The History of PCI IO Technology: 30 Years of PCI-SIG® Innovation

PCI-SIG Webinar Series

June 29, 2022

Meet the Speaker

Dr. Debendra Das Sharma

Intel Senior Fellow and co-GM Memory and I/O Technologies, Intel Corporation and PCI-SIG® Board Member and Chair of PHY Logical

Agenda

- Introduction to PCI-SIG® and its technologies: PCI and PCI Express® (PCIe®) technology
- PCI – the age of bus-based architectures
- PCI Express technology – the BIG transition
- I/O Virtualization – the enterprise (and cloud) play by PCI Express infrastructure
- PCIe 2.0 specification – the backwards-compatible bandwidth doubling journey starts
- PCIe 3.0 specification – navigating the fork in the road; PCIe technology integrated in CPU sockets!
- Low-power L1 sub-states – PCI Express technology in Smart Phones and Hand-held Devices
- PCIe 4.0 specification – Overcoming the channel challenges to get to 16 GT/s
- PCIe 5.0 specification – the bandwidth doubling continues with Alternate Protocol support
- PCIe 6.0 specification – Can we really achieve low-latency with PAM4 and FEC?
- Conclusions and Call to Action

PCI-SIG®: An Open Industry Consortium

PCI-SIG: Organization that defines the PCI Express® (PCIe®) specifications and related form factors

(PCI-SIG: Peripheral Components Interconnect Special Interest Group)

Established in 1992 – 30 years anniversary and growing stronger – THANK YOU!!

