Introduction to PCI Express

Paolo Durante
(CERN EP-LBC)
Where can you find PCI Express?

PCI (Peripheral Component Interconnect) Express is a popular standard for high-speed computer expansion overseen by PCI-SIG (Special Interest Group).

• PCIe interconnects can be present at all levels of your DAQ chain...
  • Readout boards
  • Storage media
  • Network interfaces
  • Compute accelerators (GPUs, FPGAs...)

• ...and may be even more so in the future

• Understanding your data acquisition system requires (some) level of understanding of PCI Express
What is this presentation about?

- History and evolution of PCIe
- PCIe concepts
- PCIe layers
- PCIe performance
- PCIe future roadmap
18/01/2019

ISOTDAQ 2020 - Introduction to PCIe

4

PCI ("conventional PCI")

- 1992
- **Peripheral Component Interconnect**
- Parallel Interface
- Bandwidth
  - 133 MB/s (~1.0 Gb/s) (32-bit@33 MHz)
  - 533 MB/s (~4.2 Gb/s) (64-bit@66 MHz)
- Plug-and-Play configuration (BARs)
PCI example: ATLAS FILAR

- ~2003
- 4 optical channels
  - 160 MB/s (1.28 Gb/s)
- S-LINK protocol
  - 2 Altera FPGAs
- Burst-DMA over PCI
  - 3rd Altera FPGA
- 64-bit@66MHz PCI
PCI-X ("Extended PCI")

- 1998
- PCI compatible
  - Hardware and software
  - Half-duplex bidirectional
- Higher bus efficiency
  - Split-responses
  - Message Signaled Interrupts
- Bandwidth
  - \( \leq 1066 \text{ MB/s} \) (~8.5 Gb/s) (64-bit@133 MHz)
  - 2133 MB/s (~17 Gb/s) (PCI-X 266)
  - 4266 MB/s (~34 Gb/s) (PCI-X 533)
PCI-X example: CMS FEROL

• ~2011
• 4 SFP+ cages
  • 1x 10 Gb/s Ethernet
  • 3x SlinkXpress
• PCI-X interface to legacy FE (Slink64)
• Altera FPGA
• Simplex TCP-IP
PCI Express (PCIe)

• 2004
• PCI “inspired”
  • software, topology
• Serial interface
• Full-duplex bidirectional
• Bandwidth
  • x1: ≤2 GB/s (16 Gb/s) (in each direction)
  • x16: ≤32 GB/s (256 Gb/s) (in each direction)
• Still evolving
  • 1.0, 2.0, 3.0, 4.0, 5.0, 6.0...
PCle example: ALICE C-RORC

• ~2014
• 3x QSFP
  • 36 channels
  • up to 6.6Gb/s/channel
• 2x DDR SO-DIMM
• XilinX FPGA
• PCle Gen2 x8

• Also used by ATLAS
PCIe example: LHCb TELL40

- Introduced for LHC Run3
- Currently in production
- ≤ 48 duplex optical links
  - GBT (3.2 Gb/s)
  - WideBus (4.48 Gb/s)
  - GWT (5.12 Gb/s)
- Altera Arria10 FPGA
- 110 Gb/s DMA
- PCIe 3.0 x16
- Also used by ALICE
PCIe example: ATLAS FELIX

- Introduced for LHC Run3
- \( \leq 48 \) duplex optical links
- Xilinx Ultrascale FPGA
- 2x DDR4 SO-DIMM
- PCIe 3.0 x16
- Wupper DMA (Open Source)
- Also used by DUNE
What is this presentation about?

- History and evolution of PCIe

- **PCIe concepts**

- PCIe layers

- PCIe performance

- PCIe future roadmap
PCIe concepts – Packets

• Point-to-point connection
• “Serial” “bus” (fewer pins)
• Scalable link: x1, x2, x4, x8, x12, x16, x32
• Packet encapsulation
PCle concepts – Root complex

• Connects the processor and memory subsystems to the PCle fabric via a *Root Port*
• Generates and processes transactions with *Endpoints* on behalf of the processor
PCle concepts – Topology

Relative to root – up is towards, down is away

“UPSTREAM”

“DOWNSTREAM”

