OCP SUMMIT

March 20-21
2018
San Jose, CA

OPEN. FOR BUSINESS.

PCI Express: Delivery Bandwidth for OCP

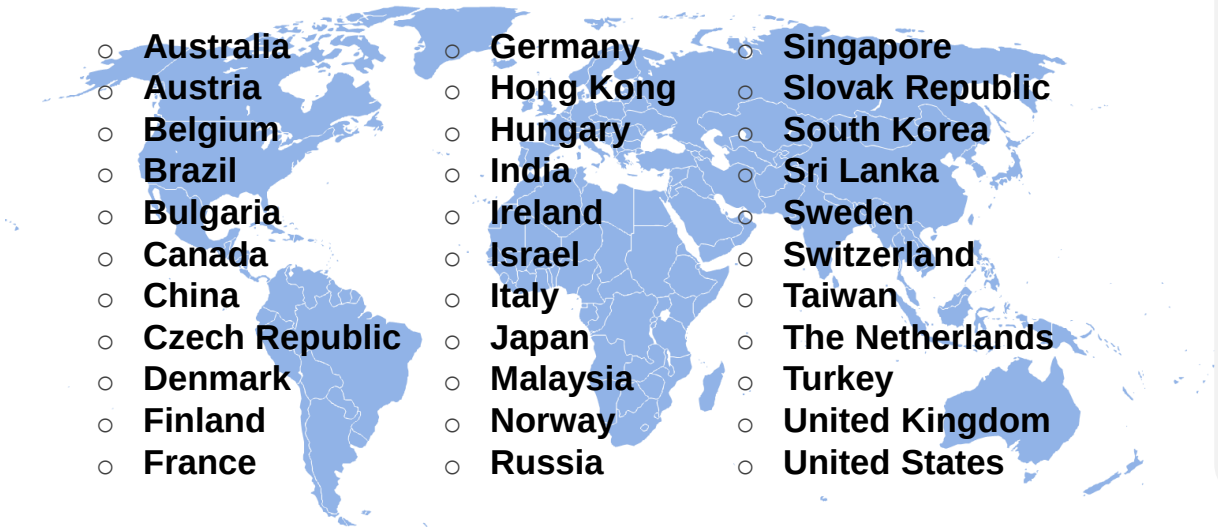
Al Yanes / President / PCI-SIG

PCI-SIG® Snapshot

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

- **750+** member companies located worldwide

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



- Australia
- Austria
- Belgium
- Brazil
- Bulgaria
- Canada
- China
- Czech Republic
- Denmark
- Finland
- France
- Germany
- Hong Kong
- Hungary
- India
- Ireland
- Israel
- Italy
- Japan
- Malaysia
- Norway
- Russia
- Singapore
- Slovak Republic
- South Korea
- Sri Lanka
- Sweden
- Switzerland
- Taiwan
- The Netherlands
- Turkey
- United Kingdom
- United States

BOARD OF DIRECTORS 2017-2018

AMD
DELL
Hewlett Packard Enterprise
IBM
intel
lenovo
NVIDIA
QUALCOMM®
SYNOPSYS®
Silicon to Software™

PCI Express 4.0

PCI-SIG continues its solid reputation of delivering **low cost, high-performance, low-power specifications** for **multiple applications** and **markets.**

- **PCI Express 4.0 Specification – (16GT/S)**
  - Finalized and published October 2017
  - Includes new performance enhancements
  - Maintains position as the interconnect of choice for the expansive storage market and the backbone for the fast growing cloud ecosystem