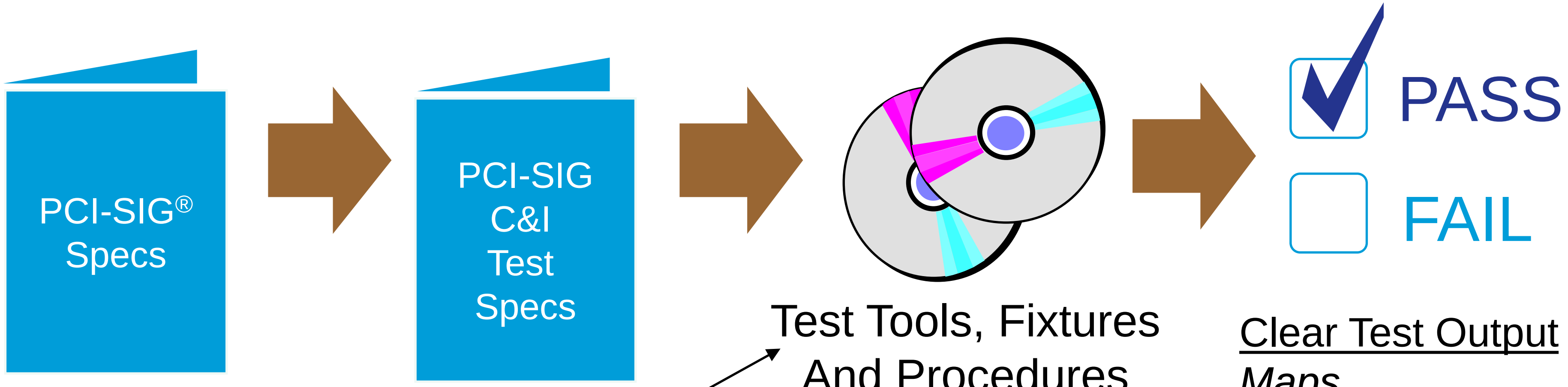
900+ member companies worldwide

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

Board of Directors
2021 –2022

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

PCI-SIG®: From Spec to Compliance



PCI-SIG® Specs

Describes

Requirements

Base specification and Form-Factor specs

C&I (Compliance and Interop) Test Specs

Describes

Test criteria based on spec requirements

- Test Definitions
- Pass/Fail Criteria

Test Tools, Fixtures And Procedures

Test H/W & S/W

Validates

Test criteria

- Compliance
- Interoperability

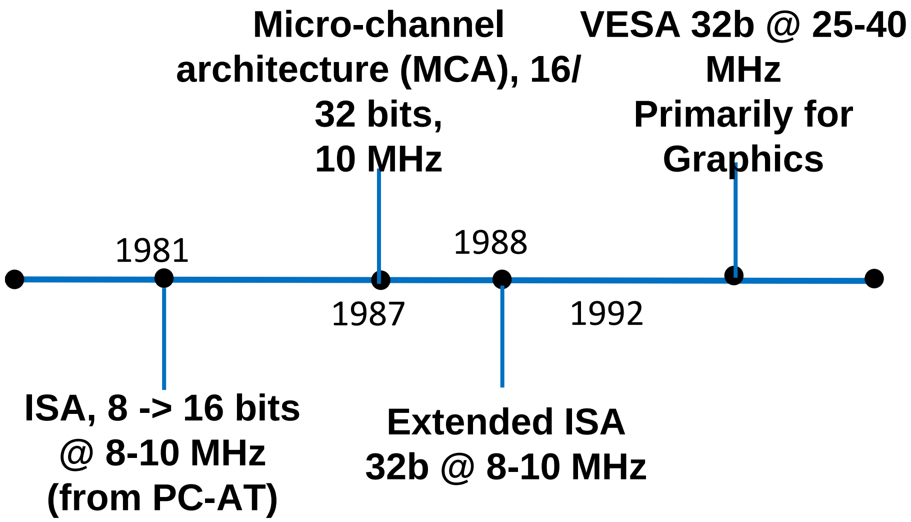
Clear Test Output

Maps

- Directly to Test Spec

Predictable path to design compliance

PCI – Debuts in 1992



Micro-channel architecture (MCA), 16/ 32 bits, 10 MHz

VESA 32b @ 25-40 MHz Primarily for Graphics

1981

1987

1988

1992

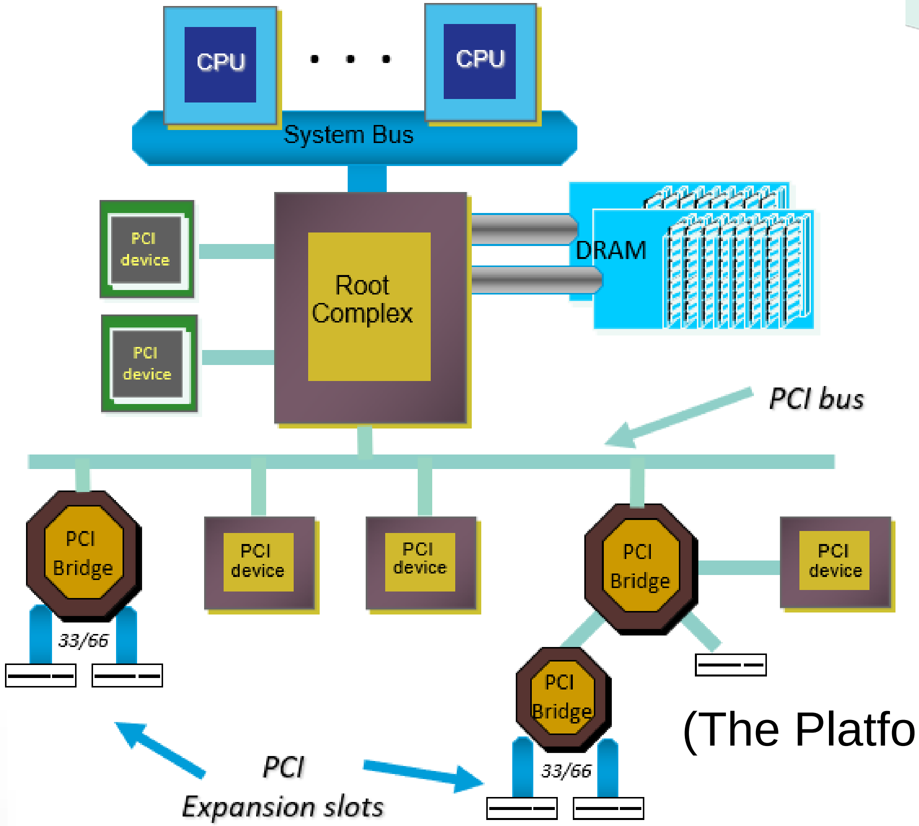
ISA, 8 -> 16 bits @ 8-10 MHz (from PC-AT)

Extended ISA 32b @ 8-10 MHz

Other events in 1992

- Elvis Presley Stamp introduced by US Postal Service with a younger Elvis Presley
- 25th Olympic Games held in Barcelona, Spain
- Terminator-2 movie debuts

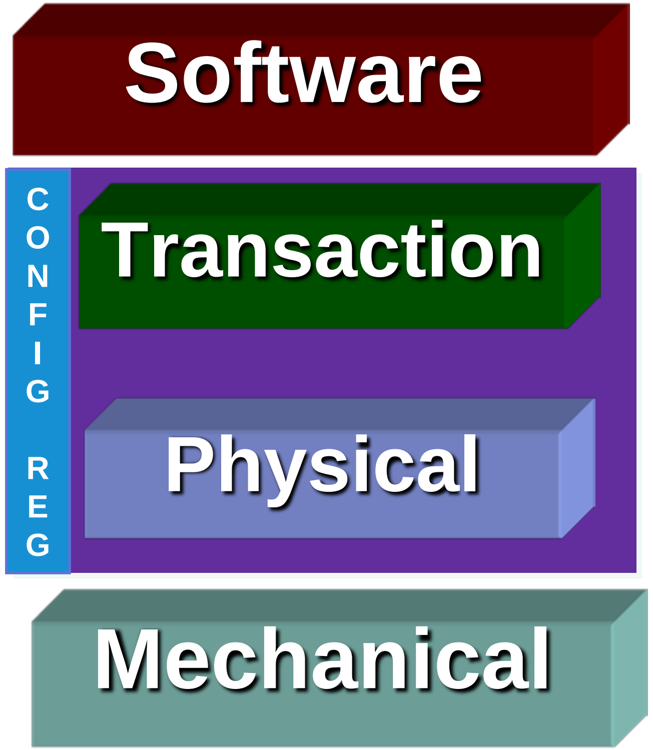
Photo by Dave Kim on Unsplash
Photo by Miquel Migg on Unsplash



(The Platform View)

- PCI successful in consolidating a fragmented industry with multiple standards to one
  - better customer experience
  - accelerated innovation through an open industry standard slot
- Primary compute: PC, Workstations

PCI and PCI-X: 1992 - 2003



⇦ Configuration Register for device discovery (BIOS/ OS), Driver model

⇦ Addressing automated through Base Address Register (BAR)

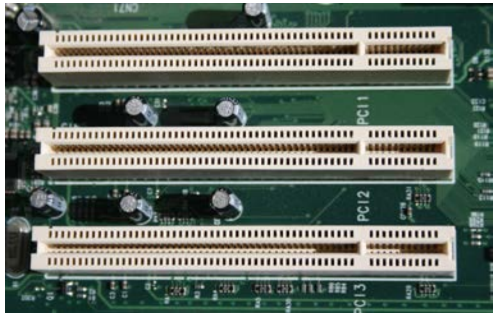
⇦ Connected -> Delayed -> Split transaction starting with PCI-X

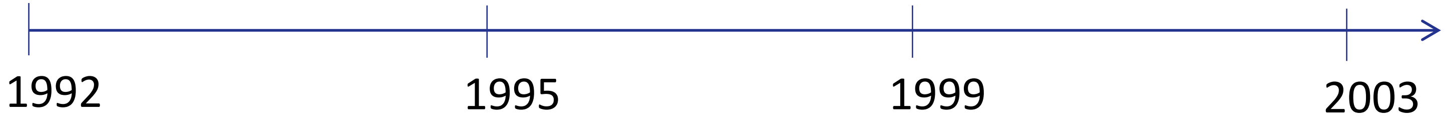
⇦ Architectural Ordering Model: Producer-Consumer based

⇦ Interrupt architected (dedicated wire)

⇦ Bus-based, multi-drop with multiple bus masters (arbitration)

