

A 10 Gigabit Ethernet TCP/IP Stack Implementation on MicroChip PolarFire for High-Speed Camera Image Transport

Missing Link Electronics

Ulrich Langenbach, Andreas Schuler

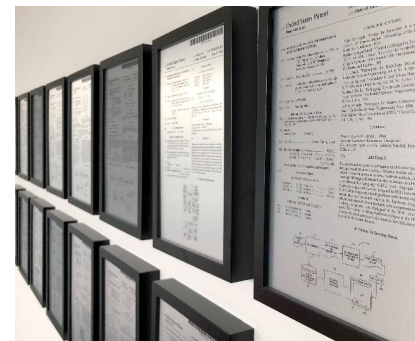
MLE - Experts for Domain-Specific Compute Architectures

Our Mission is

- to support customer projects with deep expertise and hands-on design services
- Offering pre-validate FPGA subsystems of FPGA IP blocks and open-source software
- Applying novel FPGA design methodologies for increased productivity
- Partners to

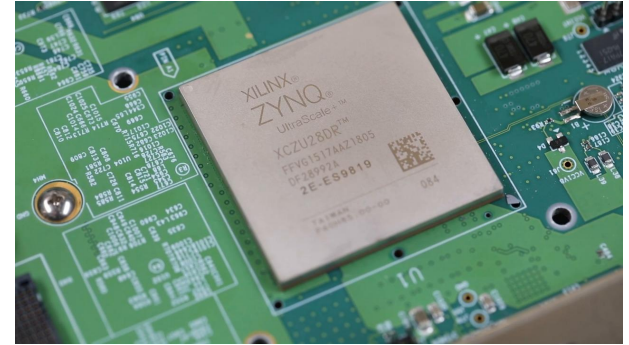
Headquartered in Silicon Valley with Design Offices in Germany

- Founded 2010, employee owned
- 17+ Certified FPGA Designers
- 50+ Presentations at Technology Conferences, 5 Patents awarded



Our Design Services Expertise

- RTL and High-Level Synthesis using Intel or Xilinx Toolflows
- Zynq-7000 SoC in designs since Q1/2012
- Zynq Ultrascale+ MPSoC in designs since Q4/2015
- Zynq UltraScale+ RFSoc in designs since Q2/2018
- Arria-10, Cyclone V SoC PCIe subsystems
- PetaLinux / Vanilla Linux and Yocto-based SW development
- Multigigabit transceiver configurations
 - PCIe Gen2/3/4/5, SATA 3/6G, SAS 6/12G, NVMe,
 - CAPI, JESD204B, DP/HDMI, MIPI CSI-2 D-PHY
 - 10/25/4050/100G Ethernet, Low Latency Ethernet
- Radar & Lidar for civil, mil/aero, automotive, industrial
- Image processing for HDMI, Displayport, SDI
- Time Sensitive Networking, Detnet, Layer-2 Switching
- Functional Safety Design Flows ISO 26262 (ASIL), IEC 61508 (SIL)
- Security & Trust (PUF, Crypto, OP-TEE)



Agenda

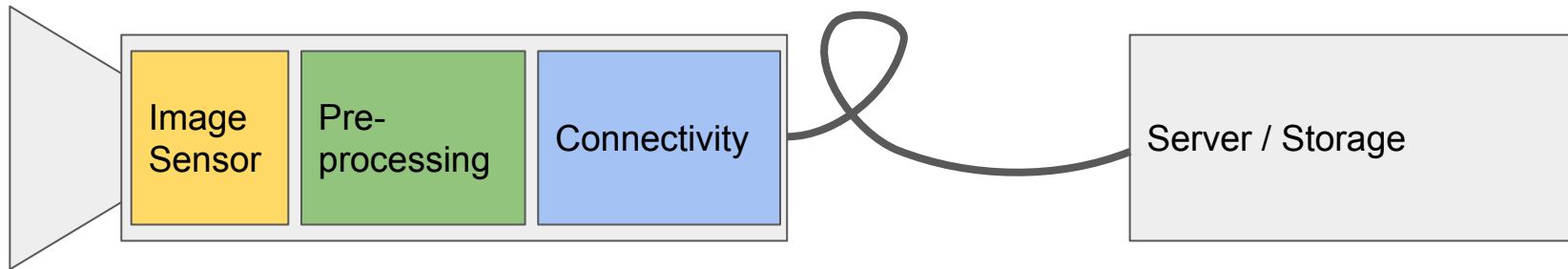
- 1) Application - Camera Image Transportation
=> Why TCP/IP?
- 2) Microchip Polarfire Overview
- 3) Protocol Overview
- 4) TCP/IP
 - a) Why TCP/IP?
 - b) How TCP/IP Works
- 5) NPAP
 - a) Overview Stack
 - b) Overview ERD
 - c) Latency
- 6) NPAP Applications

Camera Image Transport

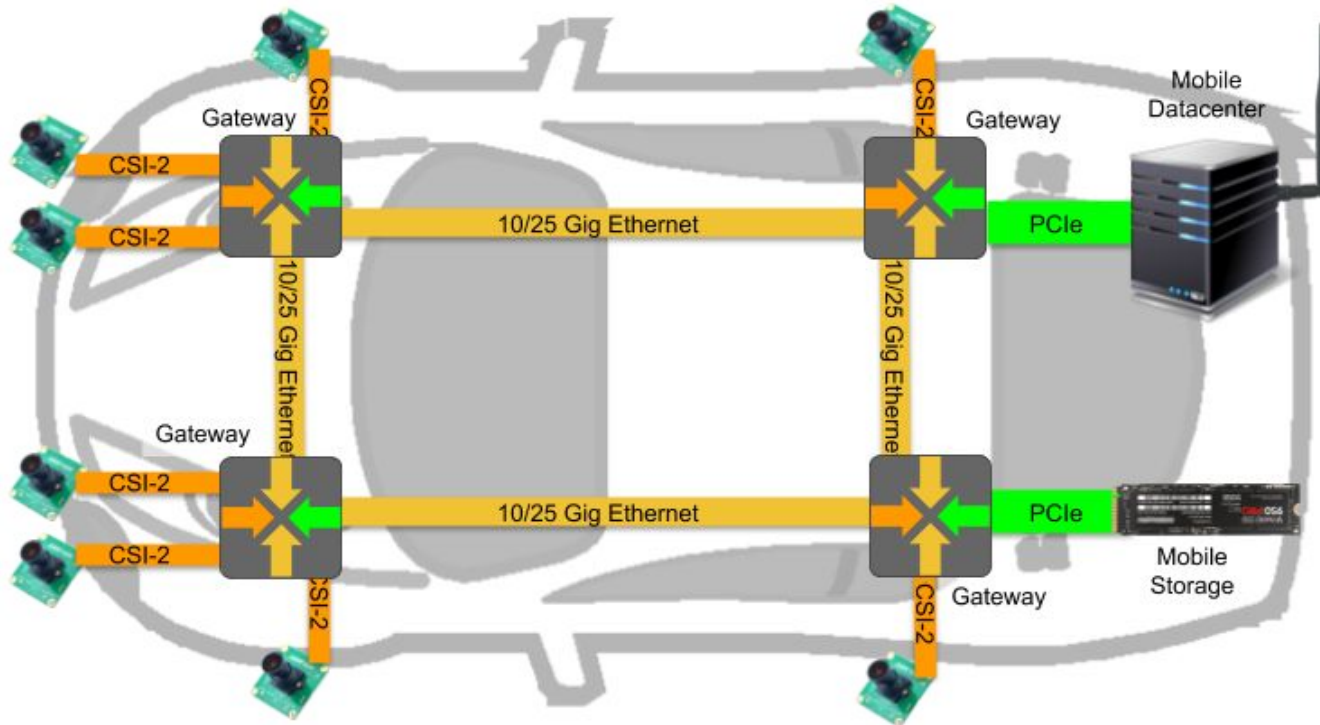
Cameras getting more demanding in regards of Bandwidth