CPU

Root complex

Memory

PCle bridge to PCI/PCI-X

PCI/PCI-X

PCle endpoint

Switch

PCle endpoint

Legacy endpoint

PCle endpoint
PCle concepts – BDF

“geographical addressing”

- **Bus : Device . Function**
  - Form a hierarchy-based address
  - Multiple logical “Functions” allowed on one physical device
  - Bridges (PCI/PCI-X) form hierarchy
  - Switches (PCle) form hierarchy

On linux: $ man lspci
Troubleshooting with lspci

• Device works but is “slow”
  • Link speed
  • Link width
  • MaxPayloadSize
  • Interrupts
  • Error flags
  • Look for bottlenecks upstream

• Device is “there” but driver fails to load
  • Unreadable config space
  • Unallocated BARs
# PCIe concepts – Address spaces

- **Address spaces**
  - Configuration (Bus/Device/Function)
  - Memory (64-bit)
  - I/O (32-bit)

- **Configuration space**
  - Base Address Registers (BARs) (32/64-bit)
  - Capabilities (linked list)

---

<table>
<thead>
<tr>
<th>Address Field</th>
<th>Subfield 1</th>
<th>Subfield 2</th>
<th>Subfield 3</th>
<th>Subfield 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Device ID</td>
<td>Vendor ID</td>
<td>Command</td>
<td>Revision ID</td>
<td></td>
</tr>
<tr>
<td>Status</td>
<td>Class Code</td>
<td>Latency Timer</td>
<td>Cache Line Size</td>
<td></td>
</tr>
<tr>
<td>BIST</td>
<td>Header Type</td>
<td>Secondary Bus Number</td>
<td>Primary Bus Number</td>
<td></td>
</tr>
<tr>
<td>Base Address Registers (BAR) 0</td>
<td>Base Address Registers (BAR) 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Secondary Longet Timer</td>
<td>Subordinate Bus Number</td>
<td>Secondary Bus Number</td>
<td>Primary Bus Number</td>
<td></td>
</tr>
<tr>
<td>Secondary Status</td>
<td>I/O Limit</td>
<td>I/O Base</td>
<td>Memory Limit</td>
<td>Memory Base</td>
</tr>
<tr>
<td>Memory Limit</td>
<td>Prefetchable Memory Limit</td>
<td>Prefetchable Memory Base</td>
<td>Prefetchable Base Upper 32 Bits</td>
<td>Prefetchable Base Limit 32 Bits</td>
</tr>
<tr>
<td>I/O Limit Upper 16 Bits</td>
<td>I/O Base Upper 16 Bits</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reserved</td>
<td>Capabilities Pointer</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expansion ROM Base Address Register (XROMBAR)</td>
<td>Bridge Control</td>
<td>Interrupt Pin</td>
<td>Interrupt Line</td>
<td></td>
</tr>
</tbody>
</table>

---

---

18/01/2019

ISOTDAQ 2020 - Introduction to PCIe
PCIe concepts – Memory & I/O

• Memory space maps cleanly to CPU semantics
  • 32-bits of address space initially
  • 64-bits introduced via Dual-Address Cycles (DAC)
    • Extra period of address time on PCI/PCI-X
    • 4DWORD header in PCI Express
  • Burstable (= Multiple DWORDs)

• I/O space maps cleanly to CPU semantics
  • 32-bits of address space
  • Non-burstable
PCle concepts - Bridges

Transparent
- Single root (or SR-IOV)
- Single address space
- Multiple downstreams (switch)
- Downstreams appear in the same topology
- Addresses are passed through unchanged

Non-Transparent
- Joins two independent topologies
- One root on each side
- Each side has its own address space
- Needs translation table
- Fault tolerance, “networking”, HPC
PCIe concepts – Bus address

This is actually not specific to PCIe, but a generic reminder:

• **Physical address**: the address the CPU sends to the memory controller

• **Virtual address**: an indirect address created by the operating system, translated by the CPU to physical

• **Bus address**: an address understood by the devices connected to a specific bus

• On Linux, see: `pci_iomap()`, `remap_pfn_range()`, ...
PCle concepts – Interrupts

• PCI
  • INTx#
    • $x \in \{A, B, C, D\}$
  • Level sensitive
  • Can be mapped to CPU interrupt number