⇦ Cards on a PCI / PCI-X Slot model





1992: 32b @ 33/ 66 MHz

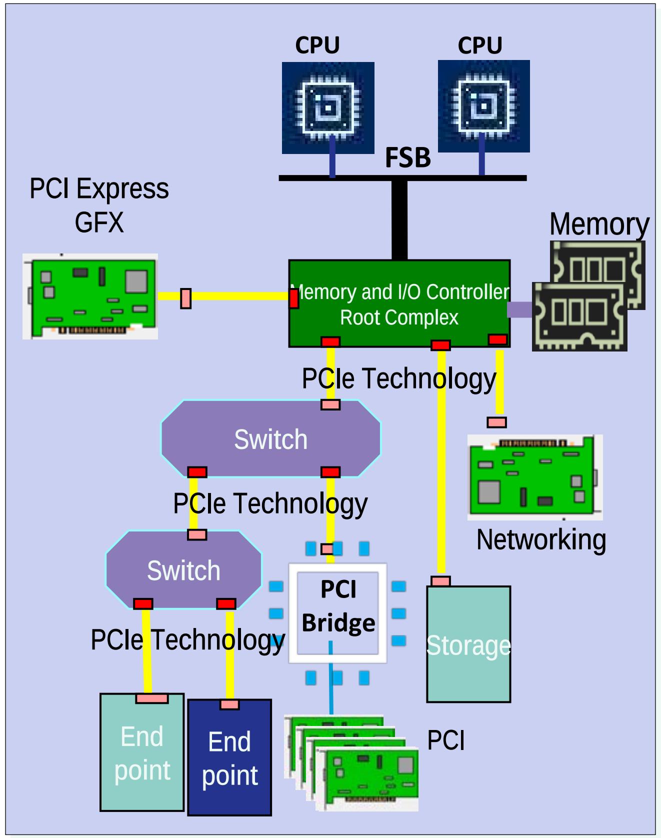
1995: PCI 2.1: 32b/ 64b @ 33/ 66 MHz

1999: PCI X: 32b/ 64b @ 33/ 66/ 133 MHz

2003: PCI X 2.0: 16b/ 32b/ 64b @ 33/ 66/ 133/ 266/ 533 MHz

The BIG Transition: PCI Express® Specification Debuts in 2003

- Problem Statement: Continued I/O bandwidth and connectivity demand makes PCI bus untenable
  - Pin-inefficiency and scaling challenges => Cost increase
  - Performance implications of bus sharing
- Solution: PCI Express architecture, a Link-based interconnect
  - Differential, full-duplex signaling at 2.5GT/s
  - Multiple widths: x1, x2, x4, x8, x12, x16, x32
- Software compatibility w/ PCI makes transition feasible
  - No hardware compatibility
  - PCIe® to PCI bridge for platform transition to PCIe technology



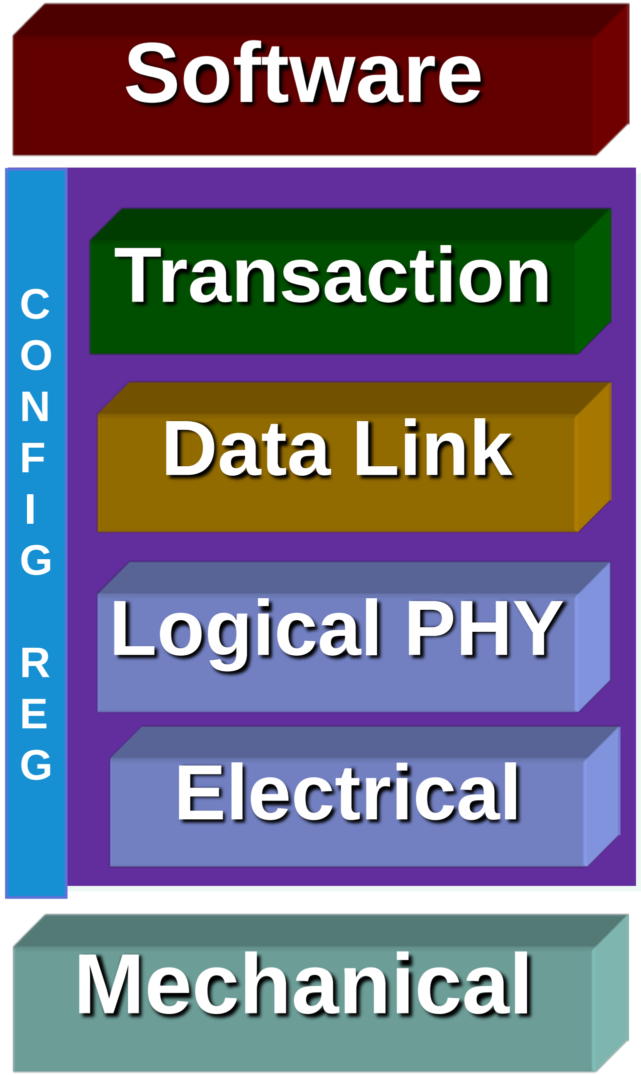
(The Platform View)

Other events in 2003

- Human Genome projected launched in 1990 completed
- Tesla, Inc. founded
- Space shuttle Columbia disaster
- International Year of Fresh Water

Photo by NASA on Unsplash
Photo by Braňo on Unsplash

PCIe® Architecture Layering for Modularity and Reuse



Software

CONFIG REG

Transaction

Data Link

Logical PHY

Electrical

Mechanical

⇦ PCI compatibility, configuration/ enhanced configuration, driver model

⇦ Advanced Error Reporting, Hot-Plug, Power Management

⇦ Split-transaction, packet-based protocol with producer-consumer ordering

⇦ Credit-based flow control, virtual channels, hierarachical timeout

⇦ Logical connection between devices

⇦ Reliable data transport services (CRC, Retry, Ack/Nak)

⇦ Physical information exchange

⇦ Interface initialization and maintenance

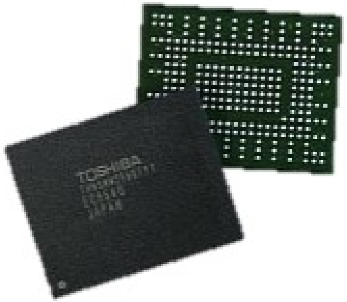
⇦ Market segment specific form factors

⇦ Evolutionary and revolutionary

PCIe technology has a long track record of being implemented in high volume manufacturing products with server-grade reliability

PCIe® Architecture: One Base Specification - Multiple Form Factors

BGA



16x20 mm small and thin platforms

M.2

Smallest footprint (22mm x 30 to 110 mm): SSDs in boot slots, data center storage, WWAN

U.2 2.5in (aka SFF-8639)

SSDs x4 or 2 x2 w/ hot-plug

CEM Add-in-card

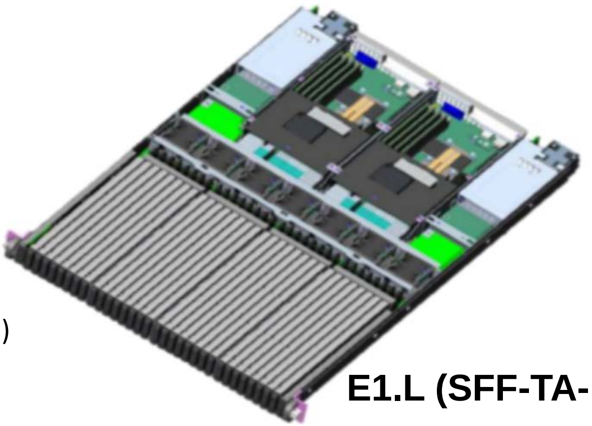
Widely used in systems w/ 4 HL options. Higher Power. Robust compliance program

High B/W: hand-held, IoT, automotive

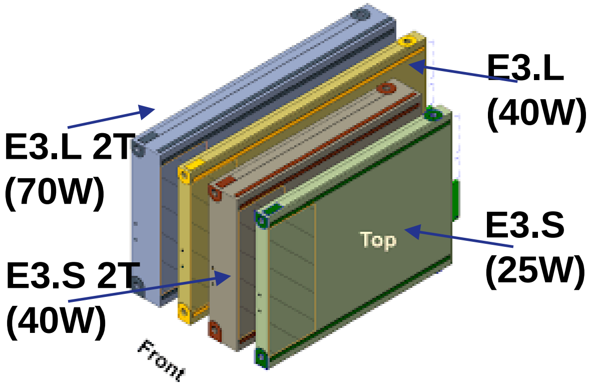
High-end still and motion cameras

(Up to 36 Modules)
(Up to 32 Modules)