Preprocessing is not always possible - raw data is required

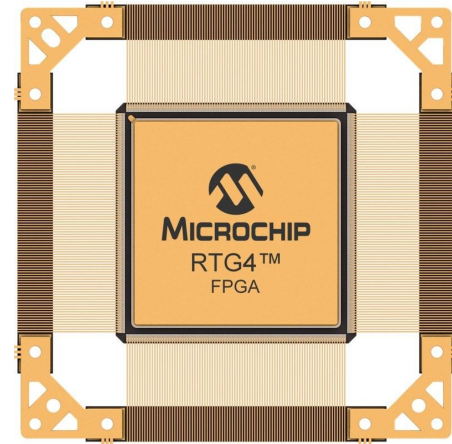
Long distance between camera and server/operator



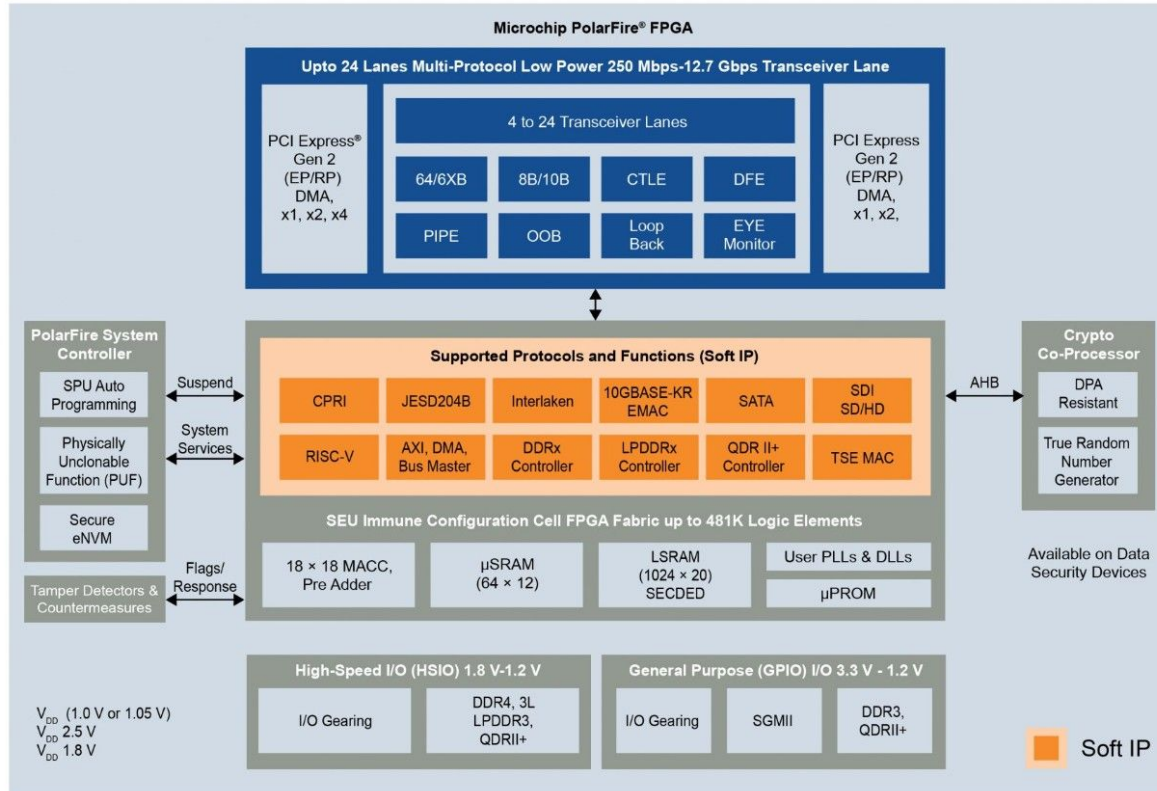
Zone-Based 10 GigE Automotive Backbone



MicroChip - PolarFire



PolarFire - What's inside?



Why Polarfire?

Non-volatile FPGA fabric

Low Power

- Low device static power
- Low inrush current
- Low power transceivers

Reliability Features

- Configuration cells single event upset (SEU) immune

Security Features

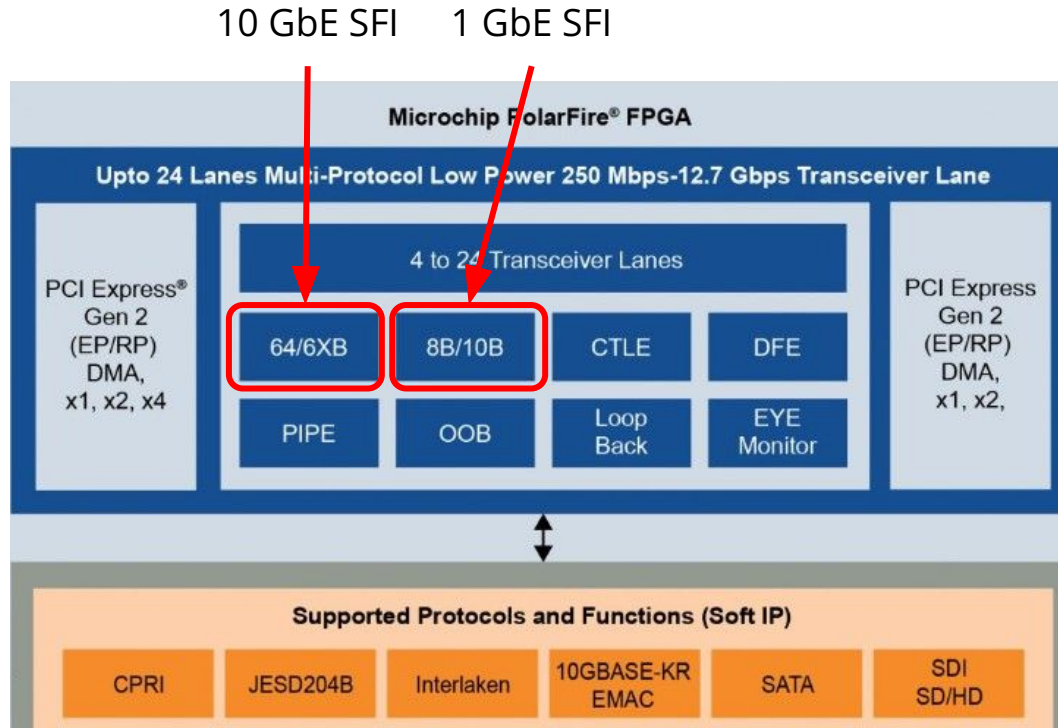
- Differential Power Analysis protection
- Physical Unclonable Function
- Secure Non-volatile Memory