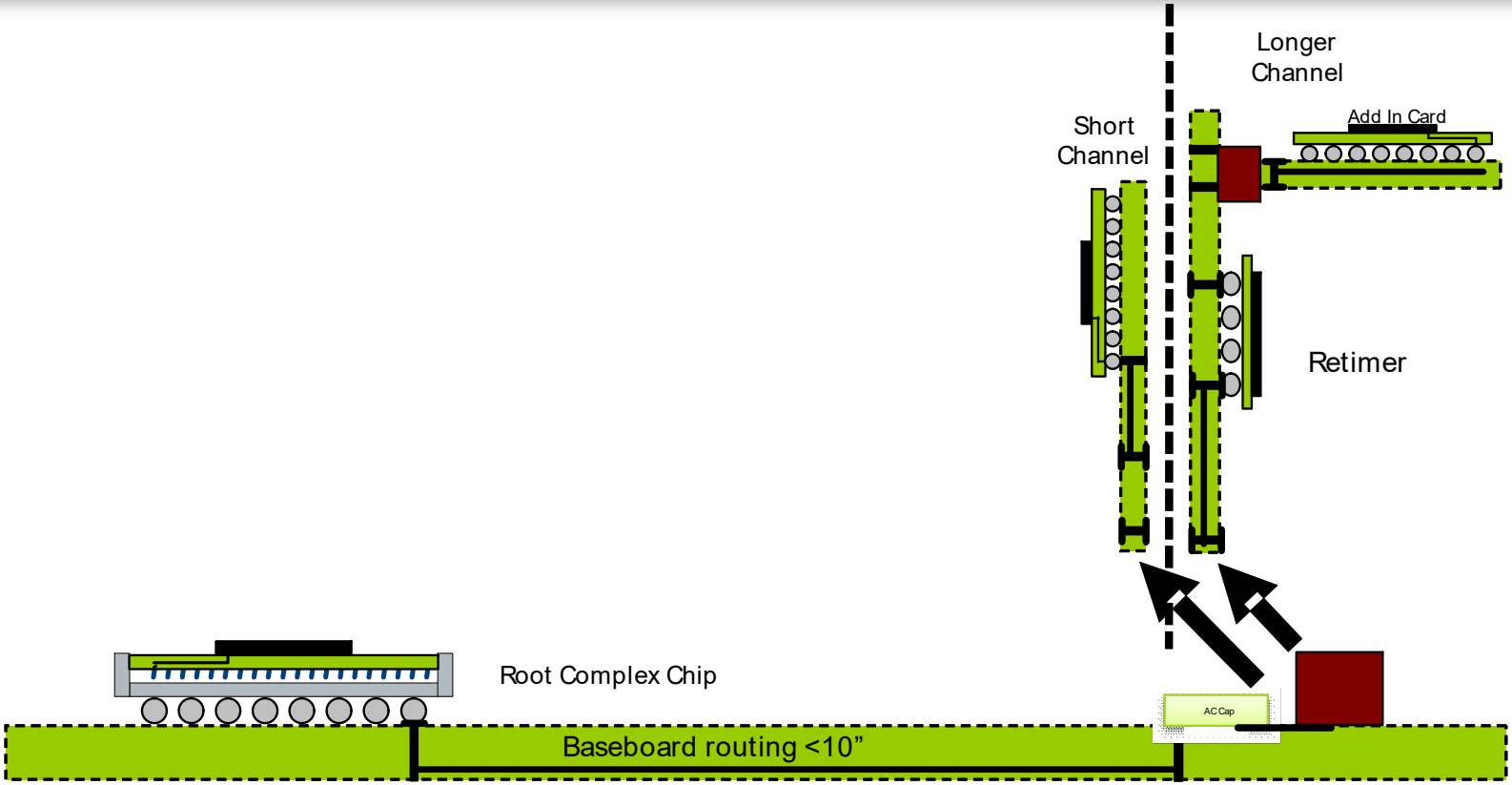
PCI Express 4.0

- **PCIe 4.0 Key Functional Enhancements**
  - Lane Margining at the Receiver - Allows systems to determine how close to “the edge” each lane is operating
  - Expanded Tag and Credits
    - Allows both tags/credits to service devices for future usage
      - 10 bit Extended Tags support upto 1024 transactions
      - Scaled Flow Control supports larger credits
  - PCIe 4.0 Electrical
    - Maintains backward compatibility with installed base of PCIe devices
    - Limited channel reach: approx. 12” one connector (including 4” add-in card)
    - Longer channels require retimers or lower loss channel
- **PCIe 4.0 Adoption**
  - Numerous vendors have 16GT/s PHYs in silicon
  - Major IP vendors offering 16GT/s controllers
  - Dozen 16GT/s solutions at a recent PCI-SIG Compliance Workshop
  - Several member companies have exhibited 16GT/s demos

