PCIe Range Extension via Robust, Long Reach Protocol Tunnels

Jim Peek

Director of Technology

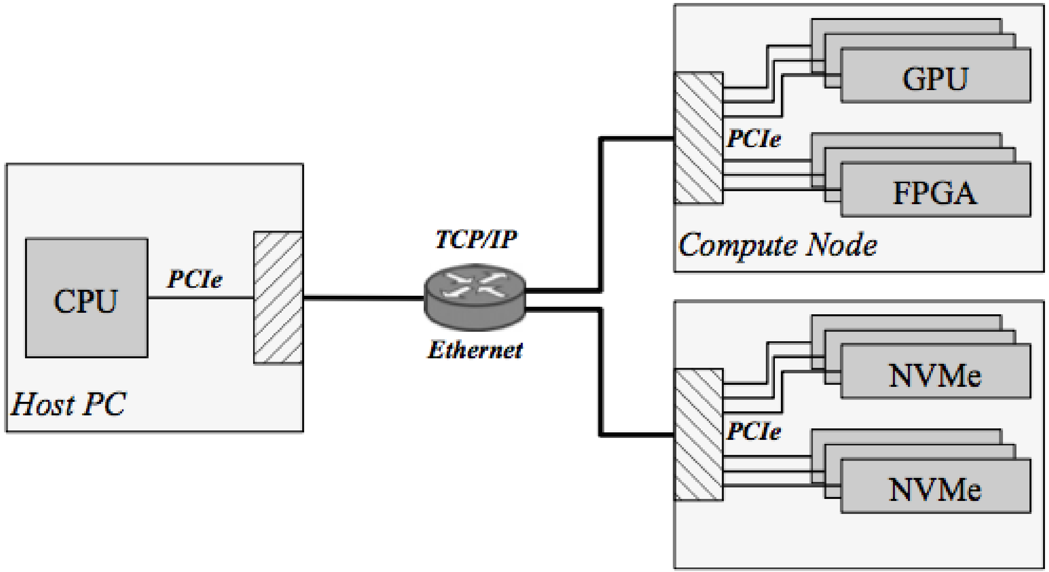
Missing Link Electronics

Disclaimer

Presentation Disclaimer: All opinions, judgments, recommendations, etc. that are presented herein are the opinions of the presenter of the material and do not necessarily reflect the opinions of the PCI-SIG®.

Motivation

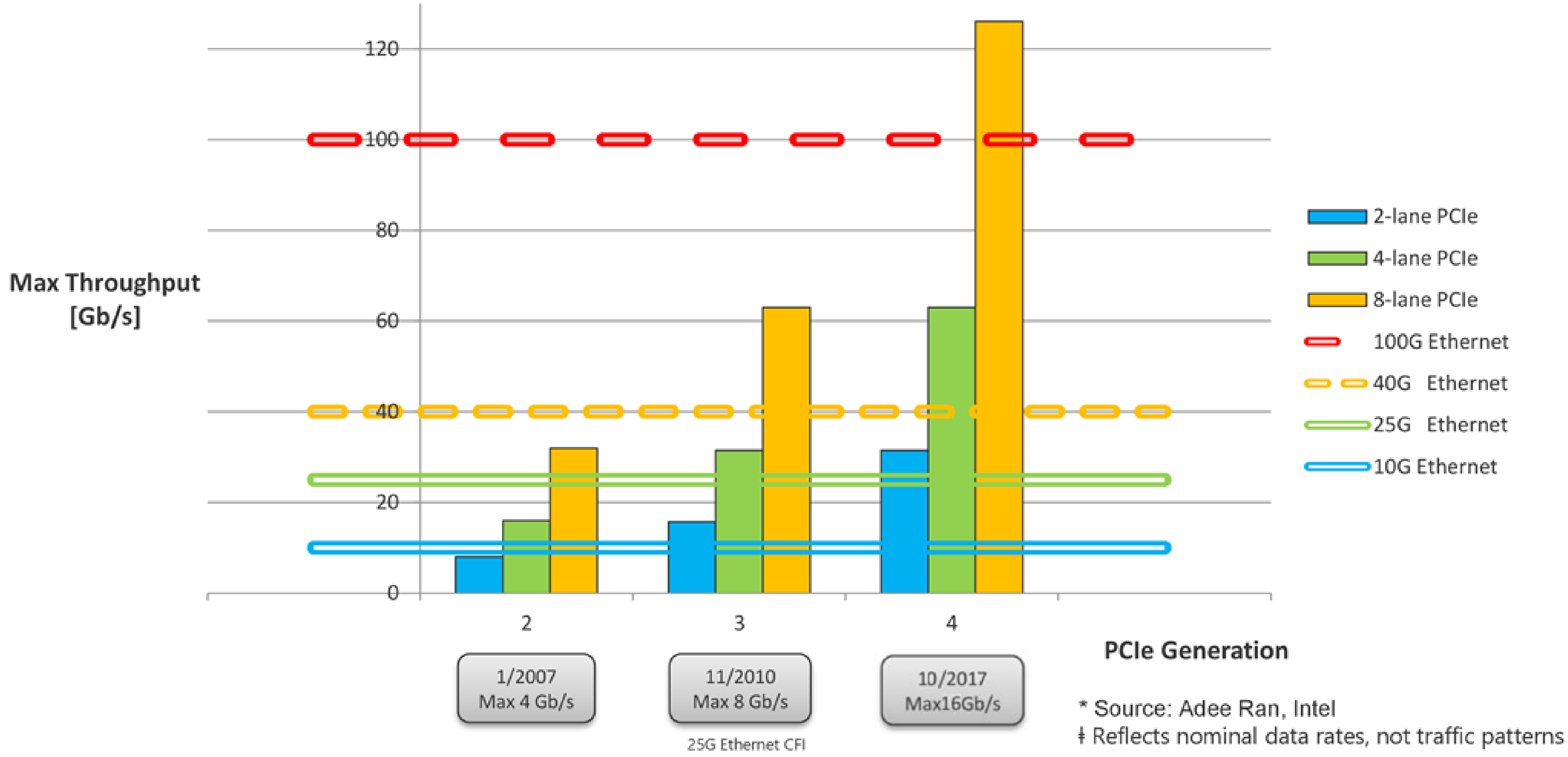
PCI Express over IP



Example: IP based from Host complex PCI to PCI end point devices

Introduction

PCIe to Ethernet Bandwidth Matching

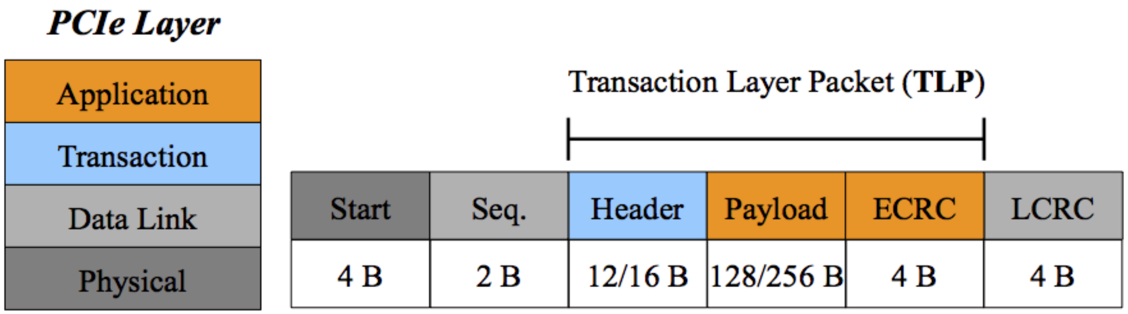


* Source: Adee Ran, Intel

ǂ Reflects nominal data rates, not traffic patterns

PCI Express

- PCIe replaces the PCI Local Bus (backwards compatible)
- Full-duplex serial transmission
- At 8GT/s line rate (Gen3) on up to 32 lanes
- Packet-based protocol with four layers



PCIe Layer

Application

Transaction

Data Link

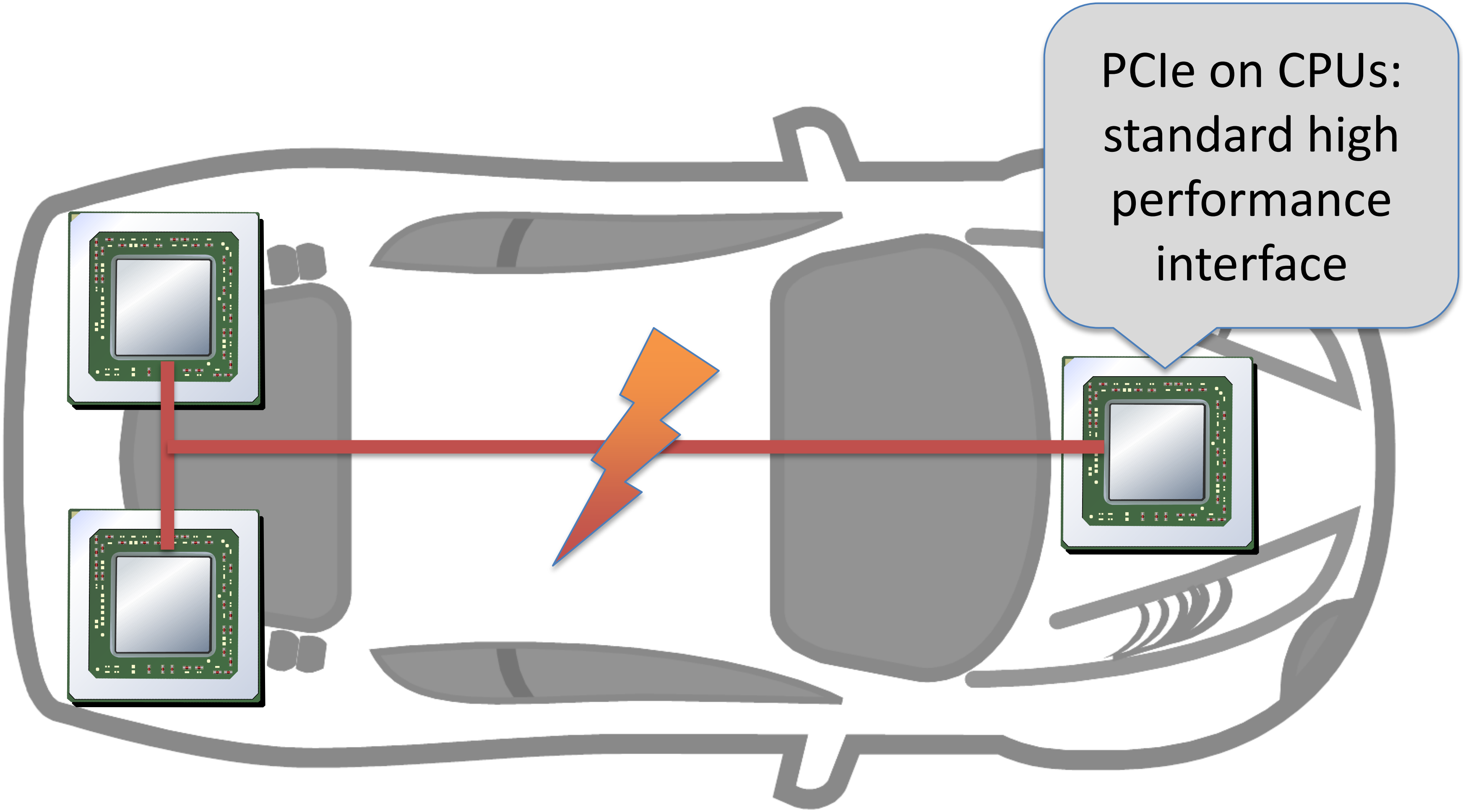
Physical

Transaction Layer Packet (**TLP**)

