Introduction to PCI Express

Paolo Durante

(CERN EP-LBC)

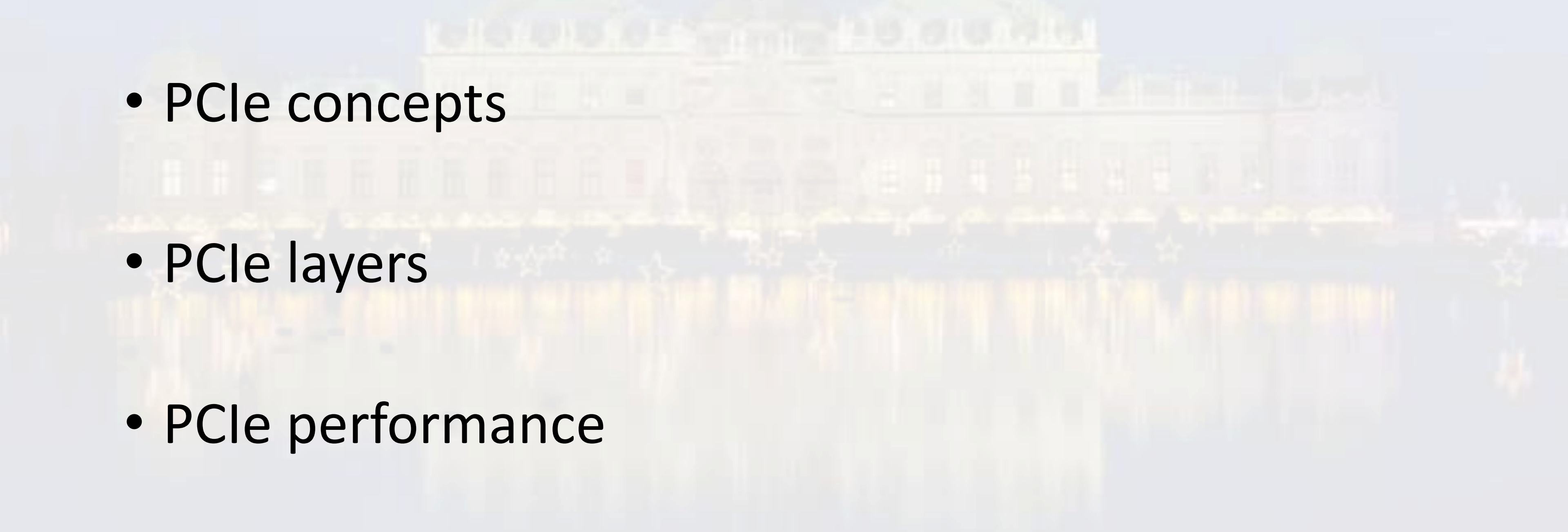
Where will you find PCI EXPRESS ?

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)

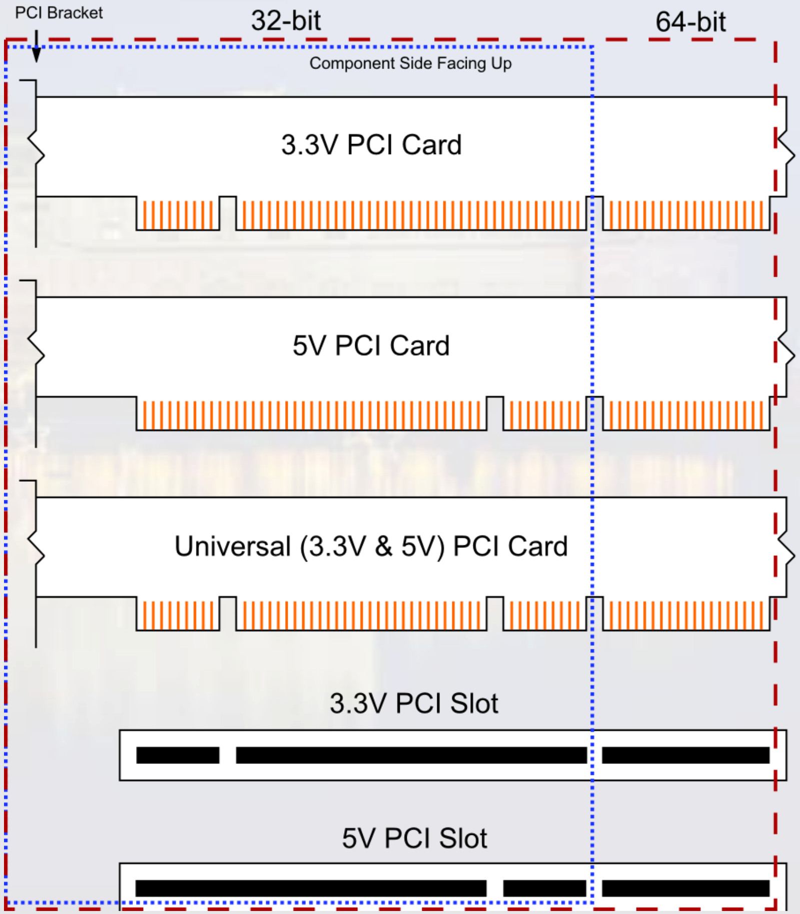
What is this presentation about?

- **History and evolution of PCIe**
- PCIe concepts
- PCIe layers
- PCIe performance
- PCIe in practice



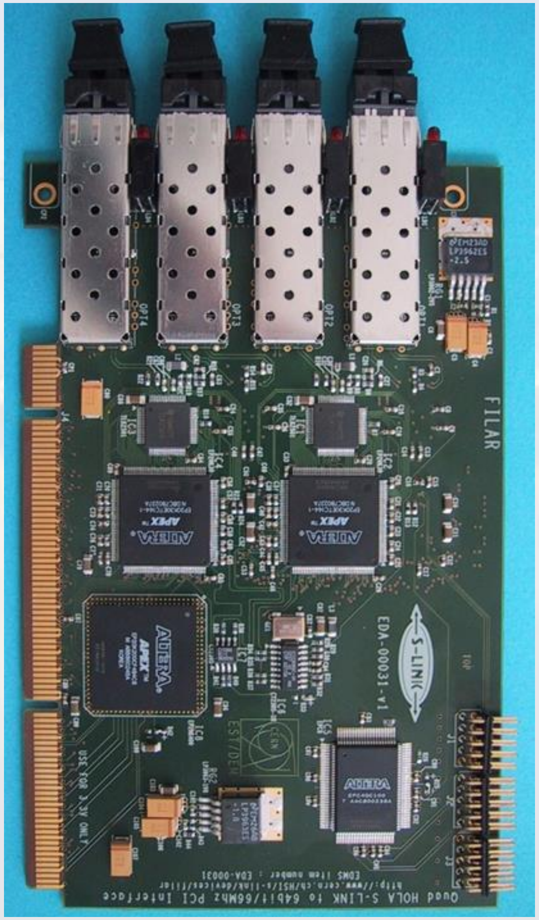
PCI ("conventional PCI")

- 1992
- ***P**eripheral **C**omponent **I**nterconnect*
- 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)



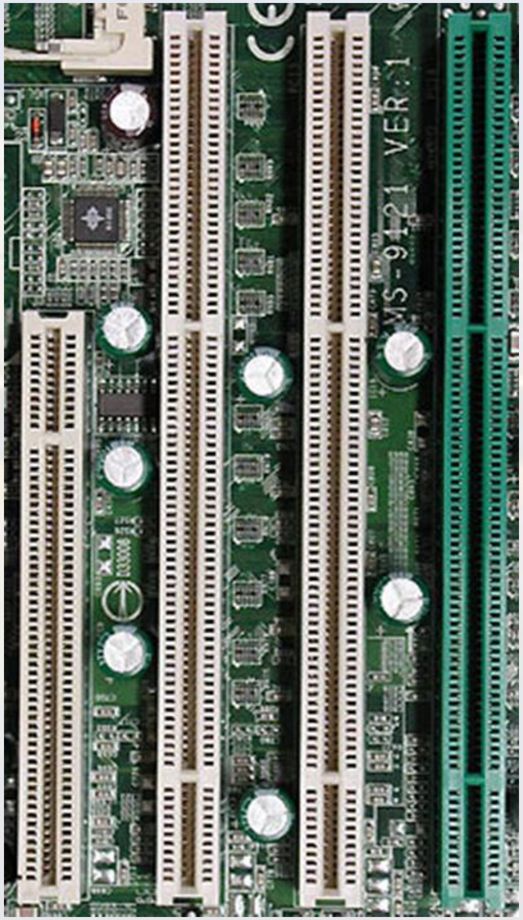
PCI example: ATLAS FILAR

- ~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


FILAR
S-LINK
EDA-00031-v1
ALTERA
APEX
Quad HOLA S-LINK to 64bit/66MHz PCI Interface
http://www.cern.ch/HSI/s-link/devices/filar
EDMS item number : EDA-00031

PCI-X ("Extended PCI")

- 1998
- PCI compatible
    - hardware and software
    - half-duplex bidirectional
- Bandwidth
    - ≤ 1066 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*)



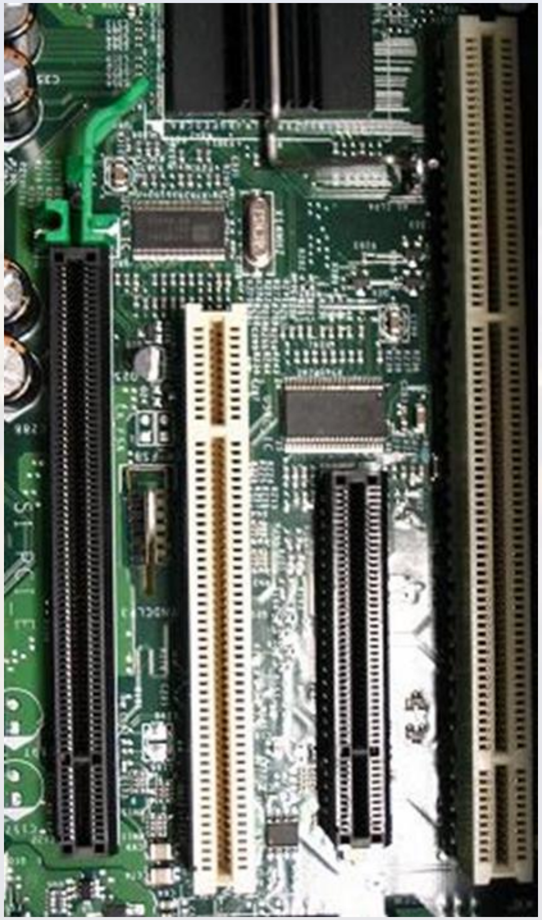
PCI-X example: CMS FEROL

- ~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: ≤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…

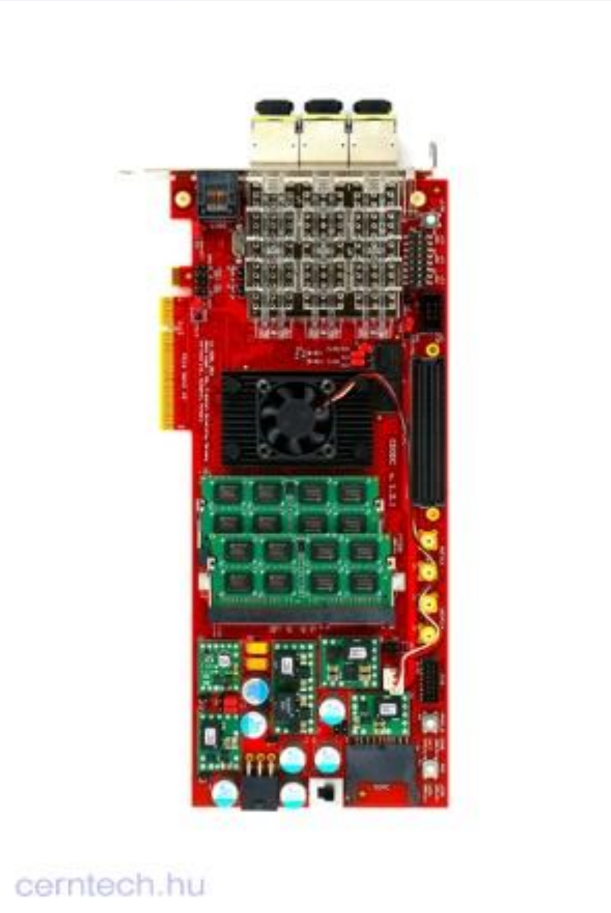
PCIe x16 | PCI | PCIe x8|PCI-X



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


cerntech.hu

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

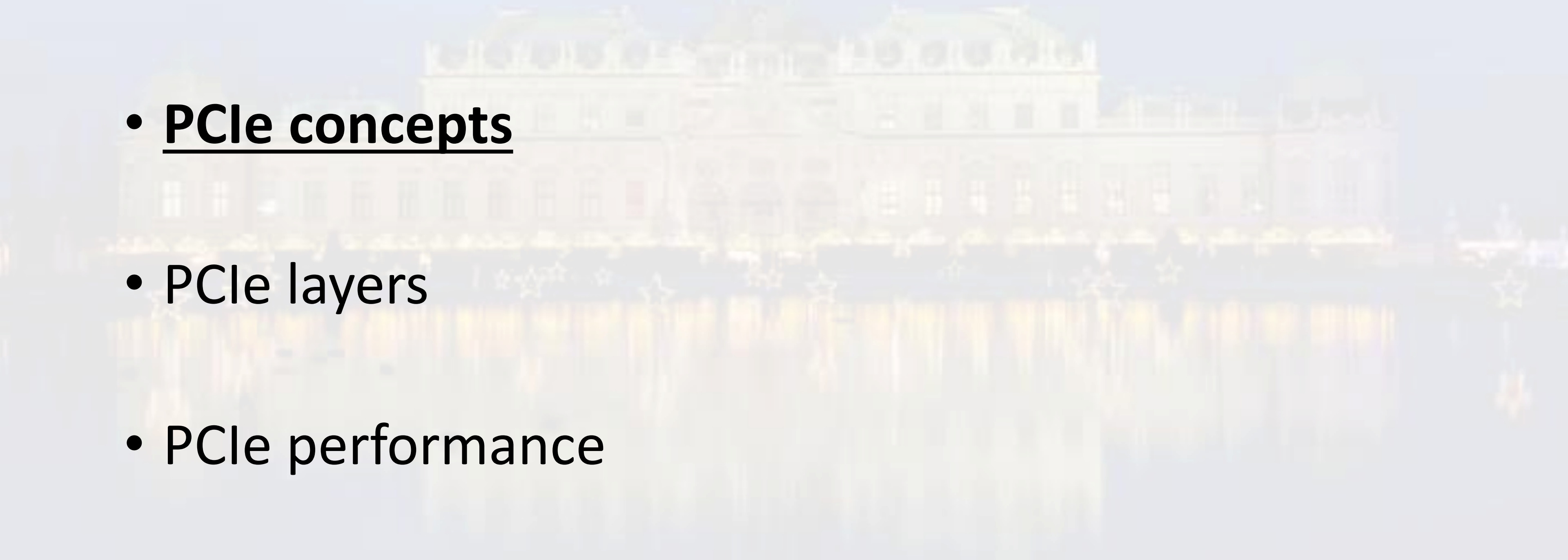
- 2016
- ≤ 48 duplex optical links
- XilinX Ultrascale FPGA
- 2x DDR4 SO-DIMM
- PCIe 3.0 x16
- Wupper DMA
  (Open Source!)