• PCle
  • “Virtual Wire” emulation
  • Assert_INTx code
  • Deassert_INTx code

```c
pci_read_config_byte(dev, PCI_INTERRUPT_PIN, &(...));
pci_read_config_byte(dev, PCI_INTERRUPT_LINE, &(...));
p pci_enable_msi(dev);
request_irq(dev->irq, my_isr, IRQF_SHARED, devname, cookie);
```
PCIe concepts – MSI & MSI-X

- Based on messages ($MWr$)
- **MSI** uses one address with a variable data value indicating which “vector” is asserting
  - $\leq 32$ per device (in theory)
- **MSI-X** uses a table of independent address and data pairs for each “vector”
  - $\leq 2048$ per device (use affinity!)
- **Vector**: interrupt id
PCle concepts – interr. coalescing
PCIe concepts – latency

Typical: ~1us
PCIe Gen1 (2003)

- Introduced at 2.5 GT/sec
- Also called 2.5 GHz, 2.5 Gb/s
- 100 MHz reference clock
  - Eases synchronization between ends
  - Can use Spread Spectrum Clocking to reduce EMI
  - Optional, but nearly universal
- 8b/10b encoding used to provide DC balance and reduce “runs” of 0s or 1s which make clock recovery difficult
- Specification Revisions: 1.0, 1.0a, 1.1
PCIe Gen2 (2006)

- Speed **doubled** from 2.5 to 5 GT/sec
- Reference clock remains at 100 MHz
  - Lower jitter clock sources required vs 2.5 GT/sec
  - Generally higher quality clock generation/distribution required
- 8b/10b encoding continues to be used
- Specification Revisions: 2.0, 2.1
- Devices choosing to implement a maximum rate of 2.5 GT/sec can still be fully 2.x compliant
PCle Gen3 (2010)

2 \times 5 = ?
PCle Gen3 (2010)

2 x 5 = 8

• Speed “doubled” from 5 GT/sec
• More efficient encoding (20% → ~1%)
• 8 GT/sec electrical rate
  • 10 GT/sec required significant cost and complexity in channel, receiver design, etc.
• Reference clock remains at 100 MHz
• Backwards-compatible speed negotiation
PCle Gen4 (2017)

2 x 8 = ?
PCIe Gen4 (2017)

2 x 8 = 16

- Speed doubled from 8 GT/sec
- Same 128b/130b encoding
- 16 GT/sec electrical rate
  - Channel length: ≤ 10”/14”
  - Retimer mandatory for longer channels
  - More complex pre-amplification, equalization stages
- Reference clock remains at 100 MHz
- Backwards-compatible protocol negotiation and CEM spec
PCle Gen5 (2019)
What is this presentation about?

• History and evolution of PCIe

• PCIe concepts

• **PCIe layers**

• PCIe performance

• PCIe future roadmap
PCIe – Protocol stack

PCI Express Device A

- Application Layer
- PCI Express Logic Interface
- Transaction Layer
- Data Link Layer
- Physical Layer

PCI Express Device B

- Application Layer
- PCI Express Logic Interface
- Transaction Layer
- Data Link Layer
- Physical Layer

Link

TX RX

TX RX
FPGA Hardened PCIe IP

[Diagram of PCIe IP and circuitry]

18/01/2019
ISOTDAQ 2020 - Introduction to PCIe
PCIe – Transaction layer

• Four possible transaction types
  • **Memory Read | Memory Write**
    • Transfer data from or to a memory mapped location
    • Address routing
  • **IO Read | IO Write**
    • Transfer data from or to an IO location (on a legacy endpoint)
    • Address routing
  • **Config Read | Config Write**
    • Discover device capabilities, status, parameters
    • ID routing (BDF)
  • **Messages**
    • Event signaling
PCle – TLP structure

- **Transmit order**
  - STP: 1B
  - Sequence: 2B
  - Header: 3-4DW
  - Data Payload: 0-1024DW
  - ECRC: 1DW
  - LCRC: 1DW
  - End: 1B