Start	Seq.	Header	Payload	ECRC	LCRC
4 B	2 B	12/16 B	128/256 B	4 B	4 B

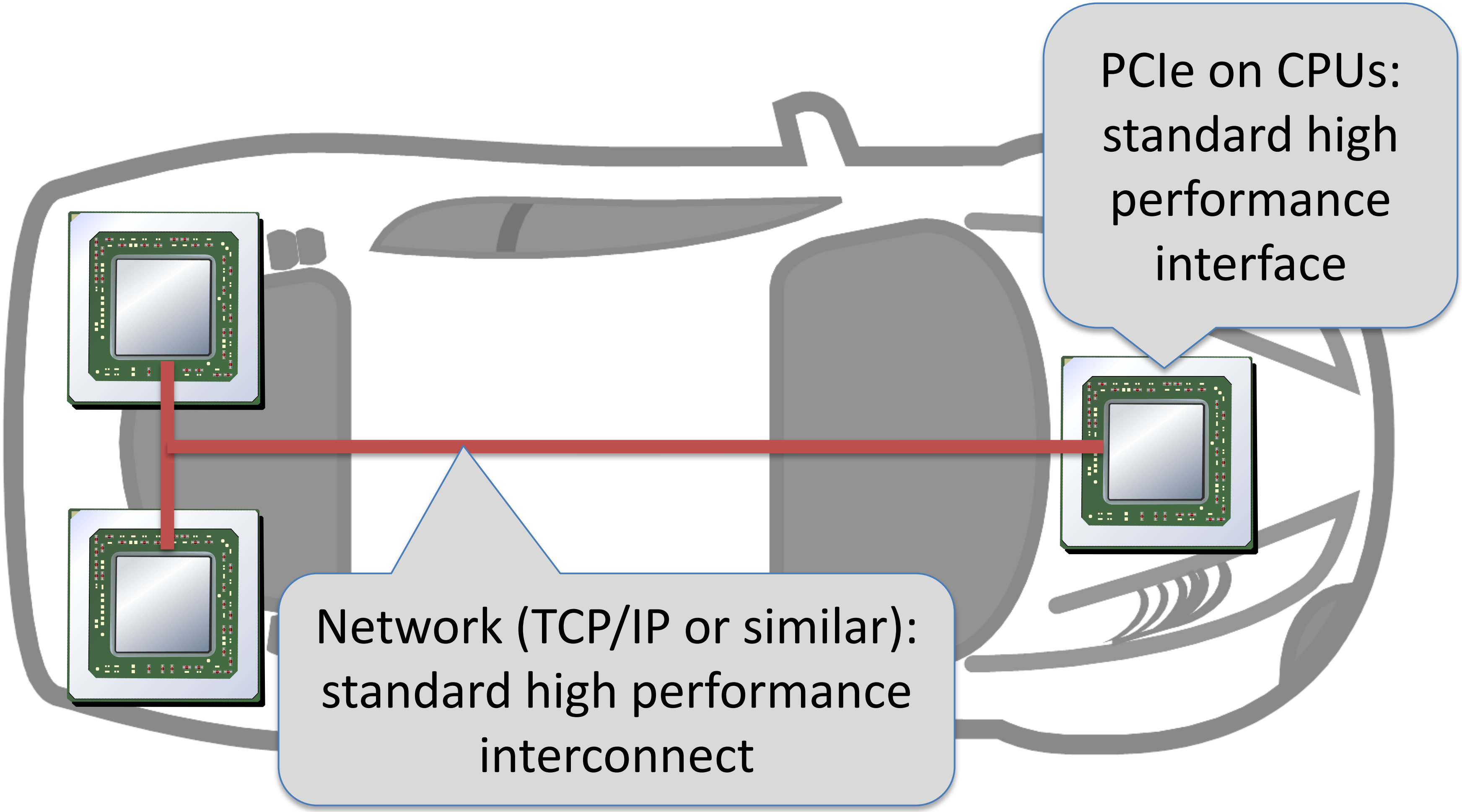
- Data Link layer, physical layer: Reliable transport on the link
- Transaction layer:
  - Transport of application data, device configuration, interrupts, Quality of service
  - TLP categories: Memory, I/O, configuration, message

Motivation



PCIe on CPUs: standard high performance interface

Motivation

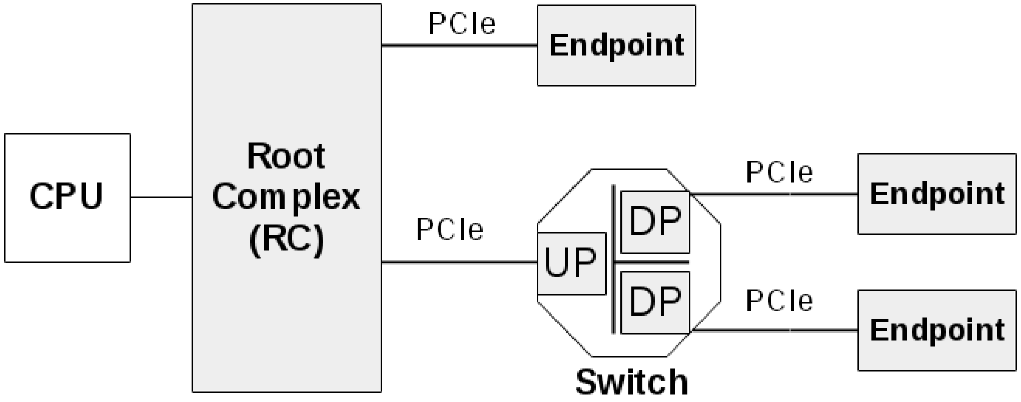


PCIe on CPUs: standard high performance interface

Network (TCP/IP or similar): standard high performance interconnect

PCI Express Topology

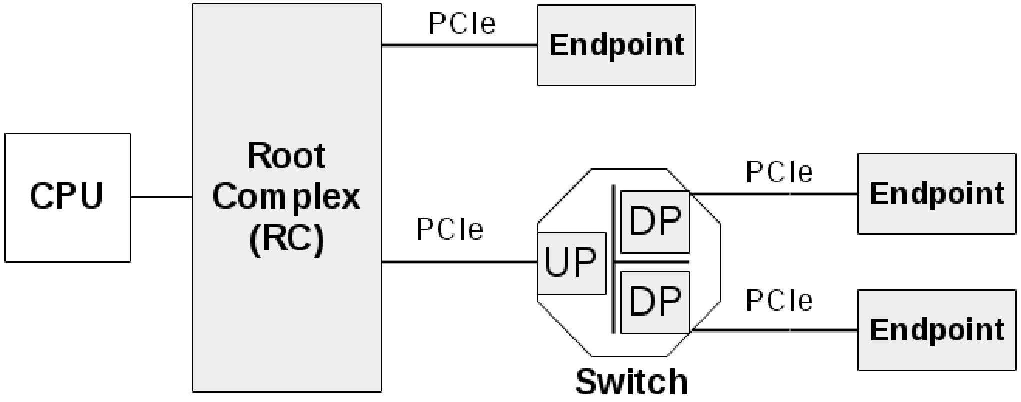
- PCIe is point-to-point. Hierarchical system topologies via switches



- ID based routing (bus/device/function number) and address based routing
- Transactions may require completion (posted or split-transaction)
- Range problem: Physical line length of PCIe on PCBs is limited to centimeter range

PCI Express Topology

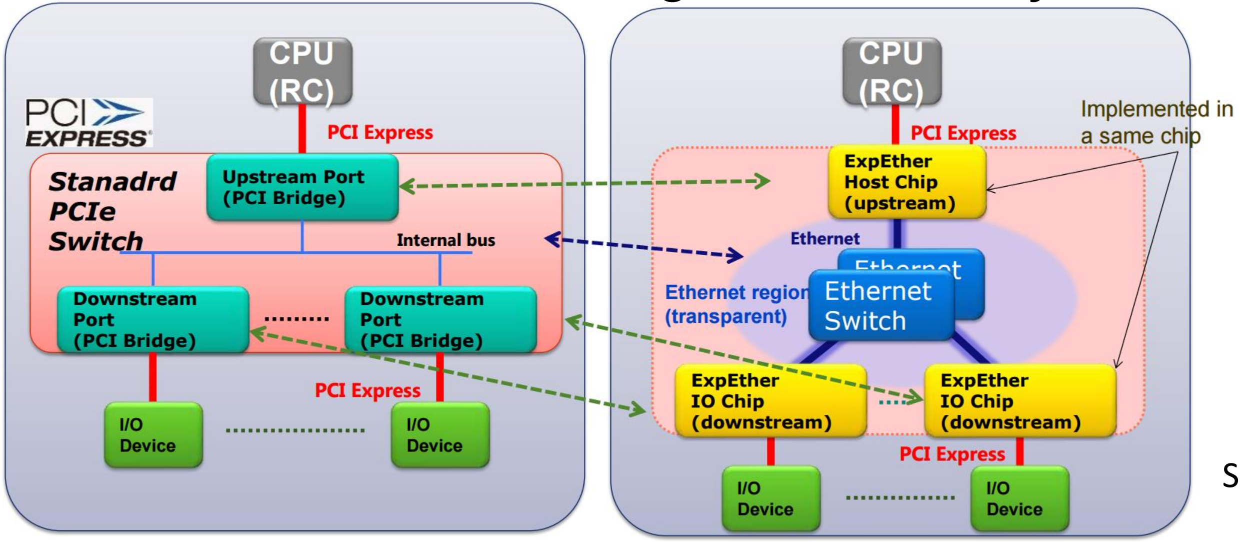
- PCIe is point-to-point. Hierarchical system topologies via switches