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

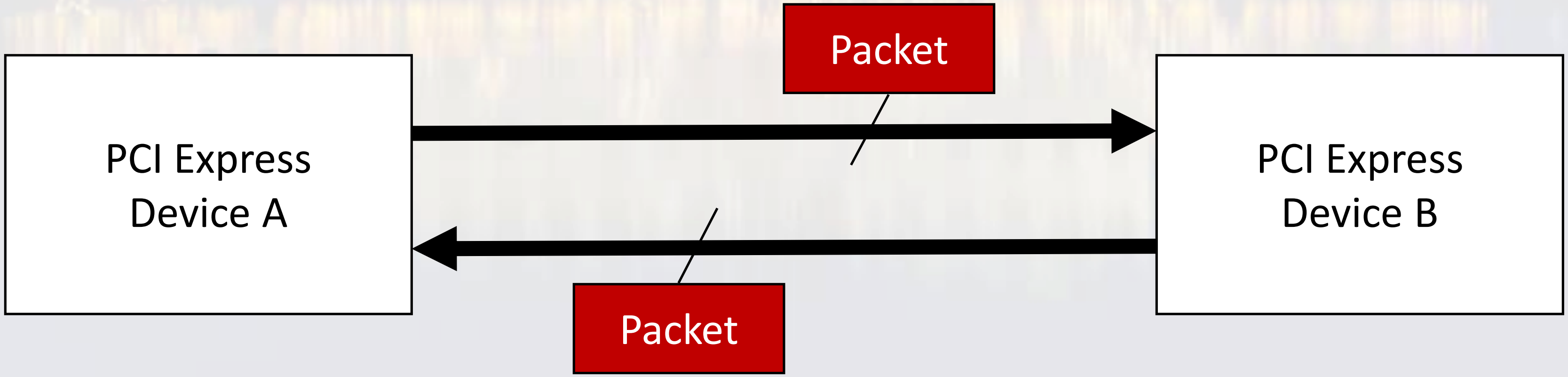
What is this presentation about?

- History and evolution of PCIe
- **<u>PCIe concepts</u>**
- PCIe layers
- PCIe performance
- PCIe in practice



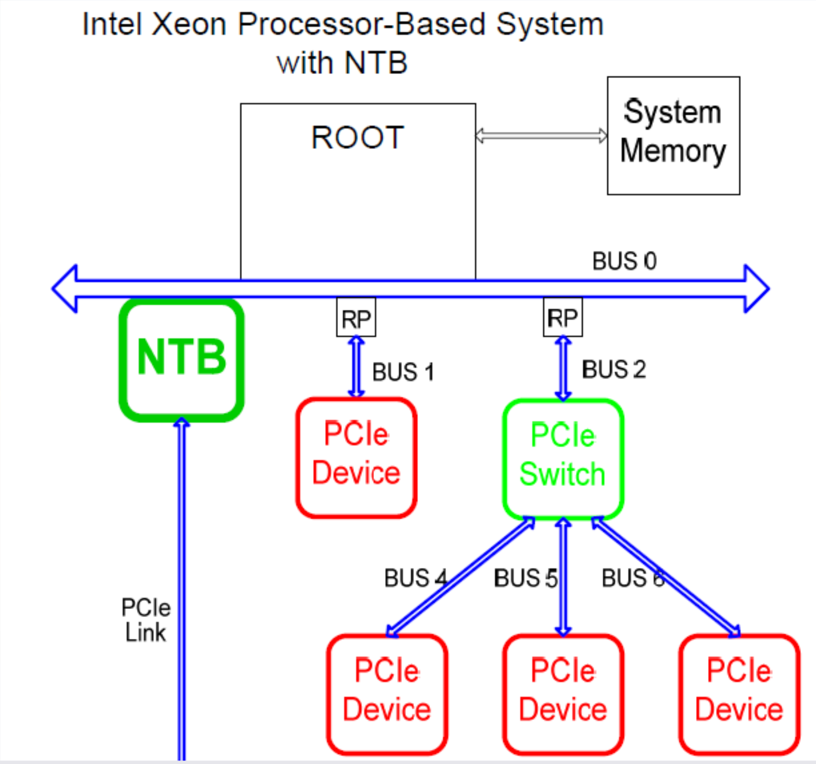
PCIe concepts – Packets

- Point-to-point connection
- “Serial” “bus” (fewer pins)
- Scalable link: **x1**, x2, **x4**, **x8**, x12, **x16**, x32
- Packet encapsulation

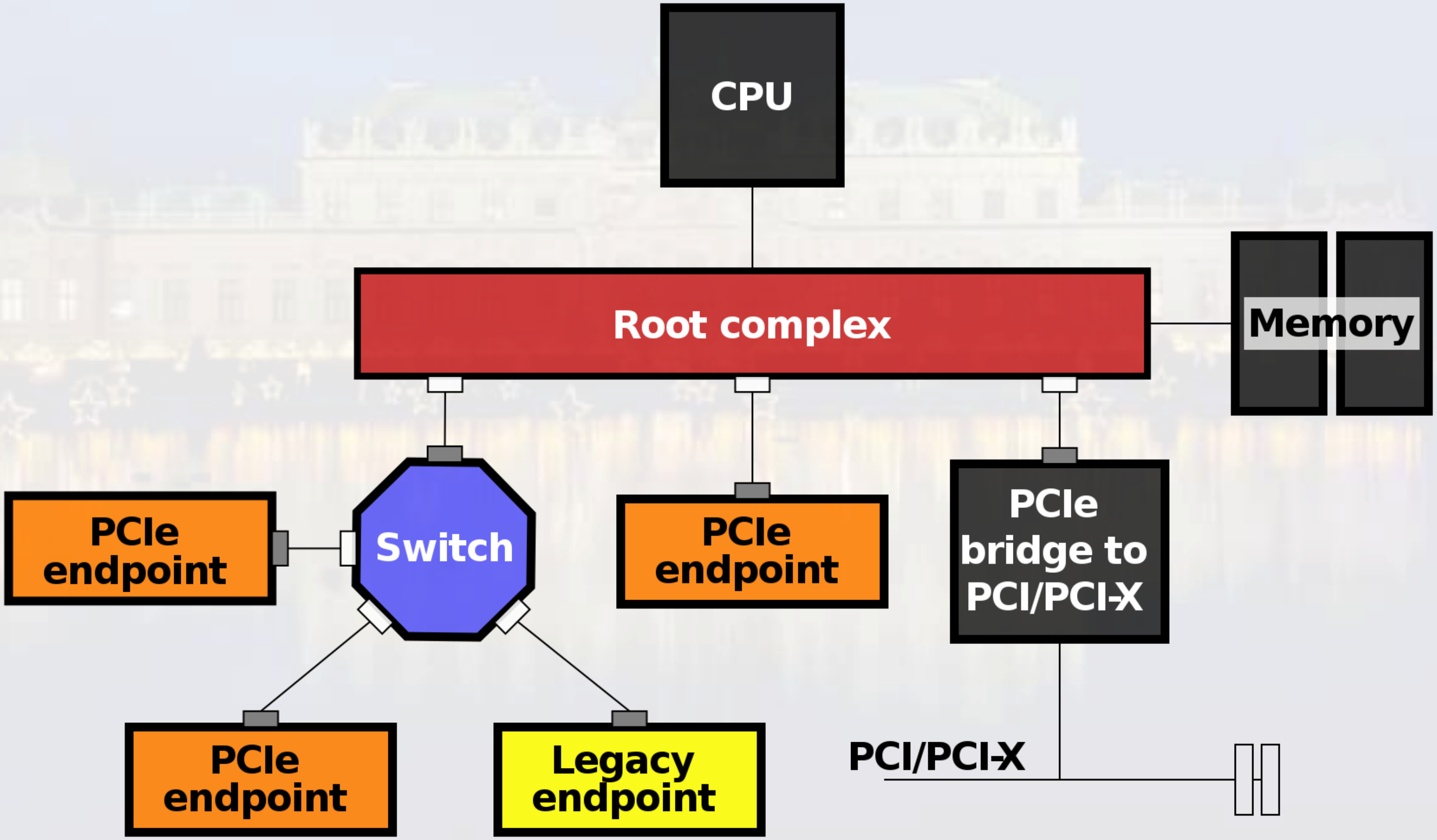


PCIe concepts – Root complex

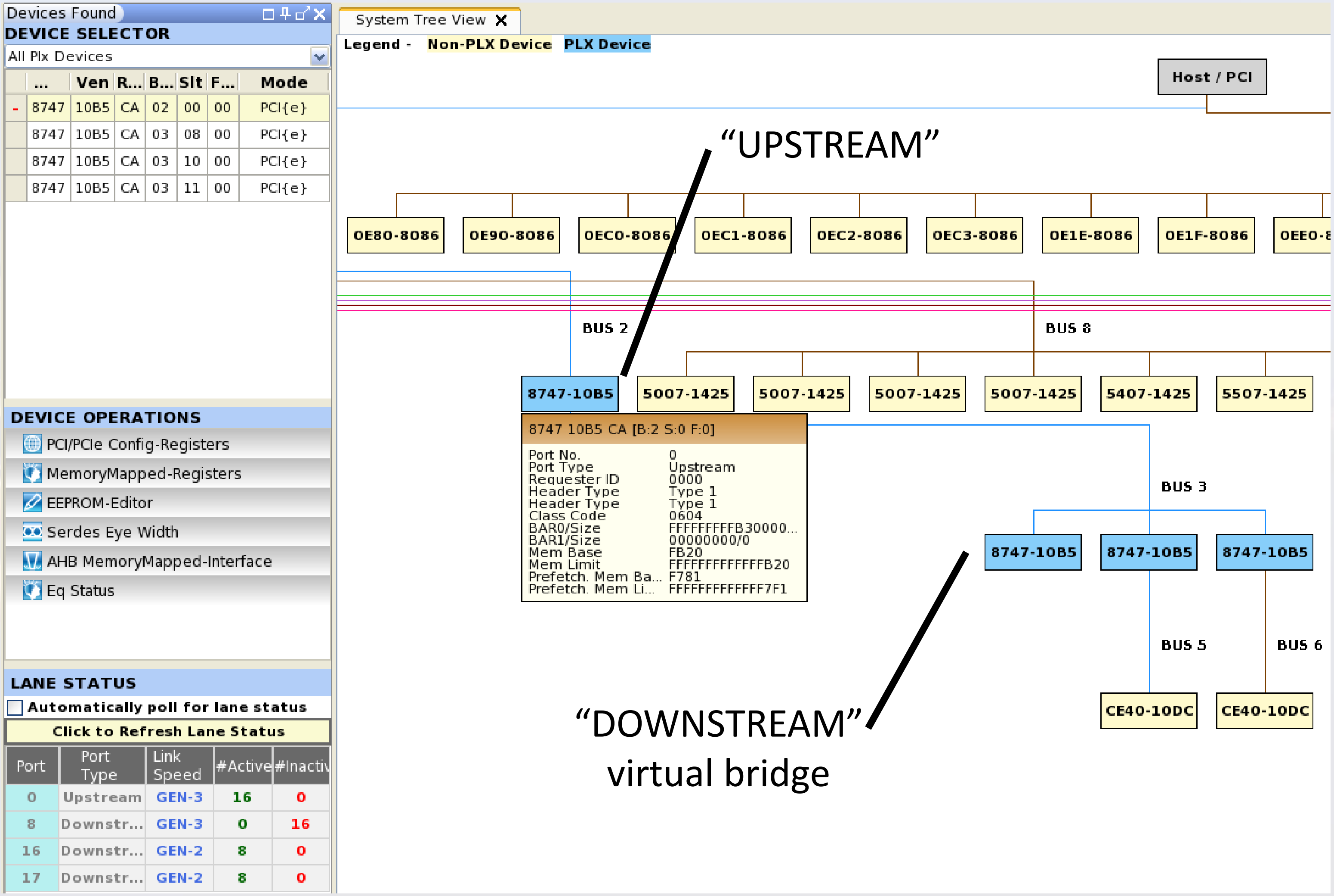
- Connects the processor and memory subsystems to the PCIe fabric via a ***Root Port***
- Generates and processes transactions with ***Endpoints*** on behalf of the processor



PCIe concepts – Topology



Example: PCIe switch



Devices Found

DEVICE SELECTOR

All Plx Devices

	...	Ven	R...	B...	Slt	F...	Mode
-	8747	10B5	CA	02	00	00	PCI{e}
	8747	10B5	CA	03	08	00	PCI{e}
	8747	10B5	CA	03	10	00	PCI{e}
	8747	10B5	CA	03	11	00	PCI{e}

DEVICE OPERATIONS

- PCI/PCIe Config-Registers
- MemoryMapped-Registers
- EEPROM-Editor
- Serdes Eye Width
- AHB MemoryMapped-Interface
- Eq Status

LANE STATUS

□ Automatically poll for lane status

Click to Refresh Lane Status

Port	Port Type	Link Speed	#Active	#Inactiv
0	Upstream	GEN-3	16	0
8	Downstr...	GEN-3	0	16
16	Downstr...	GEN-2	8	0
17	Downstr...	GEN-2	8	0

8747 10B5 CA [B:2 S:0 F:0]

Port No.	0
Port Type	Upstream
Requester ID	0000
Header Type	Type 1
Header Type	Type 1
Class Code	0604
BAR0/Size	FFFFFFFFFB30000...
BAR1/Size	00000000/0
Mem Base	FB20
Mem Limit	FFFFFFFFFFFFFB20
Prefetch. Mem Ba...	F781
Prefetch. Mem Li...	FFFFFFFFFFFFF7F1

“UPSTREAM”

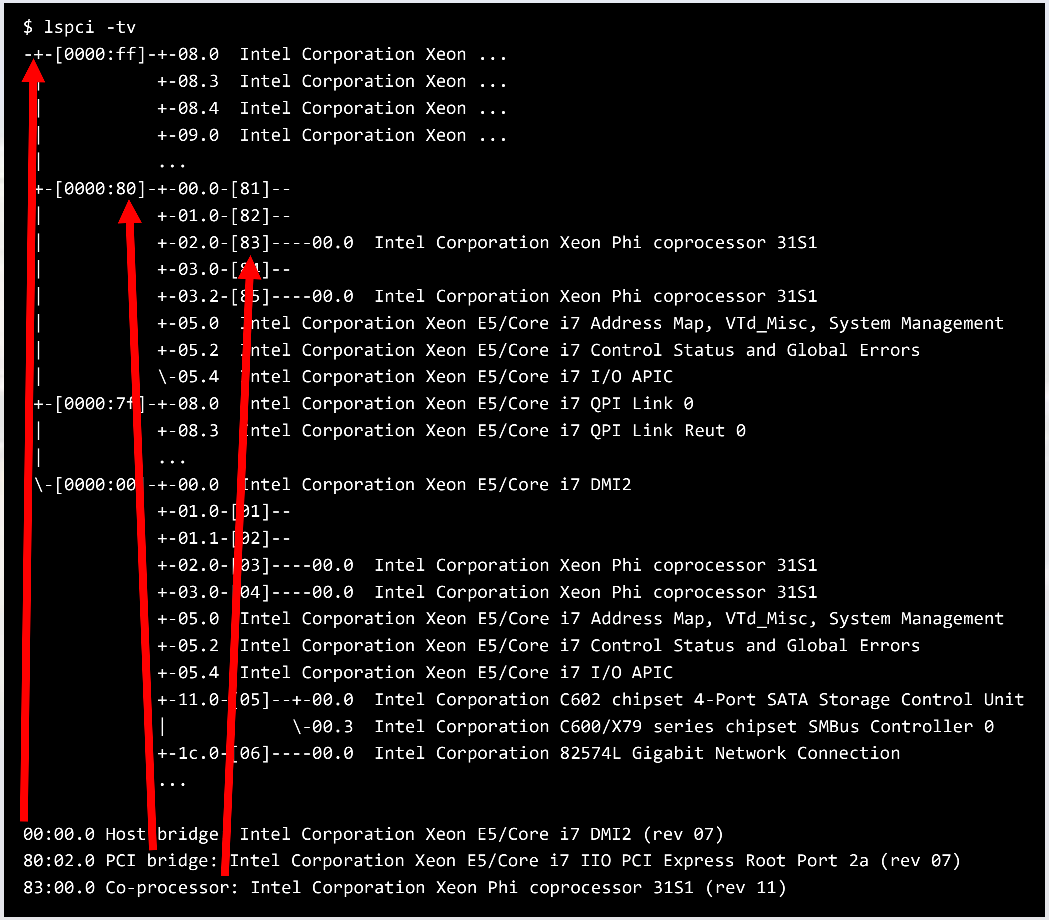
“DOWNSTREAM” virtual bridge

PCIe 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 (PCIe) form hierarchy

On linux: $ man lspci



```
$ lspci -tv
-+-[0000:ff]-+-08.0  Intel Corporation Xeon ...
 |           +-08.3  Intel Corporation Xeon ...
 |           +-08.4  Intel Corporation Xeon ...
 |           +-09.0  Intel Corporation Xeon ...
 |           ...
 +-[0000:80]-+-00.0-[81]--
 |           +-01.0-[82]--
 |           +-02.0-[83]----00.0  Intel Corporation Xeon Phi coprocessor 31S1
 |           +-03.0-[84]--
 |           +-03.2-[85]----00.0  Intel Corporation Xeon Phi coprocessor 31S1
 |           +-05.0  Intel Corporation Xeon E5/Core i7 Address Map, VTd_Misc, System Management
 |           +-05.2  Intel Corporation Xeon E5/Core i7 Control Status and Global Errors
 |           \-05.4  Intel Corporation Xeon E5/Core i7 I/O APIC
 +-[0000:7f]-+-08.0  Intel Corporation Xeon E5/Core i7 QPI Link 0
 |           +-08.3  Intel Corporation Xeon E5/Core i7 QPI Link Reut 0
 |           ...
 \-[0000:00]-+-00.0  Intel Corporation Xeon E5/Core i7 DMI2
             +-01.0-[01]--
             +-01.1-[02]--
             +-02.0-[03]----00.0  Intel Corporation Xeon Phi coprocessor 31S1
             +-03.0-[04]----00.0  Intel Corporation Xeon Phi coprocessor 31S1
             +-05.0  Intel Corporation Xeon E5/Core i7 Address Map, VTd_Misc, System Management
             +-05.2  Intel Corporation Xeon E5/Core i7 Control Status and Global Errors
             +-05.4  Intel Corporation Xeon E5/Core i7 I/O APIC
             +-11.0-[05]--+-00.0  Intel Corporation C602 chipset 4-Port SATA Storage Control Unit
             |            \-00.3  Intel Corporation C600/X79 series chipset SMBus Controller 0
             +-1c.0-[06]----00.0  Intel Corporation 82574L Gigabit Network Connection
             ...

00:00.0 Host bridge: Intel Corporation Xeon E5/Core i7 DMI2 (rev 07)
80:02.0 PCI bridge: Intel Corporation Xeon E5/Core i7 IIO PCI Express Root Port 2a (rev 07)
83:00.0 Co-processor: Intel Corporation Xeon Phi coprocessor 31S1 (rev 11)
```

