

PCI Express Basics

Dolphin Interconnect Solutions
Roy Nordstrøm

CONFIDENTIAL Copyright 2016 All rights reserved.

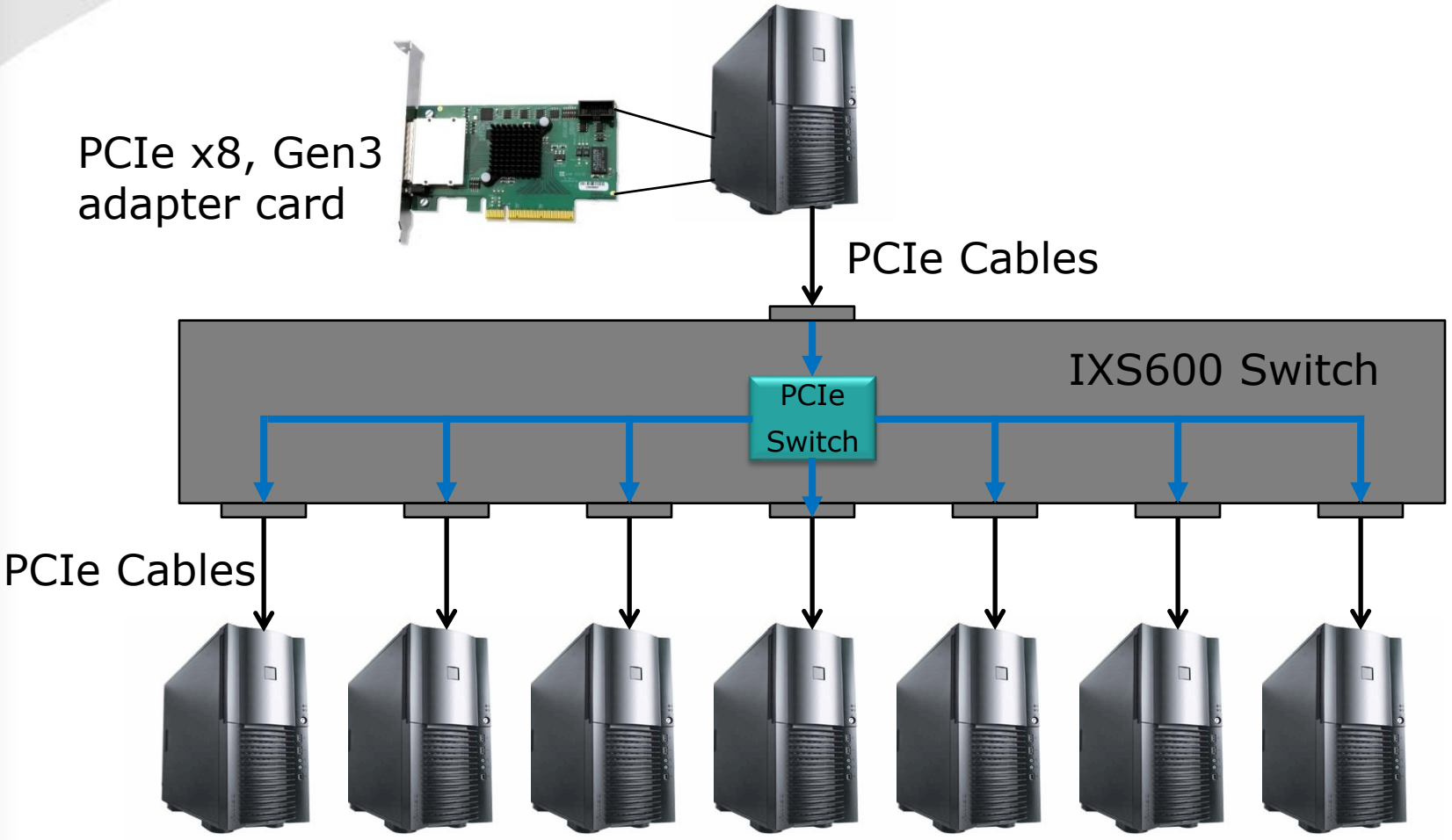
PCI Express (PCIe)

- All PCs and servers use PCI Express as an internal communication protocol
- PCI Express is a packet based protocol
- A high-speed hardware interface for connecting peripheral devices.
- Provides a high-bandwidth scalable solution for reliable data transport
- PCI Express is a serial point-to-point interconnect between two devices
- Scalable performance based on number of signal lanes implemented on the PCI Express interconnect

connect better

CONFIDENTIAL - Copyright 2016 All rights reserved.