- ID based routing (bus/device/function number) and address based routing
- Transactions may require completion (posted or split-transaction)
- Range problem: Physical line length of PCIe on PCBs is limited to centimeter range

State-of-the-Art

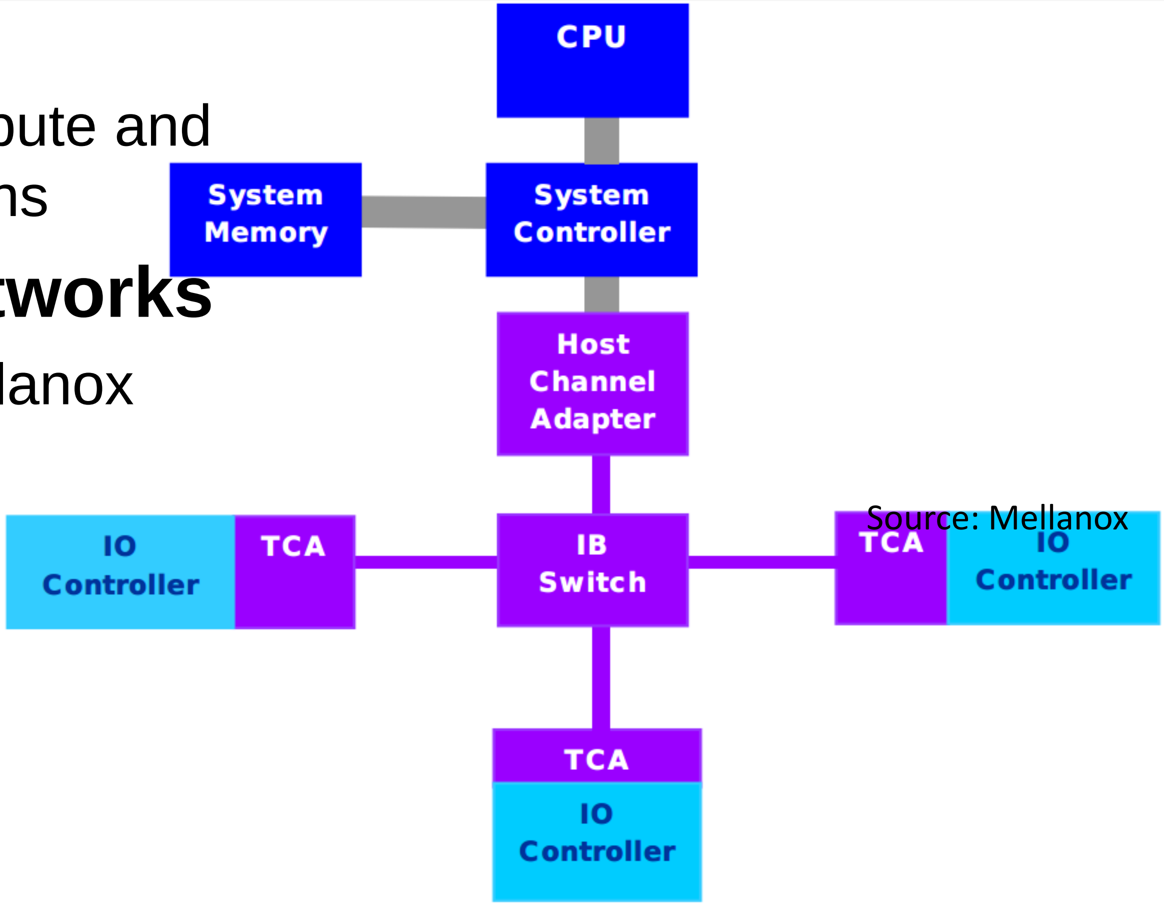
- **PCIe external cabling**
  - Standard for copper cables
- **FireFly PCIe over Fibre**
  - PCIe 3.0 x 4 over fibre by Samtec
- **ExpEther**
  - PCIe 2.0 over 40 GigE networks by NEC


CPU (RC)
PCI EXPRESS
PCI Express
Stanadrd PCIe Switch
Upstream Port (PCI Bridge)
Internal bus
Downstream Port (PCI Bridge)
....
Downstream Port (PCI Bridge)
PCI Express
I/O Device
I/O Device
CPU (RC)
PCI Express
Implemented in a same chip
ExpEther Host Chip (upstream)
Ethernet
Ethernet
Ethernet region (transparent)
Ethernet Switch
ExpEther IO Chip (downstream)
ExpEther IO Chip (downstream)
PCI Express
I/O Device
I/O Device

Source: NEC Corporation

Alternatives

- **Dolphin ICS**
  - PCIe based compute and IO cluster solutions
- **Storage / IO Networks**
  - Infiniband by Mellanox



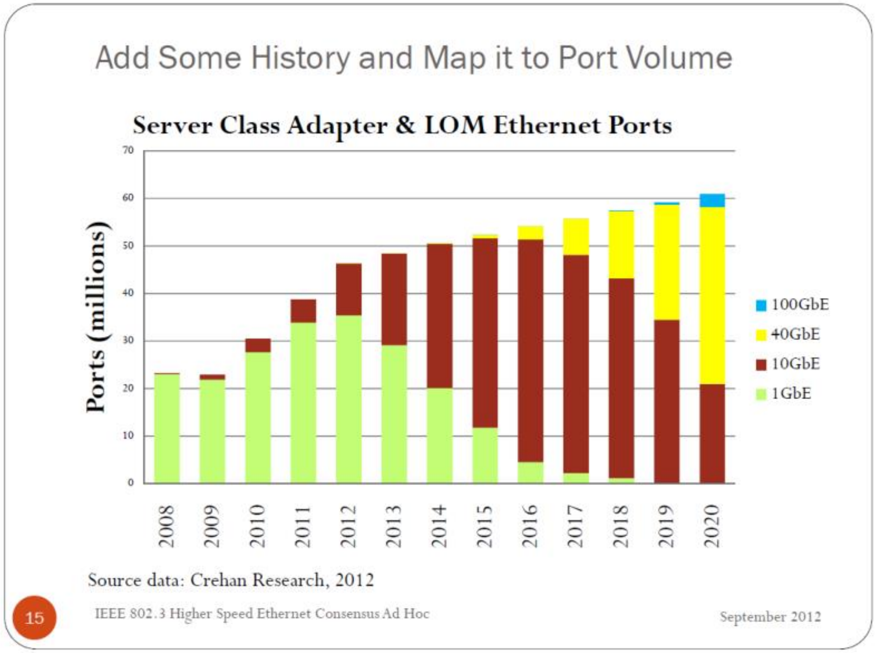
Source: Mellanox

Proposal: PCIe over TCP/IP

- **Fully transparent to network equipment**
  - Just a bunch of TCP sessions
  - No special traffic handling required
- **Fully transparent to PCIe**
  - Reliable transport via TCP
  - Congestion control via TCP
- **Based on separated and distributed upstream and downstream switch ports**
  - Easily scalable via TCP session count
  - Support for multiple ethernet ports
  - Decouples cable routing from transaction layer routing
- **Independent of lower network layers, e.g. physical layer**

Network Processing

- 10 GigE will soon push from data center into embedded markets



Add Some History and Map it to Port Volume

Source data: Crehan Research, 2012

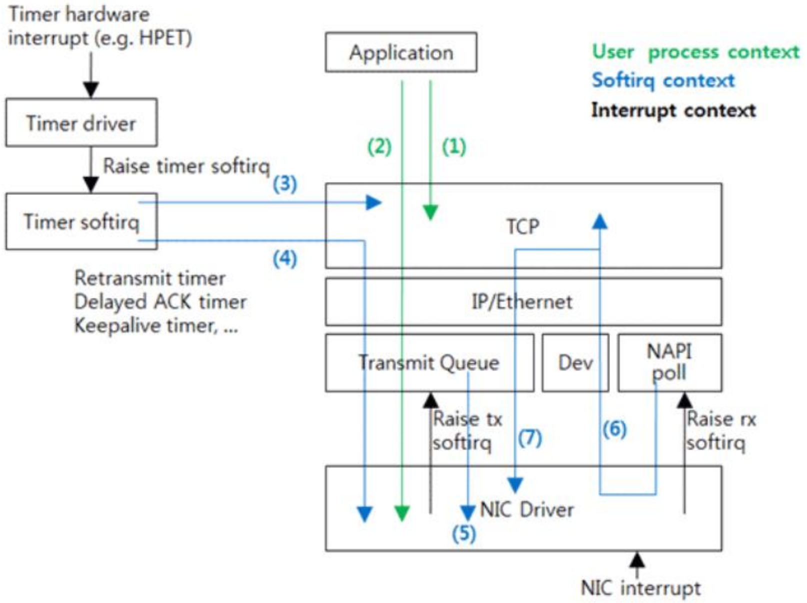
15

IEEE 802.3 Higher Speed Ethernet Consensus Ad Hoc

September 2012

Transporting 1 bit per second needs 1 Hz

- 1 GigE ➔ 1 CPU at 1 GHz
- 10 GigE ➔ 4 CPUs at 2.5 GHz



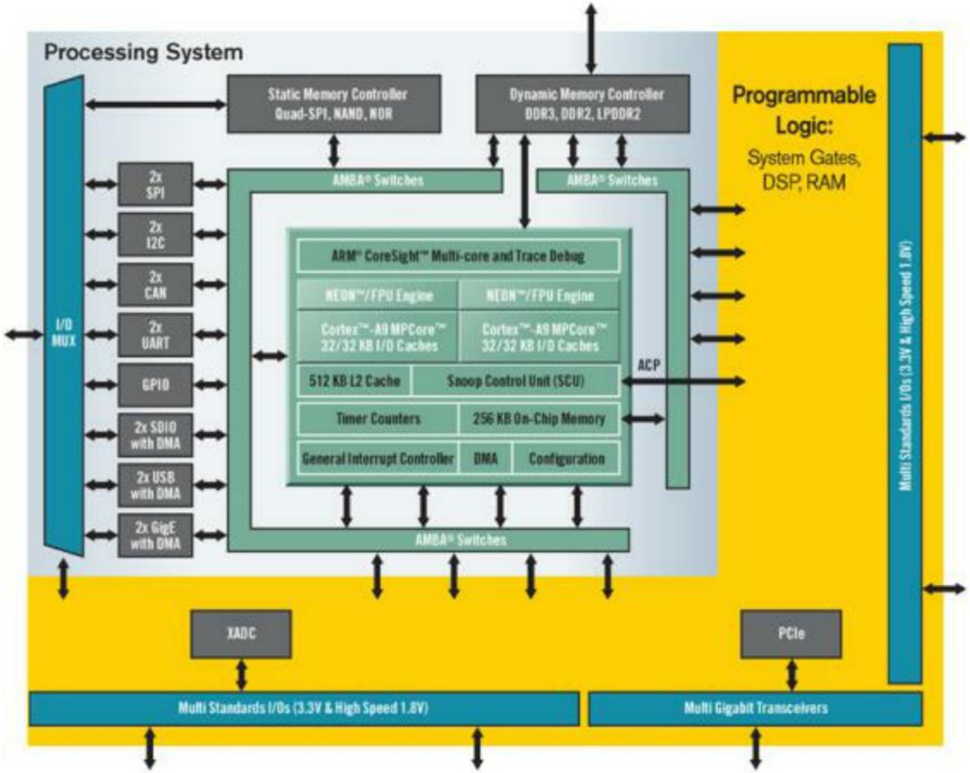
Network Processing

SoC FPGA as (yet) another computer

	Intel i7-4770	Xilinx Zynq 7045
Compute	~100 GFLOPS	5 GFLOPS (PS) 778 GFLOPS (PL)
TDP	84 W	<20 W (typ)