Understanding `lspci --tv` (1/3)

```
\-[0000:00]-+-00.0  Intel Corporation Xeon E7 v4/Xeon E5 v4/Xeon E3 v4/Xeon D DMI2
            +-01.0-[01-02]--+-00.0  Intel Corporation Ethernet Controller 10-Gigabit X540-AT2
            |               \-00.1  Intel Corporation Ethernet Controller 10-Gigabit X540-AT2
```

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)

```
00:01.0 PCI bridge: Intel Corporation Xeon E7 v4/Xeon E5 v4/Xeon E3 v4/Xeon D PCI Express Root Port 1
(rev 01)
01:00.0 Ethernet controller: Intel Corporation Ethernet Controller 10-Gigabit X540-AT2 (rev 01)
01:00.1 Ethernet controller: Intel Corporation Ethernet Controller 10-Gigabit X540-AT2 (rev 01)
```

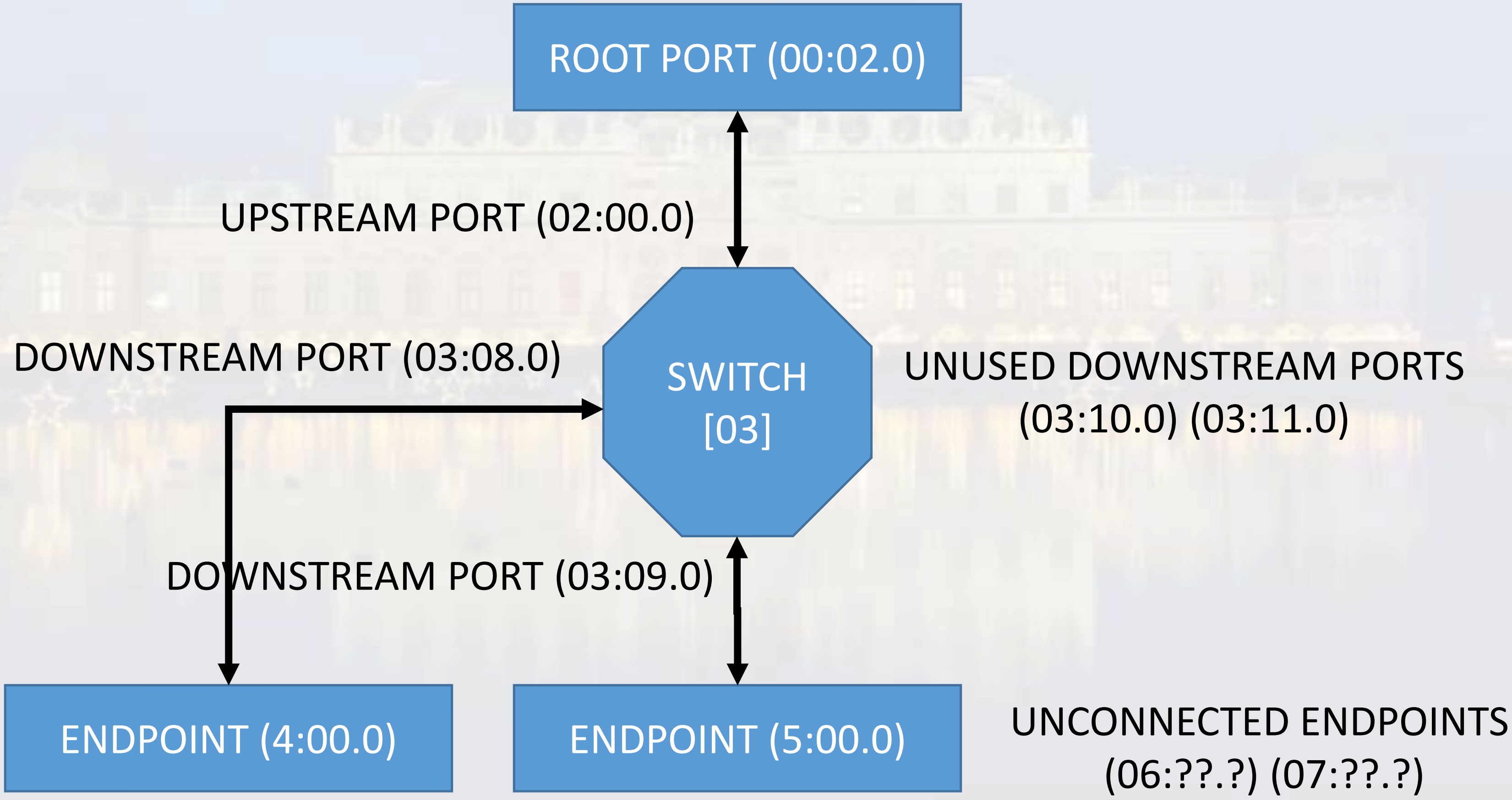
Understanding `lspci --tv` (2/3)

```
\-[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)



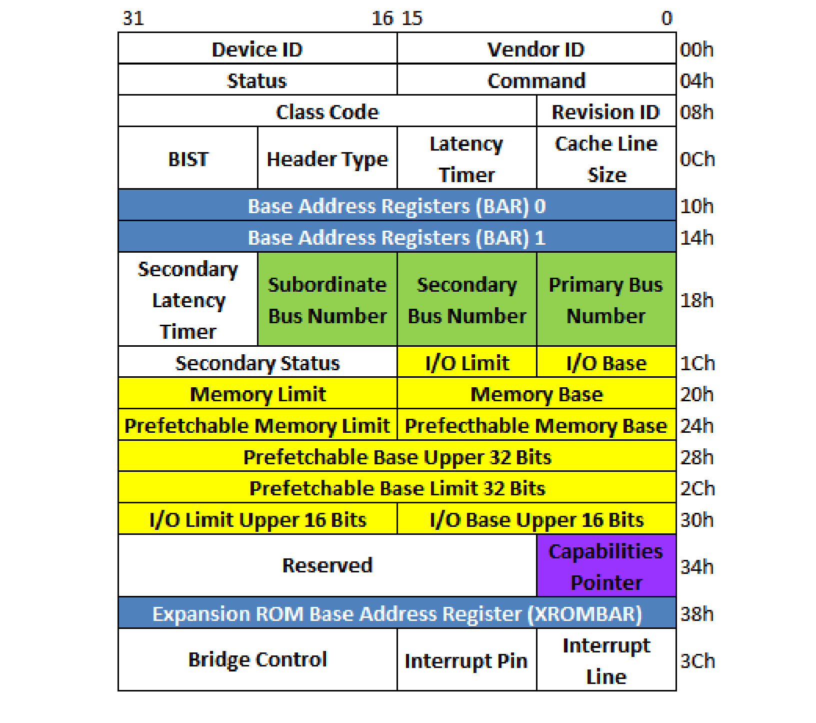
Troubleshooting with lspci

- 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)



31	16	15	0	
Device ID		Vendor ID		00h
Status		Command		04h
Class Code			Revision ID	08h
BIST	Header Type	Latency Timer	Cache Line Size	0Ch
Base Address Registers (BAR) 0				10h
Base Address Registers (BAR) 1				14h
Secondary Latency Timer	Subordinate Bus Number	Secondary Bus Number	Primary Bus Number	18h
Secondary Status		I/O Limit	I/O Base	1Ch
Memory Limit		Memory Base		20h
Prefetchable Memory Limit		Prefecthable Memory Base		24h
Prefetchable Base Upper 32 Bits				28h
Prefetchable Base Limit 32 Bits				2Ch
I/O Limit Upper 16 Bits		I/O Base Upper 16 Bits		30h
Reserved			Capabilities Pointer	34h
Expansion ROM Base Address Register (XROMBAR)				38h
Bridge Control		Interrupt Pin	Interrupt Line	3Ch

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*(), …

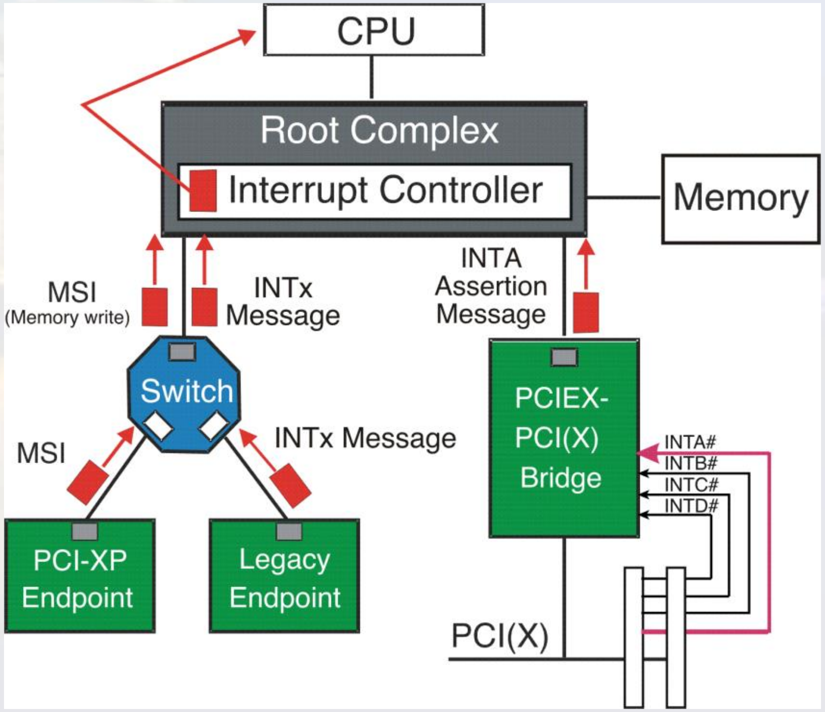
PCIe concepts – Interrupts

- PCI
  - INT*x*#
    - $x \in \{A, B, C, D\}$
  - Level sensitive
  - Can be mapped to CPU interrupt number
- PCIe
  - “Virtual Wire” emulation
  - Assert_INTx code
  - Deassert_INTx code

```
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
  - ≤ 32 per device (in theory)
- MSI-X uses a table of independent address and data pairs for each "vector"
  - ≤ 2048 per device (use affinity!)
- *Vector*: interrupt id



PCIe Gen1 (2003)

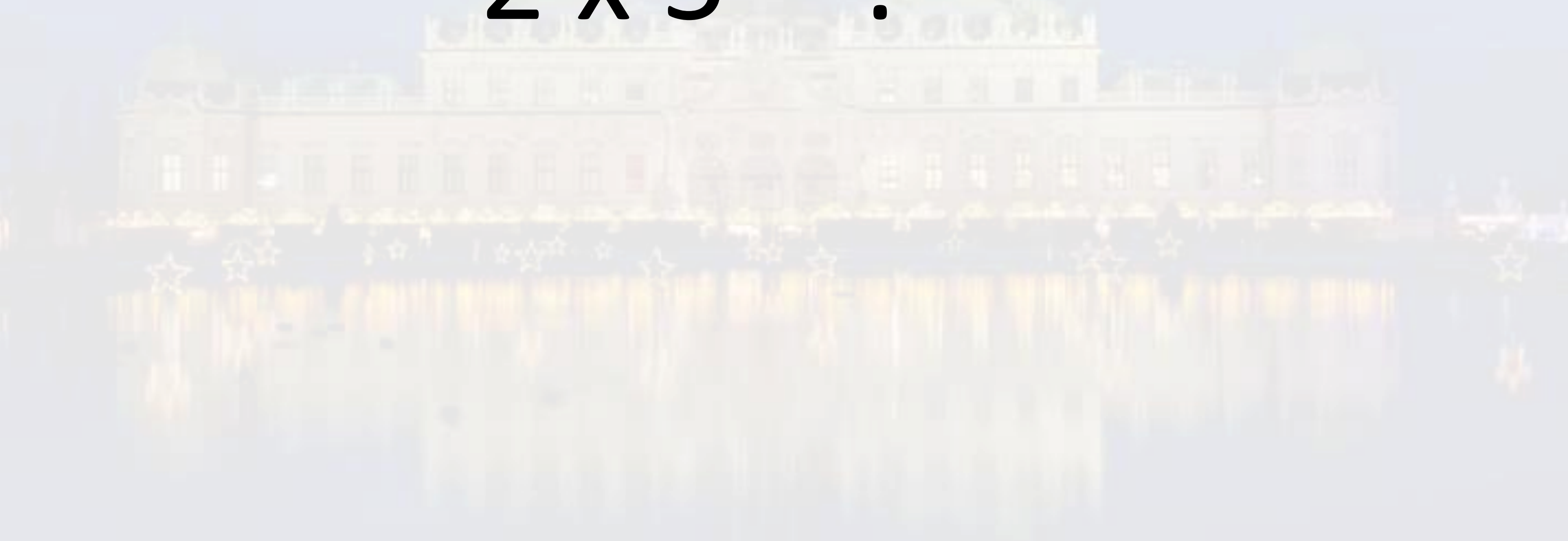
- Introduced at 2.5 GT/sec
- Also called 2.5 GHz, 2.5 Gb/s
- 100 MHz reference clock
  - Eases synchronization between ends
  - Can use Spread Spectrum Clocking to reduce EMI
  - Optional, but nearly universal
- 8b/10b encoding used to provide DC balance and reduce “runs” of 0s or 1s which make clock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

PCIe Gen2 (2007)

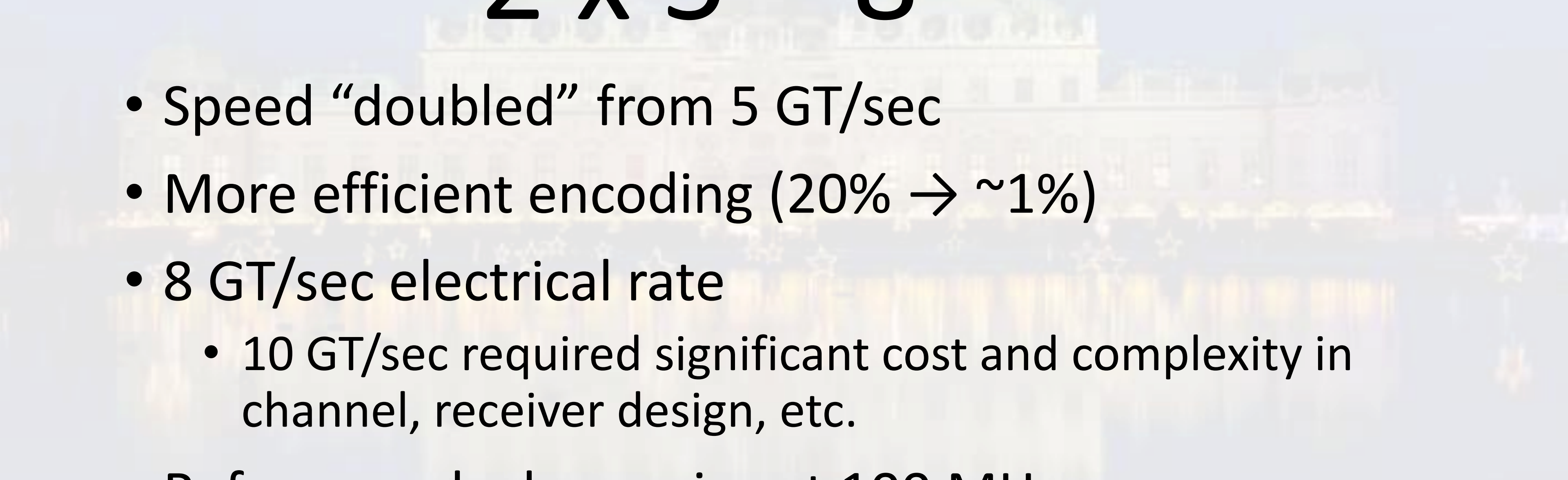
- Speed doubled from 2.5 to 5 GT/sec
- Reference clock remains at 100 MHz
  - Lower jitter clock sources required vs 2.5 GT/sec
  - Generally higher quality clock generation/distribution required
- 8b/10b encoding continues to be used
- Specification Revisions: 2.0, 2.1
- Devices choosing to implement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

PCIe Gen3

2 x 5 = ?