E1.S (SFF-TA-1006)



E1.L (SFF-TA-1007,



E3 Form-factors

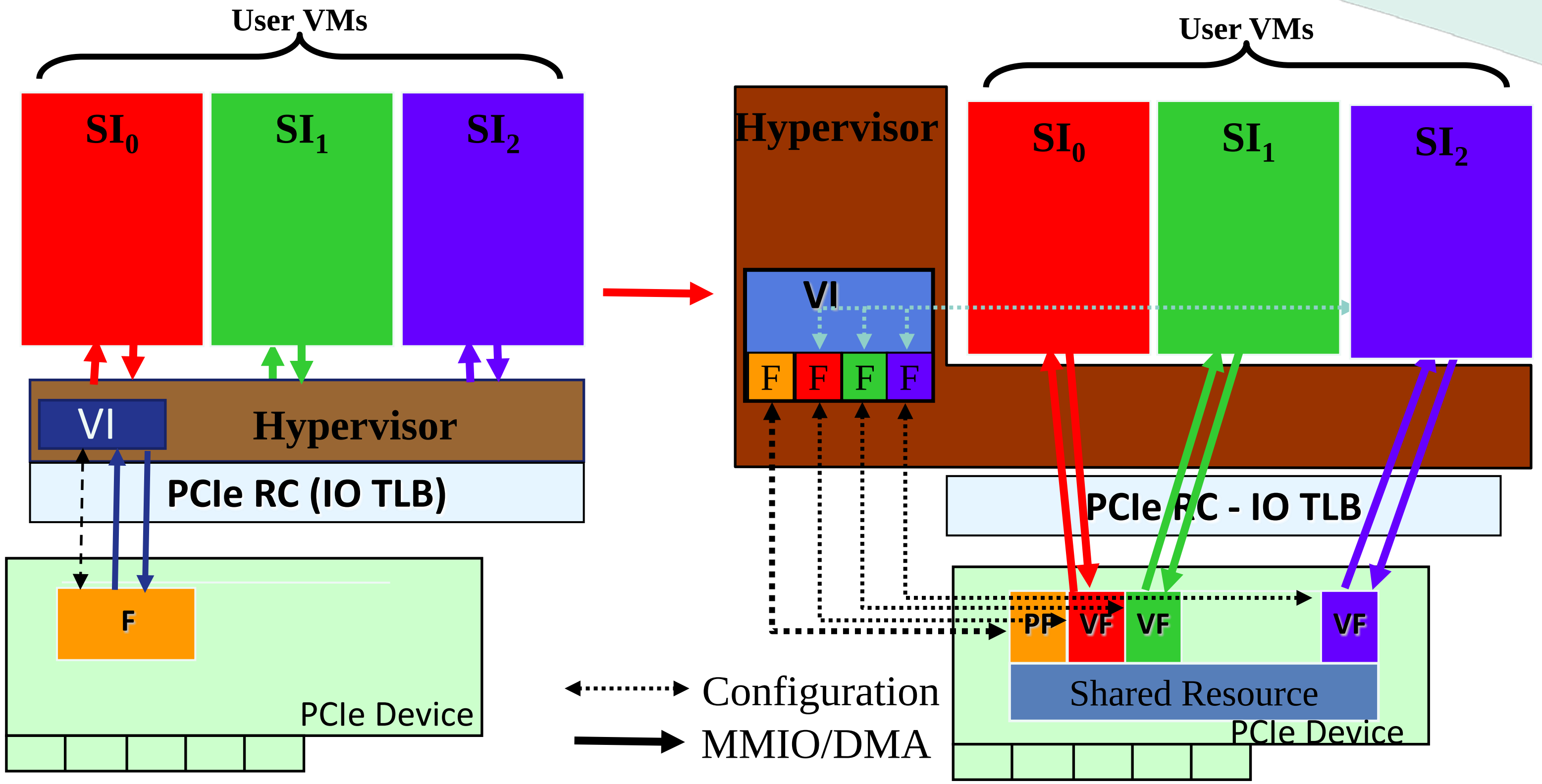
Various Proprietary FFs for HPC Applications Multi-KW cards

Enterprise and **D**atacenter **S**mall **F**orm **F**actor (EDSFF) family was designed for Enterprise and Datacenter applications and widely used for SSDs.

Multiple Form-factors from the same silicon to meet the needs of different segments

I/O Virtualization: Addressing the Enterprise Needs

- Usage: Client, Server / Cloud
- Drivers: Multi-core; better TCO
- Multiple SIs on same machine.
- Benefits: I/O Performance
- With native PCIe® IOV:
  - Each device VF mapped to one SI
  - Direct memory access
    - IOTLB translation
  - Config Cycles emulated by VI

