

PCI Express over IP - Accelerated

Dr. Endric Schubert, Univ. Ulm & Missing Link Electronics
Andreas Braun, Missing Link Electronics
Ulrich Langenbach, Fraunhofer HHI

We are

a Silicon Valley based technology company with Offices in Germany. We are Partner of leading electronic device and solution providers and have been enabling key innovators in the automotive, industrial, test & measurement markets to build better Embedded Systems, faster.

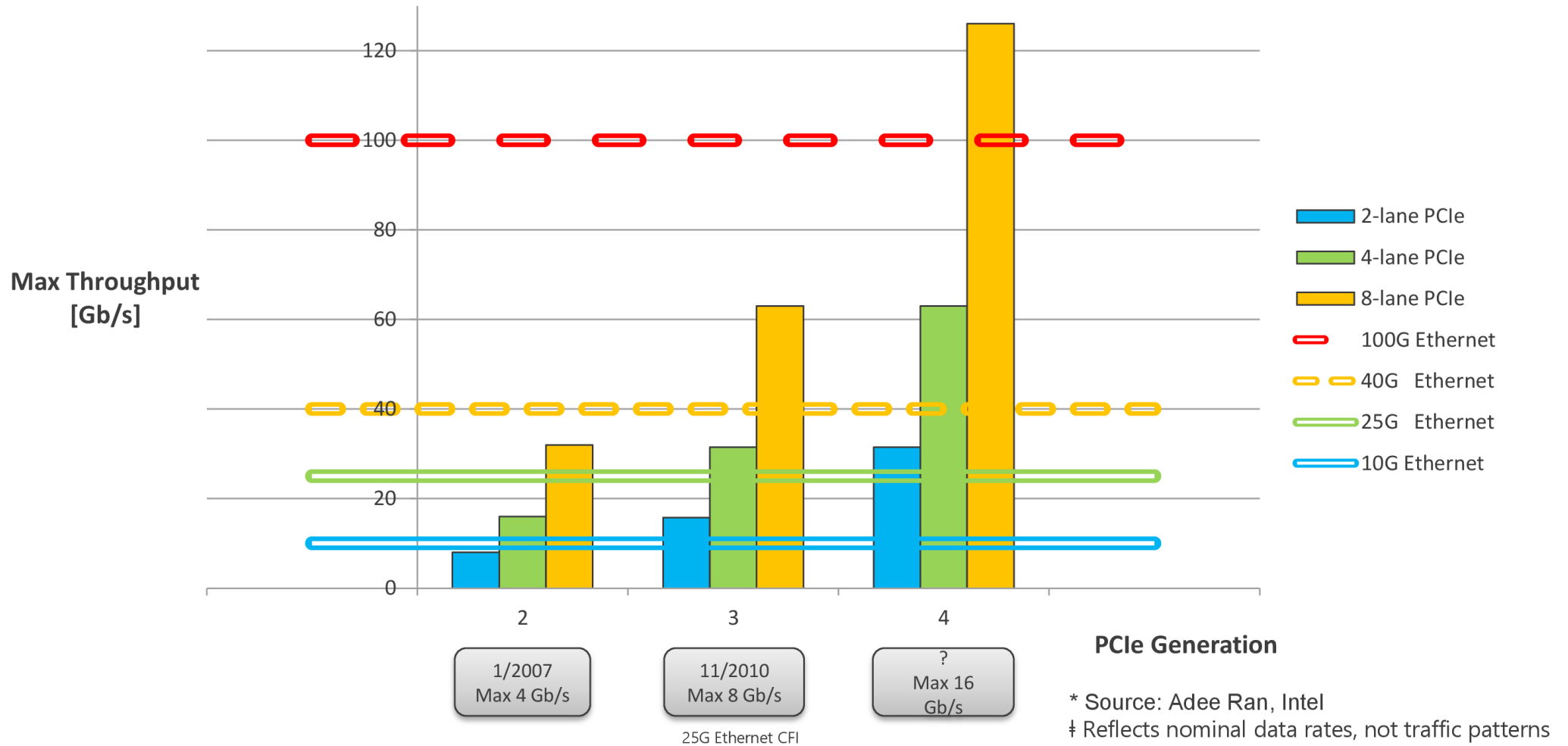
Our Mission is

To develop and market technology solutions for Embedded Systems
Realization via pre-validated IP and expert application support, and to combine off-the-shelf FPGA devices with Open-Source Software for dependable, configurable Embedded System platforms

Our Expertise is

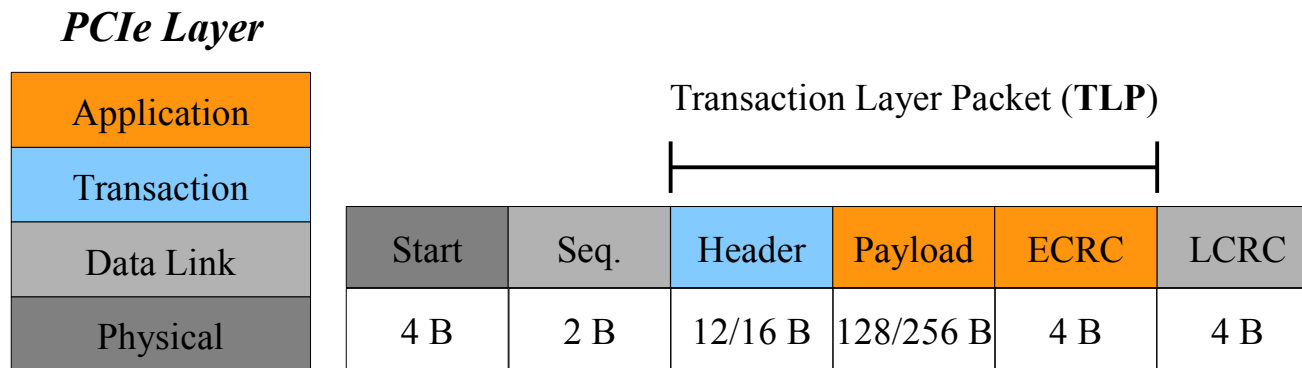
I/O connectivity and Acceleration of data communication protocols, additionally opening up FPGA technology for analog applications, and the integration and optimization of Open Source Linux and Android software stacks on modern extensible processing architectures.