Overview

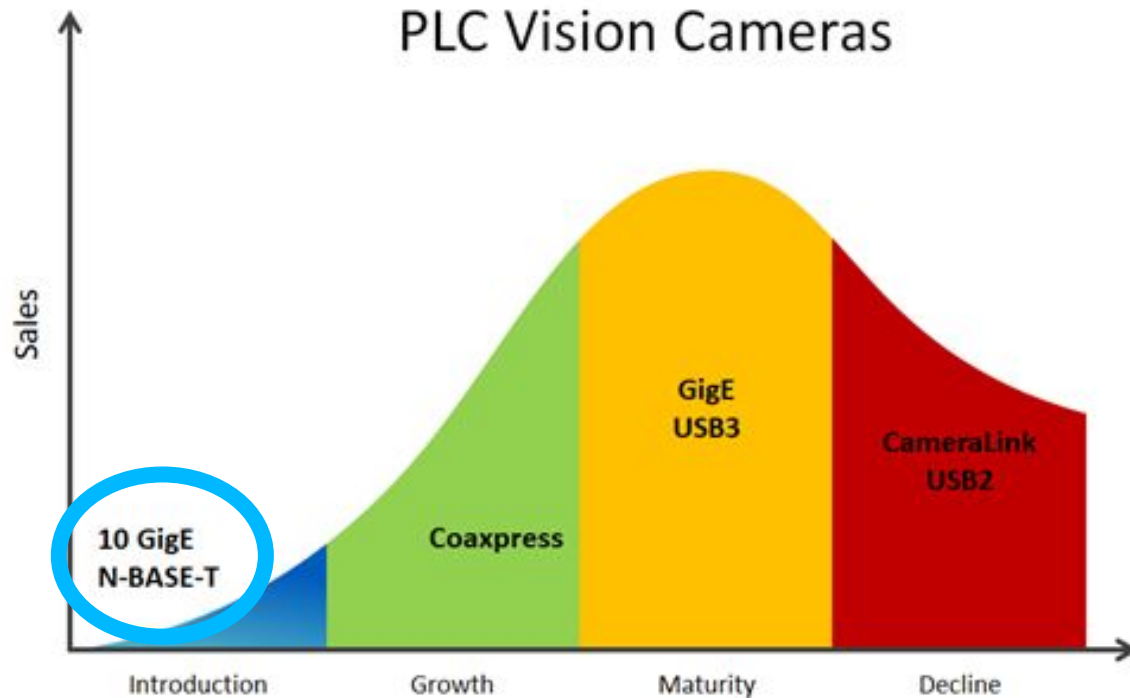
Features	MPF050T	MPF100T	MPF200T	MPF300T	MPF500T
K Logic elements (4 LUT + DFF)	48	109	192	300	481
Math blocks (18 x 18 MACC)	150	336	588	924	1480
LSRAM blocks (20 kbit)	160	352	616	952	1520
μSRAM blocks (64 x 12)	450	1008	1764	2772	4440
Total RAM (Mbits)	3.6	7.6	13.3	20.6	33
μPROM (Kbits)	216	297	297	459	513
User DLLs/PLLs	8	8	8	8	8
250 Mbps to 12.5 Gbps SERDES lanes	4	8	16	16	24
PCIe Gen2 endpoints/root ports	2	2	2	2	2
Total user I/Os	176	284	368	512	584

High-Speed Transceivers



<https://www.microsemi.com/blog/2018/04/10/polarfire-fpga-transceivers/>

Market Development of Image Transport Techn.



<https://www.get-cameras.com/How-to-select-a-machine-vision-camera-interface-USB3-GigE-5GigE-10GigE-Vision>

Protocol Overview



SDI



GMSL



TCP/IP

Protocol Overview - Wide Area > 50 m



SDI

TCP/IP

Protocol Overview - Interoperable with IT Equ.



TCP/IP

Protocol Overview - Interoperable with IT Equ.



We do **TCP/IP**

Why TCP/IP with cameras?

Mature Protocol - it is around for more than 40 years

De facto standard of the Internet

Guaranteed delivery, back pressure capability -> it's a big, distributed FIFO!

Widely available commercial off-the-shelf (cots) hardware

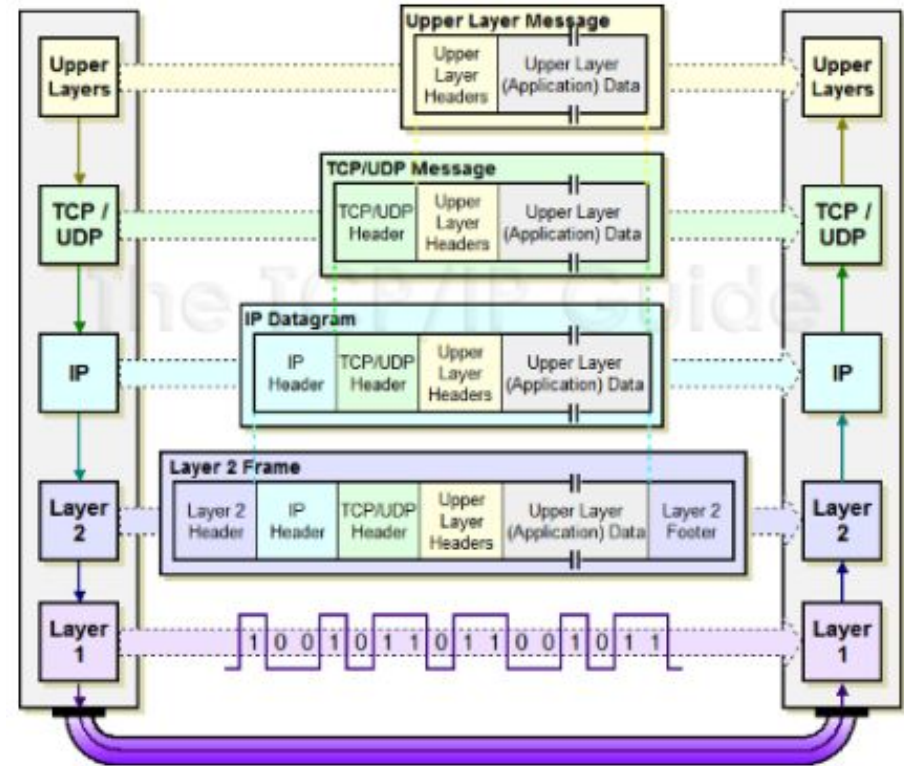
Options to add features through additional Layers, on top or below:

- Time Sensitive Network (TSN)
- Media Access Control Security (MACsec)
- Transport Layer Security (TLS)

TCP Facts

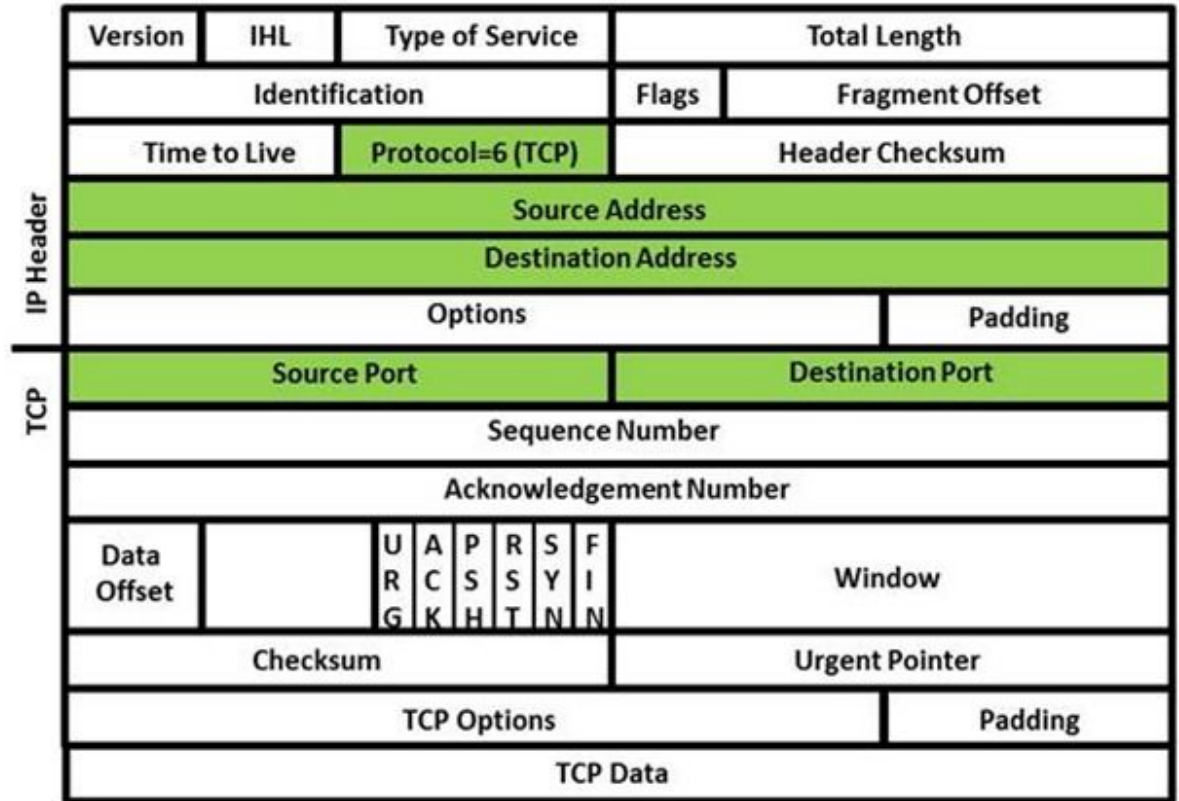
Layered architecture

- “Packet”-based with data segmented into Protocol Data Units (PDU)
- TCP message – PDU at TCP layer
- Datagram – PDU at IP layer
- Frame – PDU at link-layer
- Communication is
- Reliable
- Ordered
- Error-checked



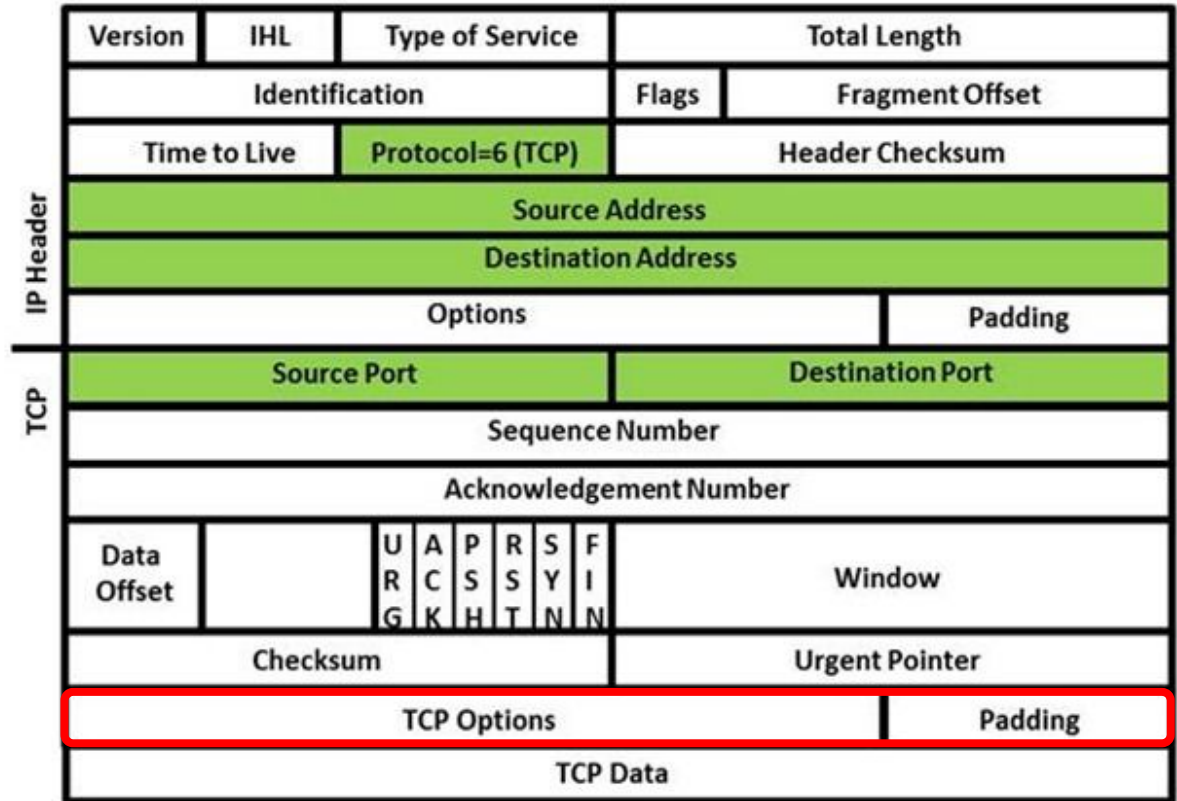
TCP/IP Header

TCP/IP Packet



TCP/IP Header Options

TCP/IP Packet



Various web application driven additions available, e.g. via TCP Options, such as *session cookies* reducing the number of 3 way handshakes required to load a single web page