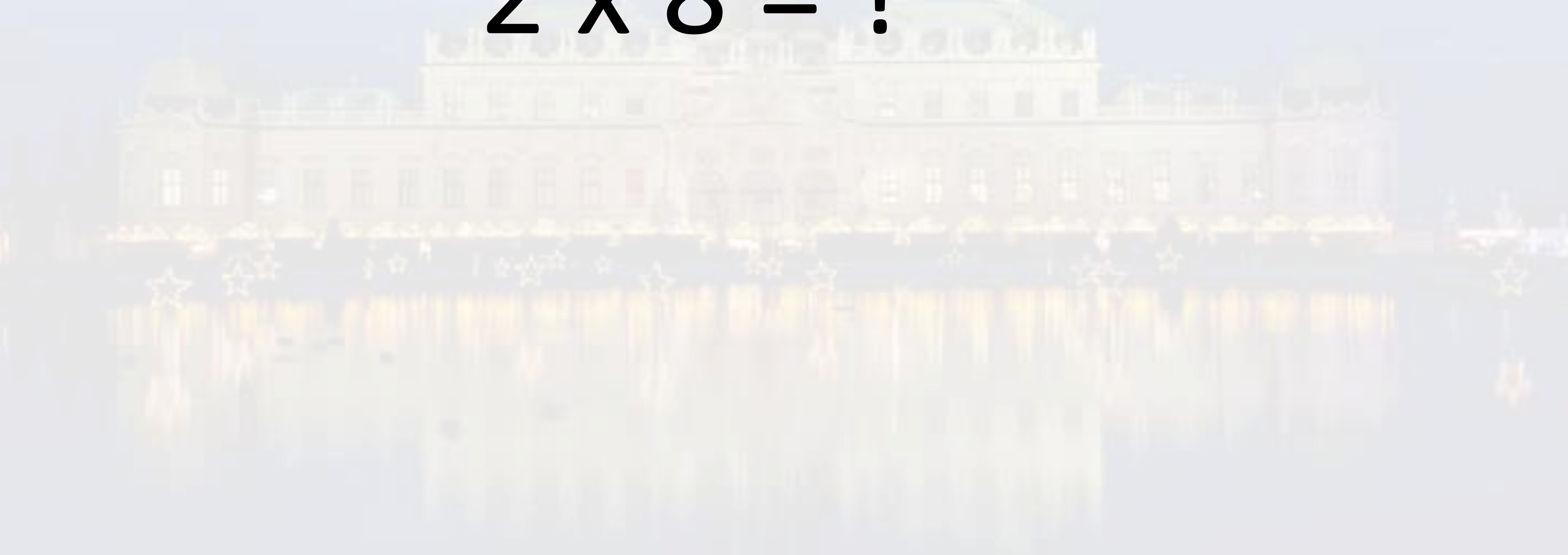
PCIe Gen3



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

PCIe Gen4



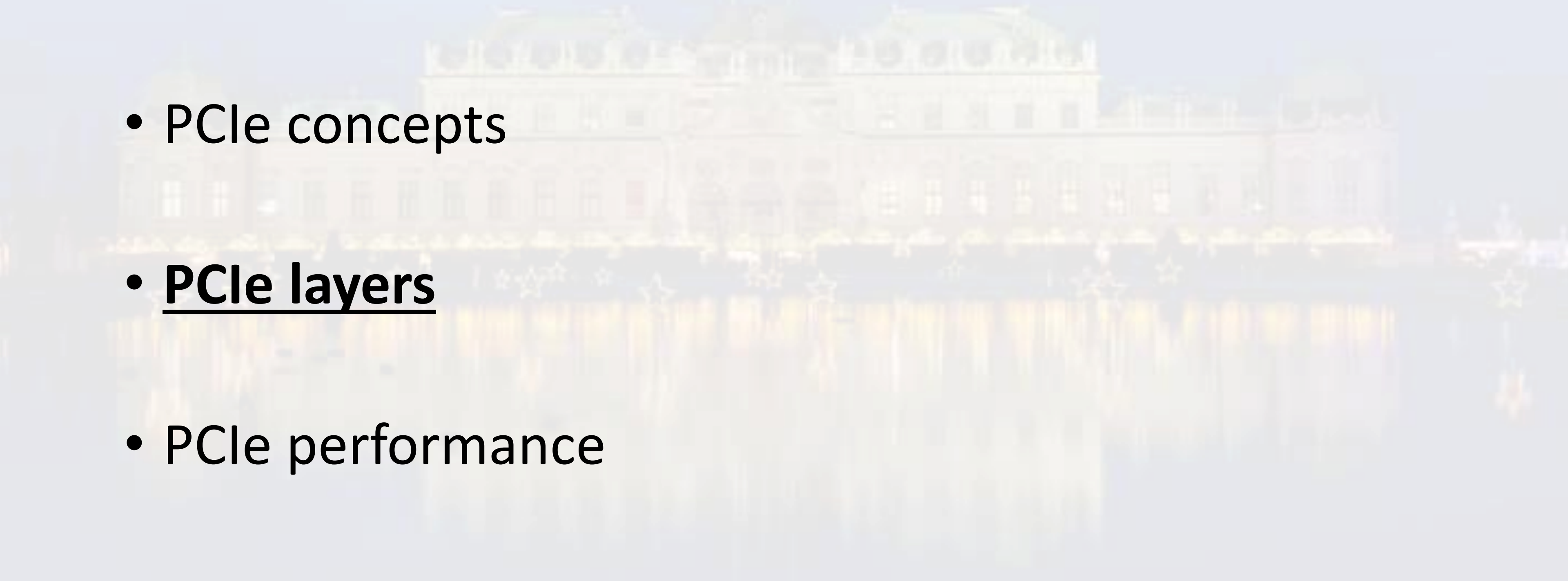
2 x 8 = ?

PCIe Gen4

2 x 8 = 16

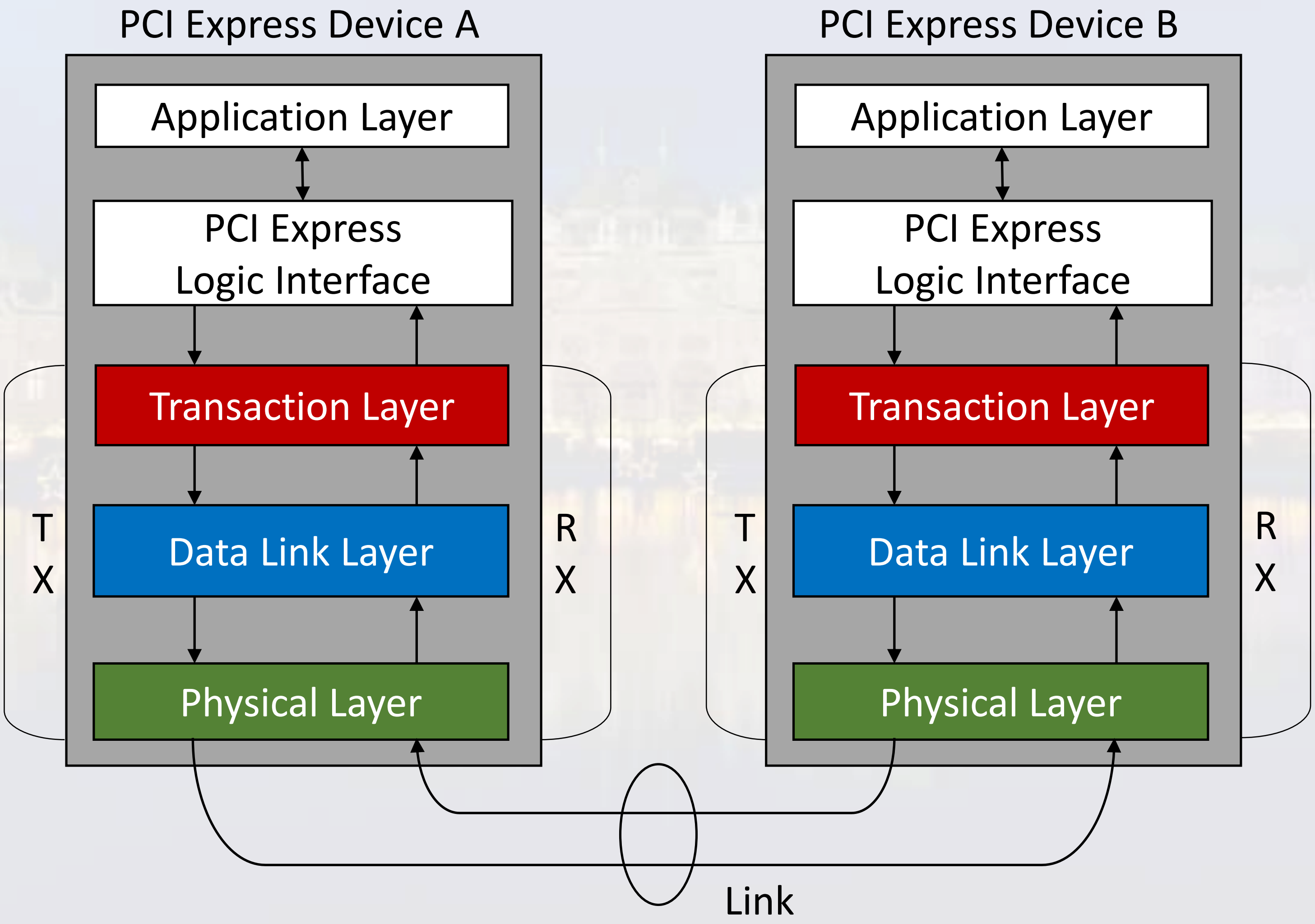
- Speed doubled from 8 GT/sec (Gen5 likely similar)
- Same 128b/130b encoding
- 16 GT/sec electrical rate
  - Channel length: ≤ 10”/14”
  - Retimer mandatory for longer channels
  - More complex pre-amplification, equalization stages
- Reference clock remains at 100 MHz
- Backwards-compatible protocol 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