Dolphin Cluster – PCIe on the cables

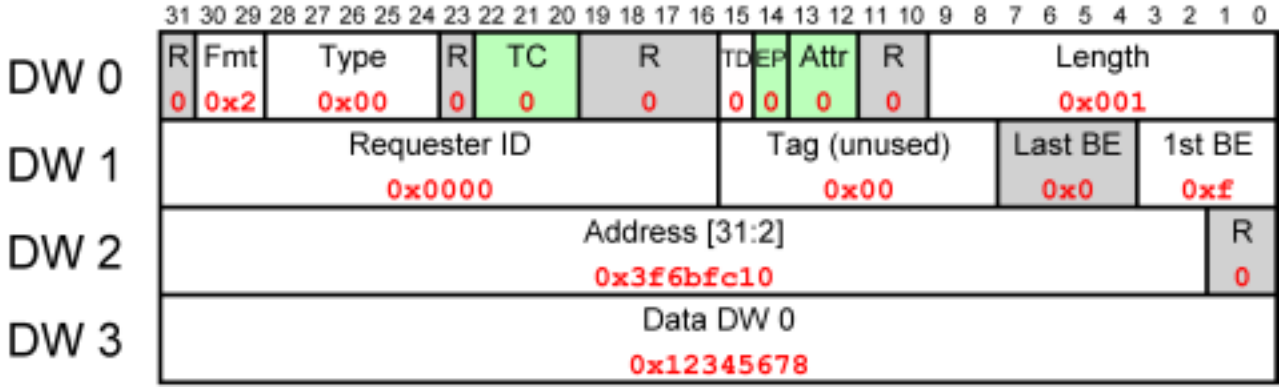


connect better

CONFIDENTIAL - Copyright 2016 All rights reserved.

PCI Express Packets

PCIe Header



PCIe Packet

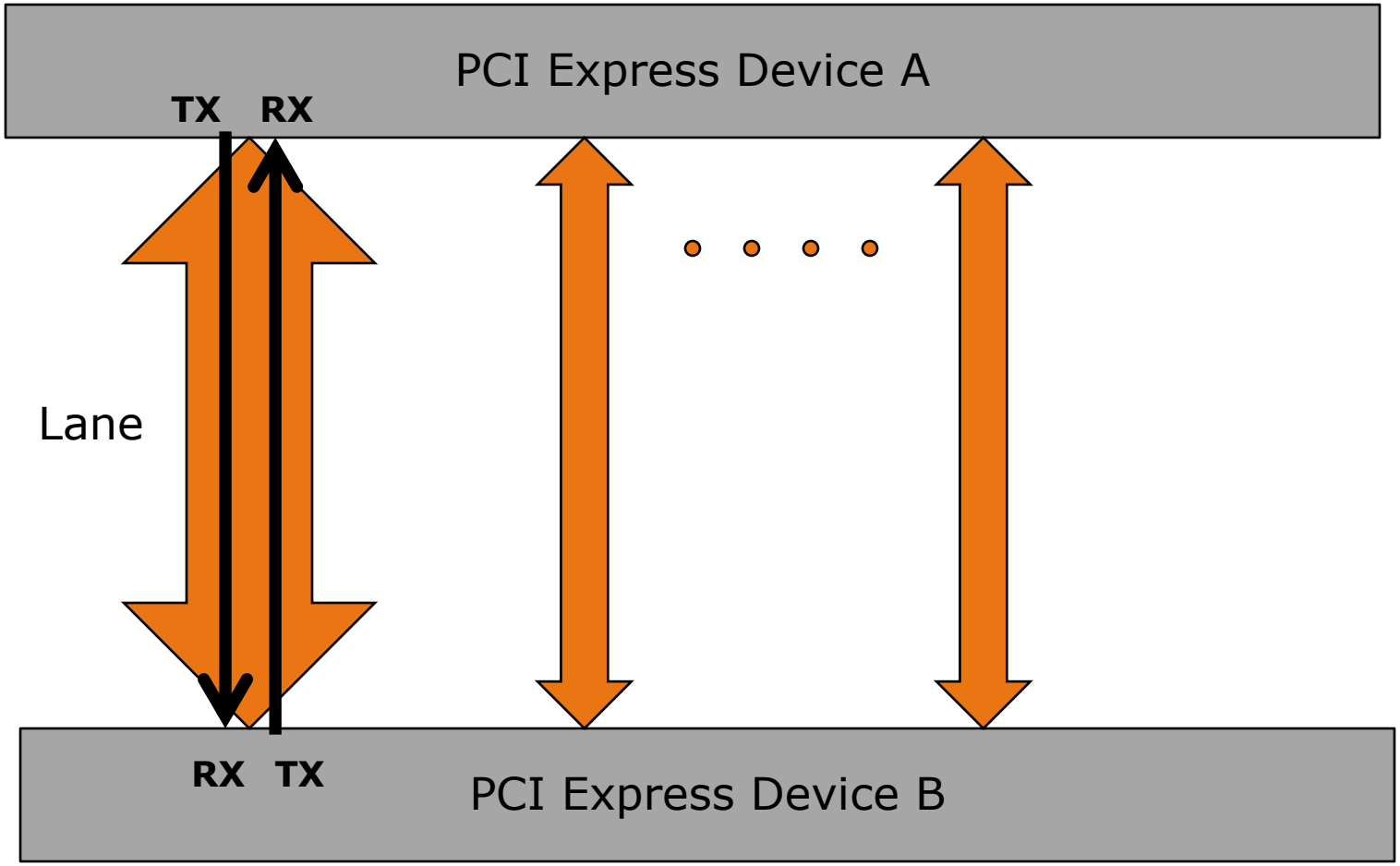
Header 3 – 4 DW	Data 0 - 1023 DW

- MPS (Max Packet Size) is 128, 256 or 512 bytes in real systems

connect better

CONFIDENTIAL - Copyright 2016 All rights reserved.

