registers, no in-band access by S/W
- Actively participates in link training, power management, ppm difference adjustment, link equalization
- Electrically separate links on either end of retimer

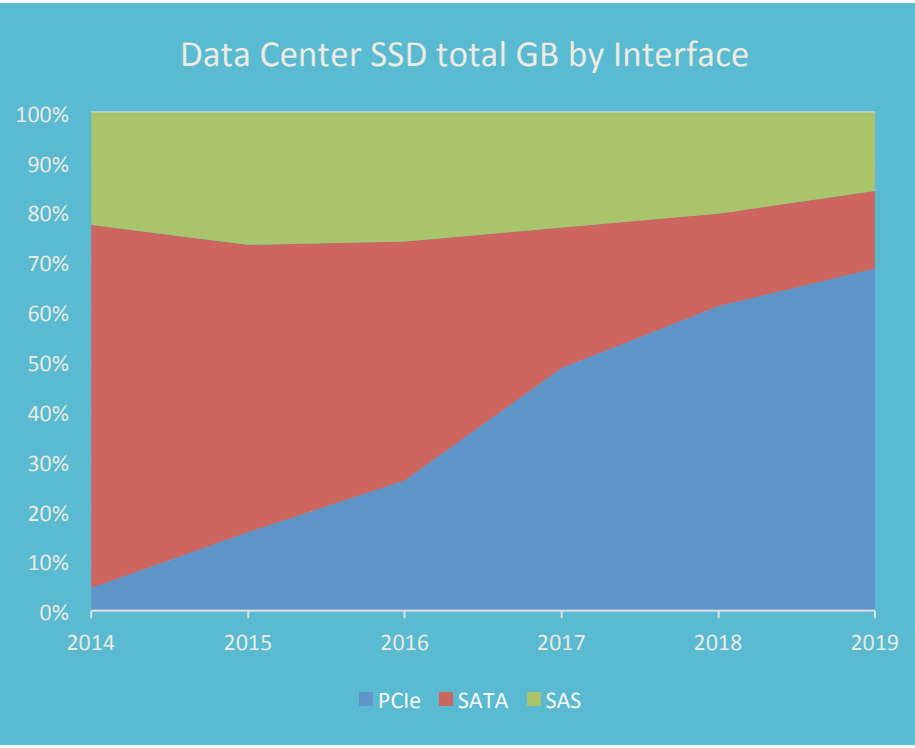


AGENDA

- Introduction
- PCI Express® (PCIe®) 4.0 Specification
- Retimers for Extending Channel Reach
- Form Factors
- Compliance
- Conclusions

NVM EXPRESS™ DRIVING PCI EXPRESS® SSDs IN THE DATA CENTER

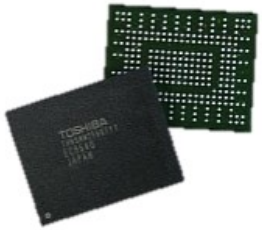




Source: Forward Insights Q1'15

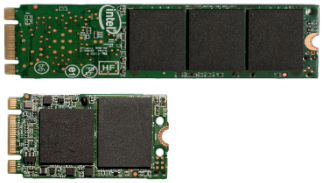
DATA CENTER FORM FACTORS FOR PCI EXPRESS®

BGA



16x20 mm ideal for small and thin platforms

M.2

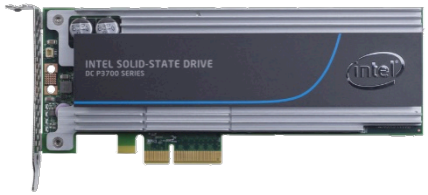


42, 80, and 110mm lengths, smallest footprint of PCI Express® (PCIe®) connector form factors, use for boot or for max storage density

U.2 2.5in (aka SFF-8639)

2.5in makes up the majority of SSDs sold today because of ease of deployment, hotplug, serviceability, and small form factor Single-Port x4 or Dual-Port x2

CEM Add-in-card



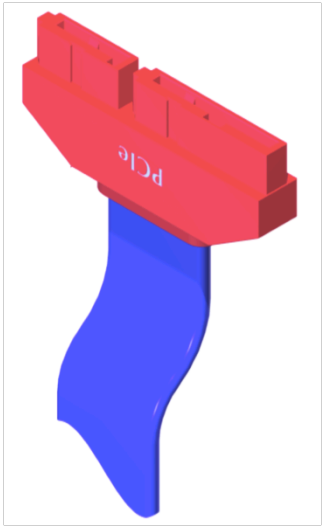
Add-in-card (AIC) has maximum system compatibility with existing servers and most reliable compliance program. Higher power envelope, and options for height and length