TX

RX

PCI Express Device B

Application Layer

PCI Express Logic Interface

Transaction Layer

Data Link Layer

Physical Layer

TX

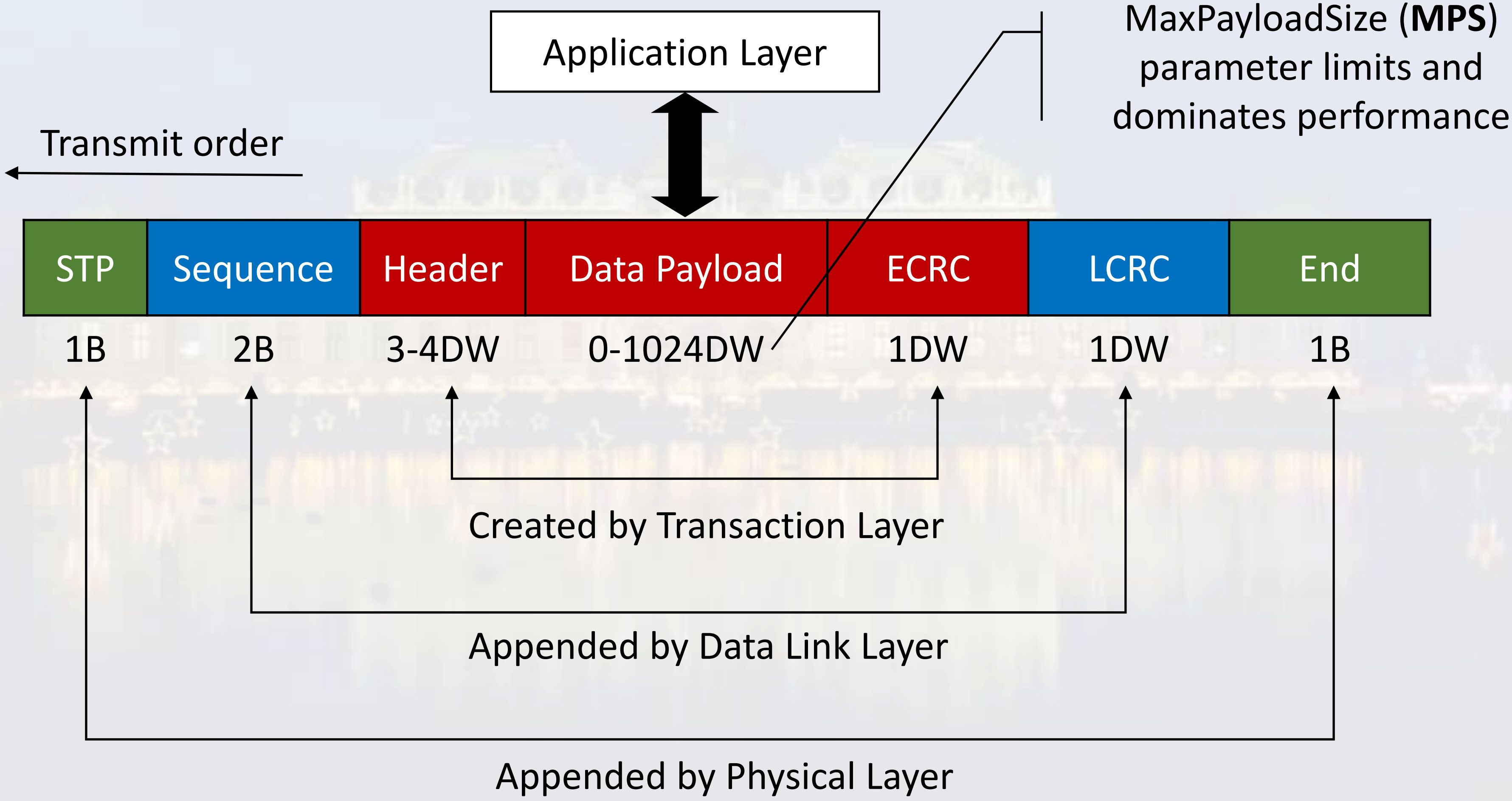
RX

Link

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

PCIe – TLP structure



MaxPayloadSize (**MPS**) parameter limits and dominates performance

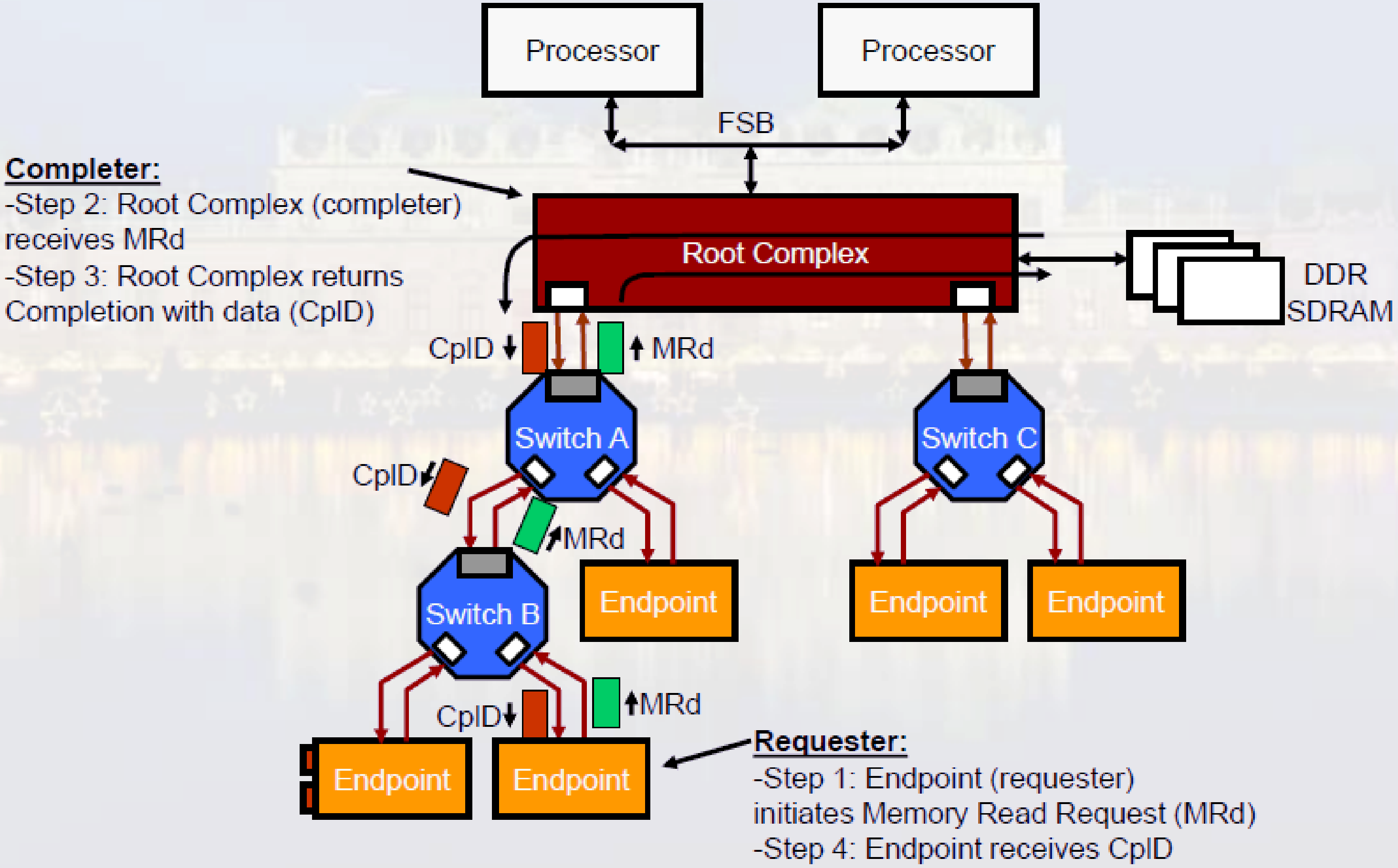
PCIe – Split transaction model

- 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



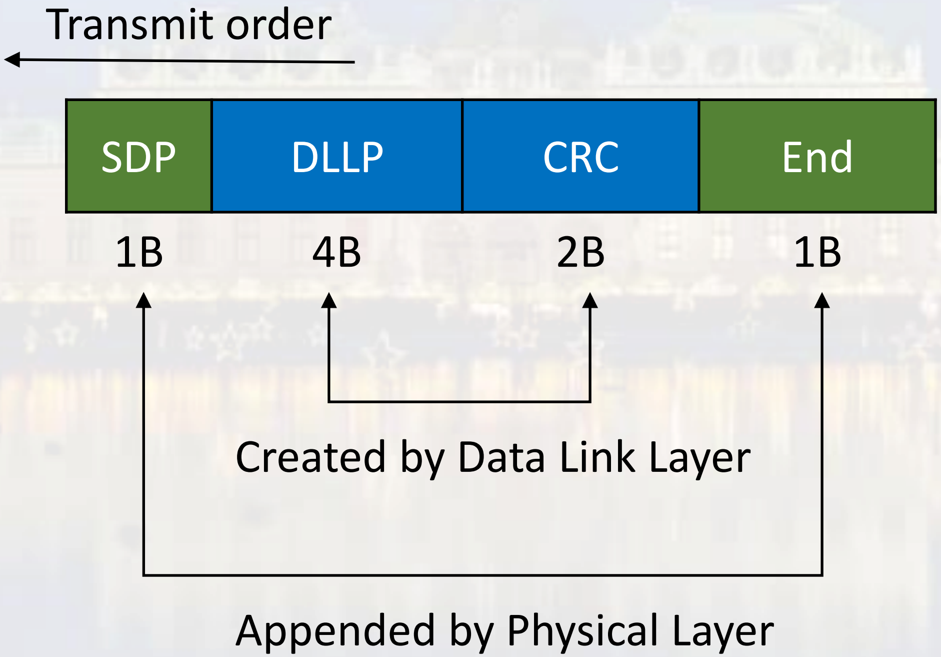
Completer:
-Step 2: Root Complex (completer) receives MRd
-Step 3: Root Complex returns Completion with data (CplD)

Requester:
-Step 1: Endpoint (requester) initiates Memory Read Request (MRd)
-Step 4: Endpoint receives CplD

PCIe – Data Link Layer

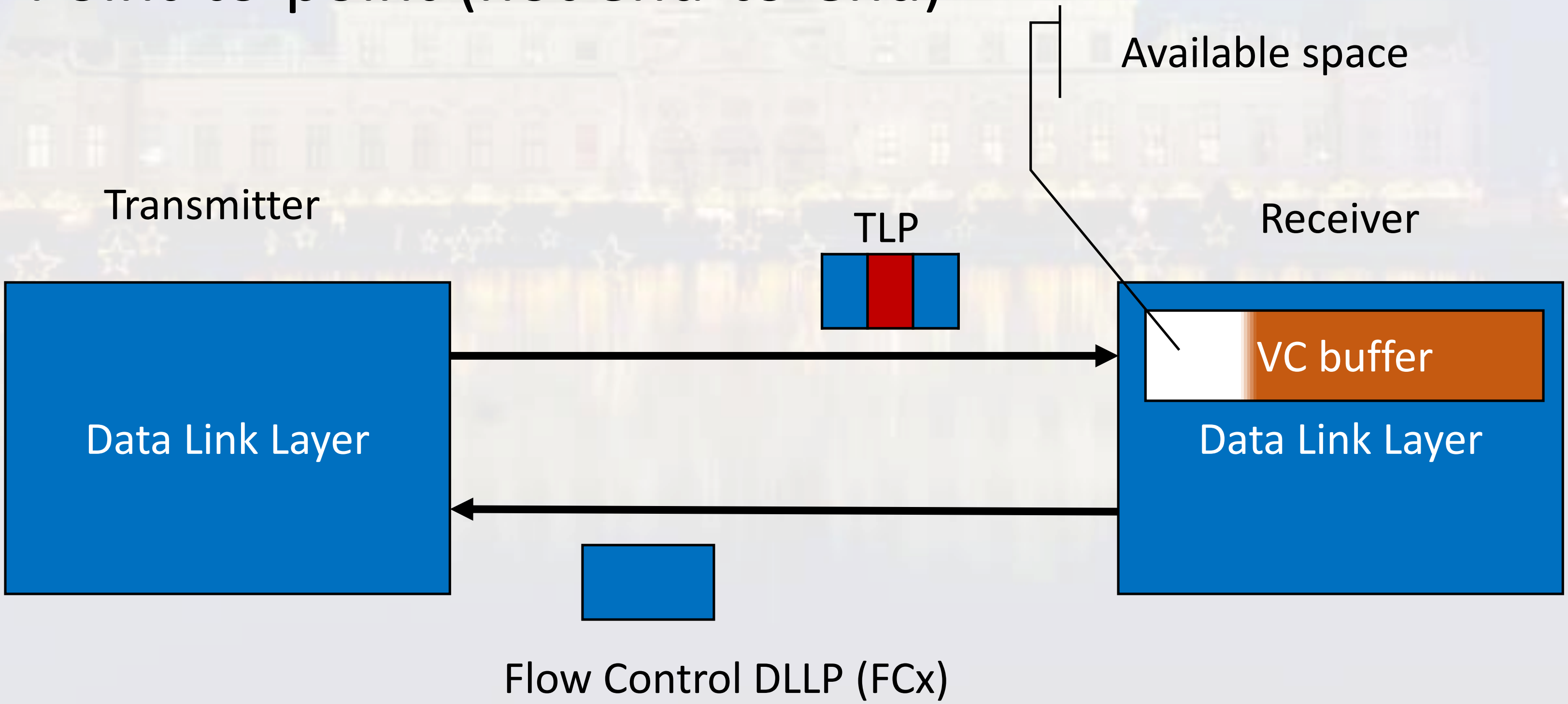
- ACK / NAK Packets
  - Error handling mechanism
- Flow Control Packets (FCPs)
  - Propagate credit allocation status
- Power Management Packets
- Vendor extensions
  - E.g.: CAPI, CCIX (memory coherency)

PCIe – DLLP structure



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

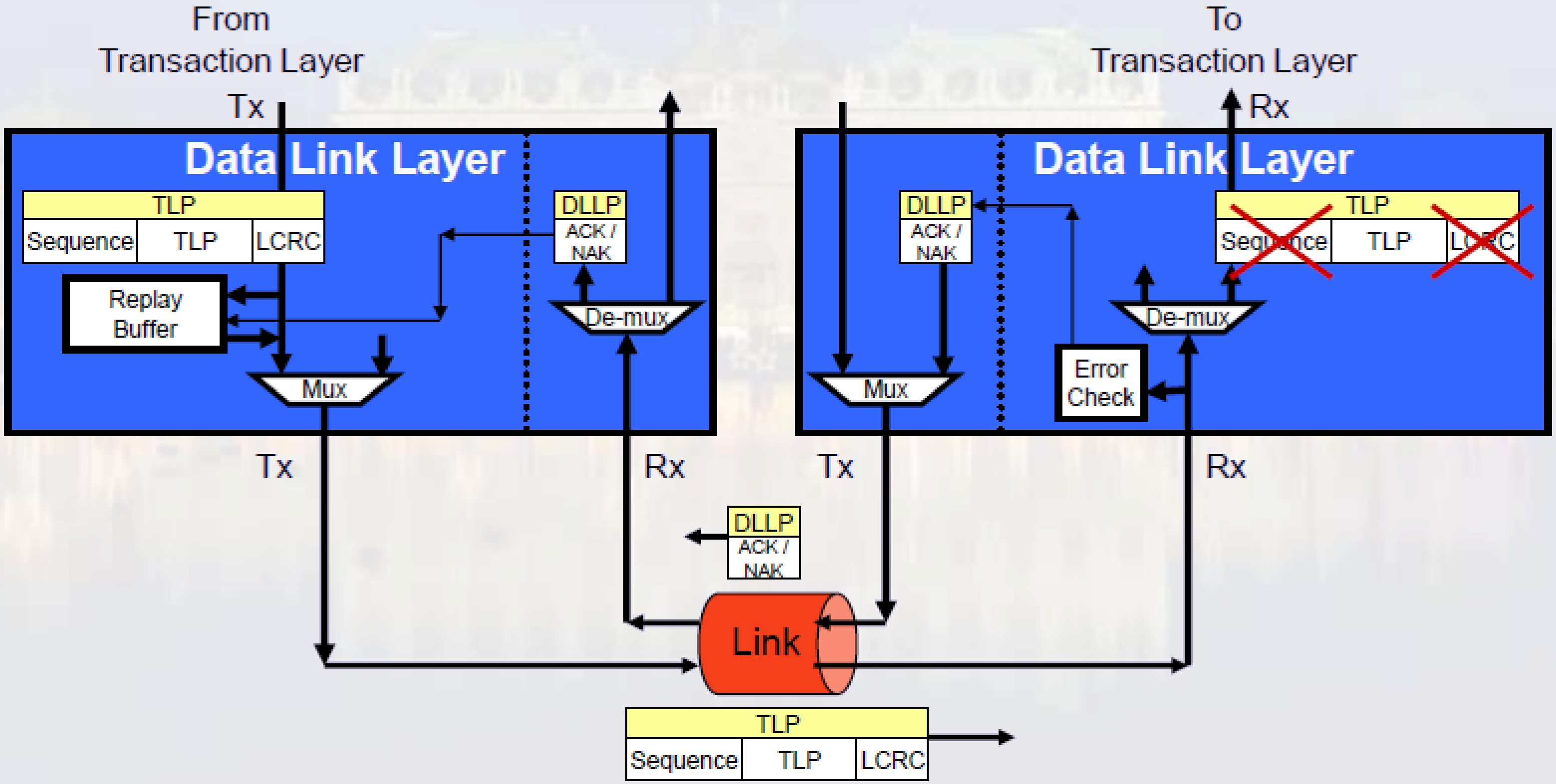
Correctable

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

Uncorrectable

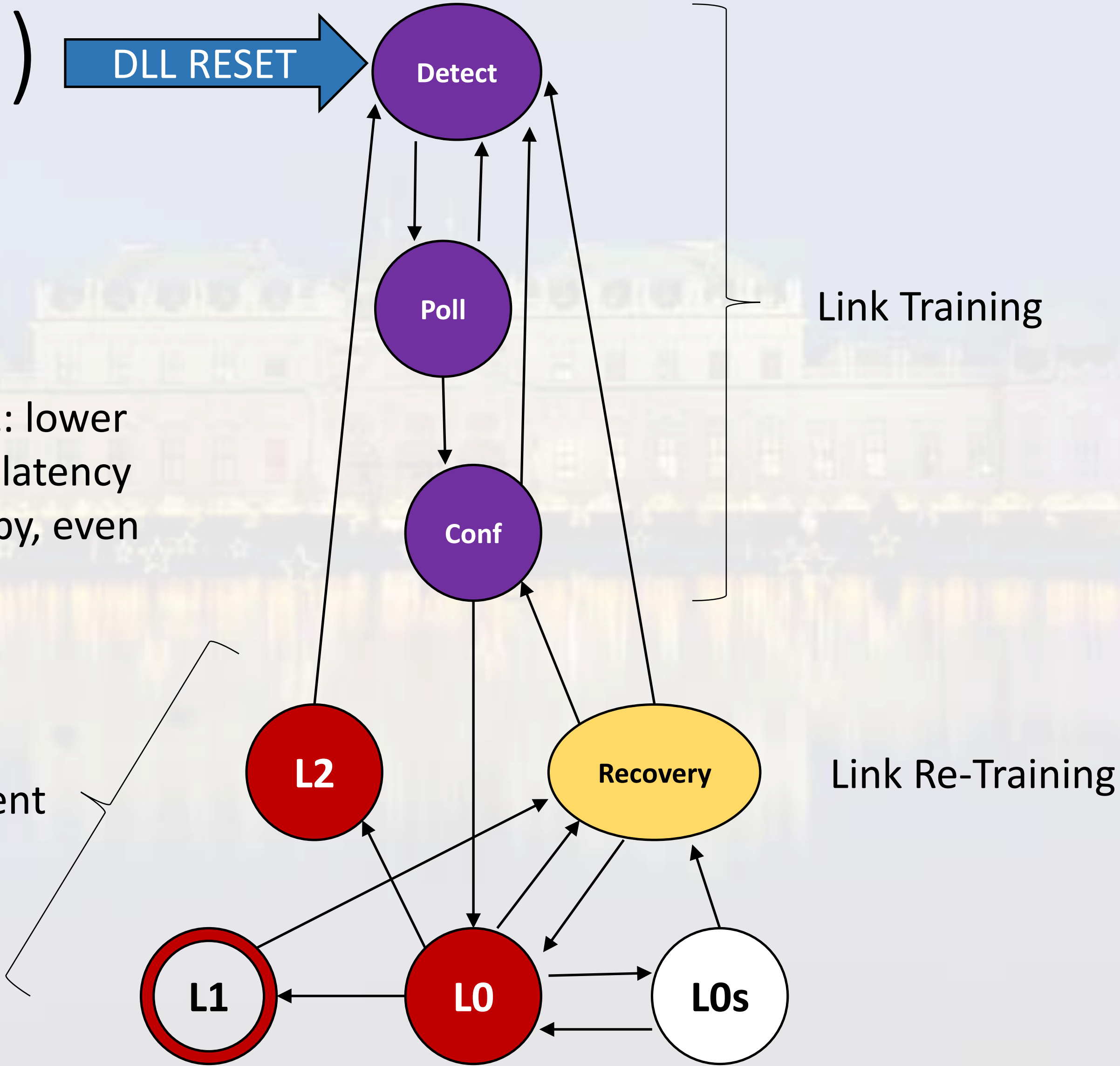
- **Fatal**
  - Platform-specific handling
- **Non-fatal**
  - Can be exposed to application layer and handled explicitly
- Can and do cause system deadlock / reset
- Recovery mechanisms are outside the spec
  - Example: failover for HA