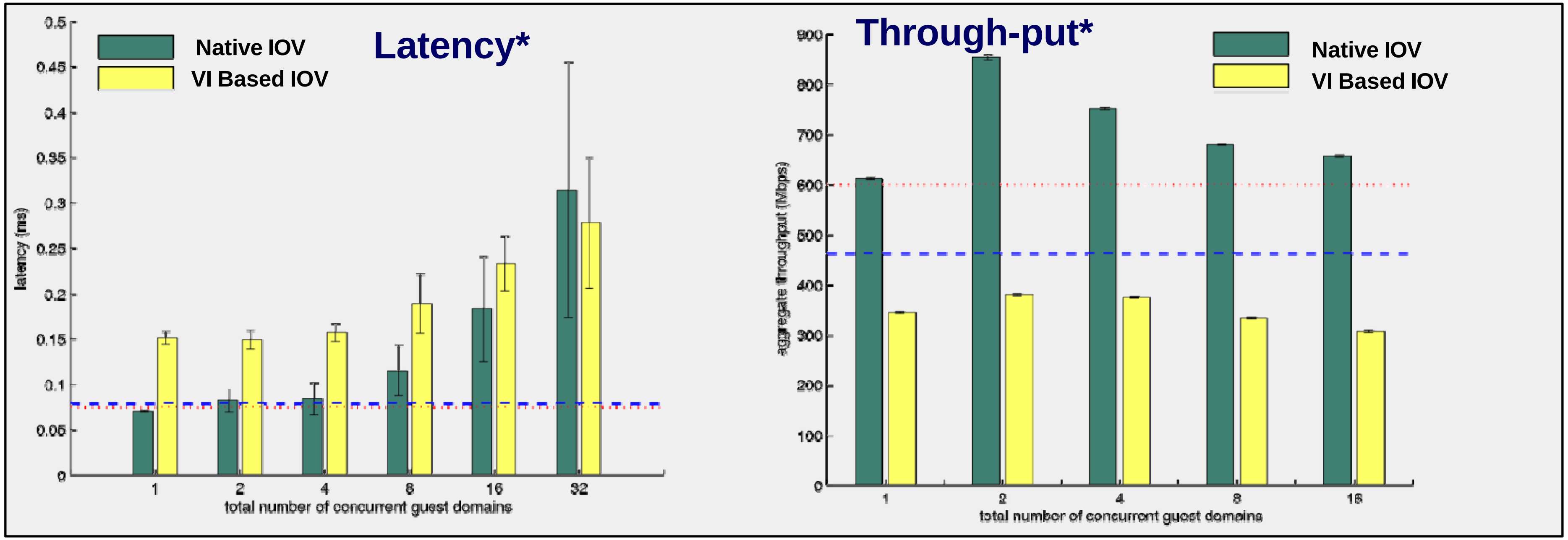


(Without PCIe IOV: All accesses go through VI – performance suffers)

(PCIe IOV: Memory accesses bypass VI)

Abbreviations – VI: Virtualization Intermediary, SI: System Image – aka Virtual Machine/ VM
F: Function, VF: Virtual Function, PF: Physical Function

IO Virtualization Performance

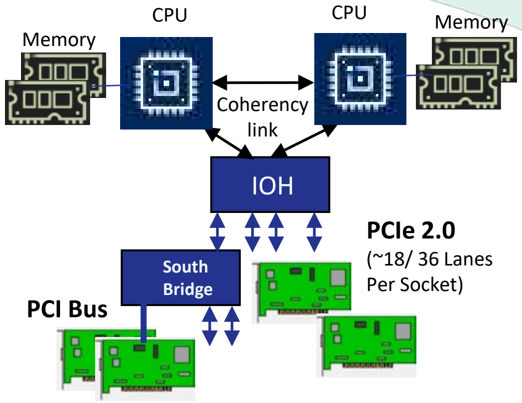


- VI based IOV adds path length on every IO operation.
- Native IOV significantly improves performance
  - ✓ Doubles throughput and reduces latency by up to half.

***Source:** Self-Virtualized I/O: High Performance, Scalable I/O Virtualization in Multi-core Systems; R. Himanshu, I. Ganev, K. Schwan - Georgia Tech and J. Xenidis - IBM

PCI Express® Technology Evolution – PCIe® 2.0 Specification in 2007

- Dynamic Speed change mechanism defined – still in use today
- PCIe specification: doubles data rate every generation with full backward compatibility
  - a x16 PCIe 5.0 interface interoperates with a x1 Gen 1!
- Ubiquitous I/O across the compute continuum
  - PC, Hand-held, Workstation, Cloud, Enterprise, HPC, Embedded, IoT, Automotive



(The Platform View)

(Memory Controller integrated to CPU in the era of multi-core computing)

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	64.0 (PAM-4, Flit)	1024 Gb/s	2022

Other events in 2007

- iPhone debuts

- NASA launches Phoenix Mars Lander
- Housing bubble bursts

- Devastating earthquake in Peru

Photo by Jared Allen on Unsplash
Photo by Breno Assis on Unsplash

PCIe® Architecture Market Applications: One Interconnect – Infinite Applications

HPC / Cloud

Data Center / Enterprise Servers

Artificial Intelligence / Machine Learning

Automotive

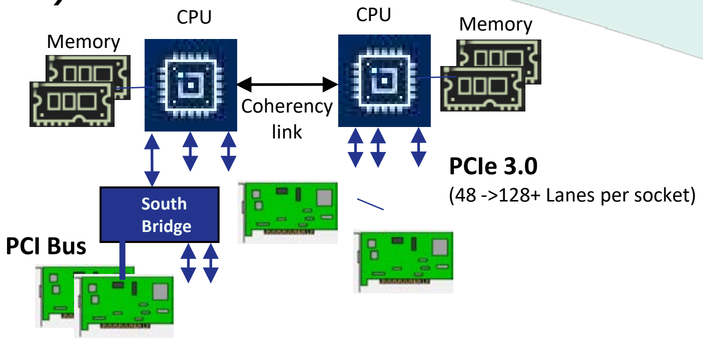
Internet of Things

Military / Aerospace

Storage

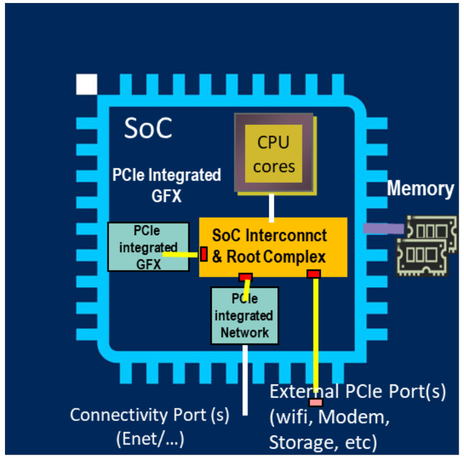
PCI Express® 3.0 Specification in 2010: The Fork in the Road

- PCIe 3.0 specification data rate analysis (cost, area, power constraints):
  - 10G not feasible server channels (20” FR4 and 2 Conn): 8G ok
  - Client/ Mobile okay at 10G – but need to take server along
- Two-pronged solution: 1.6 data rate X 1.25 encoding = 2X b/w
  - Data Rate at 8G (1.6 increase in bandwidth)
  - Use a new 128b/130b encoding instead of 8b/10b encoding (1.25x)
  - Challenges/ Solution:
    - DC wander and cross-talk (new scrambler): still in use (16G, 32G, 64G)
    - Framing Tokens w/ 128b/130b: used later (16G, 32G)
    - Equalization mechanism: still in use (16G, 32G, 64G)
- Hind-sight: One of the best decisions! One Interconnect for all!
  - Subsequent data rates easier: 16/ 32/ 64 G vs 20/ 40/ 80G



(The Scalable Platform View)

(PCIe integrated to CPU – scalable connectivity and bandwidth. From Gen 3 onwards)


SoC
CPU cores
PCIe Integrated GFX
Memory
PCIe integrated GFX
SoC Interconnct & Root Complex
PCIe integrated Network
External PCIe Port(s) (wifi, Modem, Storage, etc)
Connectivity Port (s) (Enet/…)

(Highly Integrated Platform View)