Source: Intel Corporation

INEXPENSIVE CABLING = INDEPENDENT CLOCK + SPREAD SPECTRUM (SSC) (SRIS)

- **Challenge: PCI Express® (PCIe®) specification did not support independent clock with SSC**
  - SATA* cable ~ $0.50
  - PCIe cables include reference clock > $1 for equivalent cable
- **PCIe base specification 3.0 ECNs approved**
  1) Requires use of larger elasticity buffer
  2) Requires more frequent insertion of SKIP ordered set
  3) Requires receiver changes (CDR)
  4) Second ECN updates Model CDRs
- **SRIS will create a number of new form factor opportunities for PCIe**
  - OCuLink*
  - Lower cost external/internal cabled PCIe
  - Next generation of PCI-SIG® cable specification

Example of possible PCIe cable

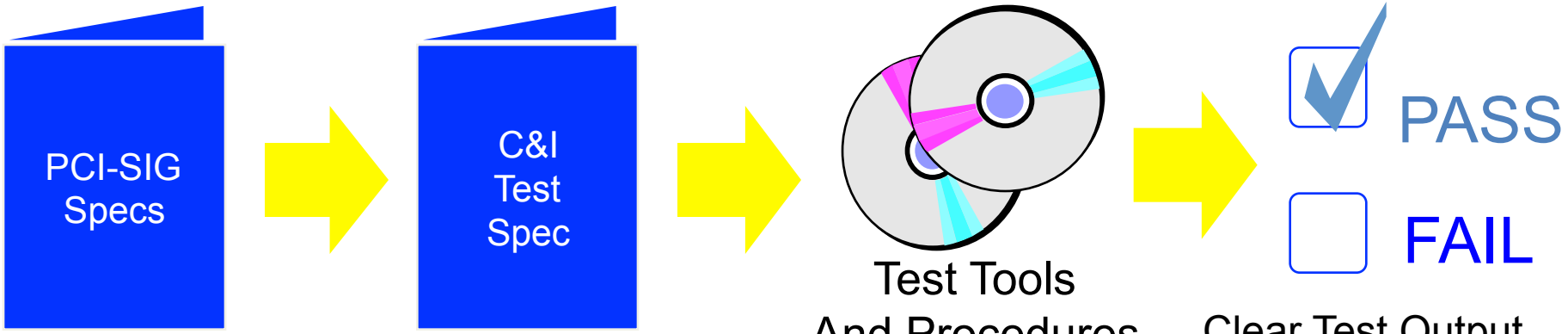


Separate Refclk Modes of Operation: 5600ppm (New - SRIS) and 600ppm (Existing - SRNS)

AGENDA

- Introduction
- PCI Express® (PCIe®) 4.0 Specification
- Retimers for Extending Channel Reach
- Form Factors
- Compliance
- Conclusions

PCI EXPRESS® COMPLIANCE PROCESS



PCI-SIG® Specs
Describes
Device requirements
- 3.0 Base and CEM specs

C&I Test Specs
Define
Test criteria based on spec requirements
- Test Definitions
- Pass/Fail Criteria

Test H/W & S/W
Validates
Test criteria
- Compliance
- Interoperability

Clear Test Output
Maps
- Directly to Test Spec

Predictable path to design compliance

AGENDA

- Introduction
- PCI Express® (PCIe®) 4.0 Specification
- Retimers for Extending Channel Reach
- Form Factors
- Compliance
- Conclusions

CONCLUSIONS

Data Center / HPC | Mobile | Embedded

Source: Intel Corporation

- Single PHY standard covering applications and form factors from handheld to data center
- Predominant direct I/O interconnect from CPU with high bandwidth
- Active development to extend PHY rate to 16 GT/s
- A variety of standard form factors covering applications from small/ light mobile to the data center
- A robust and mature compliance and interoperability program
- Low-power
- High-performance

12th ANNUAL WORKSHOP 2016

THANK YOU

Debendra Das Sharma, PhD
Senior Principal Engineer and Director I/O Technology and Standards
Data Center Group, Intel Corporation
Chair, PHY Logical Group, PCI-SIG®