PCI Express



connect better

CONFIDENTIAL - Copyright 2016 All rights reserved.

PCI Express performance

Gen	Introduced	Transfer rates	X1 GB/s	X2 GB/s	X4 GB/s	X8 GB/s	X16 GB/s
Gen1	2003	2.5 GT/s	0.25	0.5	1.0	2.0	4.0
Gen2	2007	5.0 GT/s	0.5	1.0	2.0	4.0	8.0
Gen3	2010	8.0 GT/s	0.9	1.9	4.0	7.8	15.7
Gen4	2019	16.0 GT/s	2.0	3.9	7.8	15.7	31.5
Gen5	2021	32.0 GT/s	3.0	7.8	15.7	31.5	63.0
Gen6	Planned	64.0 GT/s	7.8	15.7	31.5	63.0	126.0

connect better

CONFIDENTIAL - Copyright 2016 All rights reserved.

DMA Gen4 x16 bandwidth – Dolphin MXH930

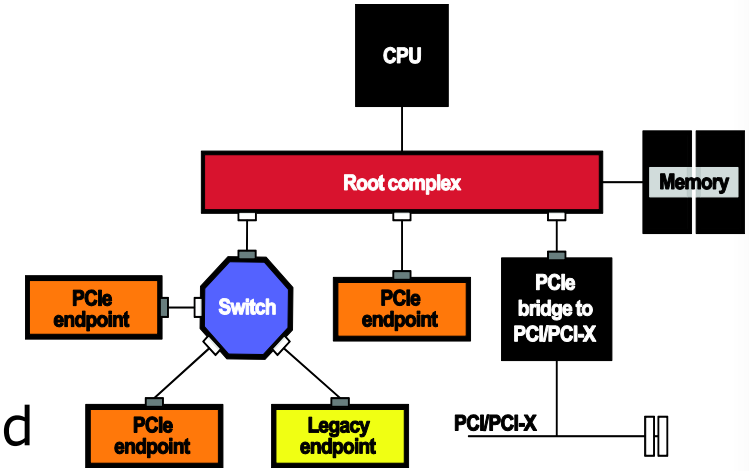
Message size	DMA bandwidth
64	630.03 MBytes/s
128	1317.57 MBytes/s
256	2589.54 MBytes/s
512	5005.61 MBytes/s
1024	10767.44 MBytes/s
2048	19233.57 MBytes/s
4096	23303.24 MBytes/s
8192	25521.49 MBytes/s
16384	26691.85 MBytes/s
32768	27358.14 MBytes/s
65536	27704.01 MBytes/s
131072	27871.21 MBytes/s
262144	27956.97 MBytes/s
524288	27977.95 MBytes/s
1048576	27988.64 MBytes/s
2097152	27989.77 MBytes/s
4194304	27990.02 MBytes/s
8388608	27990.08 MBytes/s
16777216	27990.08 MBytes/s
33554432	27990.19 MBytes/s
67108864	27990.23 MBytes/s

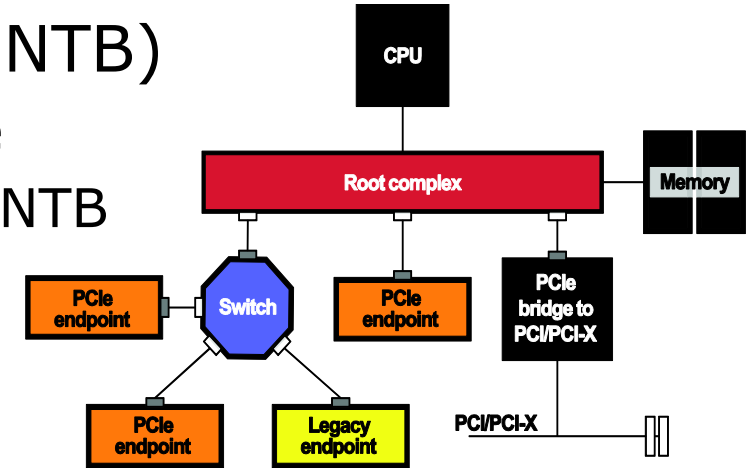
connect better

CONFIDENTIAL - Copyright 2016 All rights reserved.