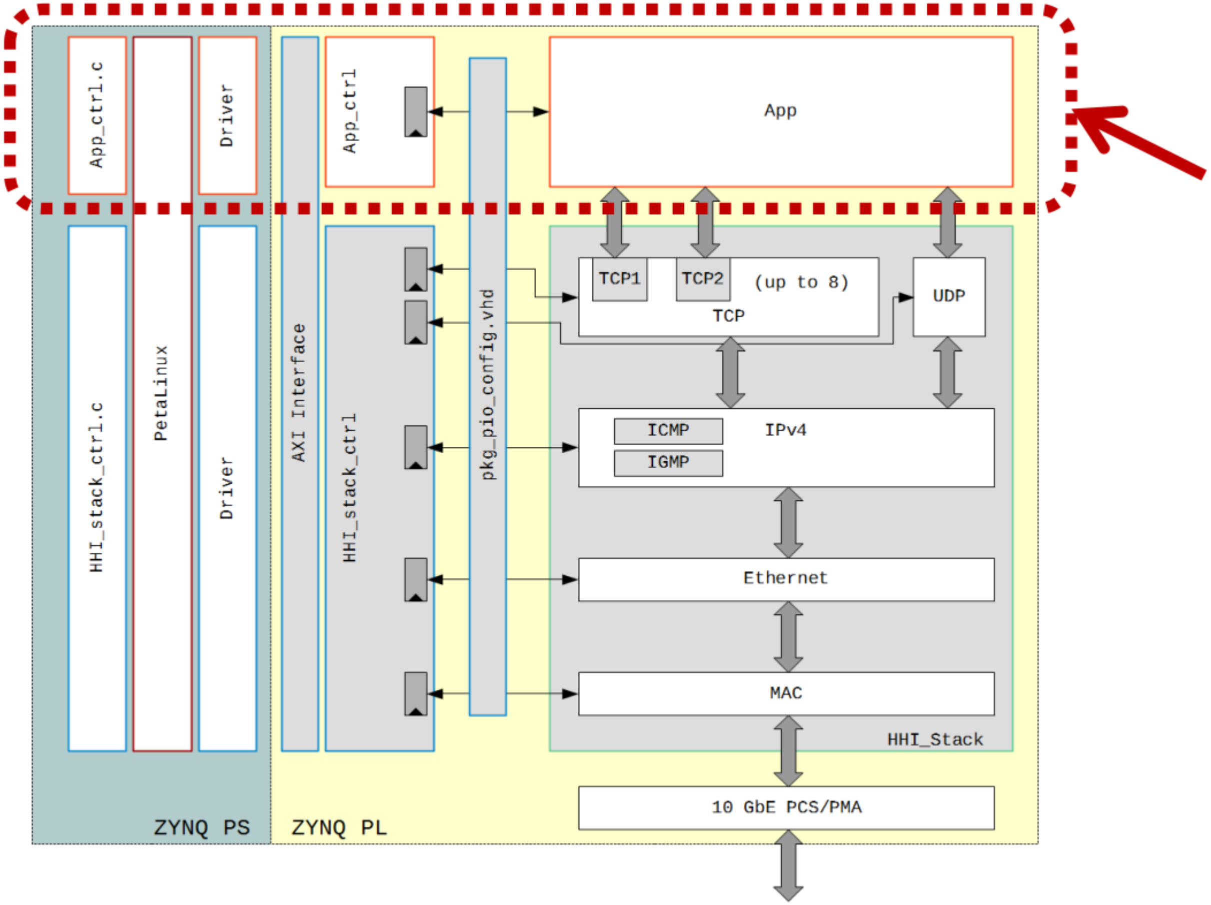
SOC FPGA has 4x more compute
With ¼ the power dissipation!

[http://www.xilinx.com/products/technology/dsp.html]



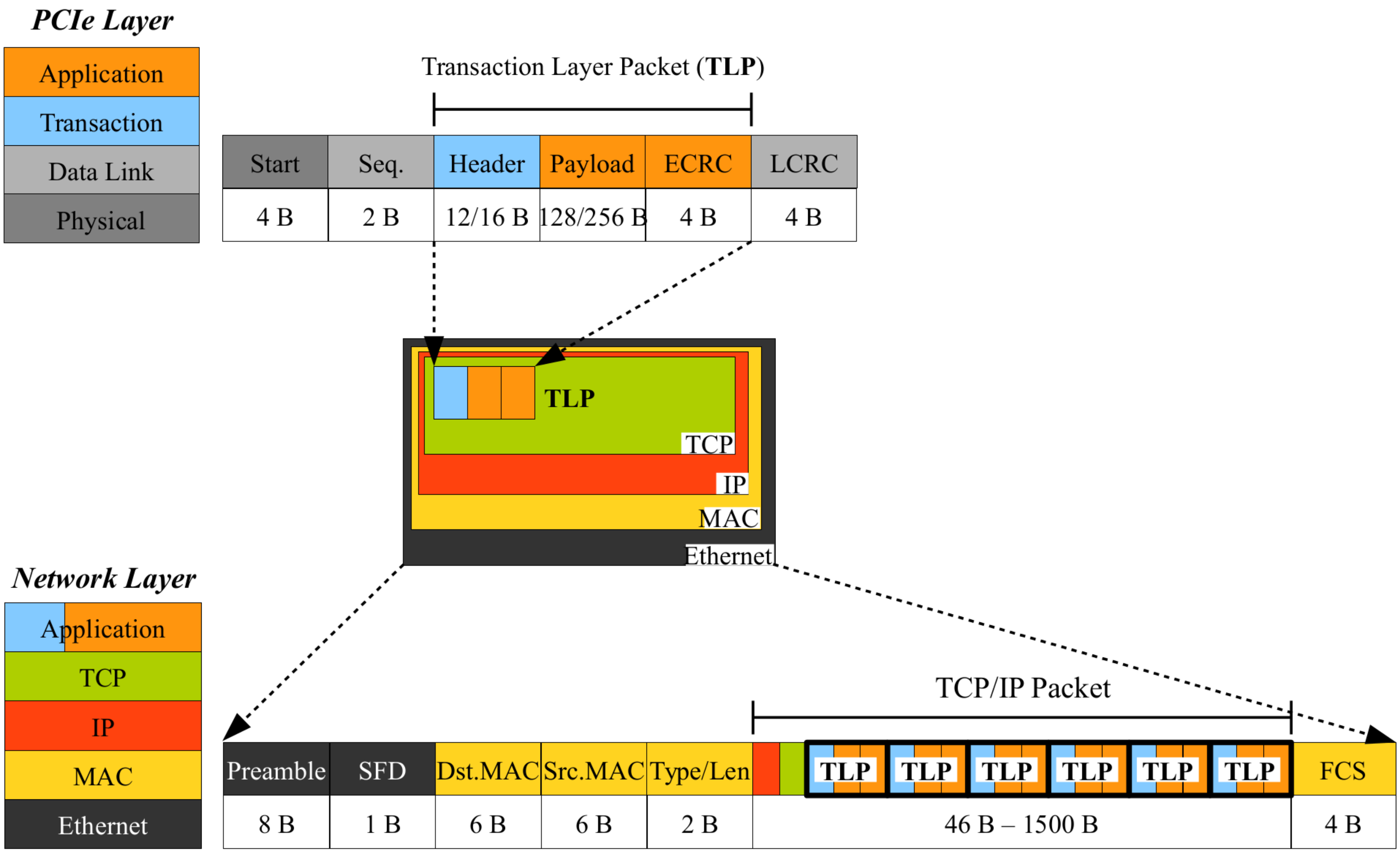
Processing System
Static Memory Controller Quad-SPI, NAND, NOR
Dynamic Memory Controller DDR3, DDR2, LPDDR2
I/O MUX
2x SPI
2x I2C
2x CAN
2x UART
GPIO
2x SDIO with DMA
2x USB with DMA
2x GigE with DMA
AMBA® Switches
AMBA® Switches
ARM® CoreSight™ Multi-core and Trace Debug
NEON™/FPU Engine
NEON™/FPU Engine
Cortex™-A9 MPCore™ 32/32 KB I/D Caches
Cortex™-A9 MPCore™ 32/32 KB I/D Caches
512 KB L2 Cache
Snoop Control Unit (SCU)
ACP
Timer Counters
256 KB On-Chip Memory
General Interrupt Controller
DMA
Configuration
AMBA® Switches
Programmable Logic:
System Gates, DSP, RAM
Multi Standards I/Os (3.3V & High Speed 1.8V)
XADC
PCIe
Multi Standards I/Os (3.3V & High Speed 1.8V)
Multi Gigabit Transceivers

Network Protocol Acceleration Platform Architecture



Network protocol processing at application layer (ISO Layer 7) can more efficiently be implemented via a programming approach (in C or C++) than by digital circuit design (in VHDL or Verilog).