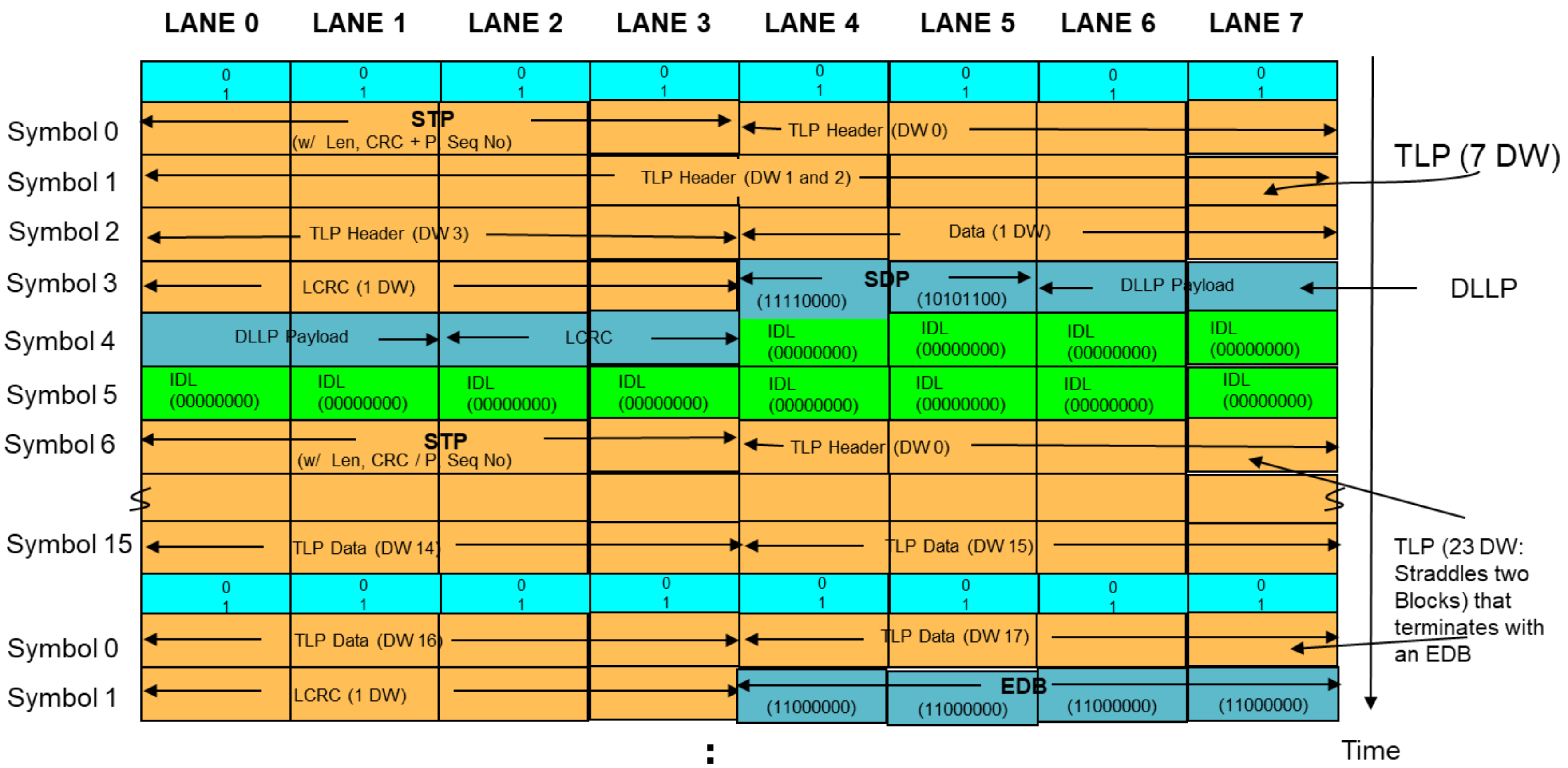
(e.g., Hand-held and thin client platforms)

Other events in 2010

- Winter Olympics in Vancouver
- Burj Khalifa opens
- Space-X: Dragon capsule returns – first successful private spacecraft

Photo by Vytautas Dranginis on Unsplash
Photo by SpaceX on Unsplash

128b/130b Encoding: x8 Example



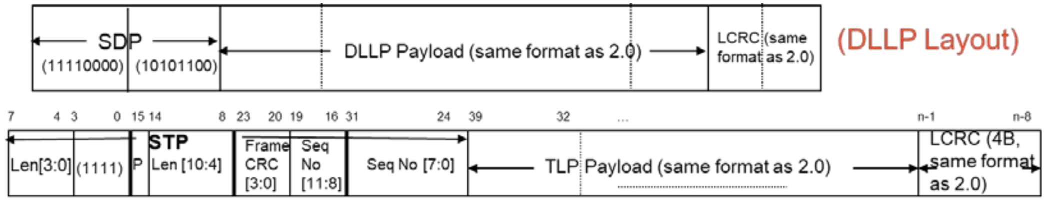
	LANE 0	LANE 1	LANE 2	LANE 3	LANE 4	LANE 5	LANE 6	LANE 7
	01	01	01	01	01	01	01	01
Symbol 0	STP (w/ Len, CRC + P Seq No)				TLP Header (DW 0)			
Symbol 1	TLP Header (DW 1 and 2)							
Symbol 2	TLP Header (DW 3)				Data (1 DW)			
Symbol 3	LCRC (1 DW)				SDP (11110000)	(10101100)	DLLP Payload	
Symbol 4	DLLP Payload		LCRC		IDL (00000000)	IDL (00000000)	IDL (00000000)	IDL (00000000)
Symbol 5	IDL (00000000)	IDL (00000000)	IDL (00000000)	IDL (00000000)	IDL (00000000)	IDL (00000000)	IDL (00000000)	IDL (00000000)
Symbol 6	STP (w/ Len, CRC / P Seq No)				TLP Header (DW 0)			
Symbol 15	TLP Data (DW 14)				TLP Data (DW 15)			
	01	01	01	01	01	01	01	01
Symbol 0	TLP Data (DW 16)				TLP Data (DW 17)			
Symbol 1	LCRC (1 DW)				EDB (11000000)	(11000000)	(11000000)	(11000000)

TLP (7 DW)

DLLP

TLP (23 DW: Straddles two Blocks) that terminates with an EDB

Time



SDP (11110000) (10101100) | DLLP Payload (same format as 2.0) | LCRC (same format as 2.0)

(DLLP Layout)

Len[3:0] | (1111) | P | STP Len [10:4] | Frame CRC [3:0] | Seq No [11:8] | Seq No [7:0] | TLP Payload (same format as 2.0) | LCRC (4B, same format as 2.0)

[Len[10:0]: length of the TLP in DWs, Frame CRC[4:0]: Check Bits covering Length[0:10], P: Frame Parity, No END]

(TLP Layout)

L1 Substates: PCI Express® Technology in Hand-Held

- Problem Statement: L1 developed for desktop/ server power source – consumed mWs when idle. For smart phone/tablet with battery source, idle power needed to be in uWatts
- Solution: L1 Low power substates for deep power savings with <10 uW power draw
- Approach: Float the differential pair vs. driving to common mode voltage, turn off PLLs and electrical idle detection circuitry, and leverage existing low-speed ClkReq for wakeup