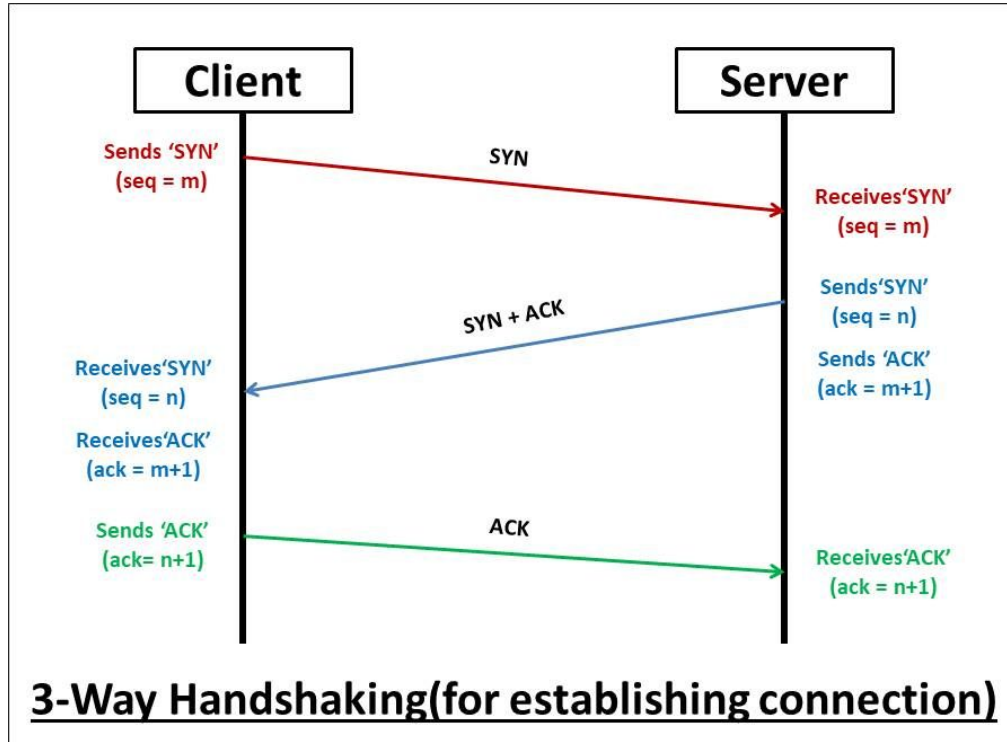
How TCP works - The Handshake(s)



No, not this one

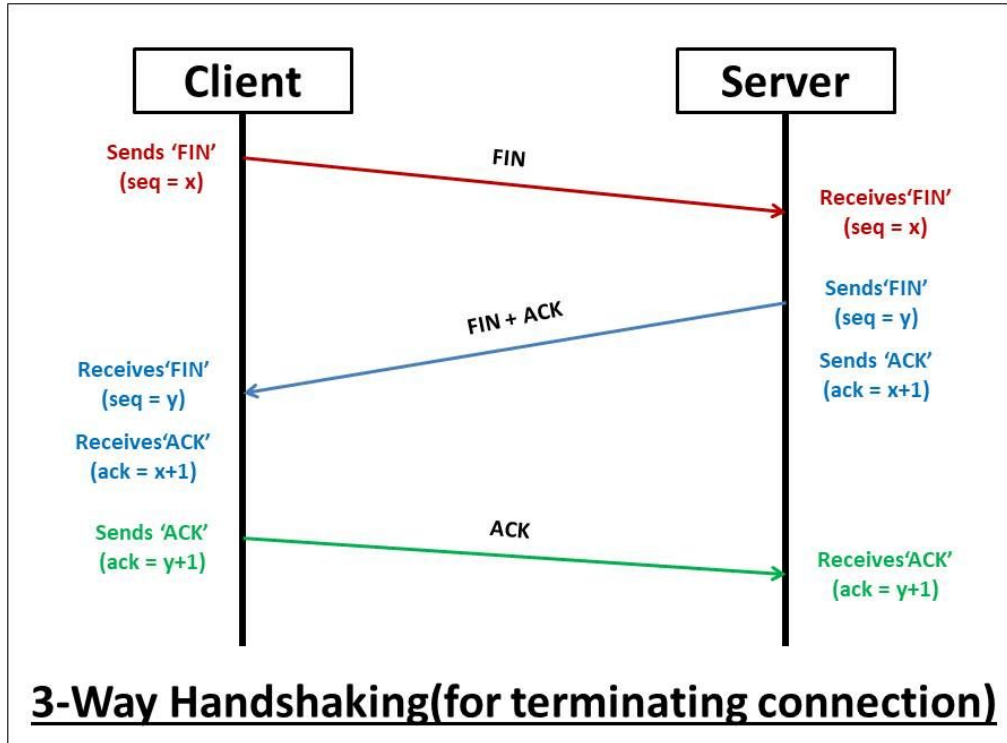
<https://www.freepik.com/vectors/corona-virus-cartoon>
created by brgfx - www.freepik.com

The 3 Way Handshake (establish connection)



<https://afteracademy.com/blog/what-is-a-tcp-3-way-handshake-process>

The 3 Way Handshake (teardown connection)

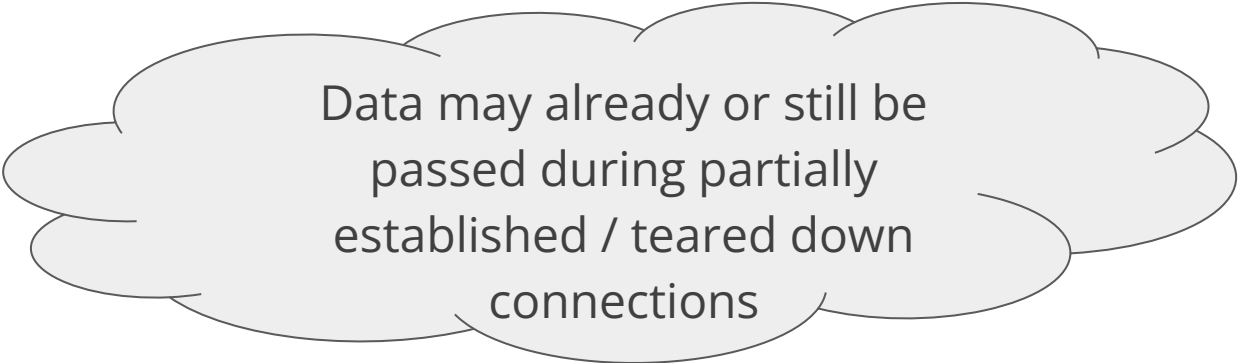


<https://www.academymy.com/blog/what-is-a-tcp-3-way-handshake-process>

The 3 Way Handshakes

1. Make sure both sides are on the same page
2. Enable both sides to detect if something got wrong
(a packet was lost)
3. Respective flags are handled as if they were a Byte of payload