PCIe over TCP/IP Tunneling



PCIe Layer

- Application
- Transaction
- Data Link
- Physical

Transaction Layer Packet (**TLP**)

Start	Seq.	Header	Payload	ECRC	LCRC
4 B	2 B	12/16 B	128/256 B	4 B	4 B

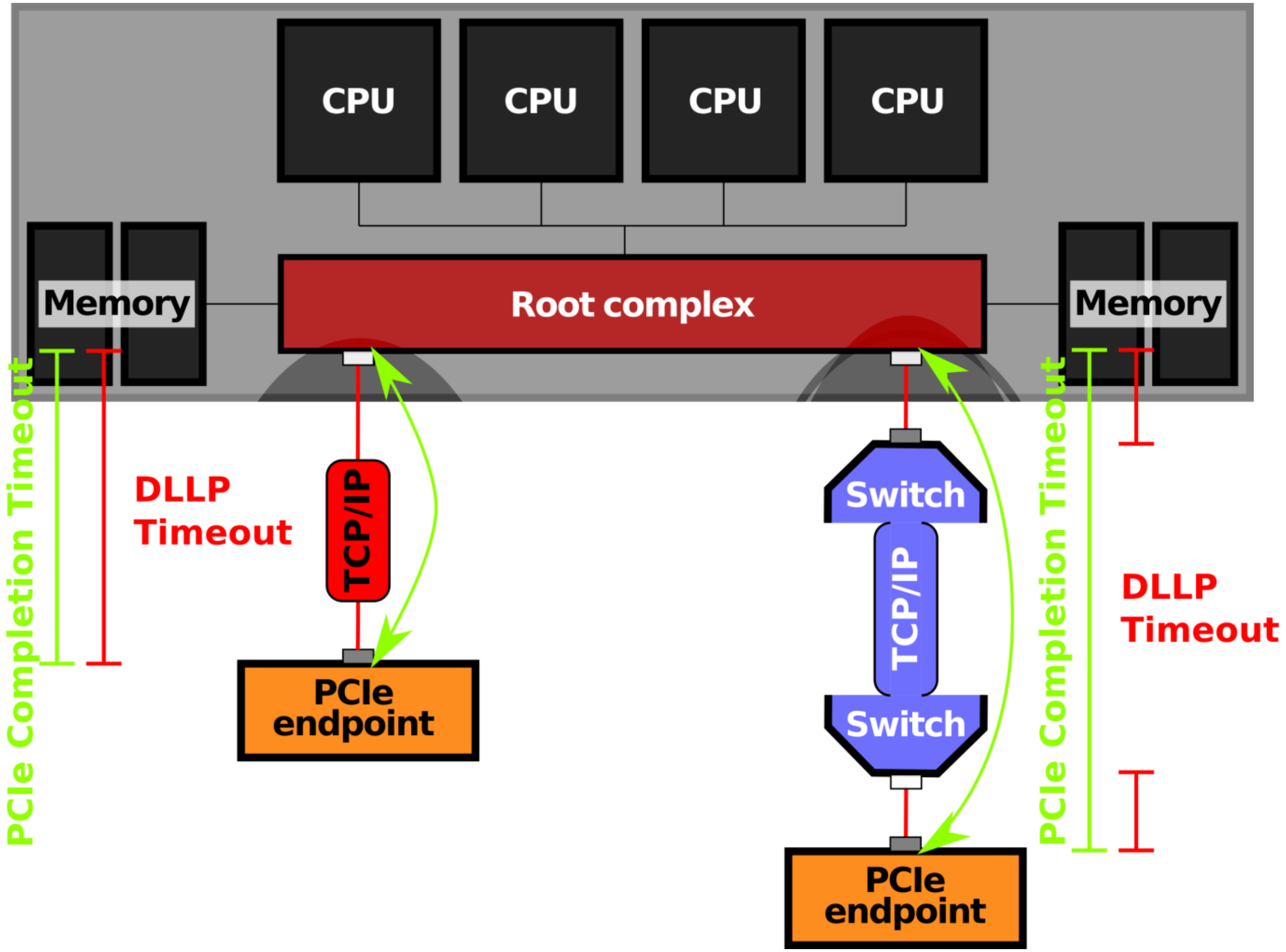
Network Layer

- Application
- TCP
- IP
- MAC
- Ethernet

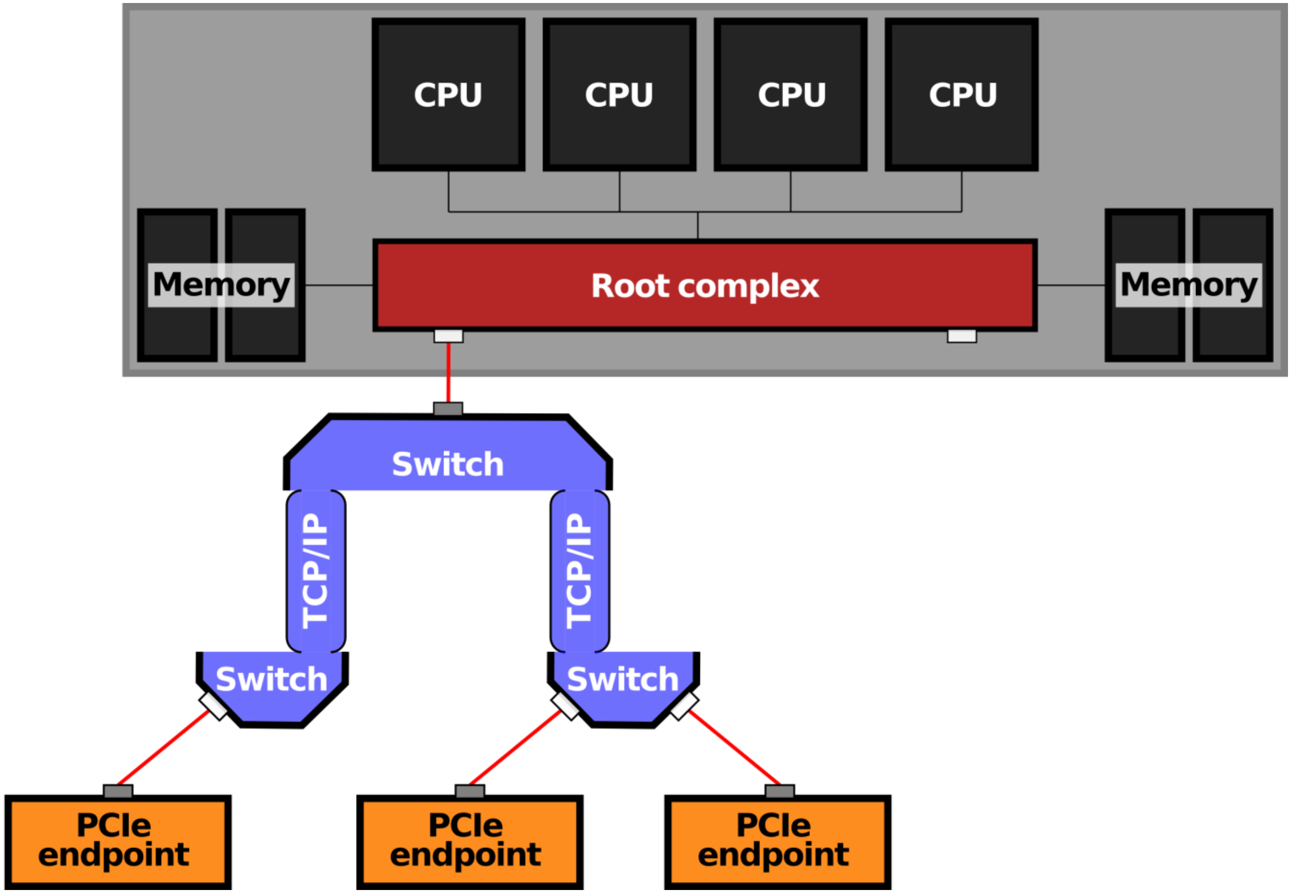
TCP/IP Packet

Preamble	SFD	Dst.MAC	Src.MAC	Type/Len	TLP TLP TLP TLP TLP TLP	FCS
8 B	1 B	6 B	6 B	2 B	46 B – 1500 B	4 B

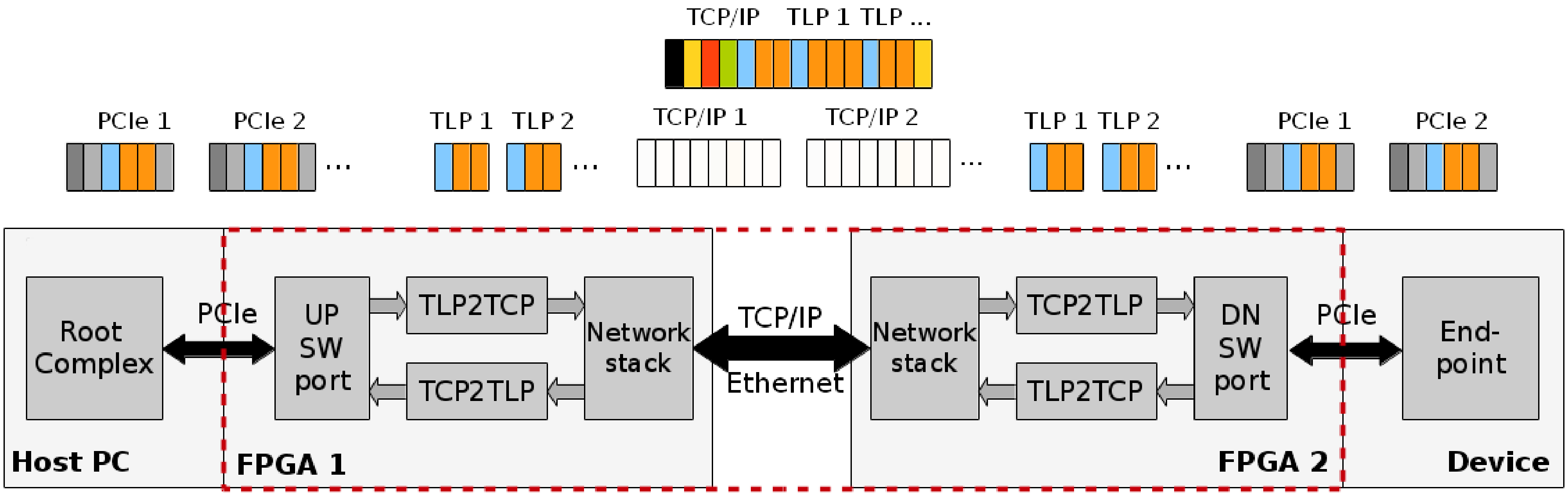
Why tunnel TLPs and not DLLPs? PCIe Timeouts