	Port Circuit Power On/Off			Target Results*	
Sub-State	**PLL**	**Rx/Tx**	**Common - Mode Keepers**	**xa1 Port Power**	**Exit Latency**
L1 (unmodified)	ON	off/idle	ON	25mW	2µs (retrain)
L1+CLKREQ (unmodified)	off	off/idle	ON	10mW	20 µs (PLL)
L1.1	off	off	ON	300 µW	20 µs (PLL)
L1.2	off	off	off	10 µW	70 µs (Common mode restore + other delays)

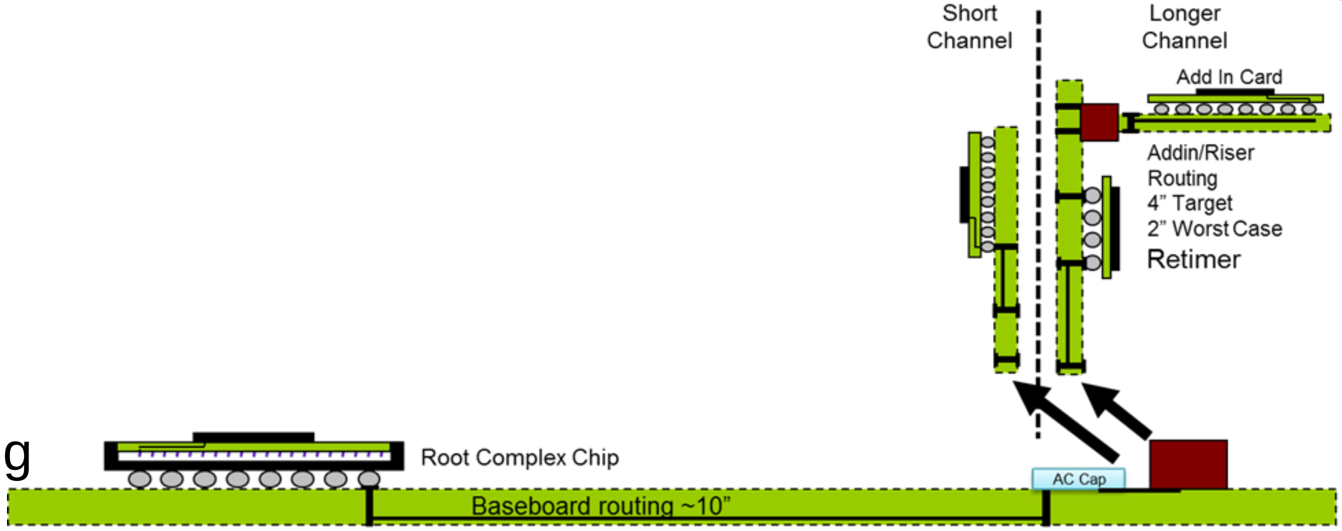
Solution: Turn circuits off

Note: Power savings will provide near linear scaling for multi-lane links.

*** These are targets for power and latency, not specified results.**

PCIe® 4.0 Specification in 2017

- Increased Lane Count w/ 8G while ecosystem develops enablers for 16G (and beyond)
  - Low-loss materials (Meg 2, 4, 6) in volume
  - Package and connector improvements
  - Improved platform volumetrics
  - Retimers for channels beyond 14", 1C
- Primarily a speed upgrade
- Protocol enhancements: performance scaling
- Common PHY for Load-Store I/O with its compelling area, latency, and power



Source: Intel Corporation

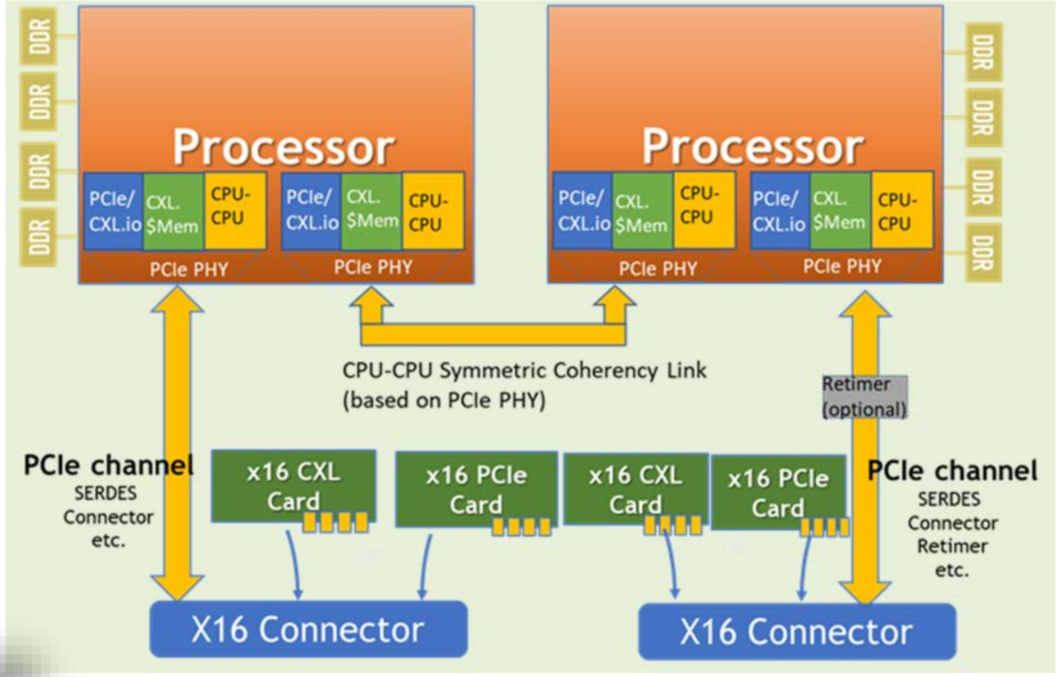
Other events in 2017

- Crypto-currencies go mainstream (Bitcoin grows 20X)
- Global growth picks up
- Brexit: Britain invokes Article 50
- Golden State warriors win NBA championship

Photo by André François McKenzie on Unsplash
Photo by Zeynep on Unsplash

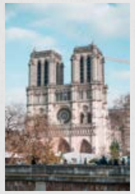
PCIe® 5.0 Specification in 2019

- 32G primarily a speed increase
  - Channel and component improvements continue
- PCIe PHY ubiquitous w/ best area, latency, power efficiency in the industry
- Alternate protocol support enables coherency and memory on PCIe PHY
  - PCIe PHY solving the memory bandwidth challenge as number of DDR channels becomes untenable in platforms
  - PCIe technology as a rack-level interconnect for resource pooling



Other events in 2019

- Fire at 850-year-old Notre-Dame Cathedral in Paris



- First all-woman spacewalk by NASA astronauts
- Covid-19 strikes!!