PCIe to Ethernet Throughput Matching



PCI Express

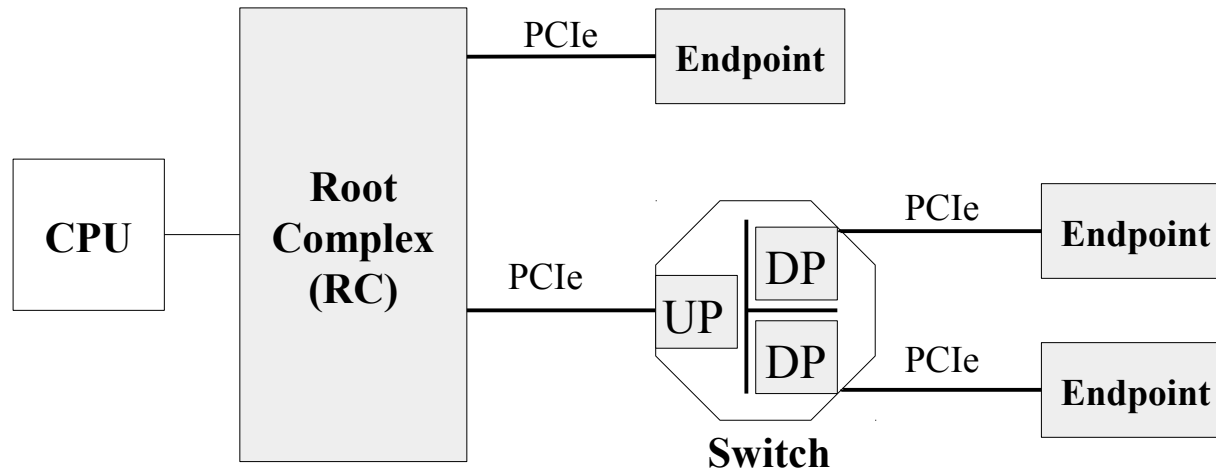
- PCIe replaces the PCI Local Bus (backwards compatible)
- Full-duplex serial transmission
- At 8GT/s line rate (Gen3) on up to 32 lanes
- Packet-based protocol with four layers



- 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

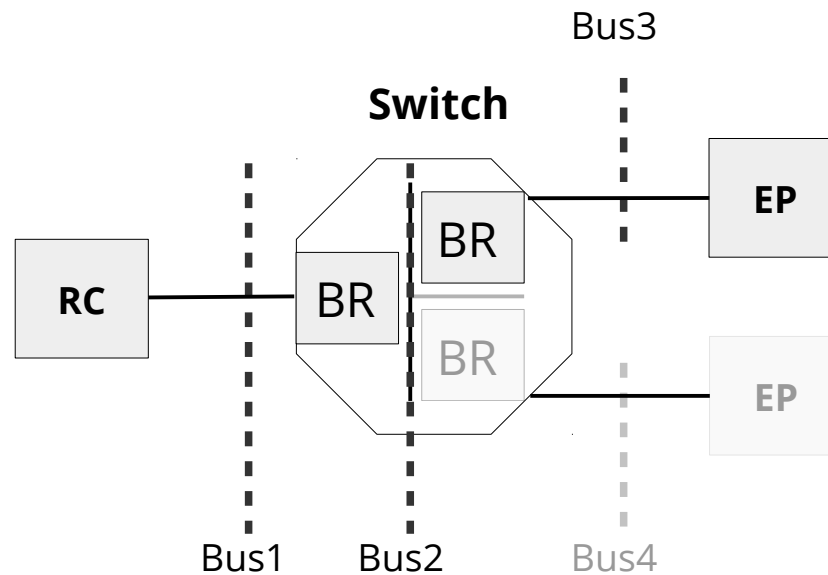
PCI Express

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



- ID based routing (bus/device/function number) and address based routing
- Transactions require completion (split-transaction) or posted transactions possible

- PCI Local Bus from software view (addressing, driver, configuration, ...)
- PCIe devices implement a set of registers (configuration space)
- PCIe topology needs to be explored at the beginning of system start-up
- Enumeration of devices by completing Configuration-TLPs



Range problem: Physical line length of PCIe on PCBs is limited to the centimeter range

State-of-the-art

PCIe external cabling:

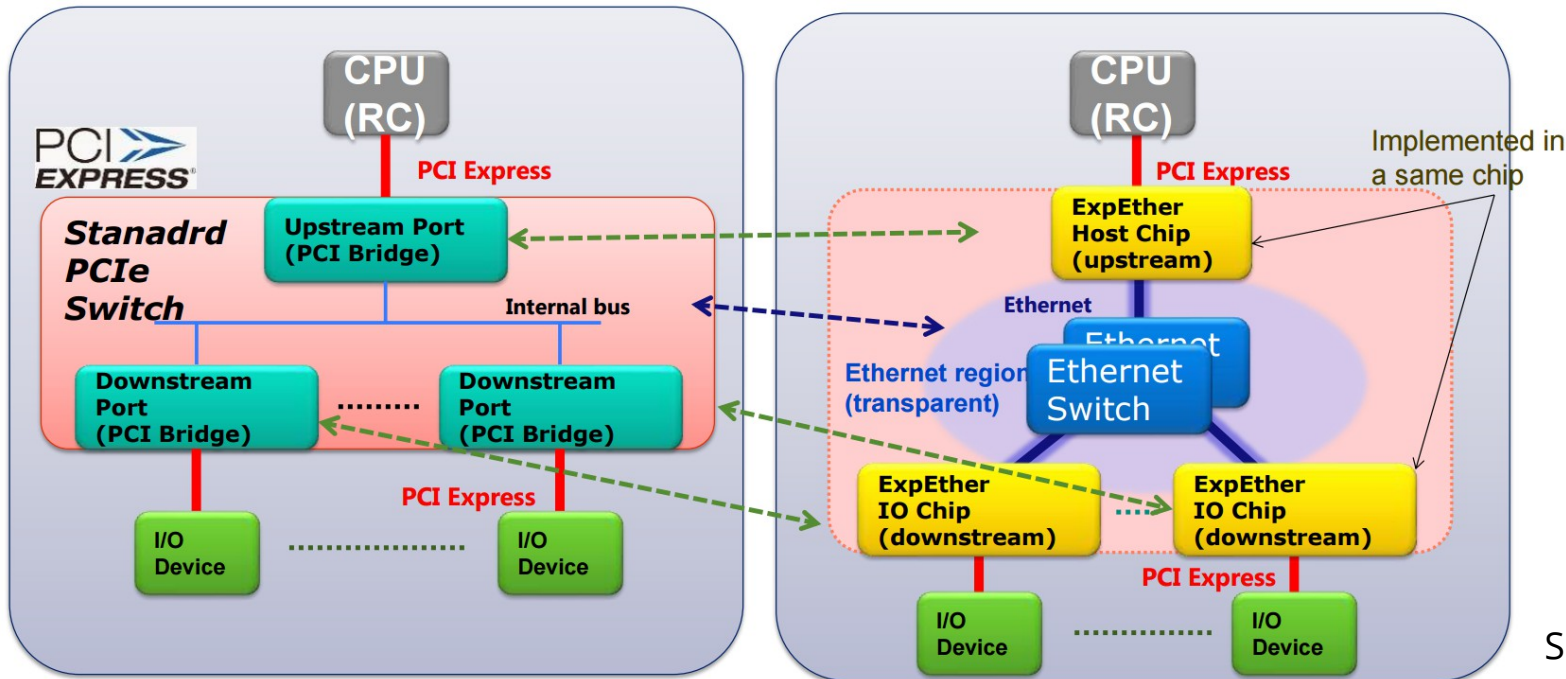
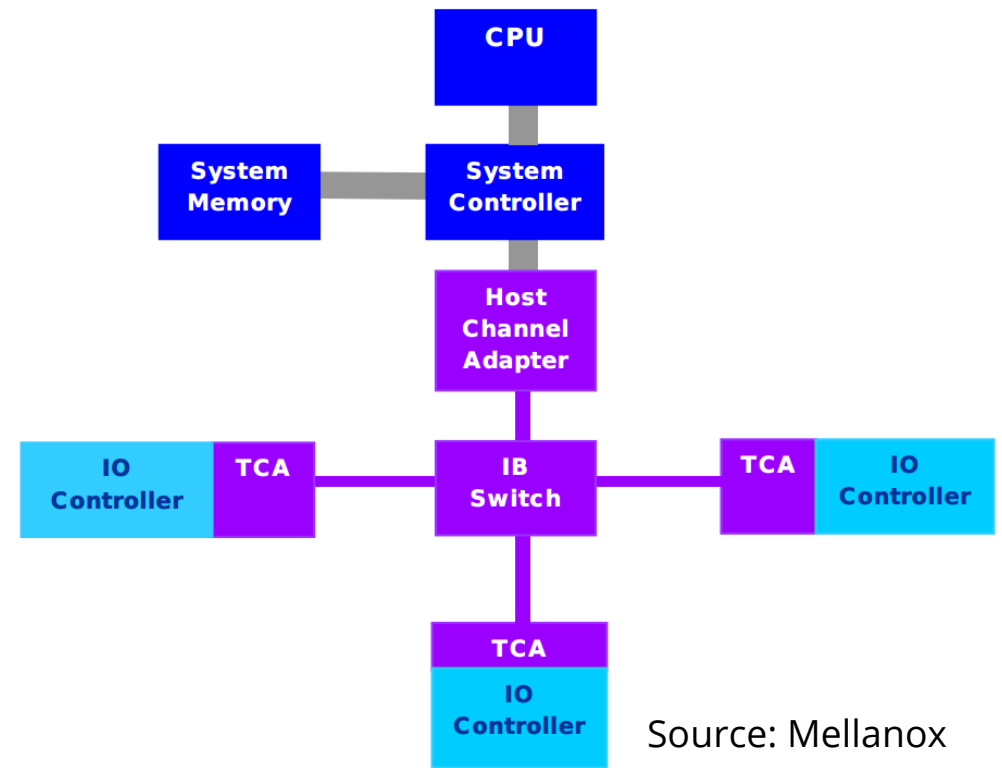
- Standard for copper cables

InfiniBand:

- Standard HPC interconnect

ExpEther:

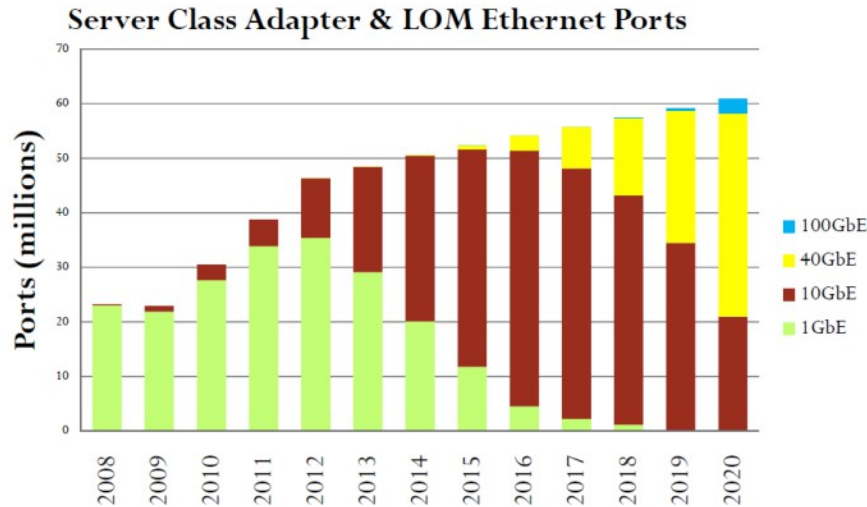
- PCIe Gen2 over 10GbE networks by NEC.



