Introduction to PCI Express

Paolo Durante
(CERN EP-LBC)
Where will 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)

- Monitoring & Readout
  - Several DAQ board at LHC and elsewhere (examples next..)

- Networking
  - Ethernet (NIC), Infiniband (HCA), Omni-Path (HFI)

- Storage
  - NVMe is the standard interface for high-end Solid-State Disks

- Computation
  - Majority of GPGPUs, CAPI & CCIX (Memory Coherency on top of PCIe)
What is this presentation about?

- **History and evolution of PCIe**
  - PCIe concepts
  - PCIe layers
  - PCIe performance
- PCIe in practice
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
- Bandwidth
  - $\leq 1066$ MB/s ($\sim 8.5$ Gb/s) (64-bit@133 MHz)
  - 2133 MB/s ($\sim 17$ Gb/s) (PCI-X 266)
  - 4266 MB/s ($\sim 34$ 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: ≤1000 MB/s (8 Gb/s) (in each direction)
  - x16: ≤16000 MB/s (128 Gb/s) (in each direction)
- Still evolving
  - 1.0, 2.0, 3.0, 4.0...
PCIe example: ALICE C-RORC

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

- Also used by ATLAS
PCIe example: LHCb TELL40

- 2015
- ≤ 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

- 2016
- \( \leq 48 \) duplex optical links
- Xilinx Ultrascale FPGA
- 2x DDR4 SO-DIMM
- PCIe 3.0 x16
- Wupper DMA (Open Source!)
DAQ using PCs? Why not ATCA or μTCA or ...?

- Depends on your requirements!

- Advantages of commodity hardware:
  - Data to the CPU (or the accelerator, network) in one hop
  - Economies of scale, less highly specialized equipment
  - Exploit HPC-grade network technologies beyond Ethernet (fundamental at high data rates)
    - InfiniBand, Omni-Path...

- Disadvantages:
  - PCs are not made to hold precision instrumentation
    - Cooling / power / mechanics...
  - Lower lifetime for most commodity hardware
What is this presentation about?

• History and evolution of PCIe

• **PCIe concepts**

• PCIe layers

• PCIe performance

• PCIe in practice
PCIe concepts – Packets

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

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

Diagram:
- **CPU**
- **Root complex**
- **PCle endpoint**
- **Switch**
- **PCle endpoint**
- **PCle bridge to PCI/PCI-X**
- **Legacy endpoint**
- **Memory**
- **PCI/PCI-X**
Example: PCIe switch

“UPSTREAM” virtual bridge

“DOWNSTREAM” virtual bridge
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
Understanding `lspci --tv` (1/3)

Things in [...] are BUS NUMBERS
A.B means DEVICE.FUNCTION
Exercise: get the BDF of each end-point from the output above
(see next slide for how switches are represented)
Understanding `lspci --tv` (2/3)

```
- [0000:00]-- 00.0  Intel Corporation Xeon E7 v3/Xeon E5 v3/Core i7 DMI2
  +-- 01.0-[01]--
  +-- 02.0-[02-07]----00.0-[03-07]---08.0-[04]----00.0  CERN/ECP/EDU Device ce40
  |                     +-- 09.0-[05]----00.0  CERN/ECP/EDU Device ce40
  |                     +-- 10.0-[06]--
  |                             \-11.0-[07]--

-[BUS RANGE (seen upstream)]--UPSTREAM PORT-...
...-[BUS RANGE (seen downstream)]-
```

00:02.0 PCI bridge: Intel Corporation Xeon E7 v3/Xeon E5 v3/Core i7 PCI Express Root Port 3 (rev 02)
02:00.0 PCI bridge: PLX Technology, Inc. PEX 8747 48-Lane, 5-Port PCI Express Gen 3 (8.0 GT/s) Switch (rev ca)
03:08.0 PCI bridge: PLX Technology, Inc. PEX 8747 48-Lane, 5-Port PCI Express Gen 3 (8.0 GT/s) Switch (rev ca)
03:09.0 PCI bridge: PLX Technology, Inc. PEX 8747 48-Lane, 5-Port PCI Express Gen 3 (8.0 GT/s) Switch (rev ca)
...04:00.0 Communication controller: CERN/ECP/EDU Device ce40 (rev 01)
05:00.0 Communication controller: CERN/ECP/EDU Device ce40 (rev 01)
Understanding `lspci -tv` (3/3)

- ROOT PORT (00:02.0)
  - UPSTREAM PORT (02:00.0)
    - SWITCH [03]
      - DOWNSTREAM PORT (03:08.0)
      - DOWNSTREAM PORT (03:09.0)
      - UNUSED DOWNSTREAM PORTS (03:10.0) (03:11.0)
      - ENDPOINT (4:00.0)
      - ENDPOINT (5:00.0)
    -UNUSED DOWNSTREAM PORTS (03:10.0) (03:11.0)
    - UNCONNECTED ENDPOINTS (06:??.?)(07:??.?)