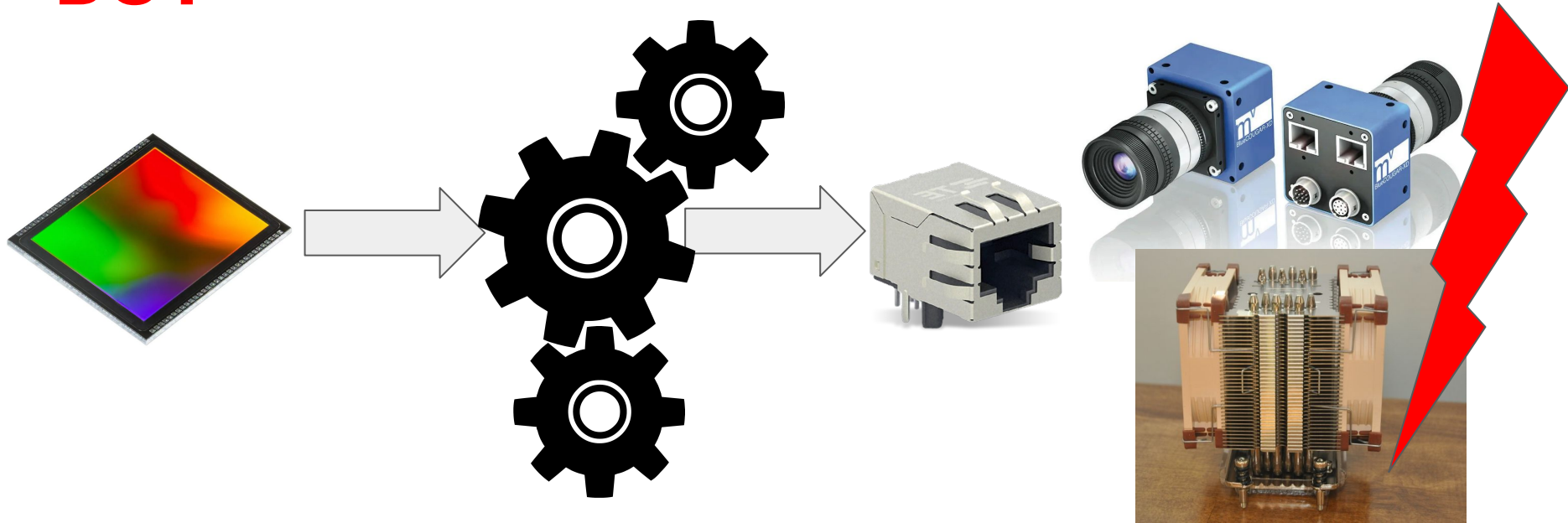
=> This actually provides integrity and consistency



Data may already or still be
passed during partially
established / teared down
connections

TCP Implementation, usually a software domain!

BUT



Dataflow processing fits best to the power, compute and space requirements!

NPAP

Network Protocol Acceleration Platform

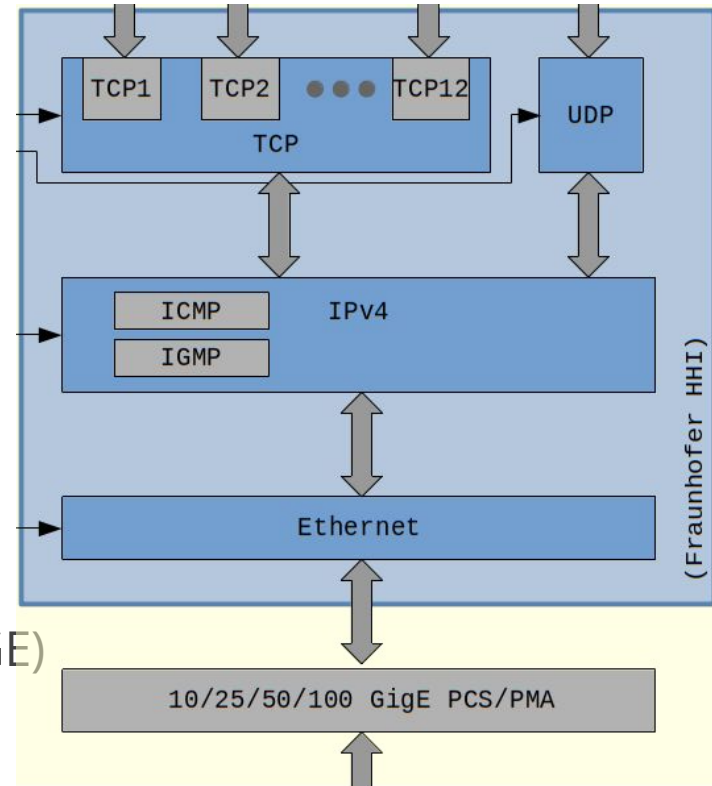
NPAP - Network Protocol Acceleration Platform

What is NPAP?

NPAP is a TCP/UDP/IP Full accelerator and is operated processor independent

Key features:

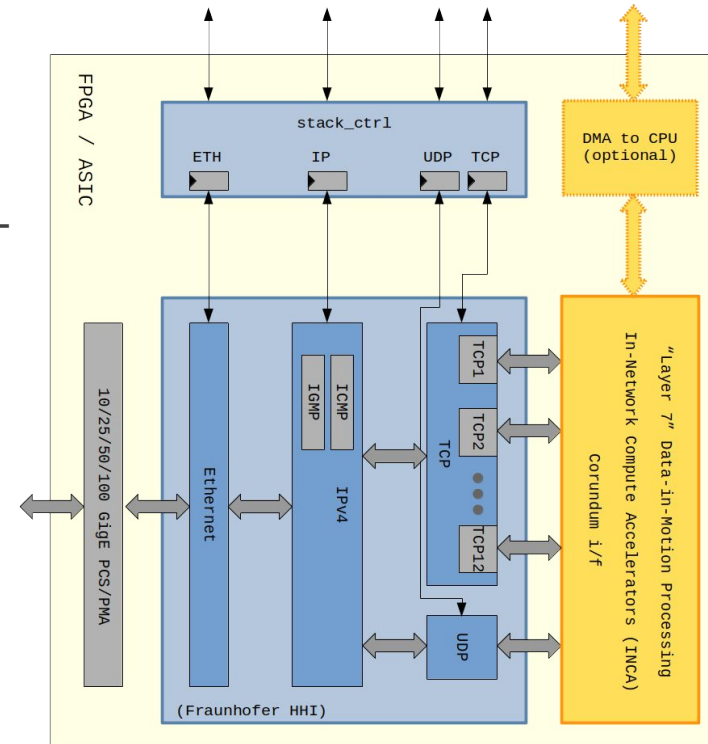
- IPv4 with ICMP and IGMP
- TCP/UDP with AXI-S interfaces
- DHCP client
- Different speeds available (10/25/40/50/100 GE)
- Jumbo frame support
- Low latency and deterministic
- Configurable buffers for each session and direction