Source: Nec Corporation

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

Add Some History and Map it to Port Volume



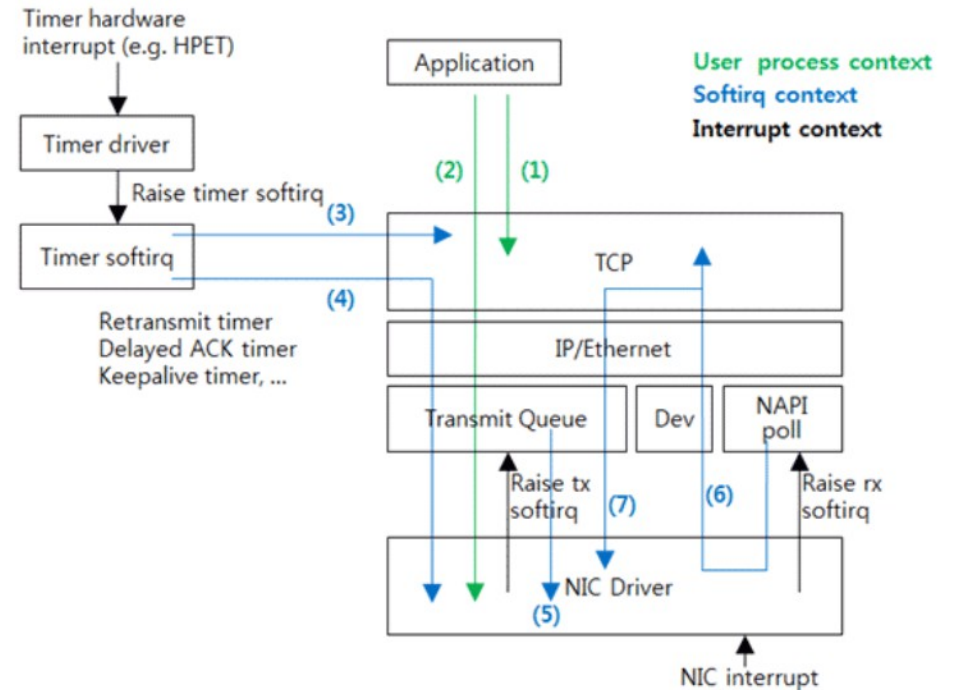
Source data: Crehan Research, 2012

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



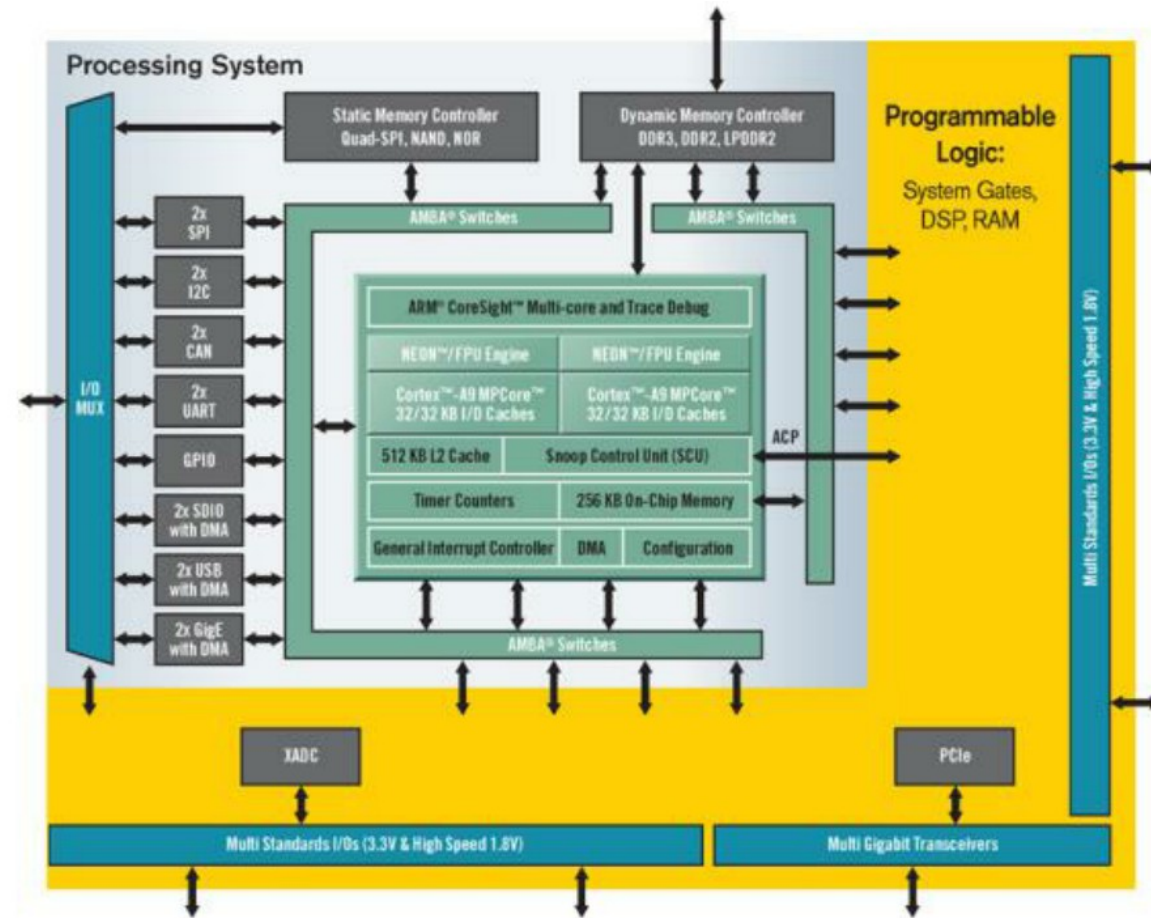
Network Processing