PCIe – ACK/NAK

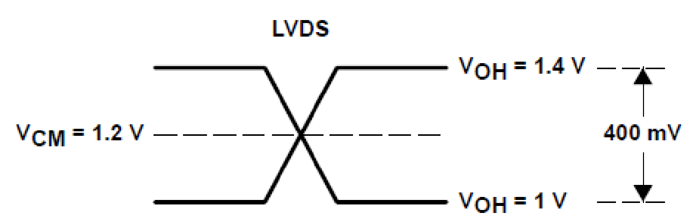

From
Transaction Layer
Tx
Data Link Layer
TLP
Sequence
TLP
LCRC
Replay
Buffer
Mux
DLLP
ACK /
NAK
De-mux
Tx
Rx
To
Transaction Layer
Rx
Data Link Layer
DLLP
ACK /
NAK
TLP
Sequence
TLP
LCRC
De-mux
Mux
Error
Check
Tx
Rx
DLLP
ACK /
NAK
Link
TLP
Sequence
TLP
LCRC

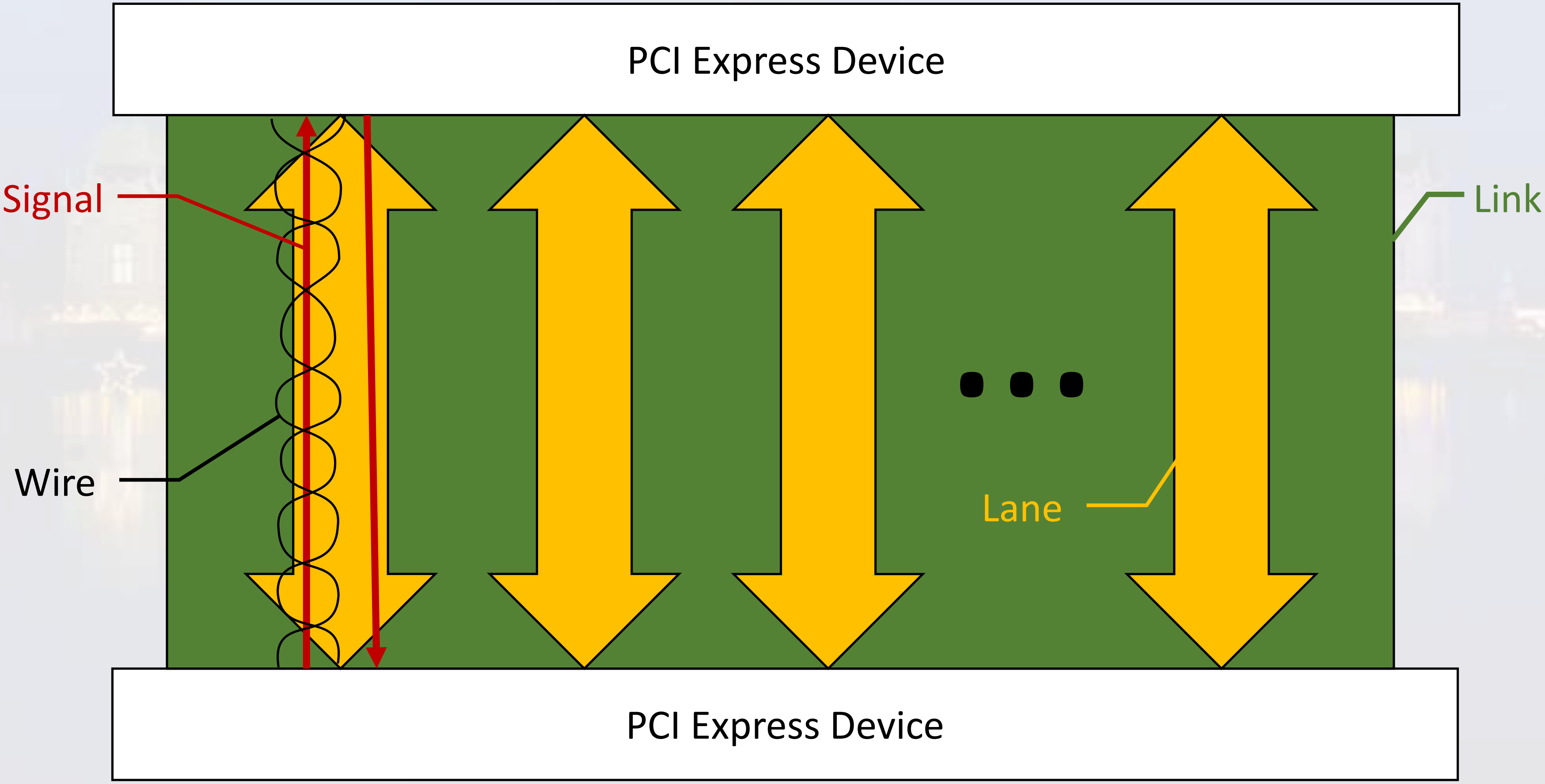
PCIe Link-Training State Machine (LTSSM)

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

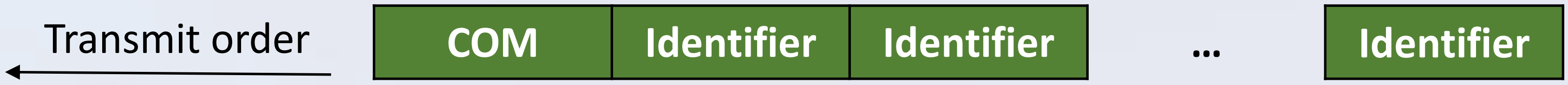


PCIe – Physical layer





PCIe – Ordered-Set Structure



Transmit order

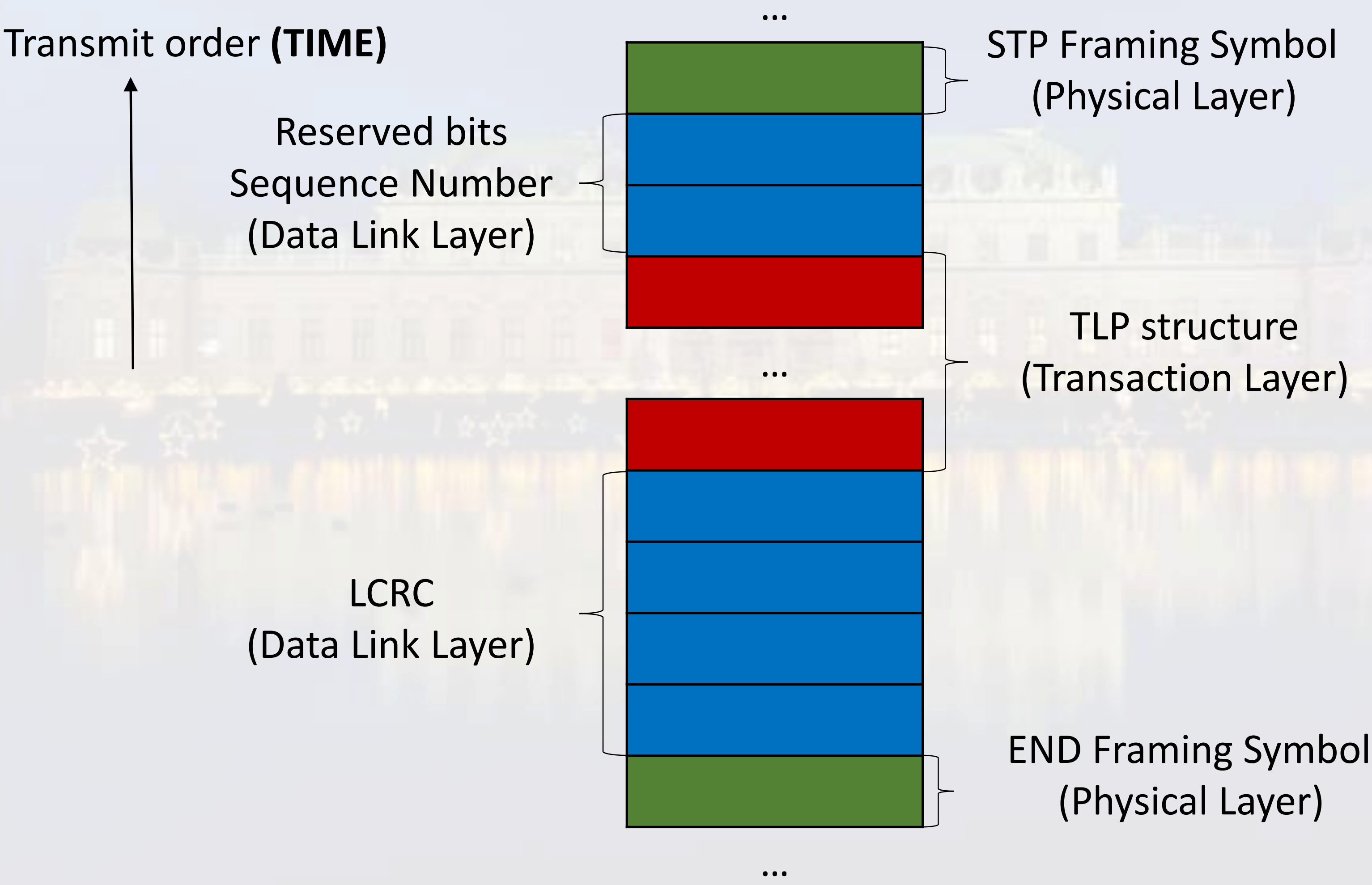
COM | Identifier | Identifier | … | Identifier

Six ordered sets are possible

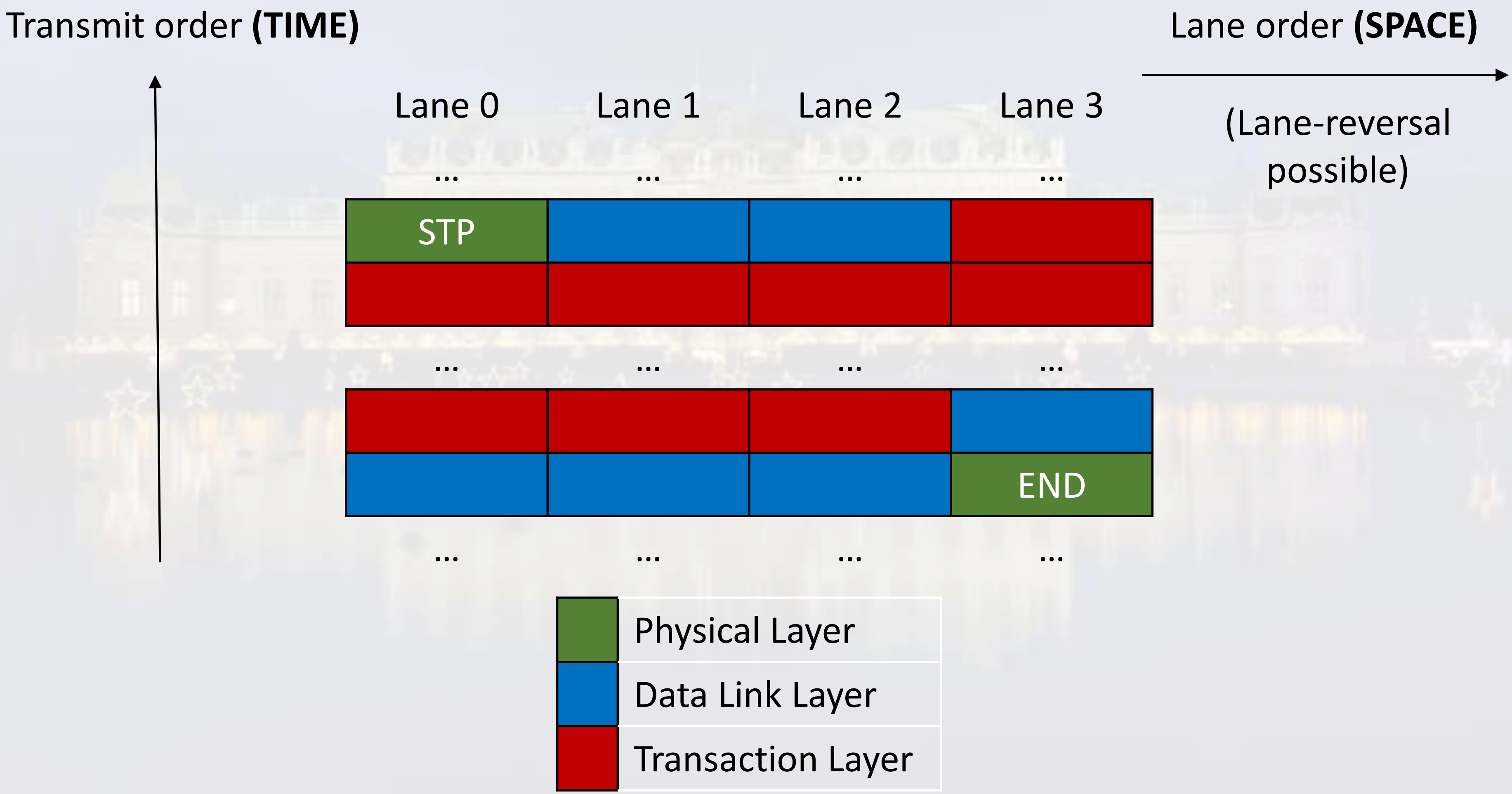
- Training Sequences (TS1, TS2): 1 COM + 15 TS
  - Used to de-skew between lanes
- SKIP: 1 COM + 3 SKP identifiers
  - Used to recalibrate receiver clock
- Fast Training Sequence (FTS): 1 COM + 3 FTS
  - Power management
- Electrical Idle (IDLE): 1 COM + 3 IDL
  - Transmitted continuously when no data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

character: 8 unscrambled bits

PCIe – Framing (x1)



PCIe – Framing (x4)