**Application Layer**
- Created by Transaction Layer

**MaxPayloadSize (MPS)**
- Parameter limits and dominates performance

**Appendeds**
- Appended by Data Link Layer
- Appended by Physical Layer
PCIe – Split transaction model

• Posted transaction
  • Single TLP, no completion

• Non-posted transaction
  • Split transaction model
    • Requester initiates transaction (Requester ID + Tag)
      • Requester and Completer IDs encode the sender BDF
    • Completer executes transaction internally
    • Completer creates completion transaction (Cpl/CplD)

• Bus efficiency of Read is different (lower) wrt Write
  • Writes are posted while Reads are not
### PCIe – DMA transaction

**Completer:**
- Step 2: Root Complex (completer) receives MRd
- Step 3: Root Complex returns Completion with data (CplD)

**Requester:**
- Step 1: Endpoint (requester) initiates Memory Read Request (MRd)
- Step 4: Endpoint receives CplD
PCIe – Data Link Layer

- ACK / NAK Packets
  - Error handling mechanism
- Flow Control Packets (FCPs)
  - Propagate credit allocation status
- Power Management Packets
- Vendor extensions
  - E.g.: CAPI, CCIX (memory coherency)
PCle – DLLP structure

Transmit order

SDP | DLLP | CRC | End
---|---|---|---
1B | 4B | 2B | 1B

Created by Data Link Layer

Appended by Physical Layer
PCle – Flow control

- Credit-based
- Point-to-point (not end-to-end)
PCIe – RAS/QoS features

• Data Integrity and Error Handling
  • PCIe is RAS (Reliable, Available, Serviceable)
  • Data integrity at
    • link level (LCRC)
    • end-to-end (ECRC, optional)

• Virtual channels (VCs) and traffic classes (TCs) to support differentiated traffic or Quality of Service (QoS)

• In theory
  • Ability to define levels of service for packets of different TCs
  • 8 TCs and 8 VCs available

• In practice
  • Rarely more than 1 VC and 1 TC are implemented
PCIe – Error handling

Correctable
• Recovery happens automatically in DLL
• Performance is degraded
• Example: LCRC error → automatic DLL retry (there is no forward error correction)

Uncorrectable
• Fatal
  • Platform-specific handling
• Non-fatal
  • Can be exposed to application layer and handled explicitly
  • Can and do cause system deadlock / reset
• Recovery mechanisms are outside the spec
  • Example: failover for HA
PCIe – ACK/NAK
PCIe – Physical layer

PCI Express Device

Signal

Wire

Link

Lane

LVDS

$V_{OH} = 1.4 \text{ V}$

$V_{CM} = 1.2 \text{ V}$

$V_{OH} = 1 \text{ V}$

400 mV
PCle – Ordered-Set Structure

Six ordered sets are possible
- Training Sequences (TS1, TS2): 1 COM + 15 TS
  - Used to de-skew between lanes
- SKIP: 1 COM + 3 SKP identifiers
  - Used to recalibrate receiver clock
- Fast Training Sequence (FTS): 1 COM + 3 FTS
  - Power management
- Electrical Idle (IDLE): 1 COM + 3 IDL
  - Transmitted continuously when no data
- Electrical Idle Exit (EIEOS): 16 characters (since 2.0)

*character*: 8 unscrambled bits
PCle – Link training

- Lane polarity
- Link width / ordering
- Link equalization
  - Dynamic equalization!
- Link speed
- ...

37181 ns EP LTSSM State: RECOVERY.RCVRLOCK
37312 ns RP PCI Express Link Status Register (1881):
37312 ns  Negotiated Link Width: x8
37312 ns    Slot Clock Config: System Reference Clock Used
37949 ns EP LTSSM State: RECOVERY.RCVRLOCK
38845 ns RP LTSSM State: RECOVERY.RCVRLOCK
41053 ns RP LTSSM State: RECOVERY.SPEED
41309 ns EP LTSSM State: RECOVERY.SPEED
43573 ns EP LTSSM State: RECOVERY.RCVRLOCK
43765 ns RP LTSSM State: RECOVERY.RCVRLOCK
43797 ns RP LTSSM State: REC_EQULZ.PHASE0
43825 ns RP LTSSM State: REC_EQULZ.PHASE1
44141 ns EP LTSSM State: REC_EQULZ.PHASE0
44673 ns EP LTSSM State: REC_EQULZ.PHASE1
44929 ns RP LTSSM State: REC_EQULZ.DONE
44949 ns RP LTSSM State: RECOVERY.RCVRLOCK
45209 ns EP LTSSM State: REC_EQULZ.DONE
45229 ns EP LTSSM State: RECOVERY.RCVRLOCK
45425 ns EP LTSSM State: RECOVERY.RCVRCFG
45581 ns RP LTSSM State: RECOVERY.RCVRCFG
45925 ns RP LTSSM State: RECOVERY.IDLE
46073 ns EP LTSSM State: RECOVERY.IDLE
46169 ns EP LTSSM State: L0
46313 ns RP LTSSM State: L0
47824 ns  Current Link Speed: 8.0GT/s
• L0: active
• L0 standby, L1: lower power, higher latency
• L2: cold standby, even lower power
• L3: power off
PCle – Framing (x1)