SoC FPGA as (yet) another computer

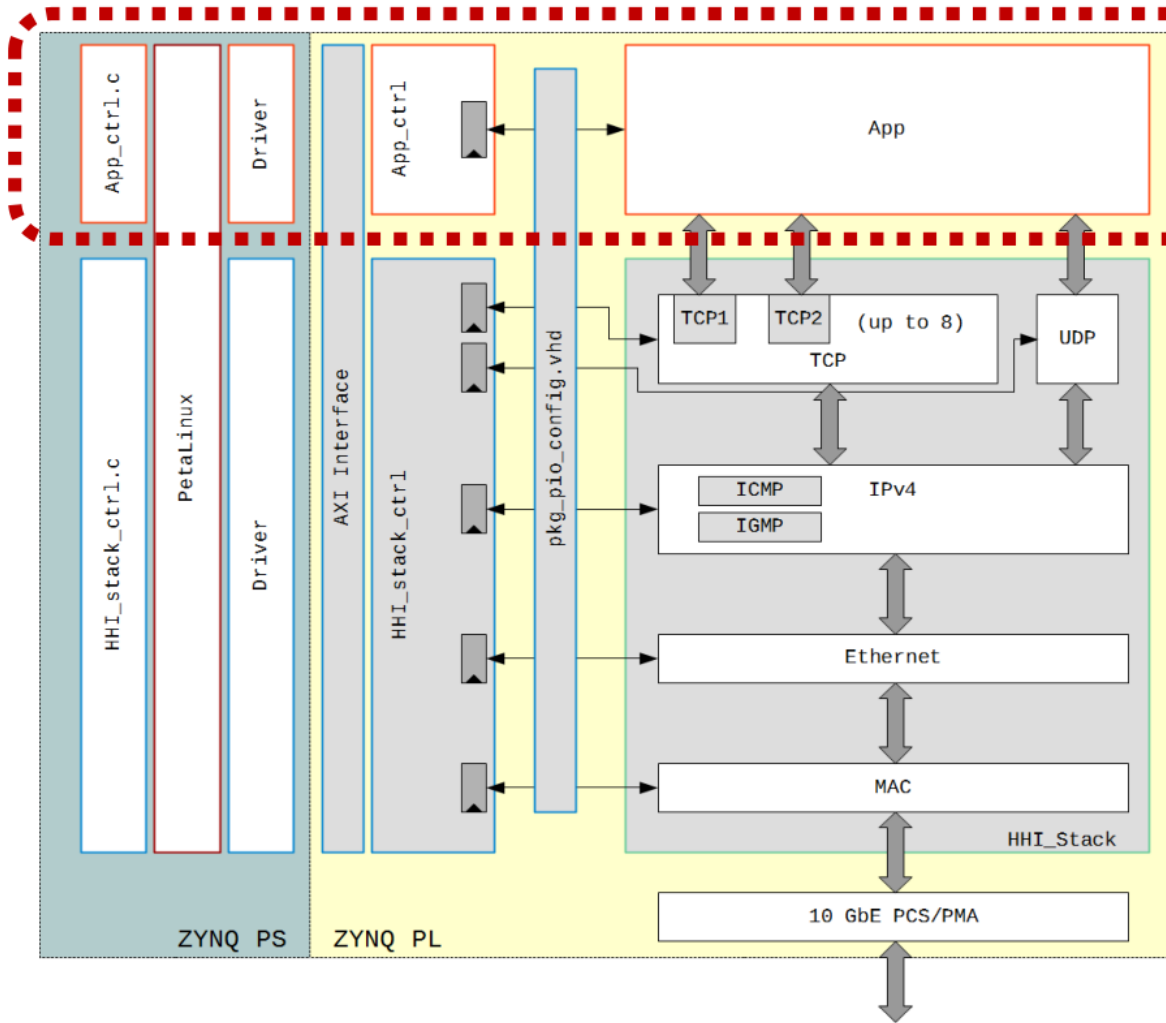
	Intel i7-4770	Xilinx Zynq 7045
Compute	~100 GFLOPS	5 GFLOPS (PS) 778 GFLOPS (PL)
TDP	84 W	<20 W (typ)

SOC FPGA has 4x more compute
With 1/4 the power dissipation!

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



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).