Why tunnel TLPs and not DLLPs? Scaling to Multiple Ports

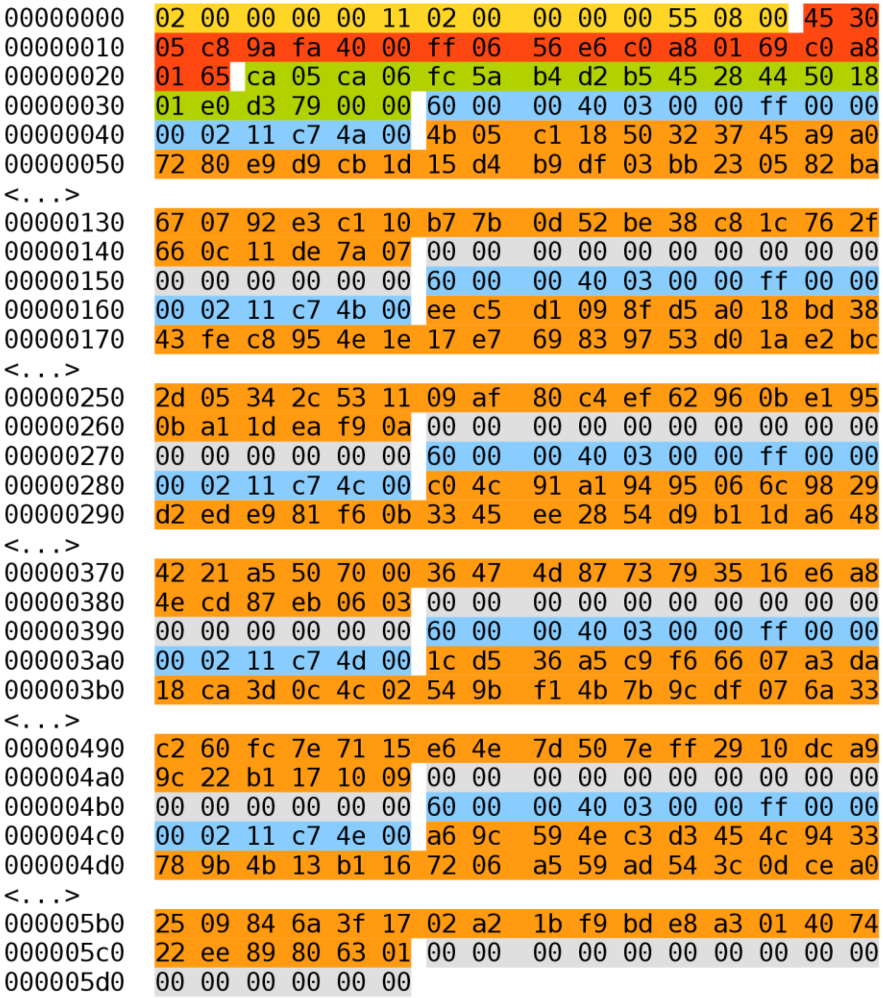


Concept of PCIe over TCP/IP



---- Distributed PCIe Switch

TLP Aggregation

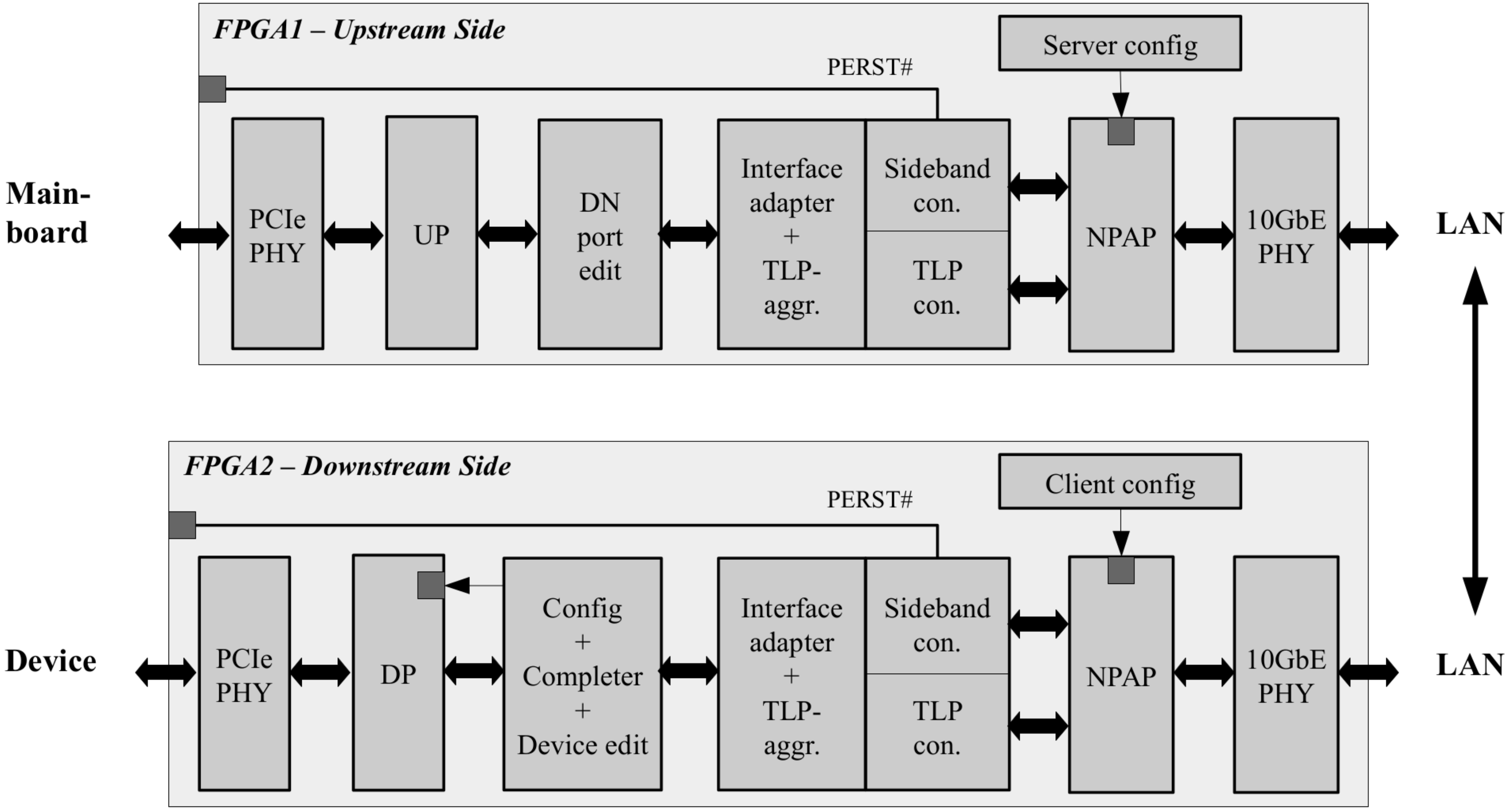


```
00000000  02 00 00 00 00 11 02 00  00 00 00 55 08 00 45 30
00000010  05 c8 9a fa 40 00 ff 06  56 e6 c0 a8 01 69 c0 a8
00000020  01 65 ca 05 ca 06 fc 5a  b4 d2 b5 45 28 44 50 18
00000030  01 e0 d3 79 00 00 60 00  00 40 03 00 00 ff 00 00
00000040  00 02 11 c7 4a 00 4b 05  c1 18 50 32 37 45 a9 a0
00000050  72 80 e9 d9 cb 1d 15 d4  b9 df 03 bb 23 05 82 ba
<...>
00000130  67 07 92 e3 c1 10 b7 7b  0d 52 be 38 c8 1c 76 2f
00000140  66 0c 11 de 7a 07 00 00  00 00 00 00 00 00 00 00
00000150  00 00 00 00 00 00 60 00  00 40 03 00 00 ff 00 00
00000160  00 02 11 c7 4b 00 ee c5  d1 09 8f d5 a0 18 bd 38
00000170  43 fe c8 95 4e 1e 17 e7  69 83 97 53 d0 1a e2 bc
<...>
00000250  2d 05 34 2c 53 11 09 af  80 c4 ef 62 96 0b e1 95
00000260  0b a1 1d ea f9 0a 00 00  00 00 00 00 00 00 00 00
00000270  00 00 00 00 00 00 60 00  00 40 03 00 00 ff 00 00
00000280  00 02 11 c7 4c 00 c0 4c  91 a1 94 95 06 6c 98 29
00000290  d2 ed e9 81 f6 0b 33 45  ee 28 54 d9 b1 1d a6 48
<...>
00000370  42 21 a5 50 70 00 36 47  4d 87 73 79 35 16 e6 a8
00000380  4e cd 87 eb 06 03 00 00  00 00 00 00 00 00 00 00
00000390  00 00 00 00 00 00 60 00  00 40 03 00 00 ff 00 00
000003a0  00 02 11 c7 4d 00 1c d5  36 a5 c9 f6 66 07 a3 da
000003b0  18 ca 3d 0c 4c 02 54 9b  f1 4b 7b 9c df 07 6a 33
<...>
00000490  c2 60 fc 7e 71 15 e6 4e  7d 50 7e ff 29 10 dc a9
000004a0  9c 22 b1 17 10 09 00 00  00 00 00 00 00 00 00 00
000004b0  00 00 00 00 00 00 60 00  00 40 03 00 00 ff 00 00
000004c0  00 02 11 c7 4e 00 a6 9c  59 4e c3 d3 45 4c 94 33
000004d0  78 9b 4b 13 b1 16 72 06  a5 59 ad 54 3c 0d ce a0
<...>
000005b0  25 09 84 6a 3f 17 02 a2  1b f9 bd e8 a3 01 40 74
000005c0  22 ee 89 80 63 01 00 00  00 00 00 00 00 00 00 00
000005d0  00 00 00 00 00 00
```

```
🟨  Ethernet II Header
    [0000-0005] Dst. MAC: (02 00 00 00 00 11) → 02:00:00:00:00:11
    [0006-0011] Src. MAC: (02 00 00 00 00 55) → 02:00:00:00:00:55
🟥  Internet Protocol Header
    [0017]      Protocol: (06) → TCP
    [001a-001d] Src. IP: (c0 a8 01 69) → 192.168.1.105
    [001e-0021] Dst. IP: (c0 a8 01 65) → 192.168.1.101
🟩  Transmission Control Protocol Header
    [0022-0023] Src. Port: (ca 05) → 51717
    [0024-0025] Dst. Port: (ca 06) → 51718
🟦  PCIe TLP Header
    [0036]      FMT/Type: (60) → 64-bit Memory Write Request
    [0038-0039] Length: (00 40) → 64 Doublewords (32-bit) → 256 Byte
    [003a-003b] Requester ID: (03 00) → 03:0.0
    [003e-0045] Address1: (00 00 00 02 11 c7 4a 00)
    [015e-0165] Address2: (00 00 00 02 11 c7 4b 00)
    [027e-0285] Address3: (00 00 00 02 11 c7 4c 00)
    [039e-03a5] Address4: (00 00 00 02 11 c7 4d 00)
    [04be-04c5] Address4: (00 00 00 02 11 c7 4e 00)
🟧  Data
⬜  Padding
```