PCI Express

- Switching
  - Switch based topologies
- Transparent Bridging
  - Root complex able to see all end point in system
- Non-Transparent Bridging (NTB)
  - PCIe endpoint is a NTB bridge
  - Software used to see beyond NTB Bridge

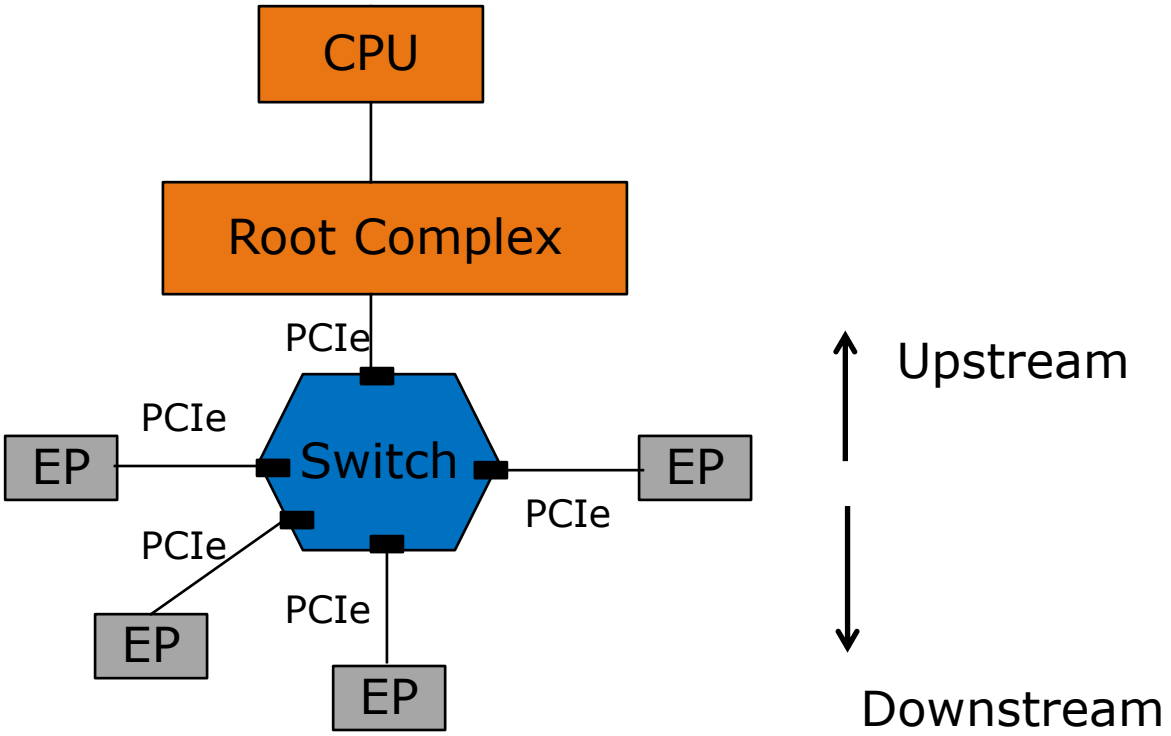




connect better

CONFIDENTIAL - Copyright 2016 All rights reserved.

PCI Express – Upstream/Downstream



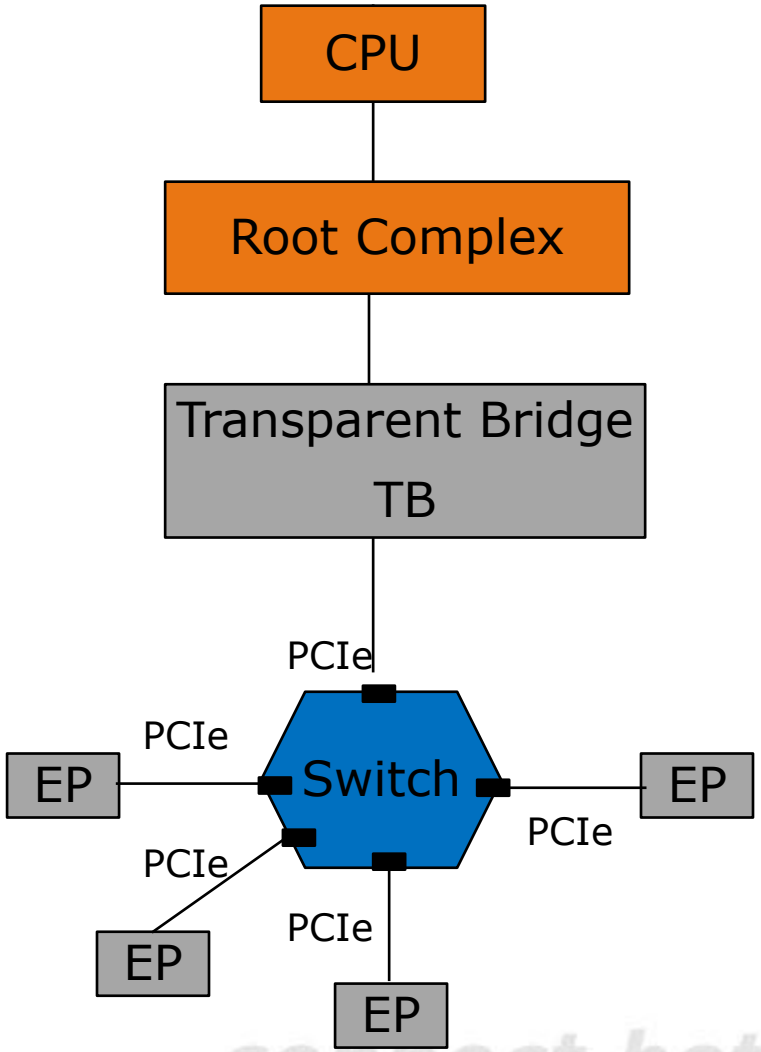
- An Upstream Port is a port that points in the direction of the root complex.
- A Downstream Port is a port that points away from the root complex.