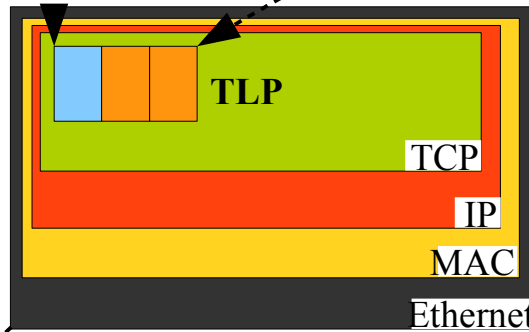
PCIe over IP

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



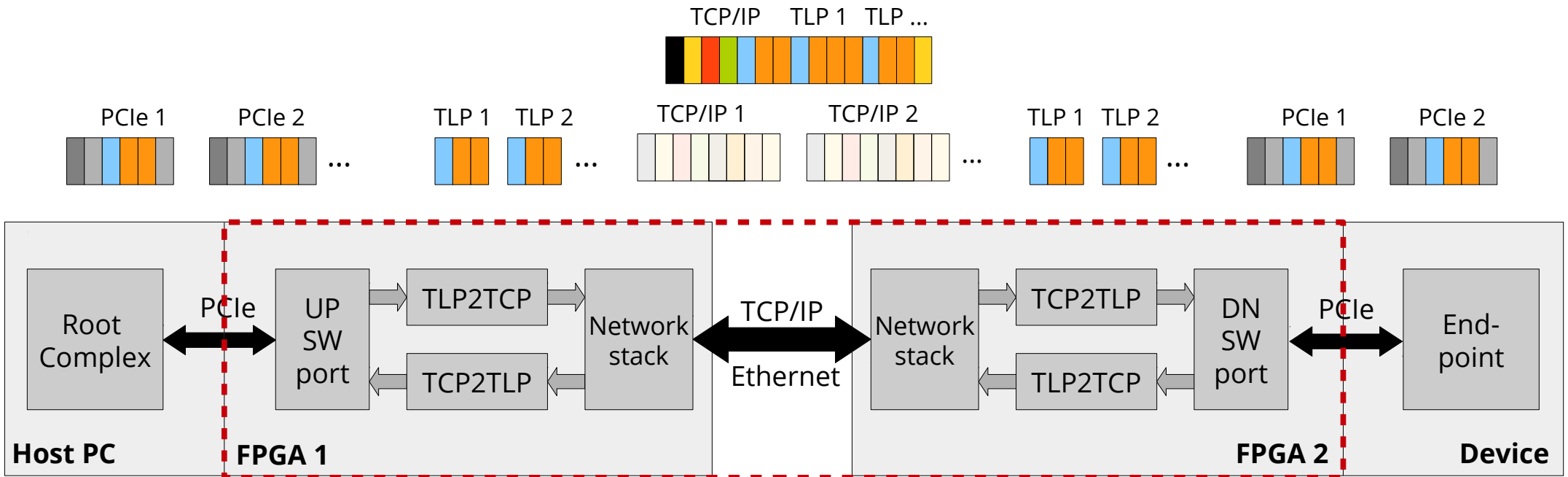
Network Layer

Application
TCP
IP
MAC
Ethernet

TCP/IP Packet

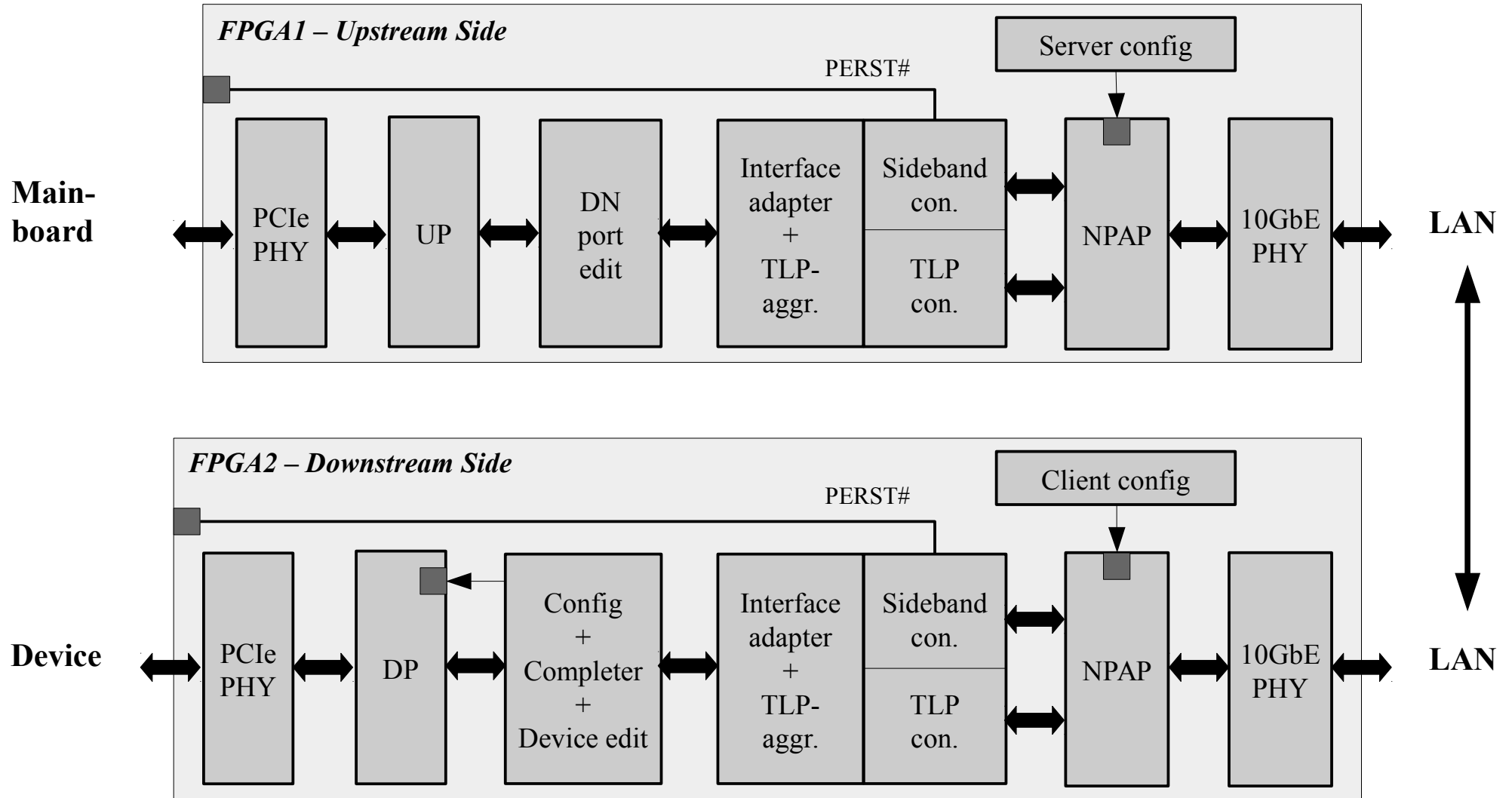
Preamble	SFD	Dst.MAC	Src.MAC	Type/Len	TLP	TLP	TLP	TLP	TLP	TLP	FCS
8 B	1 B	6 B	6 B	2 B	46 B – 1500 B						4 B