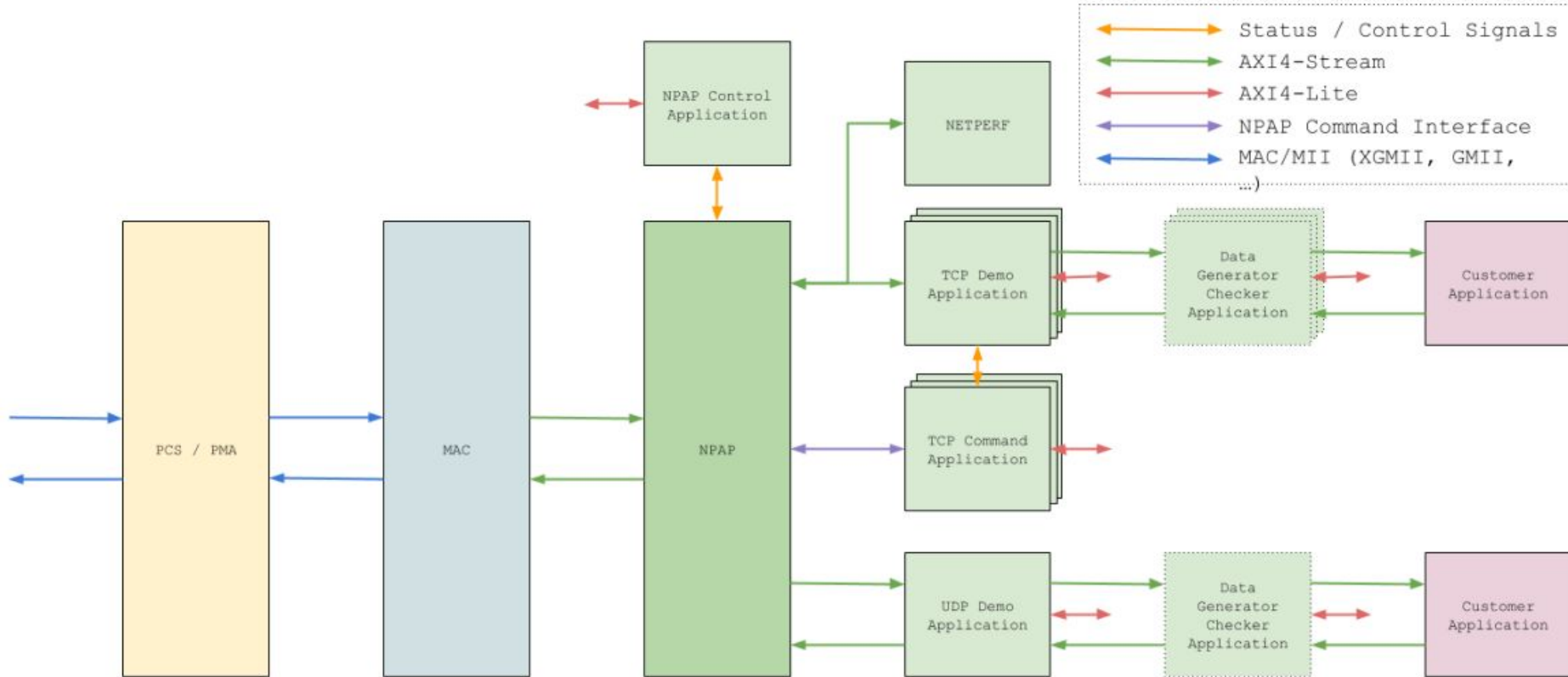
NPAP - Why Platform and not IP

It is delivered as an Evaluation Reference Design (ERD) and consists of:

- MAC (depending on speed/ FPGA technology - eval required)
- TCP/UDP/IP full accelerator
- Control Flow
- Examples for handling
 - TCP Sessions
 - UDP
 - Stack control
- Netperf
 - Open Source Network Bandwidth Measurement tool

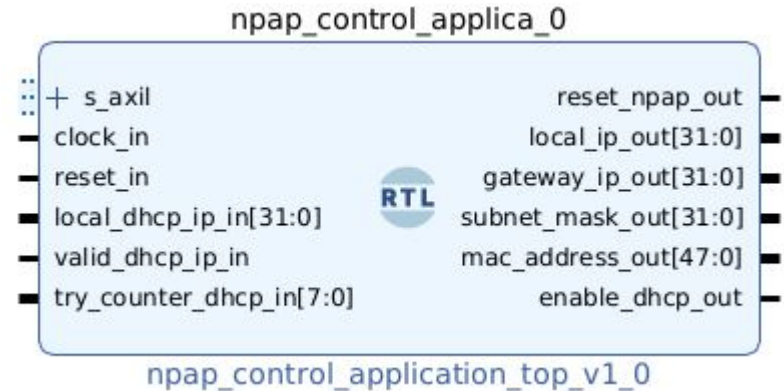


NPAP - Evaluation Reference Design (ERD)



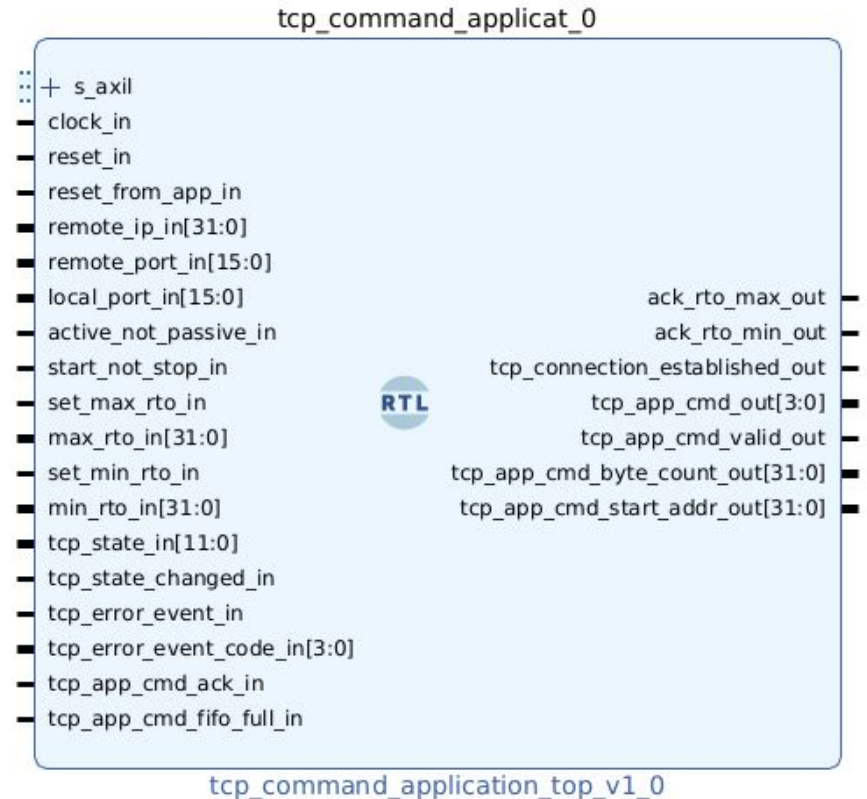
NPAP Control Application

- AXI4-Lite Interface
- NPAP Control (IP, MAC, ...)
- DHCP Control and Status
- NPAP Reset
- One IP instance per NPAP instance