Transmit order (TIME)

Reserved bits
Sequence Number
(Data Link Layer)

LCRC
(Data Link Layer)

STP Framing Symbol
(Physical Layer)

TLP structure
(Transaction Layer)

END Framing Symbol
(Physical Layer)
PCle – Framing (x4)

Transmit order (TIME)

Lane 0  Lane 1  Lane 2  Lane 3

...   ...   ...   ...

STP

...   ...   ...   ...

Lane order (SPACE)

(Lane-reversal possible)

...   ...   ...   ...

END

Physical Layer

Data Link Layer

Transaction Layer
PCle troubleshooting
Signal integrity – Environment
### Table 4-16: Tx Preset Ratios and Corresponding Coefficient Values

<table>
<thead>
<tr>
<th>Preset Number</th>
<th>Preshoot (dB)</th>
<th>De-emphasis (dB)</th>
<th>c₁</th>
<th>c₂⁺</th>
<th>Va/Vd</th>
<th>Vb/Vd</th>
<th>Vc/Vd</th>
</tr>
</thead>
<tbody>
<tr>
<td>P4</td>
<td>0.0</td>
<td>0.0</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>P1</td>
<td>0.0</td>
<td>-3.5 ± 1 dB</td>
<td>0.000</td>
<td>-0.167</td>
<td>1.000</td>
<td>0.668</td>
<td>0.668</td>
</tr>
<tr>
<td>P0</td>
<td>0.0</td>
<td>-6.0 ± 1.5 dB</td>
<td>0.000</td>
<td>-0.250</td>
<td>1.000</td>
<td>0.500</td>
<td>0.500</td>
</tr>
<tr>
<td>P9</td>
<td>3.5 ± 1 dB</td>
<td>0.0</td>
<td>-0.166</td>
<td>0.000</td>
<td>0.668</td>
<td>0.668</td>
<td>1.000</td>
</tr>
<tr>
<td>P8</td>
<td>3.5 ± 1 dB</td>
<td>-3.5 ± 1 dB</td>
<td>-0.125</td>
<td>-0.125</td>
<td>0.750</td>
<td>0.500</td>
<td>0.750</td>
</tr>
<tr>
<td>P7</td>
<td>3.5 ± 1 dB</td>
<td>-6.0 ± 1.5 dB</td>
<td>-0.100</td>
<td>-0.200</td>
<td>0.800</td>
<td>0.400</td>
<td>0.600</td>
</tr>
<tr>
<td>P5</td>
<td>1.9 ± 1 dB</td>
<td>0.0</td>
<td>-0.100</td>
<td>0.000</td>
<td>0.800</td>
<td>0.800</td>
<td>1.000</td>
</tr>
<tr>
<td>P6</td>
<td>2.5 ± 1 dB</td>
<td>0.0</td>
<td>-0.125</td>
<td>0.000</td>
<td>0.750</td>
<td>0.750</td>
<td>1.000</td>
</tr>
<tr>
<td>P3</td>
<td>0.0</td>
<td>-2.5 ± 1 dB</td>
<td>0.000</td>
<td>-0.125</td>
<td>1.000</td>
<td>0.750</td>
<td>0.750</td>
</tr>
<tr>
<td>P2</td>
<td>0.0</td>
<td>-4.4 ± 1.5 dB</td>
<td>0.000</td>
<td>-0.200</td>
<td>1.000</td>
<td>0.600</td>
<td>0.600</td>
</tr>
<tr>
<td>P10</td>
<td>0.0</td>
<td>See Note 2.</td>
<td>0.000</td>
<td>See Note 2.</td>
<td>1.000</td>
<td>See Note 2.</td>
<td>See Note 2.</td>
</tr>
</tbody>
</table>