connect better

CONFIDENTIAL - Copyright 2016 All rights reserved.

PCI Express – Transparent Bridge (TB)

- No software setup

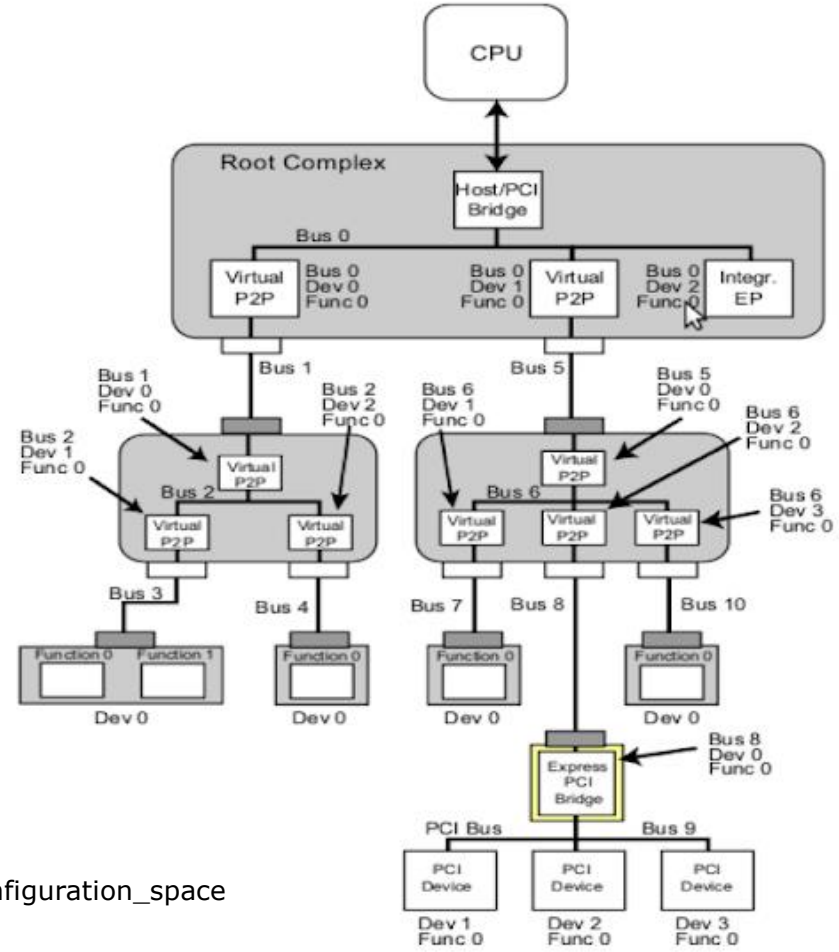


connect better

CONFIDENTIAL - Copyright 2016 All rights reserved.