Transmit order **(TIME)**

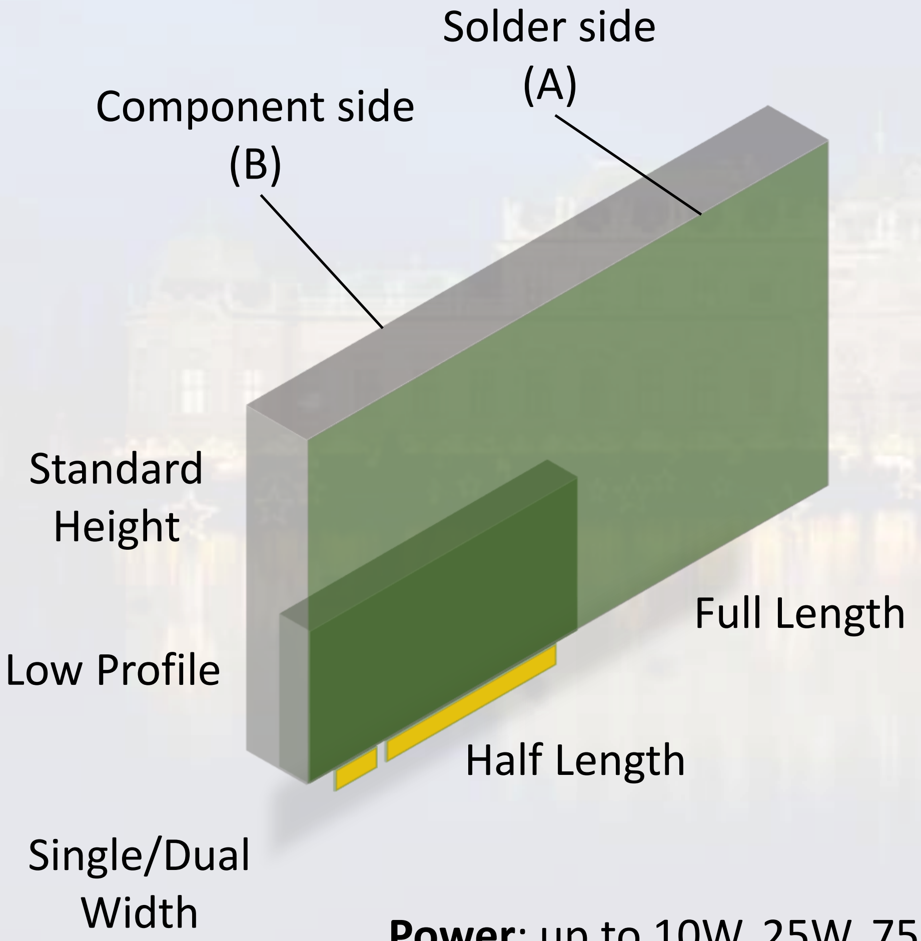
Lane order **(SPACE)**

(Lane-reversal possible)

Lane 0	Lane 1	Lane 2	Lane 3
…	…	…	…
🟩 STP	🟦	🟦	🟥
🟥	🟥	🟥	🟥
…	…	…	…
🟥	🟥	🟥	🟦
🟦	🟦	🟦	🟩 END
…	…	…	…

🟩	Physical Layer
🟦	Data Link Layer
🟥	Transaction Layer

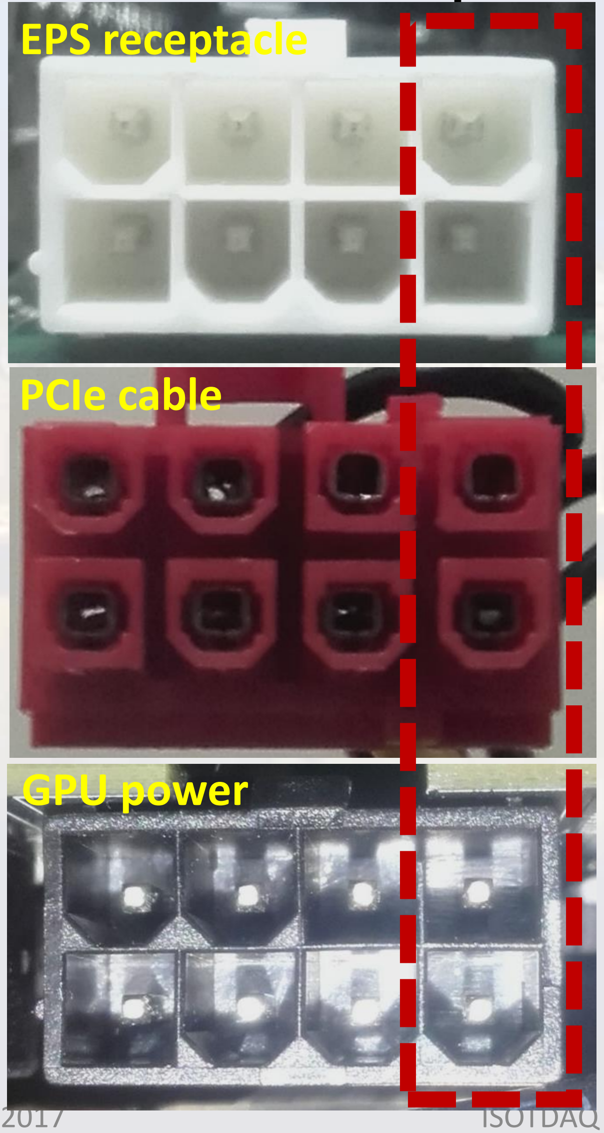
PCIe CEM Spec – AIC form factors

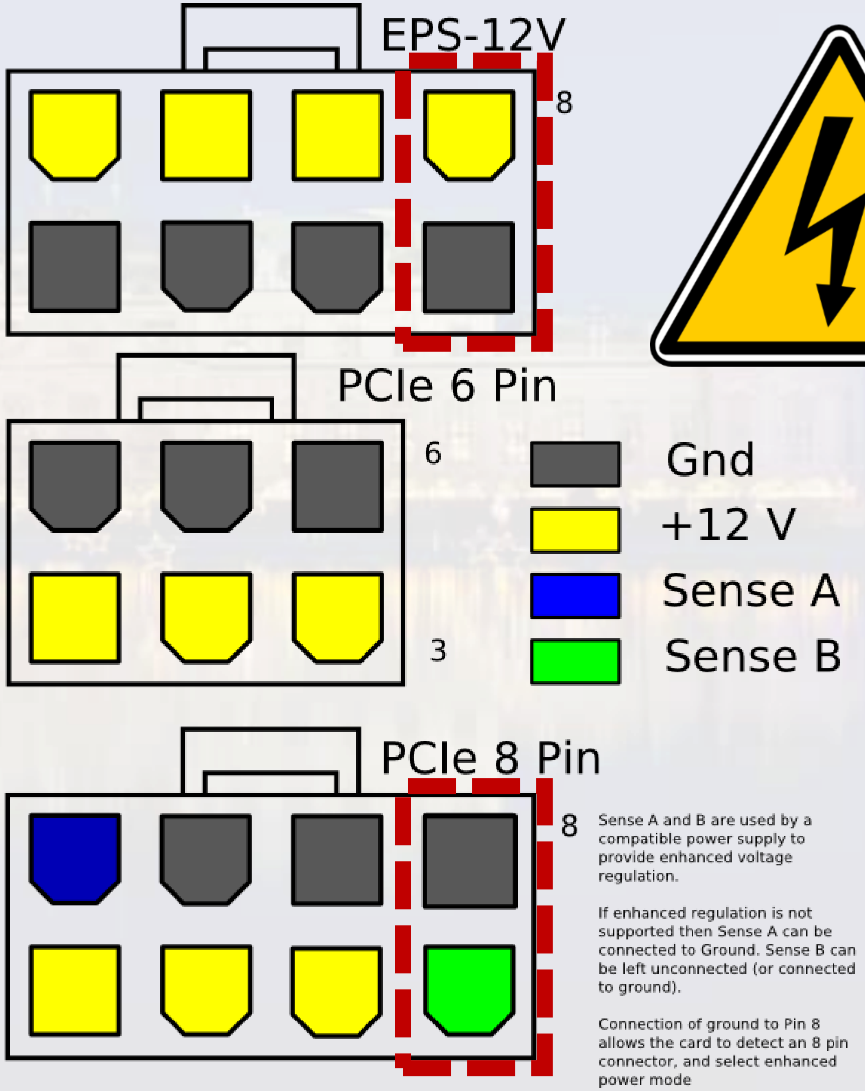


- Standard Height
  - 4.20” (106.7mm)
- Low Profile
  - 2.536” (64.4mm)
- Half Length (e.g.“HHHL”)
  - 6.6” (167.65mm)
- Full Length (e.g. “FHFL”)
  - 12.283” (312mm)

Power: up to 10W, 25W, 75W, 300W or 375W depending on form factor & optional extra power connectors

PCIe CEM Spec – Power Cables





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

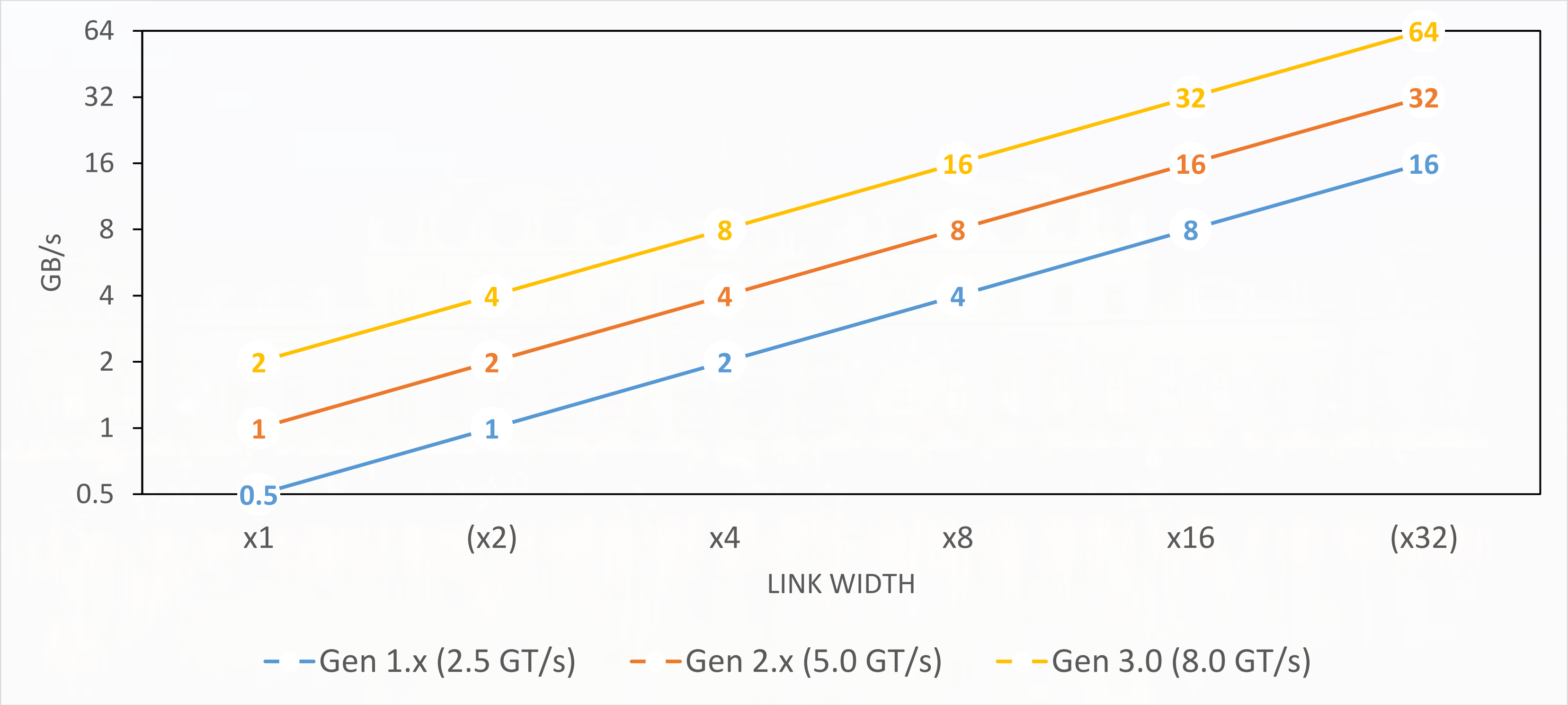
If enhanced regulation is not supported then Sense A can be connected to Ground. Sense B can be left unconnected (or connected to ground).

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

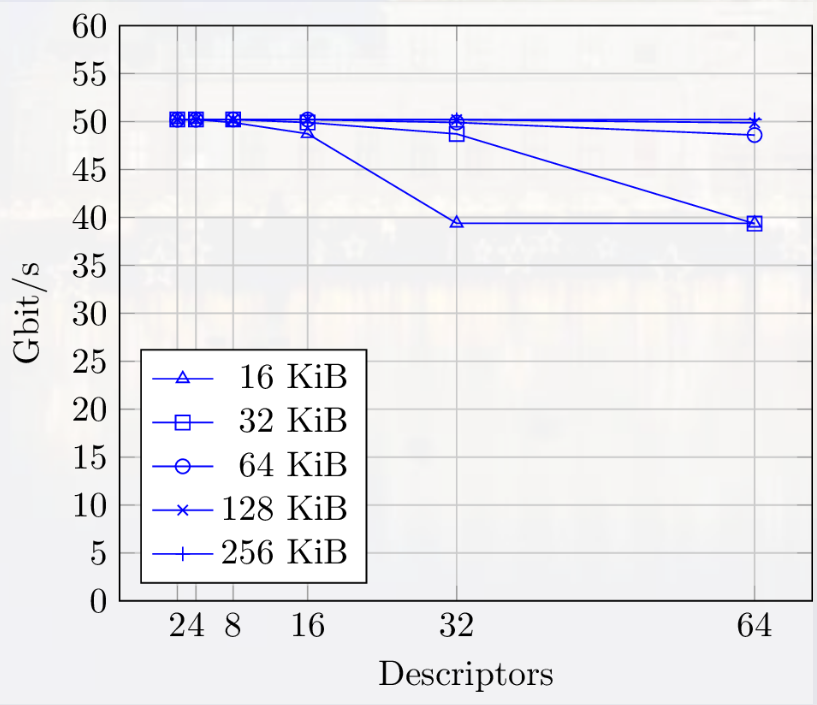
PCIe – Effective data rates

- $\rho = \overbrace{\frac{Lane\ rate \times Lane\ width}{Encoding}}^{\text{Theoretical bandwidth}} \times \overbrace{\frac{MPS}{MPS+Headers}}^{\text{Packet efficiency}}$
- Example: Gen2 x8, 128 Bytes MPS
  - $\rho = 40 \times 0.8 \times \frac{128}{128+24}$ = 32 x 0.84 = 26.9 Gb/s
- Example: Gen3 x8, 128 Bytes MPS
  - $\rho = 64 \times 0.98 \times \frac{128}{128+24}$ = 62.7 x 0.84 = 52.6 Gb/s
- Example: Gen3 x8, 256 Bytes MPS
  - $\rho = 64 \times 0.98 \times \frac{256}{256+24}$ = 62.7 x 0.91 = 57 Gb/s