Concept of PCIe-over-IP



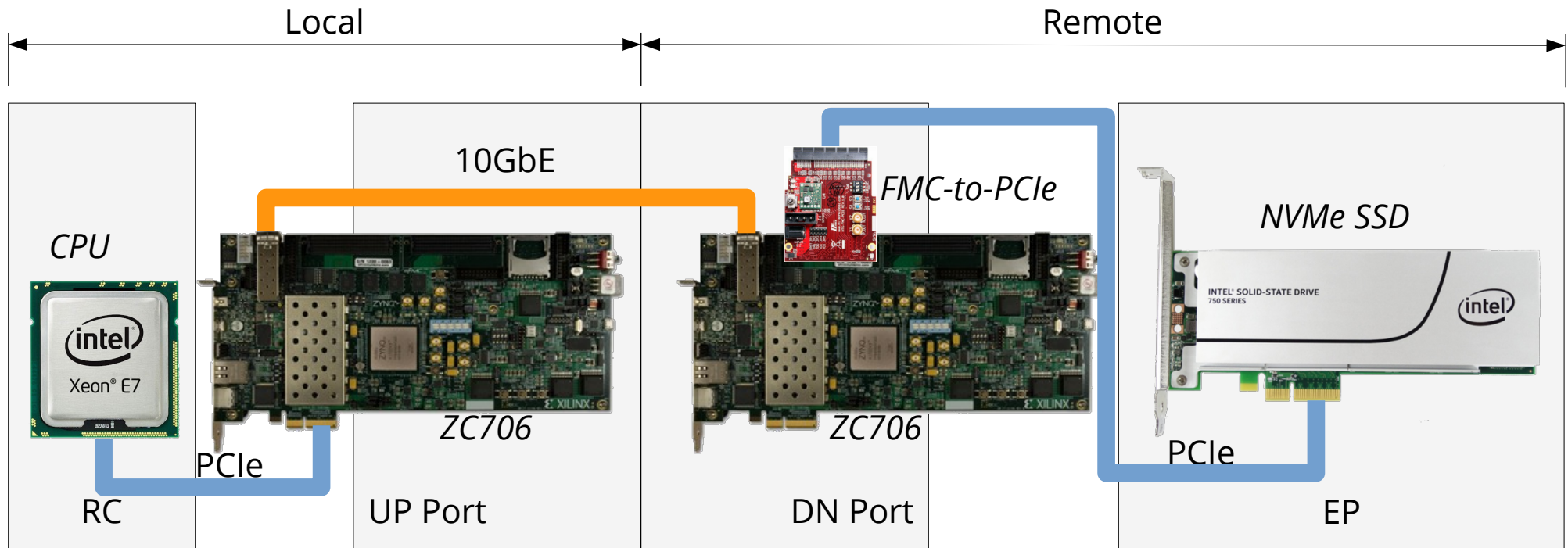
----- Distributed PCIe Switch based on
"XPressRICH3" PCIe IP Core from PLDA

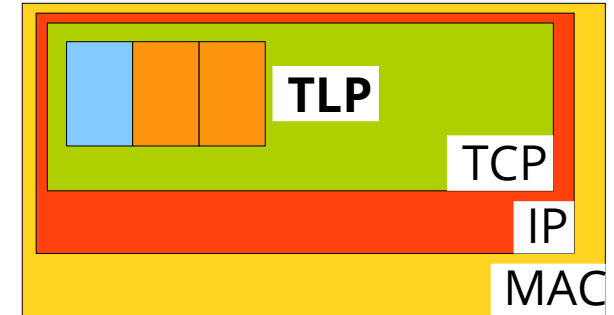
Implementation



Implementation

Setup





TLP aggregation

```

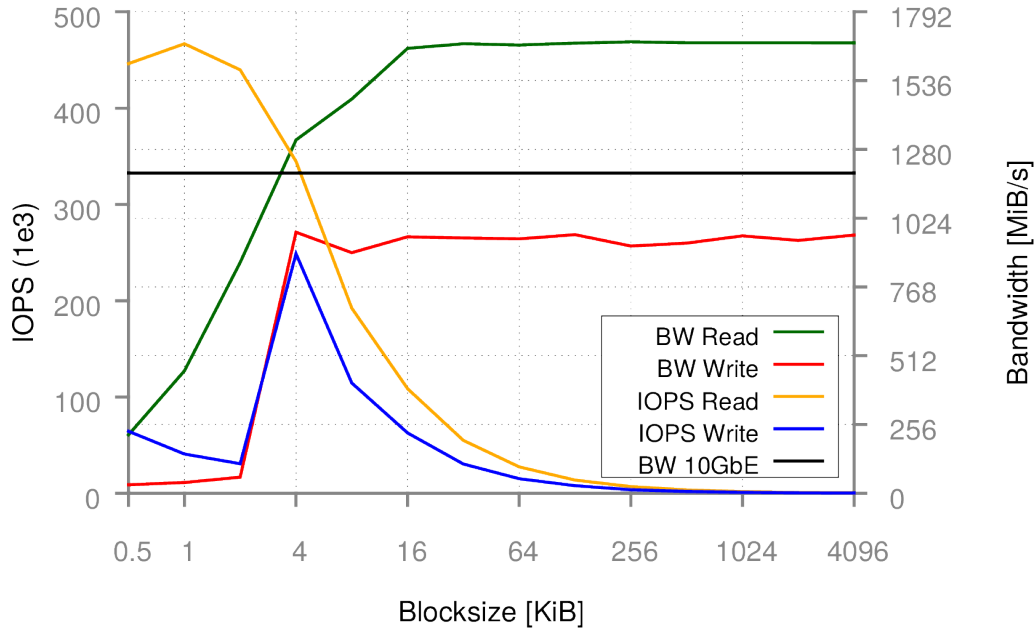
00000000 02 00 00 00 00 11 02 00 00 00 00 55 08 00 45 30
00000010 05 c8 9a fa 40 00 ff 06 56 e6 c0 a8 01 69 c0 a8
00000020 01 65 ca 05 ca 06 fc 5a b4 d2 b5 45 28 44 50 18
00000030 01 e0 d3 79 00 00 60 00 00 40 03 00 00 ff 00 00
00000040 00 02 11 c7 4a 00 4b 05 c1 18 50 32 37 45 a9 a0
00000050 72 80 e9 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
  
```

Legende:

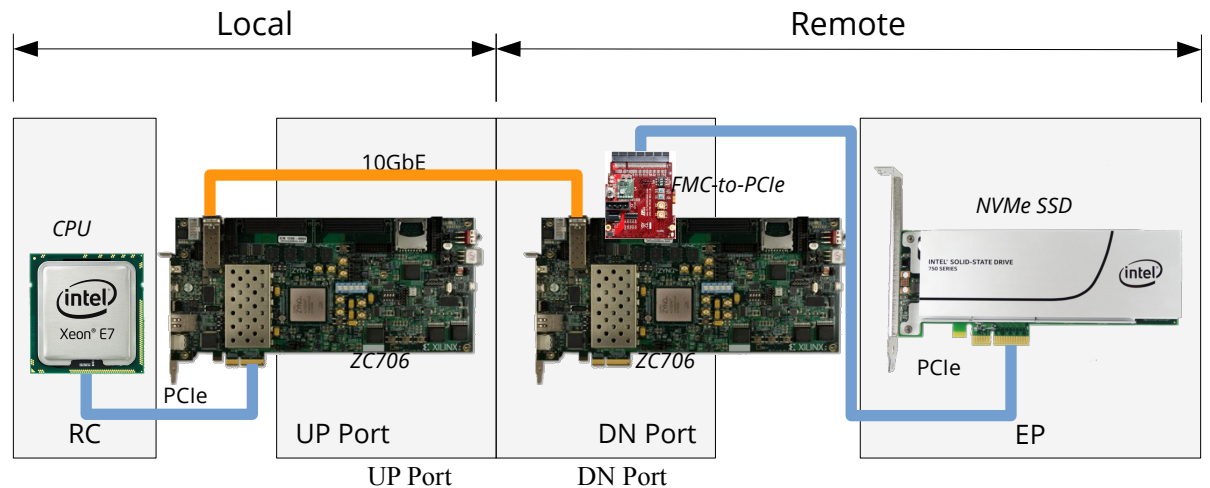
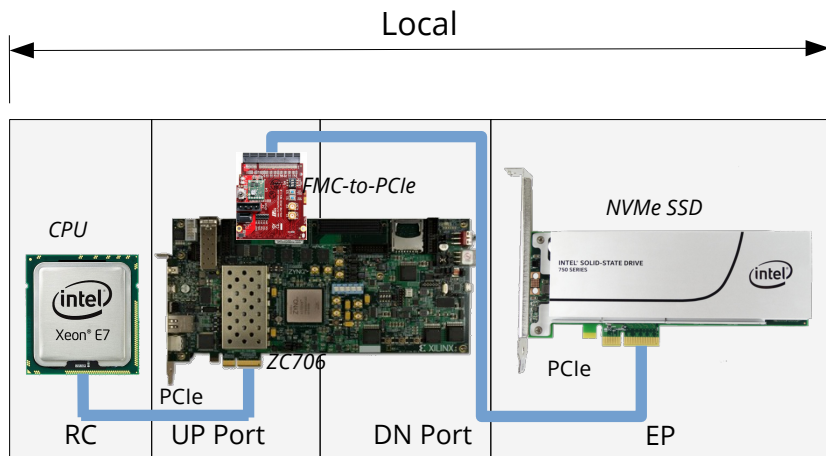
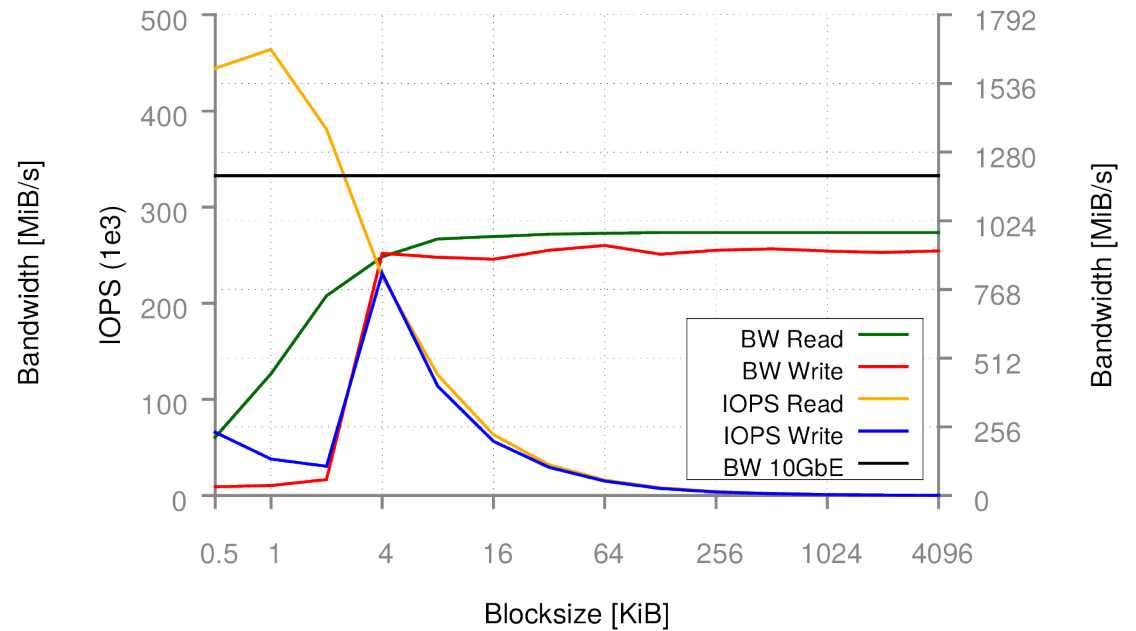
- Ethernet II Header**
[0000-0005] Dst. MAC: (02 00 00 00 00 11) → 02:00:00:00:00:11
[0006-0011] Src. MAC: (02 00 00 00 00 55) → 02:00:00:00:00:55
- Internet Protocol Header**
[0017] Protocol: (06) → TCP
[001a-001d] Src. IP: (c0 a8 01 69) → 192.168.1.105
[001e-0021] Dst. IP: (c0 a8 01 65) → 192.168.1.101
- 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**

Implementation

Local switch



Distr. switch with TLP aggregation



Conclusion

- Reliable “tunneling” of PCI Express via TCP/IP
- Fully transparent to PCIe Root Complex and Operating System
- FPGA processing enables bandwidth at 10 GigE line rates

- Based on “XPressRICH3” PCIe IP Core from PLDA



- Please visit us at the show:



HHI – Hall 4 Booth



MLE – Hall 2 Booth 2-421



Xilinx – Hall 1 Booth 1-205