PCIe Enumeration

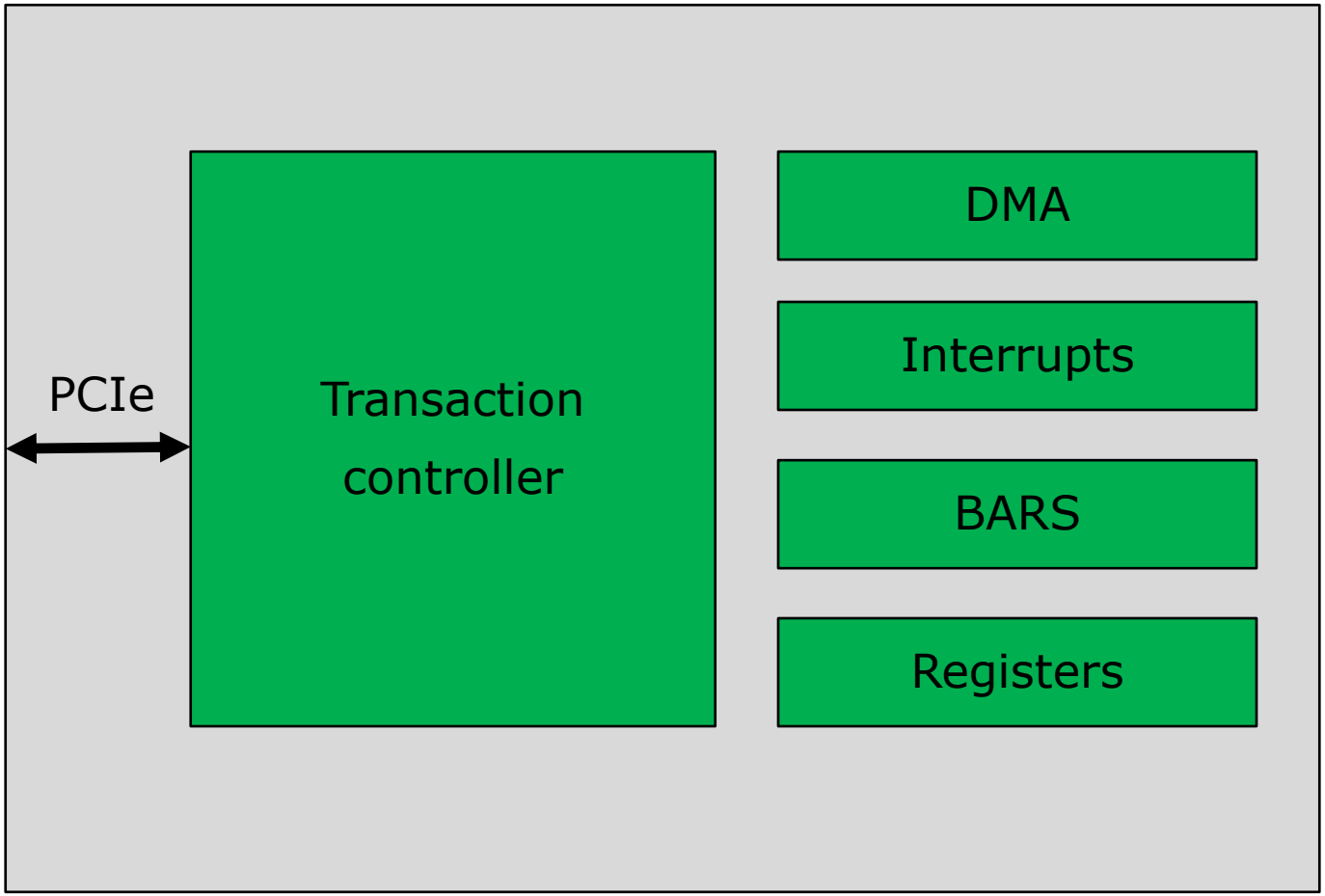
- Each endpoint is initialized with a **requester ID** which consist of a bus number, device number and function number



Reference
https://en.wikipedia.org/wiki/PCI_configuration_space

CONFIDENTIAL - Copyright 2016 All rights reserved.

Endpoint (EP)

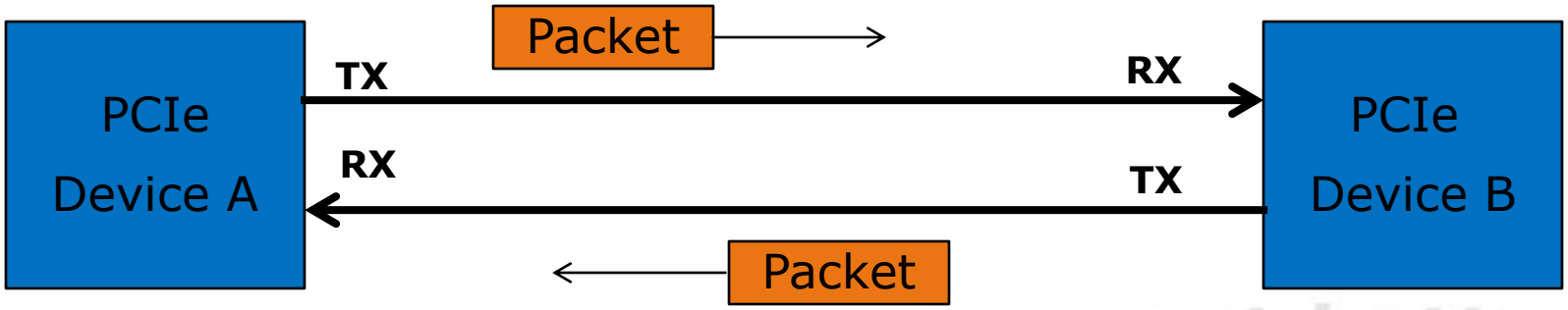


connect better

CONFIDENTIAL - Copyright 2016 All rights reserved.