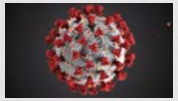
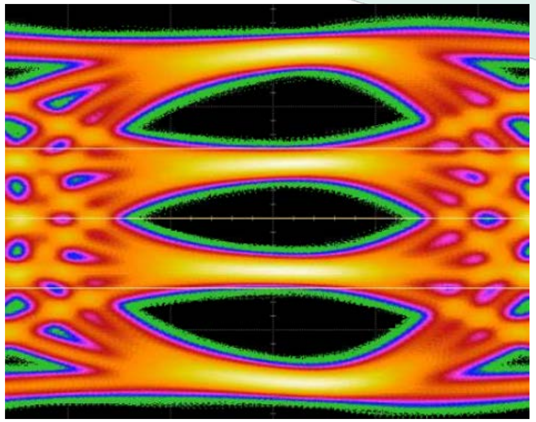


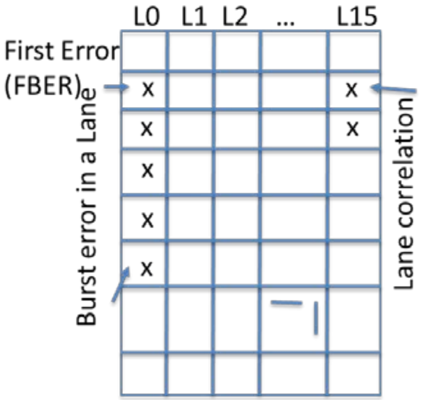
Photo by Bastien Nvs on Unsplash
Photo by CDC on Unsplash

PCIe® 6.0 Specification in 2022: Delivering power-efficient performance with PAM-4 signaling

Metrics	Requirements
Data Rate	64 GT/s, PAM4 (double the bandwidth per pin every generation)
Latency	<10ns adder for Transmitter + Receiver (including Forward Error Correct FEC) for PCIe (Ld/St can not afford the 100ns FEC latency of networking)
Bandwidth Inefficiency	<2 % adder over 32.0 GT/s across all payload sizes and protocols
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 32.0 GT/s. L0p: power proportionate to b/w consumed
Low Power	Similar entry/ exit latency for L1 low-power state
Others	HVM-ready, cost-effective, scalable to hundreds of Lanes in a platform, Fully backward-compatible



(PAM-4 Signaling: Helps Channel reach but increases errors)



(Higher Error rate + Correlated errors)

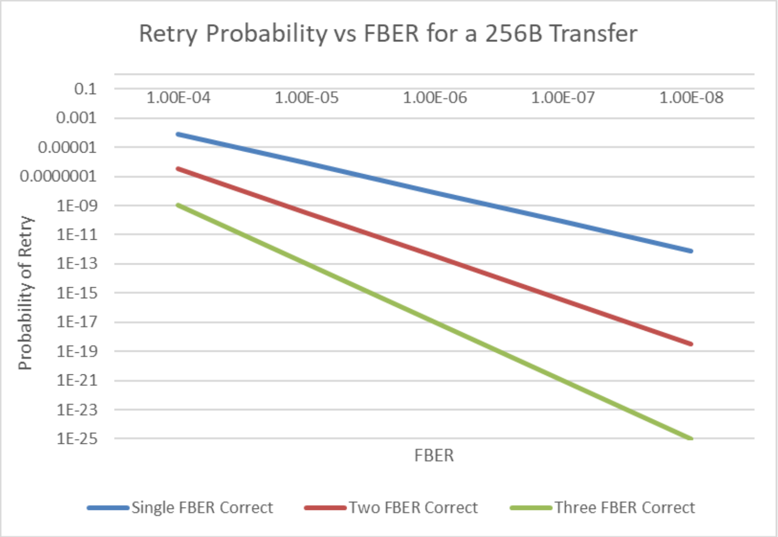
PAM-4 not new to the industry but the latency constraints require unique solutions

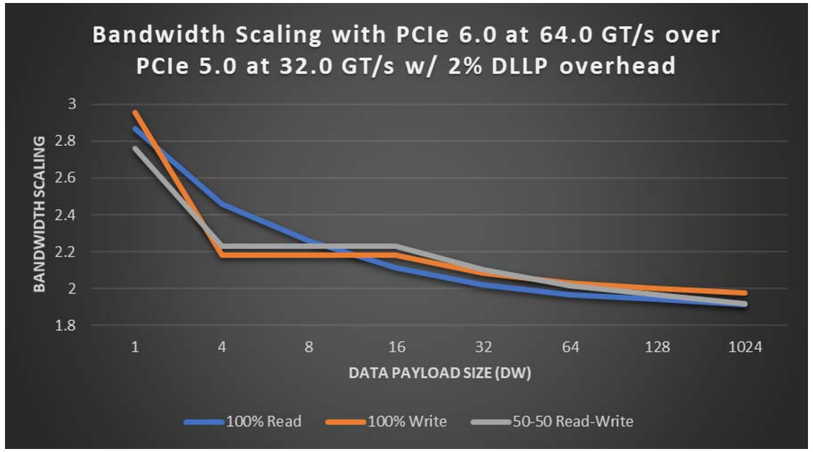
Golden State Warriors win the NBA Championship!
PCI-SIG celebrates 30-year anniversary in-person on June 21 as the ravages of Covid-19 pandemic subsides.

PCIe® 6.0 Specification Approach and Results

- Light-weight FEC & Link level replay
- 10^{-6} FBER w/ mitigations (constrained taps, precoding, Gray Coding)
- Spec defined mechanisms for low-latency replay and FEC
- 256 B Flit (Flow-Control Unit) mode
  - 236B for TLP, 6B for DLP
  - 8B CRC (strong CRC for low FIT)
  - 6B FEC (3-way FEC x 2B per FEC Group – single symbol correct for low-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 x 10^{-7}	4 x 10^{-7}	4 x 10^{-7}	4 x 10^{-4}





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, low-power, >2X bandwidth

PCIe 6.0 webinar: https://www.youtube.com/watch?v=jhehXwnu0Ss&feature=youtu.be

Conclusions and Call to Action

- Six Generations of doubling bandwidth w/ backwards compatibility – Impressive!
  - Keeping the latency flat while power efficiency improves generationally
- No signs of slowing down – PCI-SIG® has the expertise to continue to deliver
- PCIe® 7.0 specification has started – 128GT/s reusing same encoding as 64 GT/s!
- Need to look at protocol enhancements to deliver performance
- Need to comprehend fabric style multi-ported connectivity with high bisection bandwidth to deliver better performance and resource utilization across nodes

- The journey continues …
  - Consider joining PCI-SIG if you have not done so!

Q&A

Thank you for attending the PCI-SIG® Webinar 2022

For more information, please visit www.pcisig.com