Power Efficient Performance

- **Delivers Scalable Performance**
  - Width scaling: x1, x2, x4, x8, x12, x16,
  - Frequency scaling: Five generations
    - 2.5 and 5 GT/s with 8b/10b encoding
    - 8 and 32 GT/s with 128b/130b encoding
- **Low Power (Active/Idle)**
  - Rich set of Link and Device States
    - L0s, L1, L1-substates, L2/L3
    - D0, D1, D2, D3_hot/cold
    - Platform-level power optimization hooks: Dynamic Power Allocation, Optimized Buffer, Flush Fill, Latency Tolerance Reporting
    - Active power – 5pJ/b, Standby power: 10uW/Lane
- **Vibrant ecosystem with IP Providers**

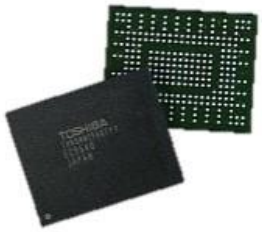
PCI Express 4.0 Channels



- End to end loss target ~= 28 dB
- Root Package loss ~= 5 dB
- Add-in Card Package loss ~= 3 dB
- Total Add-in Card ~= 8.0 dB
- Connector < 1 dB

Form Factors for PCI Express

BGA



16x20 mm ideal for small and thin platforms

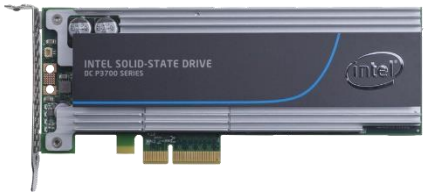
M.2

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

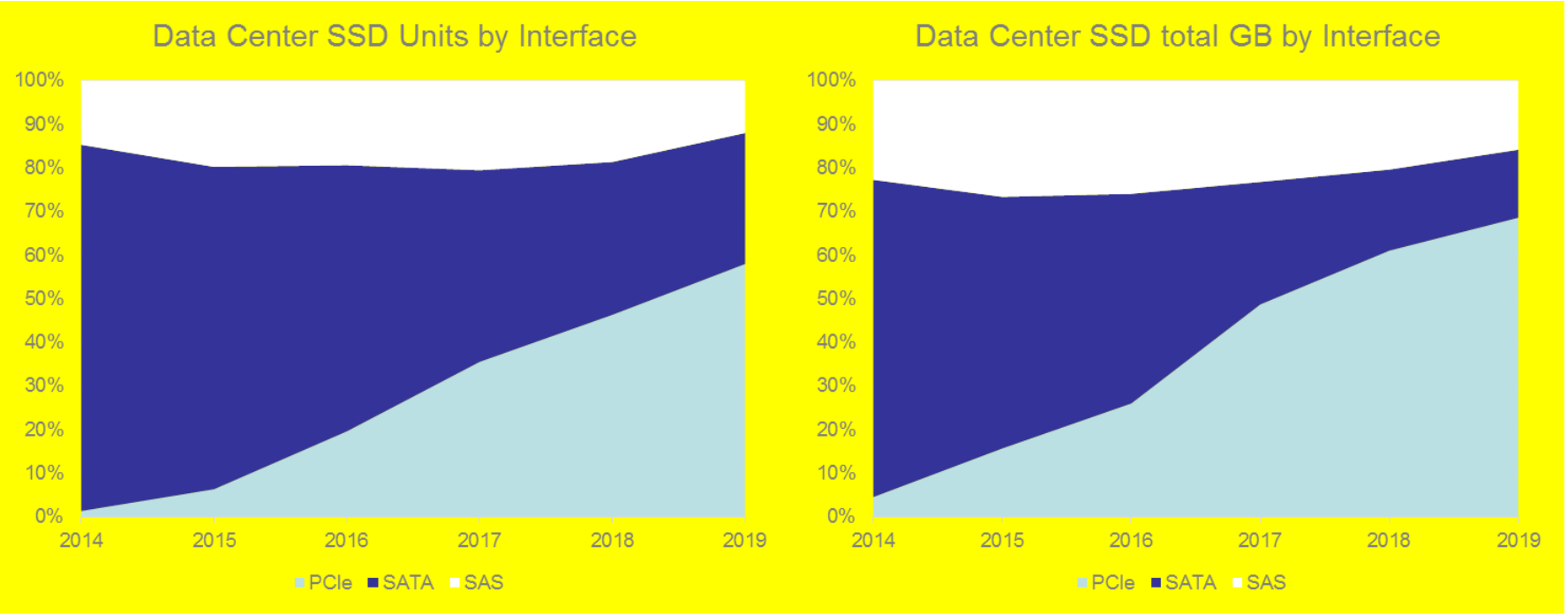
SRIS (Separate Refclk Independent SSC Architecture)