PCI Express features

- Point to point connection
- Scalable: x1, x2, x4, x8, x16, x32
- Packet based transaction protocol
- Differential Signaling

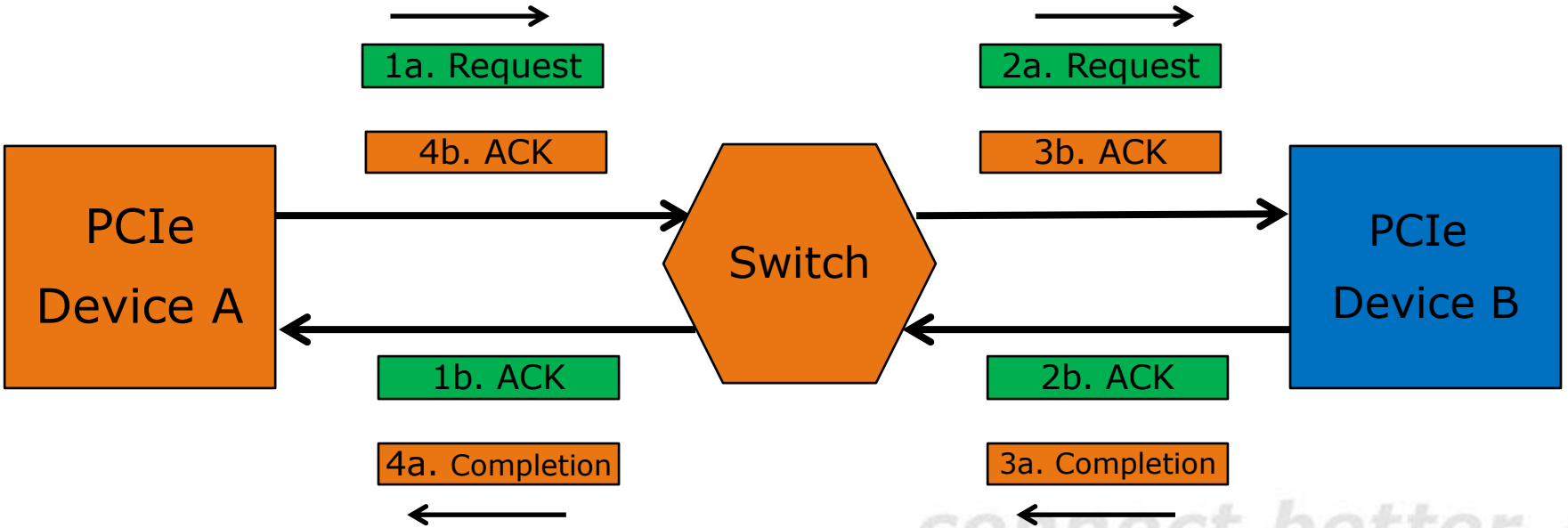


connect better

CONFIDENTIAL - Copyright 2016 All rights reserved.

ACK/NAK Protocol: Point-to-Point

- Point to point connection
- ACK returned for good reception of Request or Completion
- NACK returned for error reception of Request or Completion



connect better

CONFIDENTIAL - Copyright 2016 All rights reserved.

Link training

- The link negotiate to find the appropriate link speed
- The devices send known, ordered sets of symbols to each other and the hardware works its way up from 2.5GT/s.
- Commands to change the link speed is sent to each other
- If the link is unstable, the link can train down again

connect better

CONFIDENTIAL - Copyright 2016 All rights reserved.