14/02/2017
ISOTDAQ 2018 - Introduction to PCIe
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
PCle 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)
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
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: the address of memory as seen by other devices, not the CPU
• On Linux, see: *pci_iomap()* , *remap_pfn_range()* , ...
PCIe concepts – Interrupts

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

• PCIe
  • “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, &(...));
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
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 (2007)

• 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
PCIe Gen3

2 \times 5 = ?
PCIe Gen3

\[2 \times 5 = 8\]

- 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
PCIe Gen4

2 x 8 = ?
PCIe Gen4

$2 \times 8 = 16$

- Speed doubled from 8 GT/sec (Gen5 likely similar)
- 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
What is this presentation about?

- History and evolution of PCIe
- PCIe concepts
- PCIe layers
- PCIe performance
- PCIe in practice
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
PCle – 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
- Appended by Data Link Layer
- Appended by Physical Layer

**MaxPayloadSize (MPS)**

- Parameter limits and dominates performance
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 (CplID)

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)
PCIe – DLLP structure

Transmit order

SDP  DLLP  CRC  End

1B  4B  2B  1B

Created by Data Link Layer

Appended by Physical Layer
PCIe – 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
PCle – ACK/NAK
PCIe Link-Training State Machine (LTSSM)

- **L0**: active
- **L0 standby, L1**: lower power, higher latency
- **L2**: cold standby, even lower power
- **L3**: power off

**Power Management**

**Link Training**

**Link Re-Training**
PCIe – Physical layer

PCI Express Device

LVDS

\[ V_{OH} = 1.4 \text{ V} \]
\[ V_{CM} = 1.2 \text{ V} \]
\[ V_{OH} = 1 \text{ V} \]

Signal

Wire

Link

Lane

PCI Express Device
PCIe – 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
PCIe – Framing (x1)

Transmit order (TIME)

- Reserved bits
- Sequence Number (Data Link Layer)

LCRC (Data Link Layer)

TLP structure (Transaction Layer)

- STP Framing Symbol (Physical 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
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
PCle CEM Spec – Power Cables

**EPS receptacle**

**PCle cable**

**GPU power**

---

**EPS-12V**

**PCle 6 Pin**

**PCle 8 Pin**

- **Gnd**
- **+12 V**
- **Sense A**
- **Sense B**

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 in practice
**PCIe – Theoretical data rates**

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

Theoretical bandwidth

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

Packet efficiency

• 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%
PCIe FPGA – latency

TYP: ~1us
What is this presentation about?

• History and evolution of PCIe

• PCIe concepts

• PCIe layers

• PCIe performance

• **PCIe in practice**
FPGA Hardened PCIe IP
PCIe scalability today

- Intel Xeon Broadwell
  - PCIe Gen3 x 40 lanes
- Intel Xeon Skylake
  - PCIe Gen3 x 48 lanes
- IBM Power 8
  - PCIe Gen3 x 48 lanes
- AMD Threadripper
  - PCIe Gen3 x 64 lanes
- AMD Epyc
  - PCIe Gen3 x 128 lanes (!)
- Counting Northbridge lanes only
- Excludes additional lanes from Southbridge
- More density possible using switches
  - Shared bandwidth
- Mostly driven by storage market (dense NVMe)
PCle Gen4 – On paper

Mellanox and Synopsys Demonstrate Industry's First PCle 4.0 Interoperability
Mutual Technology Leadership Lowers Risk for Designers Implementing16GT/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 1:34AM 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

PCI-SIG Administration <administration@pcisig.com>
to PCI-SIG

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

Mellanox ConnectX®-5

- Dual-Port 100 Gbit/s
- Available!

IBM Power AC922 (2018?)

- 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

LnkCap: Port #0, Speed 16GT/s, Width x16, ASPM L0s L1
Danke für Ihre Aufmerksamkeit