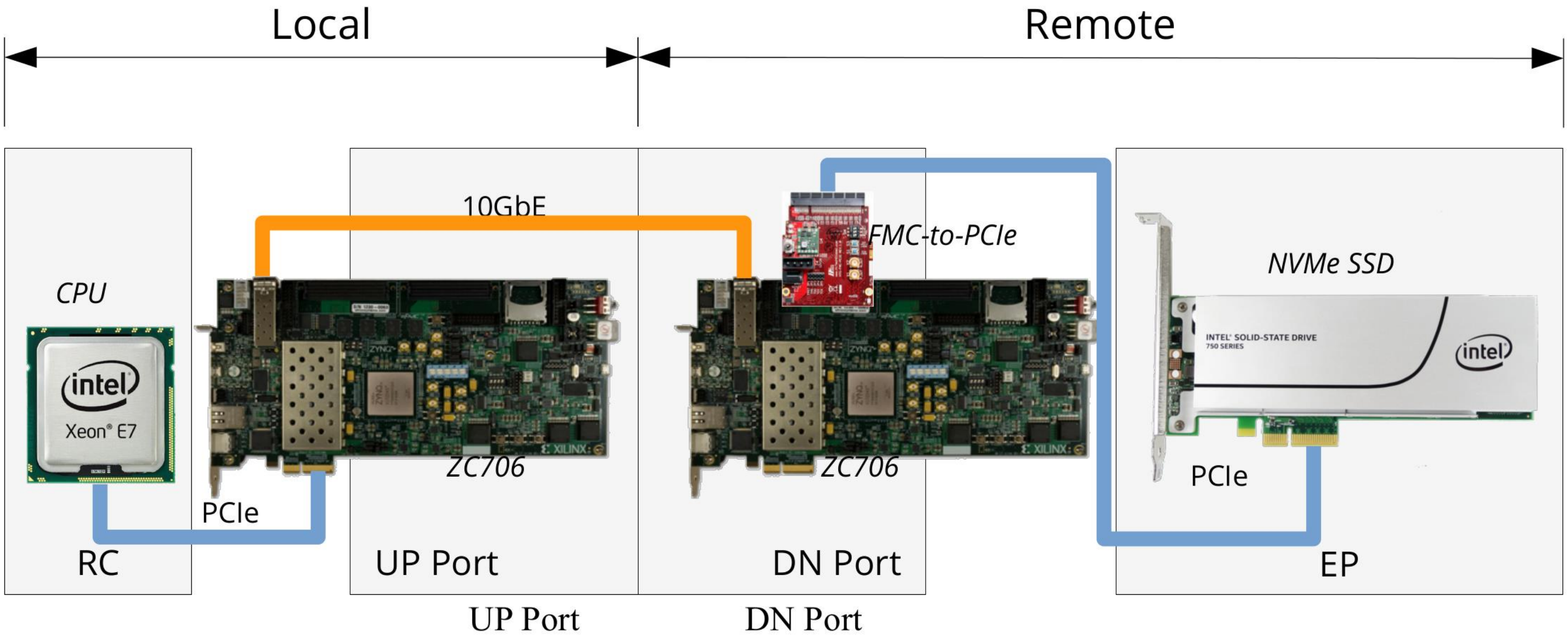
- Send multiple TLPs per TCP/IP segment
- Aggregating TLPs has minor impact on latency
- TLP aggregation reduces protocol overhead
- Up to 20 % bandwidth gain with aggregation

Implementation: FPGA Design



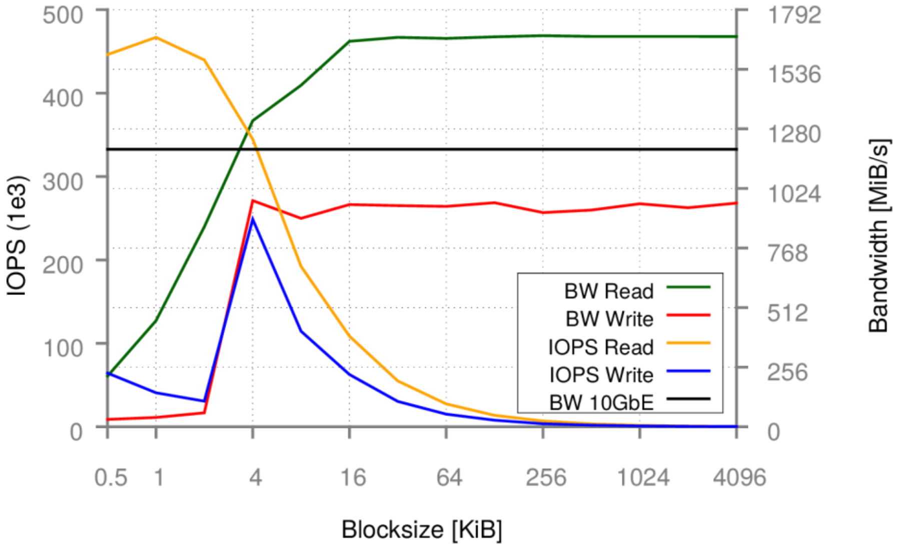
- Based on „XPressRICH3“ PCIe IP Core from PLDA
- Based on Fraunhofer HHI NPAP