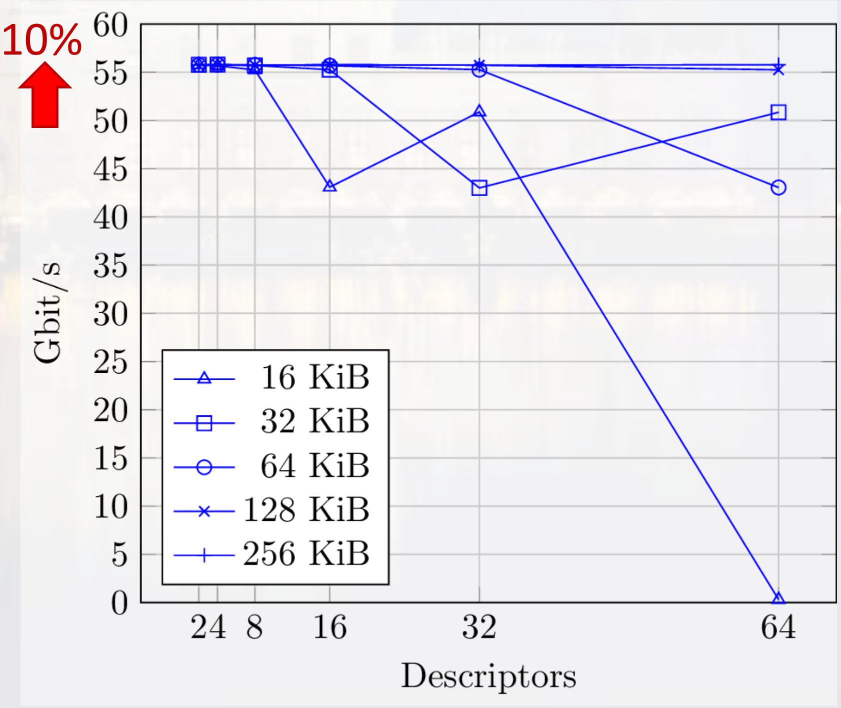
PCIe 3.0 x8 – DMA Performance

MPS = 128 Bytes

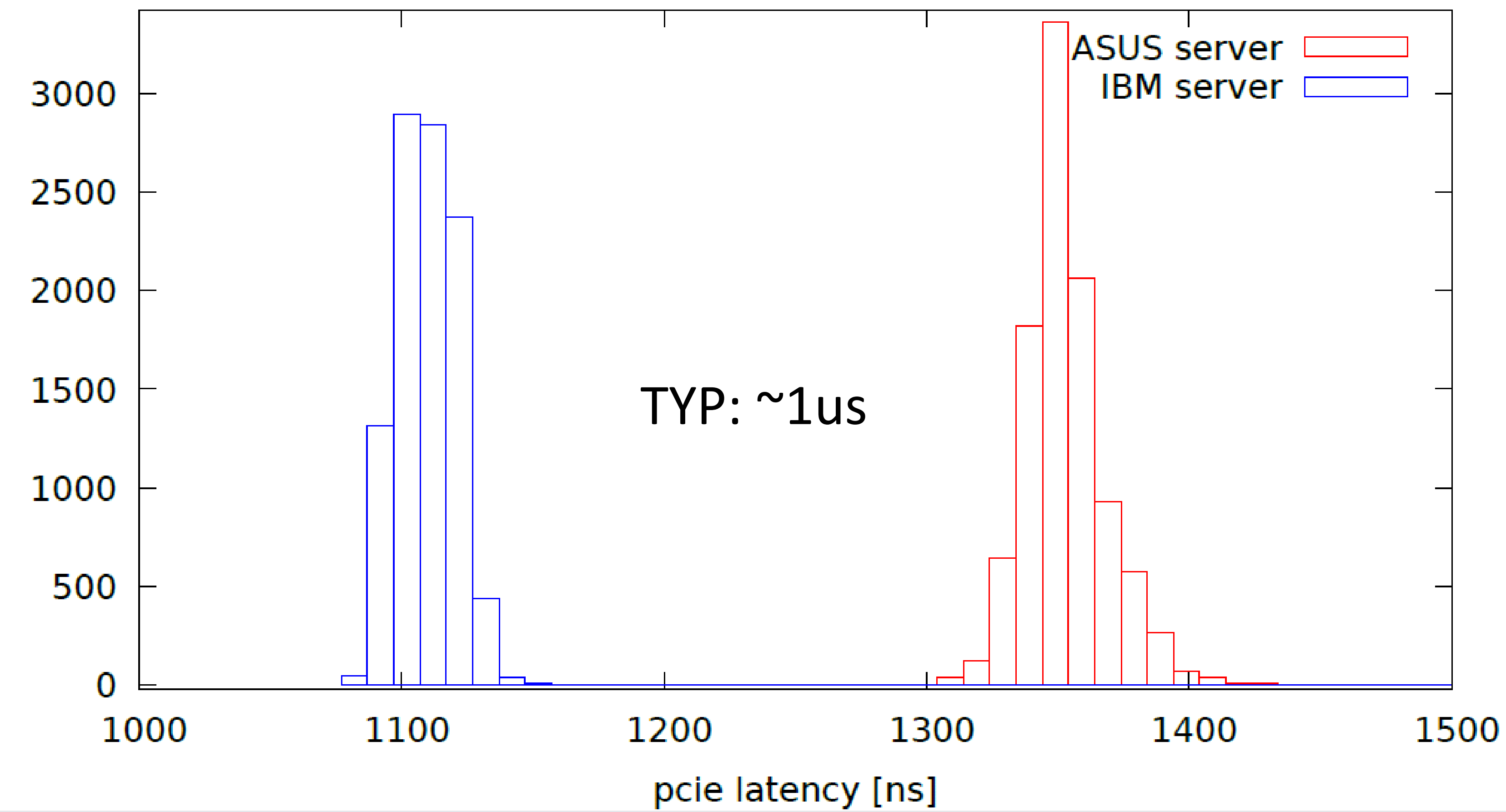
MPS = 256 Bytes



10%



PCIe FPGA – latency

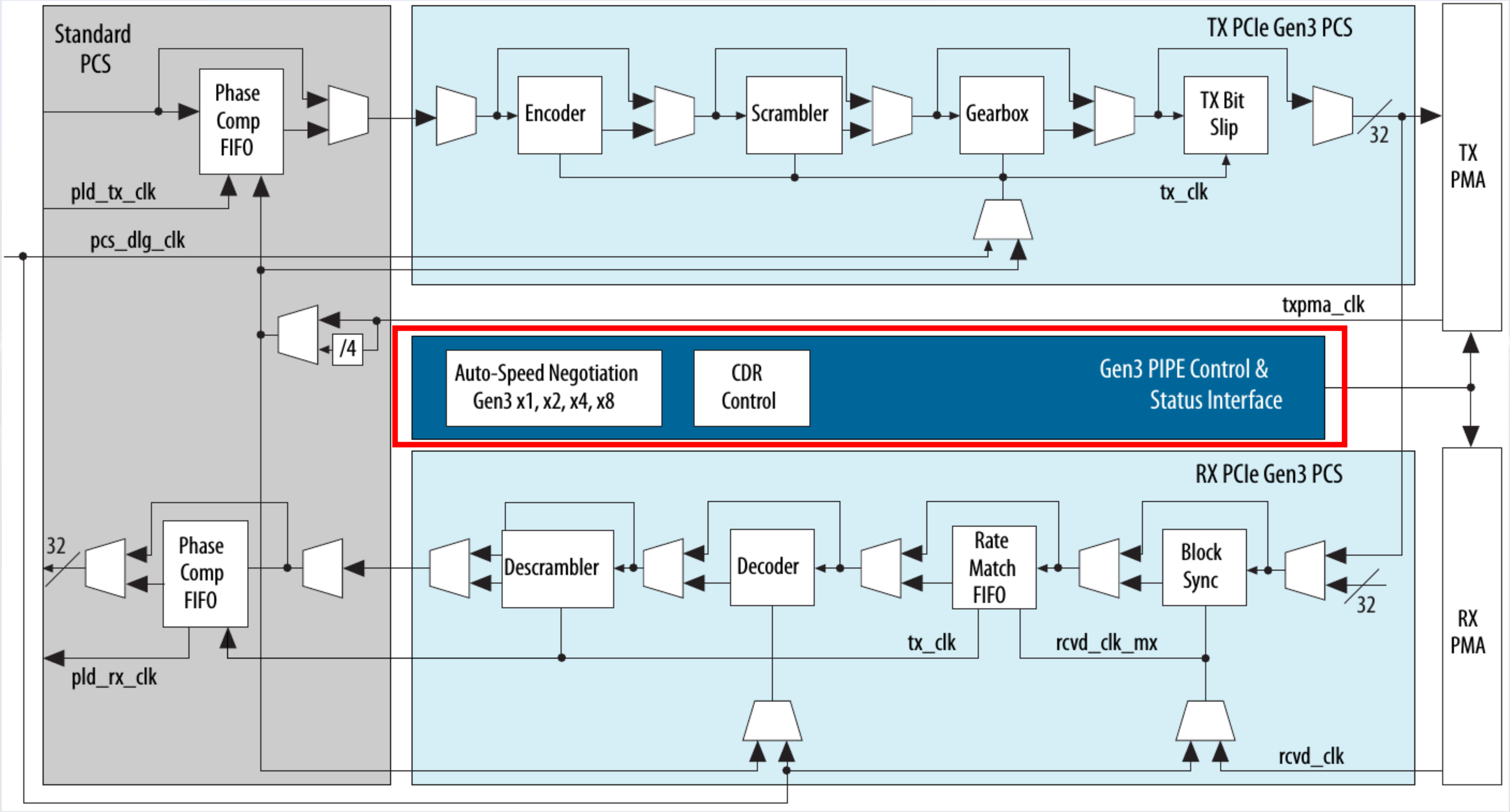


TYP: ~1us

What is this presentation about?

- History and evolution of PCIe
- PCIe concepts
- PCIe layers
- PCIe performance
- **PCIe in practice**

FPGA Hardened PCIe IP

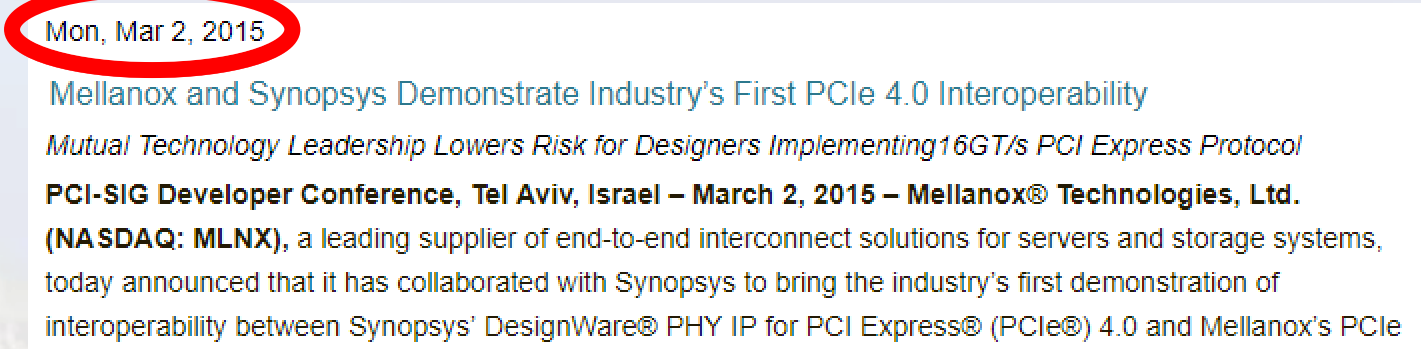


PCIe scalability today

- 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)

PCIe Gen4 – On paper

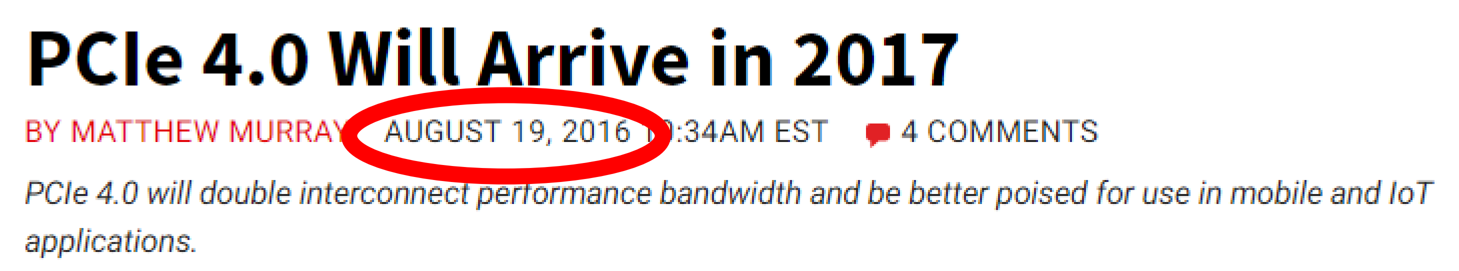


Mon, Mar 2, 2015

Mellanox and Synopsys Demonstrate Industry's First PCIe 4.0 Interoperability

Mutual Technology Leadership Lowers Risk for Designers Implementing16GT/s PCI Express Protocol

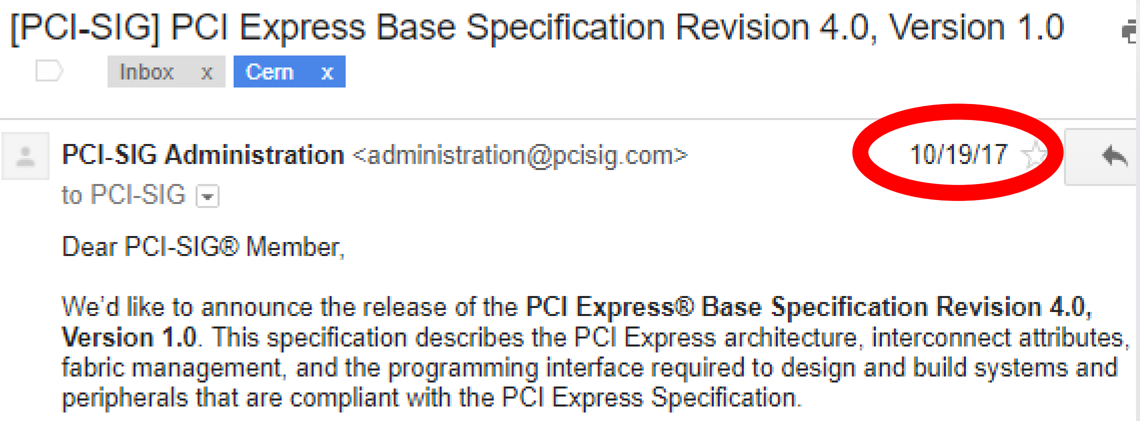
PCI-SIG Developer Conference, Tel Aviv, Israel – March 2, 2015 – Mellanox® Technologies, Ltd. (NASDAQ: MLNX), a leading supplier of end-to-end interconnect solutions for servers and storage systems, today announced that it has collaborated with Synopsys to bring the industry's first demonstration of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



PCIe 4.0 Will Arrive in 2017

BY MATTHEW MURRAY AUGUST 19, 2016 10:34AM EST 4 COMMENTS

PCIe 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> 10/19/17

to PCI-SIG

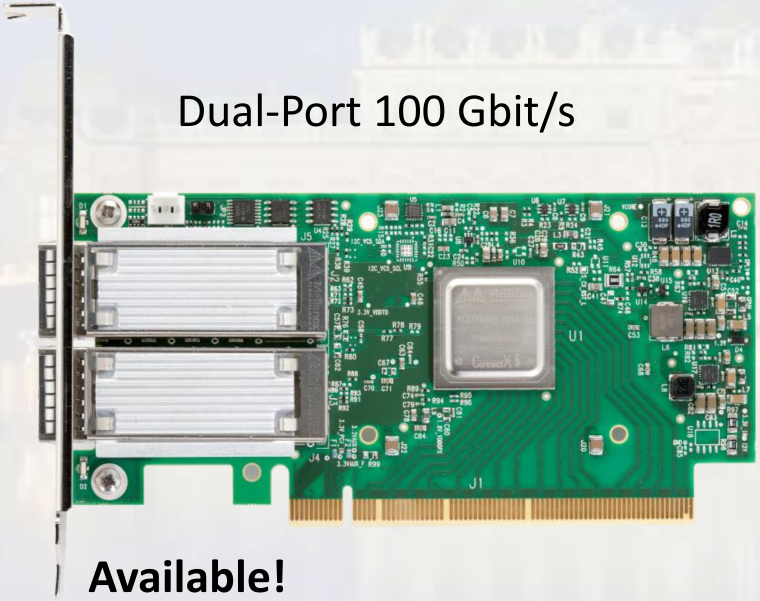
Dear PCI-SIG® Member,

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

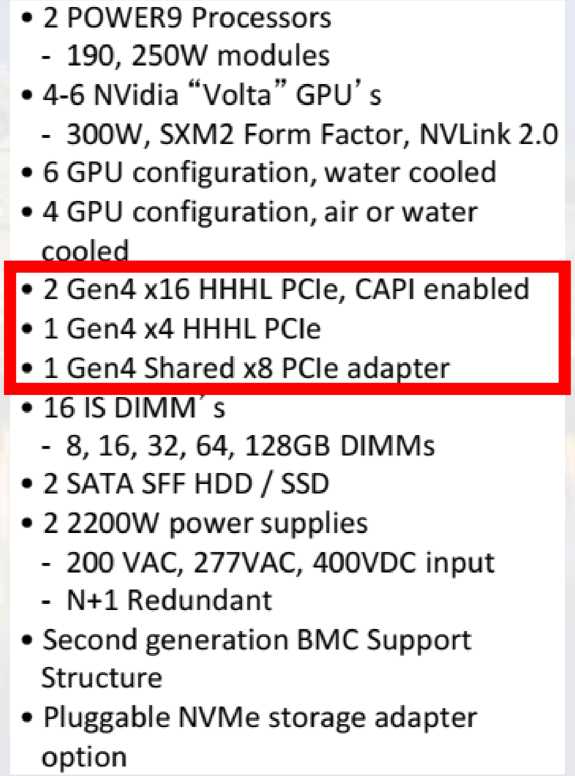
Dual-Port 100 Gbit/s



Available!

```
LnkCap:     Port #0, Speed 16GT/s, Width x16, ASPM L0s L1
```

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

Danke für Ihre Aufmerksamkeit