- **Challenge: PCIe specification did not support independent clock with SSC**
  - SATA* cable ~ $.50
  - PCIe cables include reference clock > $1 for equivalent cable
- **PCIe base specification 3.0 ECNs approved**
  - Requires use of larger elasticity buffer
  - Requires more frequent insertion of SKIP ordered set
  - Requires receiver changes (CDR)
  - 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

RAS Features

- **PCIe® architecture supports very high-level set of Reliability, Availability, Serviceability (RAS) features**
  - All transactions protected by CRC-32 and Link level Retry, covering even dropped packets
  - 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

NVM Express™ Driving PCIe SSDs in Data Center

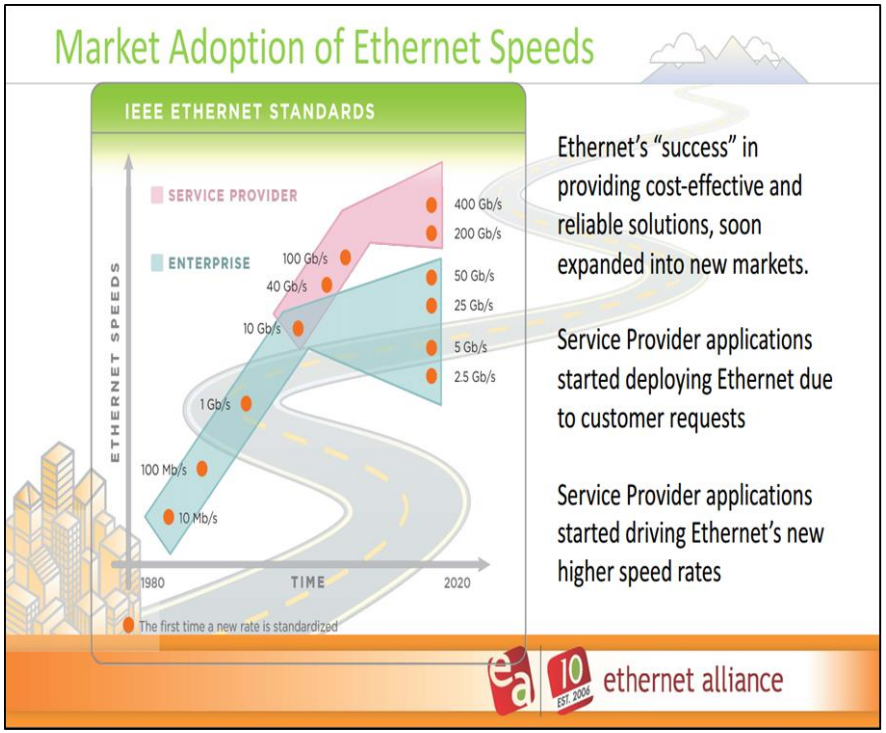


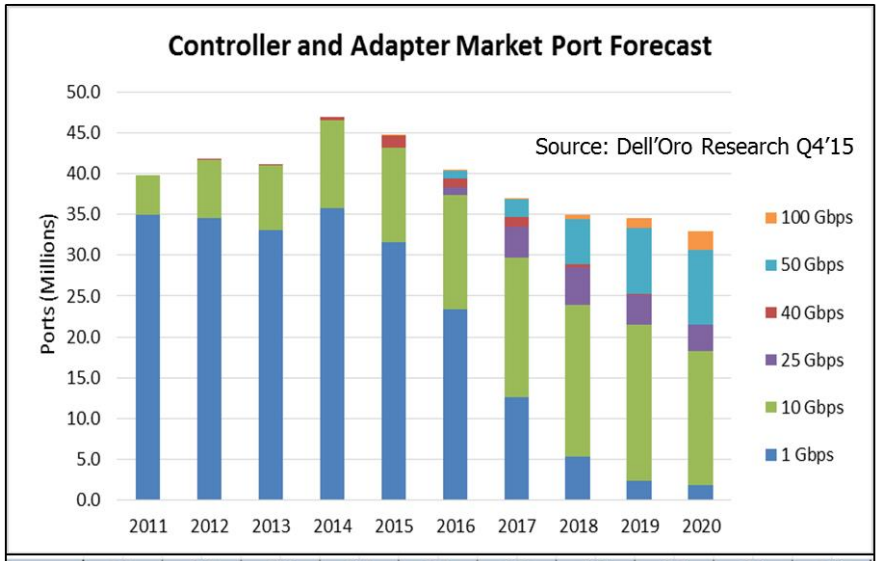
Source: Forward Insights Q1'15

PCI Express – 5.0 Specification

- **PCI Express 5.0 Specification 32GT/S NRZ:**
  - Applications such as artificial intelligence, machine learning, gaming, visual computing, storage and networking
  - High-end networking solutions (i.e. 400Gb Ethernet and dual 200Gb/s InfiniBand solutions)
  - Accelerator and GPU attachments for high-bandwidth solutions
  - Constricted form factor applications that cannot increase width and need higher frequency to achieve performance
  - Continued use of L1 Sub-states to constrain power consumption during transmission idle periods