Implementation: Test Setup

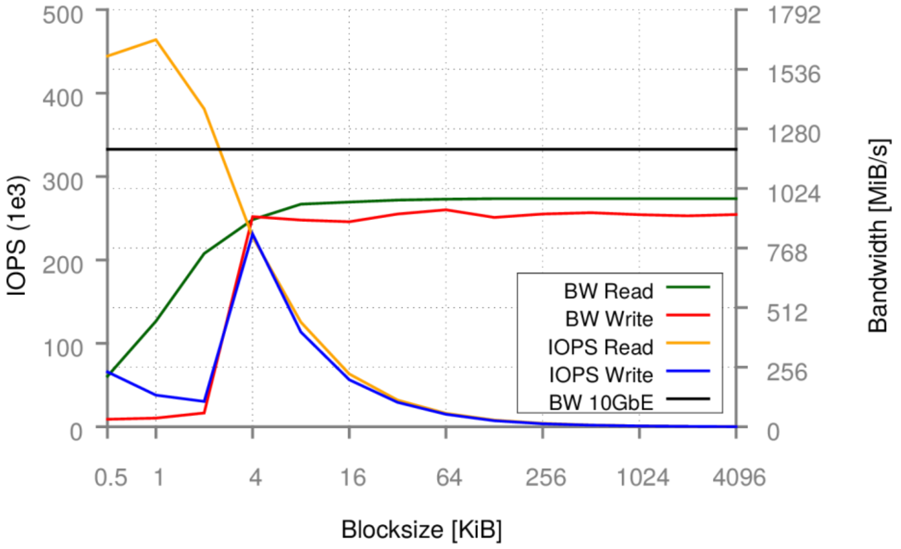


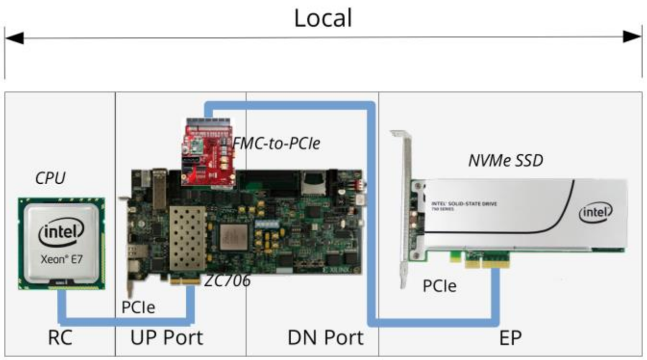
NVMe Performance Results

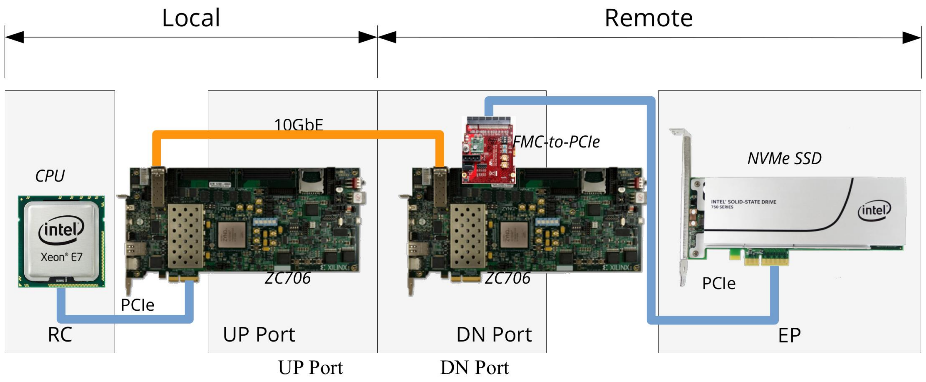
Local PCIe Switch



Distributed PCIe Switch with TLP Aggregation

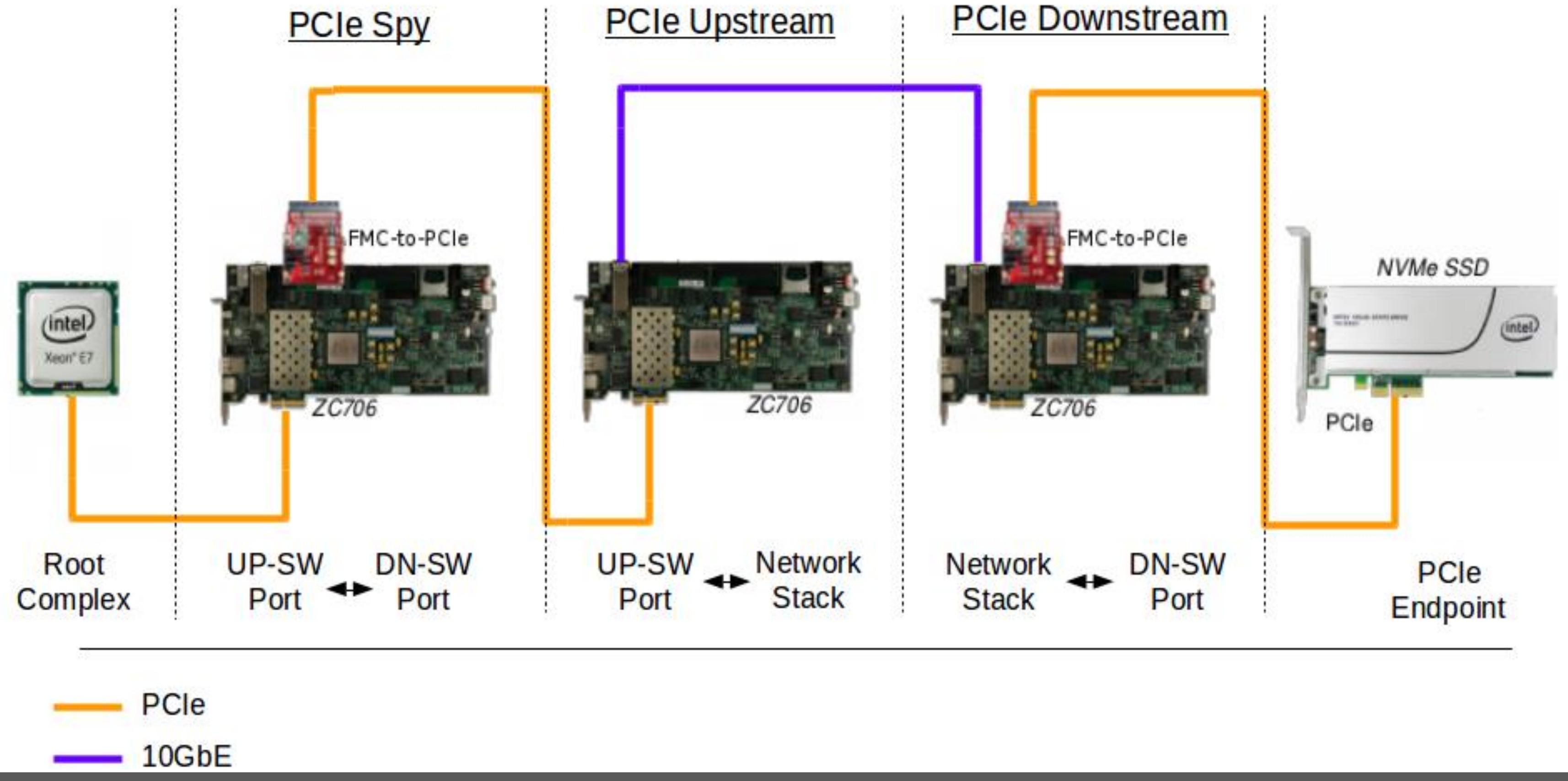




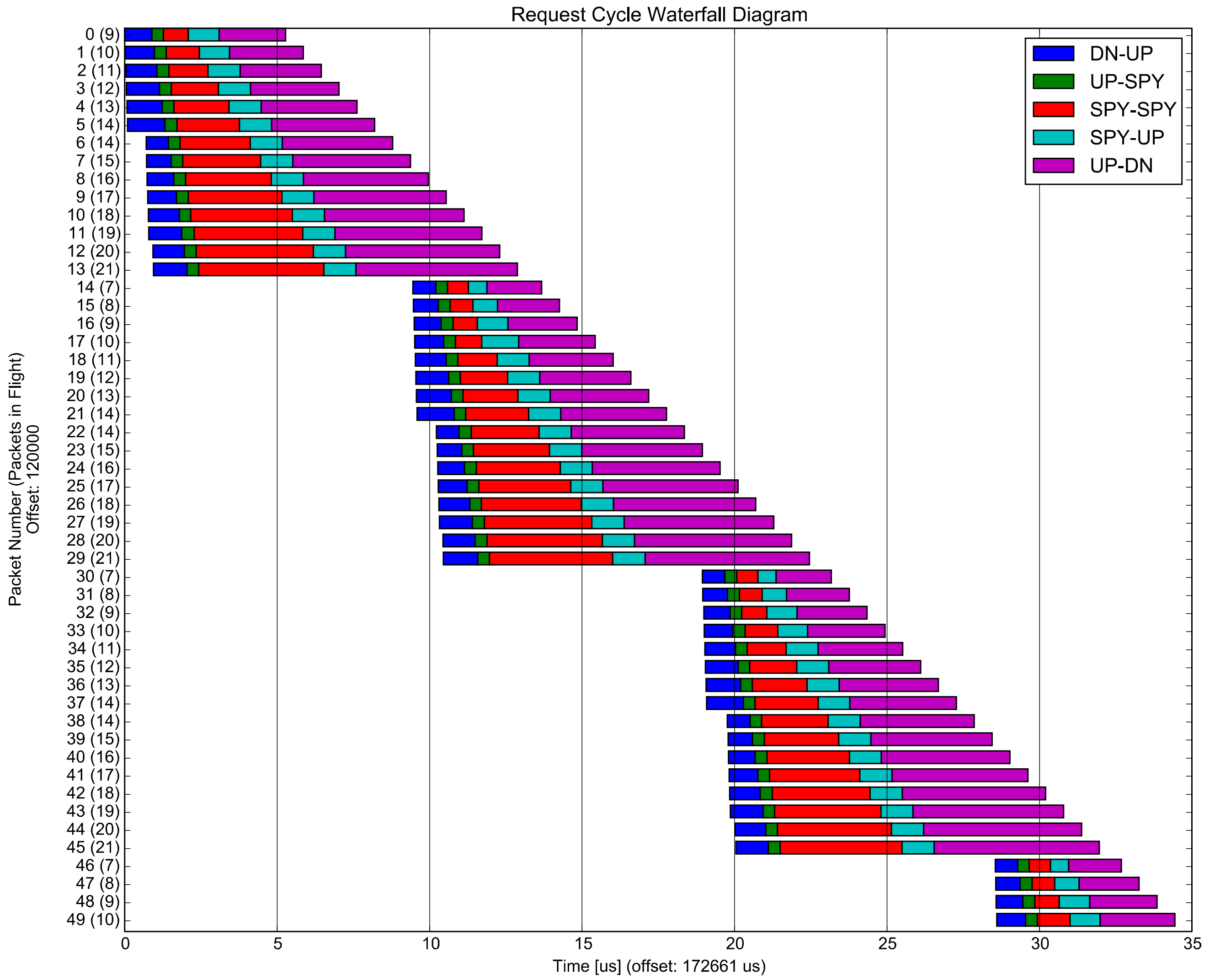


Visibility for monitoring

- **Using PCIe over TCP/IP also opens PCIe for simple (performance) monitoring via network traces**



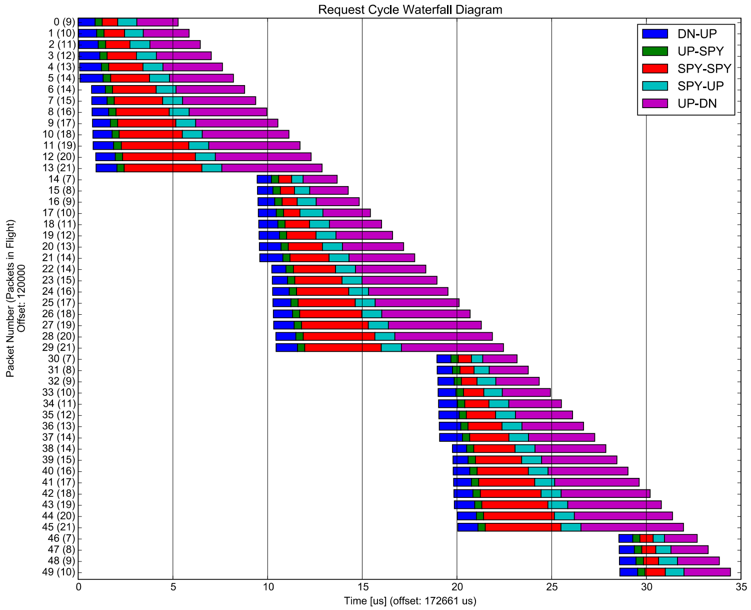
Visibility for monitoring

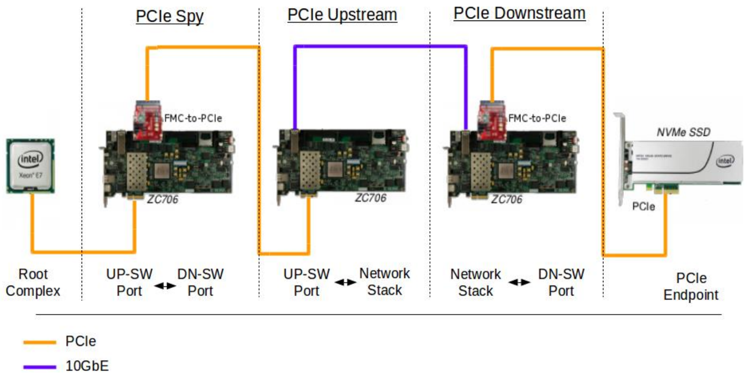


Visibility for monitoring

- A row is a request cycle of a device DMA engine
  - Request leaves PCIe device flowing upstream, enters DN node, then UP, then SPY, then host node
  - Completions flow downstream
    - Multiple completions per request
    - Max read requrest size > max payload size
- All TLPs of a transaction are captured, timestamped and correlated
- SPY-SPY (red part) is the time a host needs to complete a request (PCIe to DRAM to PCIe latency)
- DN-UP, UP-DN are network transitions in upstream or downstream direction respectively
- Network bandwidth is lower than PCIe bandwidth, so this hop needs more time

- Observation: PCIe DMA engine requests bursts





Conclusions

- Reliable “tunneling” of PCI Express via TCP/IP
- Fully transparent to PCIe Root Complex and PCIe Devices
- No additional host software for compliant hosts ist needed as PCIe switch management is natively implemented
- The solution inter operates with all compliant software and add-in cards out of the box
- Scalability based on FPGA processing for bandwidths of 1, 10 , 25 GigE line rates and beyond
- Re-use existing network infrastructure
- Technology extendable with NTB and MRA concepts in a distributed system

Thank you for attending the PCI-SIG Developers Conference 2018.

For more information please go to www.pcisig.com