**Notes:**

1. Reduced swing signaling must implement presets #4, #1, #9, #5, #6, and #3. Full swing signaling must implement all the above presets.
2. P10 boost limits are not fixed, since its de-emphasis level is a function of the LF level that the Tx advertises during training. The allowable P10 boost range is defined by the coefficient space lying between the two diagonal lines in Figure 4-45. This approach permits both full and reduced swing transmitters to use P10 for testing to their respective boost limits.
PCIe CEM Spec – AIC form factors

- **Standard Height**
  - 4.20” (106.7mm)
- **Low Profile**
  - 2.536” (64.4mm)
- **Half Length** (e.g. “HHHL”)
  - 6.6” (167.65mm)
- **Full Length** (e.g. “FHFL”)
  - 12.283” (312mm)

**Power**: up to 10W, 25W, 75W, 300W or 375W depending on form factor & optional extra power connectors
PCIe storage – More form factors

M.2 ≤ 4 lanes

U.2 ≤ 4 lanes

“ruler” (EDSFF, NGSFF) ≤ 8 lanes
PCle CEM Spec – Power Cables

- **EPS receptacle**
- **PCle cable**
- **GPU power**

**PCle 8 Pin**
- Gnd
- +12 V
- Sense A
- Sense B

**PCle 6 Pin**
- 6
- 3

**PCle 12V**
- 8

**EPS-12V**
- 8

Sense A and B are used by a compatible power supply to provide enhanced voltage regulation.

If enhanced regulation is not supported then Sense A can be connected to Ground. Sense B can be left unconnected (or connected to ground).

Connection of ground to Pin 8 allows the card to detect an 8 pin connector and select enhanced power mode.
What is this presentation about?

• History and evolution of PCIe

• PCIe concepts

• PCIe layers

• **PCIe performance**

• PCIe future roadmap
PCle – Theoretical data rates

- “Aggregate” bandwidth in both directions
- Considering 20% encoding overhead in 1.x and 2.x
PCIe – Effective data rates

\[ \rho = \frac{\text{Lane rate} \times \text{Lane width}}{\text{Encoding}} \times \frac{\text{MPS}}{\text{MPS+Headers}} \]

- Example: Gen2 x8, 128 Bytes MPS
  \[ \rho = 40 \times 0.8 \times \frac{128}{128+24} = 32 \times 0.84 = 26.9 \text{ Gb/s} \]

- Example: Gen3 x8, 128 Bytes MPS
  \[ \rho = 64 \times 0.98 \times \frac{128}{128+24} = 62.7 \times 0.84 = 52.6 \text{ Gb/s} \]

- Example: Gen3 x8, 256 Bytes MPS
  \[ \rho = 64 \times 0.98 \times \frac{256}{256+24} = 62.7 \times 0.91 = 57 \text{ Gb/s} \]
PCIe 3.0 x8 – DMA Performance

MPS = 128 Bytes

MPS = 256 Bytes

10%
What is this presentation about?

• History and evolution of PCIe

• PCIe concepts

• PCIe layers

• PCIe performance

• PCIe future roadmap
PCIe scalability – As of 2020

- Intel Xeon Broadwell
  - PCIe Gen3 x 40 lanes
- Intel Xeon Skylake / Cascade Lake
  - PCIe Gen3 x 48 lanes
- IBM Power 8
  - PCIe Gen3 x 48 lanes
- AMD Epyc Threadripper
  - PCIe Gen3 x 64 lanes
- AMD Epyc Naples
  - PCIe Gen3 x 128 lanes (!!)
- IBM Power 9
  - PCIe Gen3 x 64 lanes
  - PCIe Gen4 x 64 lanes
- AMD Epyc Rome / Ryzen (Zen2)
  - PCIe Gen4 x 128-160 lanes (!!)