Ethernet Evolution



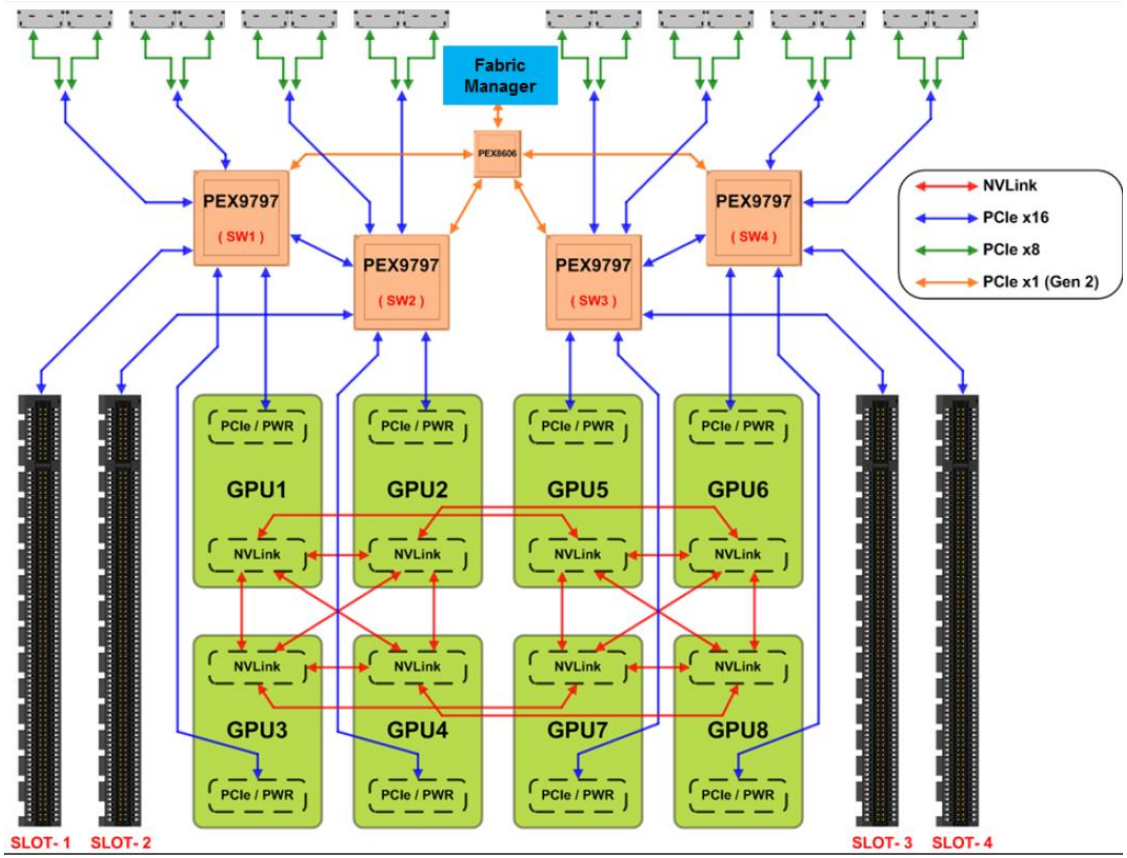


	2011		2012		2013		2014		2015		2016		2017		2018		2019		2020	
	Ports	BW	Ports	BW	Ports	BW	Ports	BW	Ports	BW	Ports	BW	Ports	BW	Ports	BW	Ports	BW	Ports	BW
1	35	35	34	34	34	34	35.5	35.5	31	31	23	23	12	12	5	5	2	2	1.5	1.5
10	5	50	8	80	8	80	10	100	11	110	14	140	18	180	23	230	20	200	18	180
25	0	0	0	0	0	0	0.5	12.5	2	50	1	25	4	100	5	125	3	75	3	75
40	0	0	0	0	0	0	0	0	0	0	1	40	1	40	1	40	0.5	20	0	0
50	0	0	0	0	0	0	0	0	0	0	1	50	2	100	6	300	8	400	9	450
100	0	0	0	0	0	0	0	0	0	0	0	0	0.1	10	0.5	50	1	100	2	200
Total Gbps		85		114		114		148		191		278		442		750		797		906.5
YOY Ratio		N/A		1.3		1.0		1.3		1.3		1.5		1.6		1.7		1.1		1.1
3 Year Ratio								**1.7**						**3.0**						**2.1**

GPUs and PCIe Bandwidth

Open Compute Project's Olympus Hyperscale GPU Accelerator chassis

- Flexible PCIe topology
- GPGPU-to-Host via high-BW PCIe links
- Peer-to-peer without Host interaction
  - GPGPU peer-to-peer via NVLink
  - GPGPU peer-to-peer to IB NICs via x16 PCIe



Source: Open Compute Project

PCIe 5.0: Near Future

- **Timeline for PCIe 5.0 specification by 1H/2019**
  - Changes limited to primarily speed upgrade
    - Protocol already supports higher speed via extended tags and credits
    - Existing PHYs in the industry already run at 28GHz / 56GHz
  - Specification process enhanced to accelerate development
    - PCIe 5.0 version 0.5 released Q4 2017
      - Max pad to pad loss target is expected to be around 35 dB
      - Focus on connector studies
      - Pre-encoding to reduce DFE burst errors
      - SRIS – potentially mandatory
      - EIEOS definition at 32 GT/s change to sequences of 32 ones and zeros
    - **PCIe 5.0 version 0.7 target release April 2018**