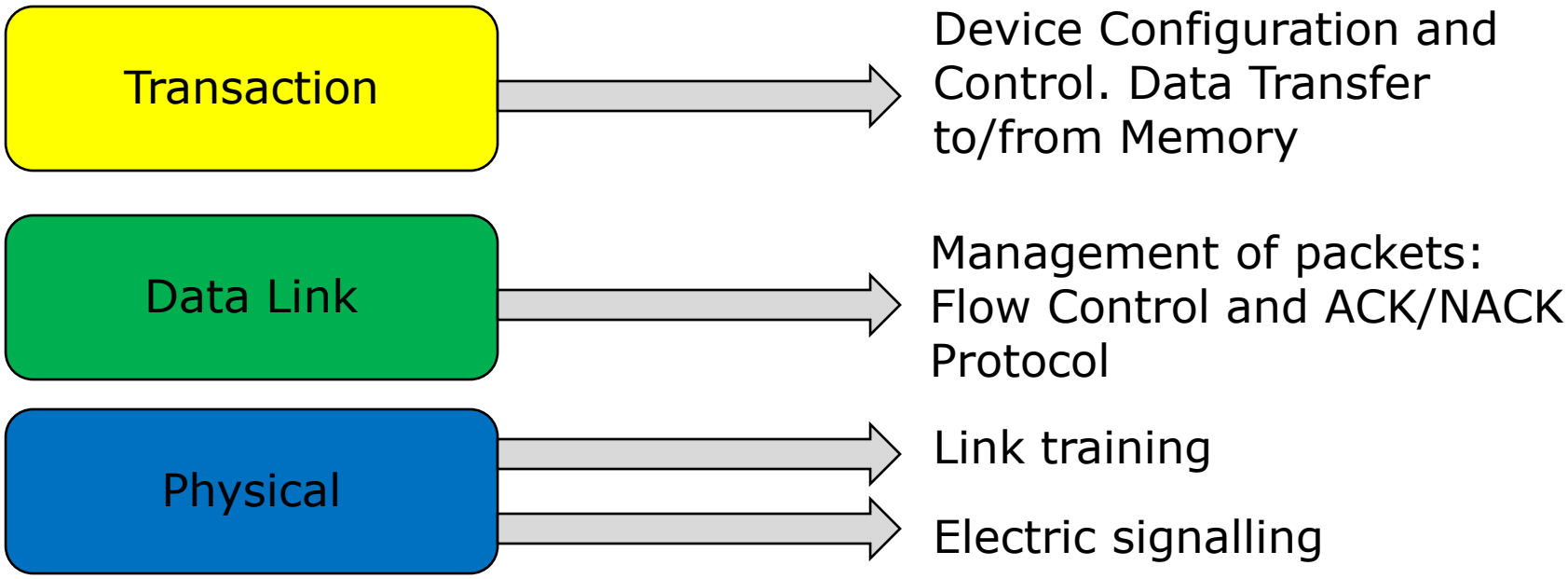
Transaction Layer Packet (TLP)

- A PCI Express system transfers data in the payload of TLPs.
- Memory data is transferred in Memory Write/Read TLPs
- Completion with Data TLPs are responses to memory read operations.

connect better

CONFIDENTIAL - Copyright 2016 All rights reserved.

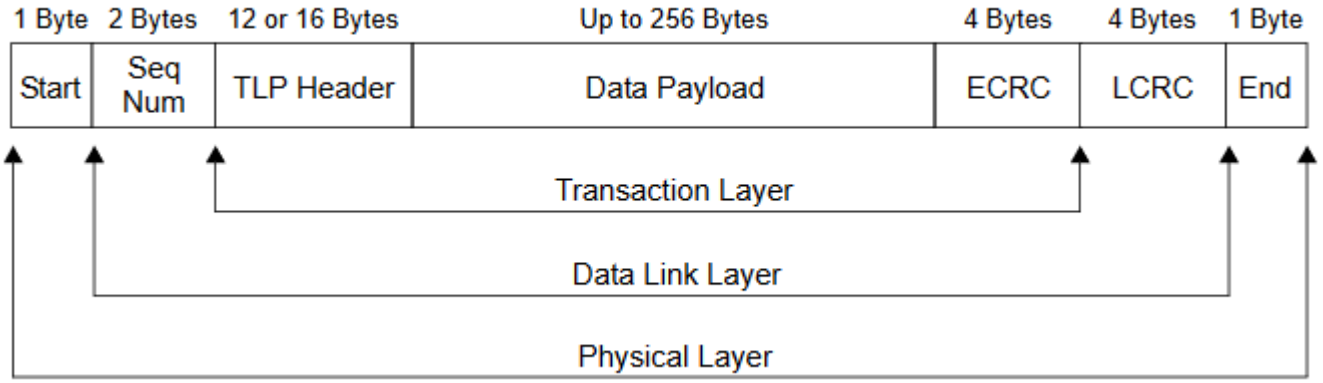
PCI Express Layered Model



connect better

CONFIDENTIAL - Copyright 2016 All rights reserved.

PCI Express Layered Model



connect better

CONFIDENTIAL - Copyright 2016 All rights reserved.

Correctable and uncorrectable link errors

- Link errors can occur on the PCIe link
- There a three different error classifications
  - Correctable errors
  - Uncorrectable Non-fatal errors
  - Uncorrectable Fatal errors

connect better

CONFIDENTIAL - Copyright 2016 All rights reserved.

Correctable errors

- Correctable errors are handled by the hardware
- Such errors may have inpact on performance
- No information is lost as a result of the error
- Retransmitted by the hardware

connect better

CONFIDENTIAL - Copyright 2016 All rights reserved.

Uncorrectable Non-Fatal Errors

- The information has been lost
- Non-fatal errors means that transaction has been corrupted
- PCI Express hardware cannot correct the errors
- The PCI Express fabric continues to function correctly and other transactions are unaffected
- Is handled by device software

connect better

CONFIDENTIAL - Copyright 2016 All rights reserved.

Uncorrectable Fatal Errors

- Indicates that a link in a PCI Express fabric is no longer reliable
- Data has been lost and every attempt to recover data under software control can fail
- Fatal errors may be resolved by resetting the link
- On the port link, the software will retransmit the lost data

connect better

CONFIDENTIAL - Copyright 2016 All rights reserved.