- Counting Northbridge lanes only
- Excludes additional lanes from Southbridge
- More density possible using switches
  - Shared bandwidth

- Mostly driven by storage market (dense NVMe)
  ➢ Plus OpenCAPI & NVLINK lanes
  ➢ Depending on XGMII setup
PCle Gen4 – On paper

Mellanox and Synopsys Demonstrate Industry’s First PCle 4.0 Interoperability

Mutual Technology Leadership Lowers Risk for Designers Implementing 16GT/s PCI Express Protocol

PCI-SIG Developer Conference, Tel Aviv, Israel – March 2, 2015 – Mellanox® Technologies, Ltd. (NASDAQ: MLNX), a leading supplier of end-to-end interconnect solutions for servers and storage systems, today announced that it has collaborated with Synopsys to bring the industry’s first demonstration of interoperability between Synopsys’ DesignWare® PHY IP for PCI Express® (PCle®) 4.0 and Mellanox's PCle

PCle 4.0 Will Arrive in 2017

BY MATTHEW MURRAY

AUGUST 19, 2016 10:34AM EST  4 COMMENTS

PCle 4.0 will double interconnect performance bandwidth and be better poised for use in mobile and IoT applications.

[PCI-SIG] PCI Express Base Specification Revision 4.0, Version 1.0

Dear PCI-SIG® Member,

We’d like to announce the release of the PCI Express® Base Specification Revision 4.0, Version 1.0. This specification describes the PCI Express architecture, interconnect attributes, fabric management, and the programming interface required to design and build systems and peripherals that are compliant with the PCI Express Specification.
PCIe Gen4 – On silicon (IBM)

IBM Power AC922

- 2 POWER9 Processors
  - 190, 250W modules
- 4-6 NVidia “Volta” GPU’s
  - 300W, SXM2 Form Factor, NVLink 2.0
- 6 GPU configuration, water cooled
- 4 GPU configuration, air or water cooled

- 2 Gen4 x16 HHHL PCIe, CAPI enabled
- 1 Gen4 x4 HHHL PCIe
- 1 Gen4 Shared x8 PCIe adapter
- 16 IS DIMM’s
  - 8, 16, 32, 64, 128GB DIMMs
- 2 SATA SFF HDD / SSD
- 2 2200W power supplies
  - 200 VAC, 277VAC, 400VDC input
  - N+1 Redundant
- Second generation BMC Support Structure
- Pluggable NVMe storage adapter option
PCIe Gen4 – On silicon (Mellanox)

Mellanox ConnectX®-5

- Dual-Port 100 Gbit/s

Mellanox ConnectX®-6

- Single-Port 200 Gbit/s

LnkCap: Port #0, Speed 16GT/s, Width x16, ASPM L0s L1
PCle Gen4 – On silicon (AMD)

- Zen 2 (EPYC™ Rome) architecture
PCIe Gen4 – On silicon (Xilinx)

- Xilinx Virtex Ultrascale+™ VU37P
- Bifurcated (2x) Gen4x8 PCIe, ~208 Gbps

PCle Gen4 – On silicon (Intel)

- No Gen4 CPU until IceLake (2020?)
- No Gen4 FPGA until AgileX (2020?)
- Gen4 SSDs starting to sample now

[PCI-SIG] PCI Express Base Specification Revision 5.0, Version 0.7 for 30 Day Member Review

PCI-SIG Administration <administration@pcisig.com> Thu, May 3, 2018, 5:44 AM

to PCI-SIG

Dear PCI-SIG® Member,

The 30 day member review period for the PCI Express Base Specification Revision 5.0, Version 0.7 will end at 5 p.m. (PT) on Friday, June 1, 2018.

**PCI Express® Base Specification Revision 5.0, Version 0.7**
This specification describes the PCI Express architecture, interconnect attributes, fabric management, and the programming interface required to design and build systems and peripherals that are compliant with the PCI Express Specification.
PCIe Gen5 + coherency = CXL

(Compute eXpress Link)
Announced by Intel in March 2019
Conclusions

• PCI Express interconnects can be present at all levels of your DAQ chain...

• ...and may be even more so in the future
  • Lindy Effect: “The longer a technology has been around, the longer it's likely to stay around.”

• Our DAQ systems keep increasing their bandwidth requirements
  • PCIe has a track record of delivering “2x” improvements
  • It is up to us to implement and interconnect our systems in ways that actually deliver this performance