PCIe 5.0 Delivering 32GT/s

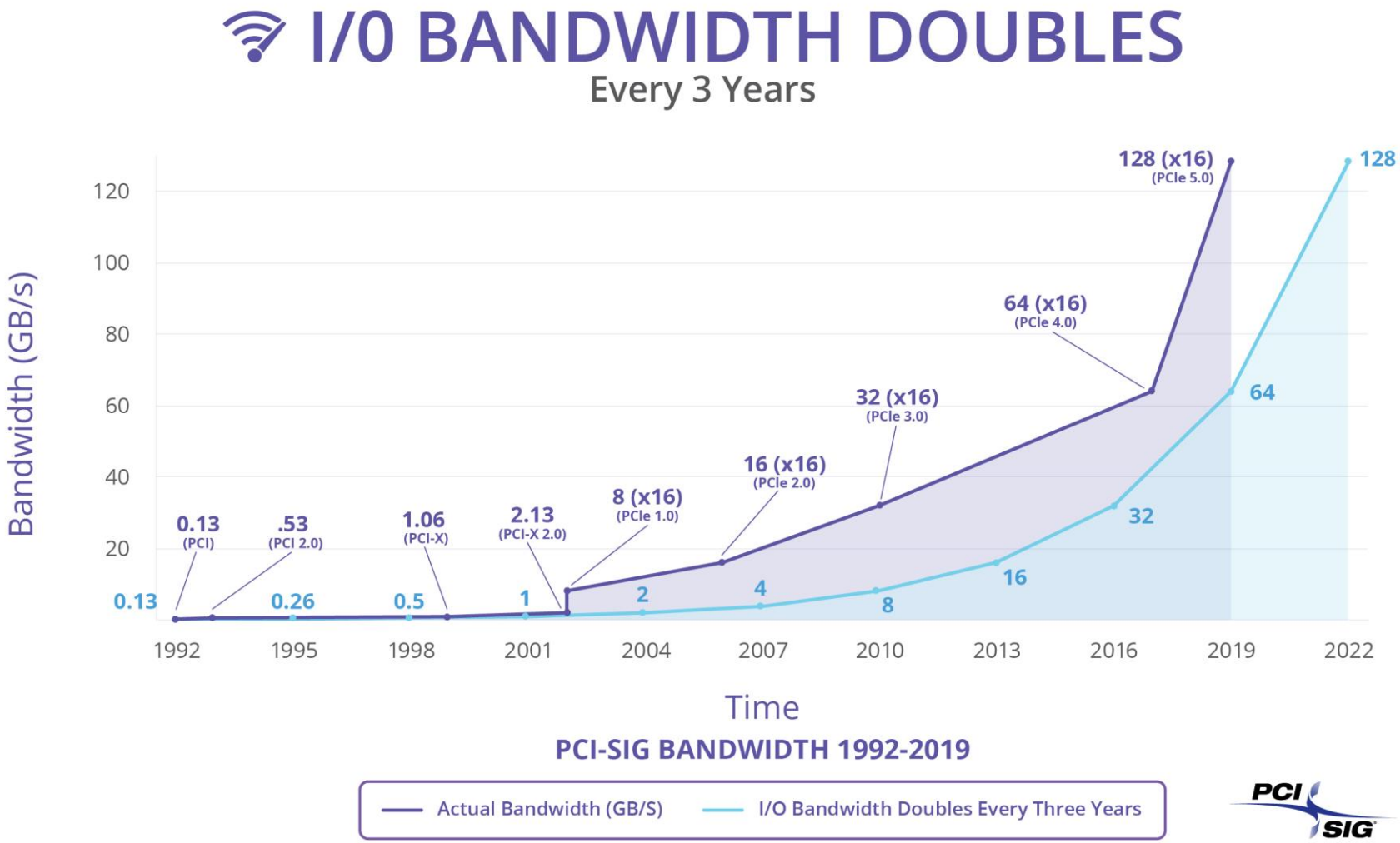
- ✓ **Supports 400Gb Ethernet Solutions**
- 400Gb = 50GB
- 50GB in both directions

- ✓ **Full duplex**
- 128/130 bit encoding with 1.5% overhead
- x16 ~64GB/s sufficient to support 400Gb Ethernet solutions (64GB > 50 GB)
- Total Full Duplex = ~128GB

- CEM connector targeted to be backwards compatible for add-in cards
- **Targeted Release in 1H 2019**

	RAW BIT RATE	LINK BW	BW/ LANE/WAY	TOTAL BW X16
PCIe 1.x	2.5GT/s	2Gb/s	250MB/s	8GB/s
PCIe 2.x	5.0GT/s	4Gb/s	500MB/s	16GB/s
PCIe 3.x	8.0GT/s	8Gb/s	~1GB/s	~32GB/s
PCIe 4.0	16GT/s	16Gb/s	~2GB/s	~64GB/s
PCIe 5.0	32GT/s	32Gb/s	~4GB/s	~128GB/s

PCI-SIG History



Summary

- PCIe technology continues to evolve to exceed industry bandwidth requirements
  - PCIe 5.0 with 32GT/s ideal for artificial intelligence, machine learning, gaming, visual computing, storage and networking
  - Growing need for increased bandwidths in GPUs attachments and accelerators

OCP SUMMIT