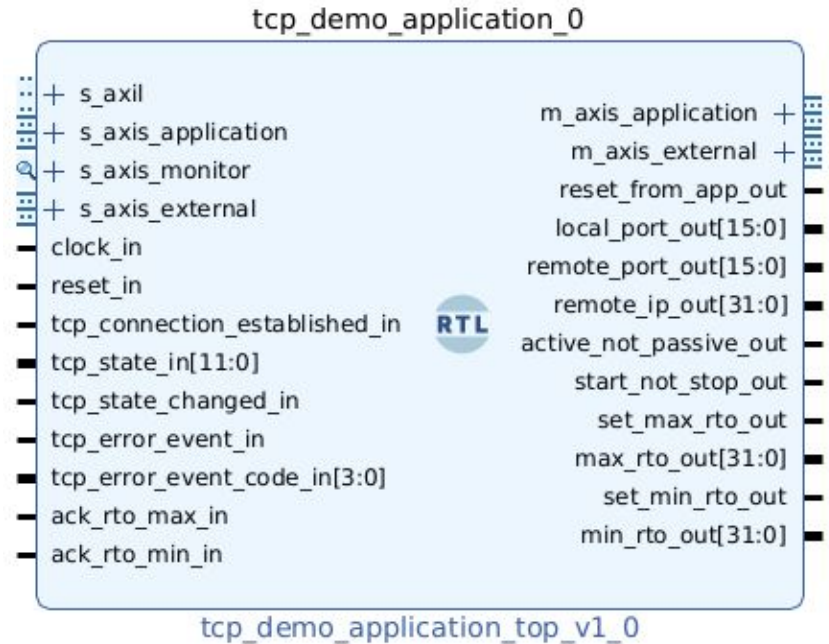
TCP Command Application

- AXI4-Lite Interface
- Example TCP Command Interface implementation
- One IP instance per TCP session
- Controlled by TCP Demo Application or standalone usage



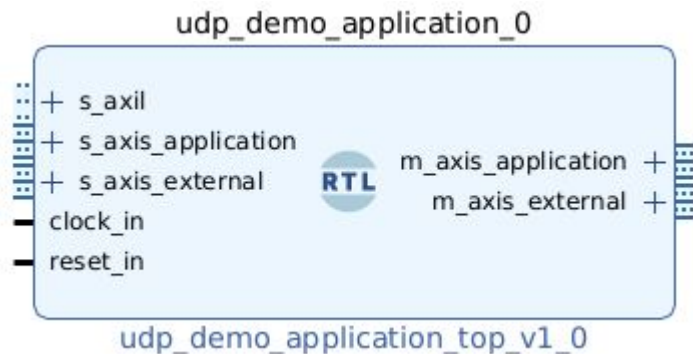
TCP Demo Application

- AXI4-Lite Interface
- Data stream control (loopback, discard, external)
- TCP Session Reset
- One IP instance per TCP session
- Controls TCP Command Application



UDP Demo Application

- AXI4-Lite Interface
- Data stream control (loopback, discard, external)
- TUSER setting (per Datagram meta-data, e.g. source + destination ports)
- One IP instance per UDP port



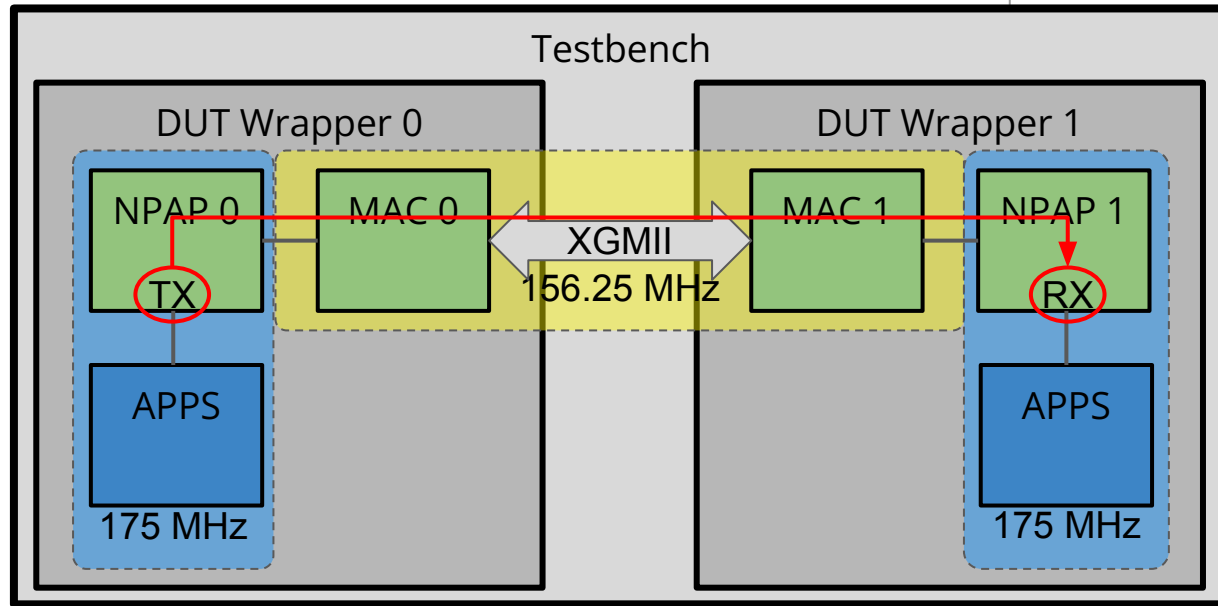
Design Flows

- Vivado Block Diagram Flow
 - Based on IPXACT packages IP cores
 - Allows for quick design generation
- Classic RTL Based Flow
 - File inclusion into project
 - Library assignment for files
 - Traditional Verilog or VHDL module instantiations
- Tool specific IP integration Flow(s)
 - Usually TCL script based
 - Adds sources and may provide interface bundles



NPAP - Performance & Metrics

Simulation Latency Results



TCP Payload Size [Byte]	Latency [ns]
1	462,8
10	457,4
16	485,8
64	520,0
160	656,0
448	1092,1
720	1502,7
960	1868,9
1216	2251,3
1456	2622,7

NPAP - Performance & Metrics

Round Trip Time and Throughput

Symbol	Parameter	Condition	Value	Units
RTT _{avg}	Average round-trip time	BL= 1	10,10	µs
		BL=10	14,20	µs
		BL=100	24,54	µs
		BL=1k	26,75	µs
TPR _{avg}	Average network throughput	BL = 1	4,17	Gbps
		BL = 10	7,21	Gbps
		BL = 100	8,97	Gbps
		BL = 1k	9,23	Gbps
BDP _{avg}	Average bandwidth-delay product	BL = 1	44	kbit
		BL = 10	101	kbit
		BL = 100	255	kbit
		BL = 1k	246	kbit

¹ Burst length BL=1 (=8 kByte)

² Measurement setup = HHI Stack-to-HHI Stack

³ The values are valid for XILINX Virtex 5XC5VFX130T, speed grade - 2 FPGA. The reference test hardware is the HHI 10G EthEval board.

NPAP - Performance & Metrics

The Bandwidth-Delay-Product

- Is a metric for network system performance
- Provides an estimate for buffer sizing

=> Let's have a closer look!

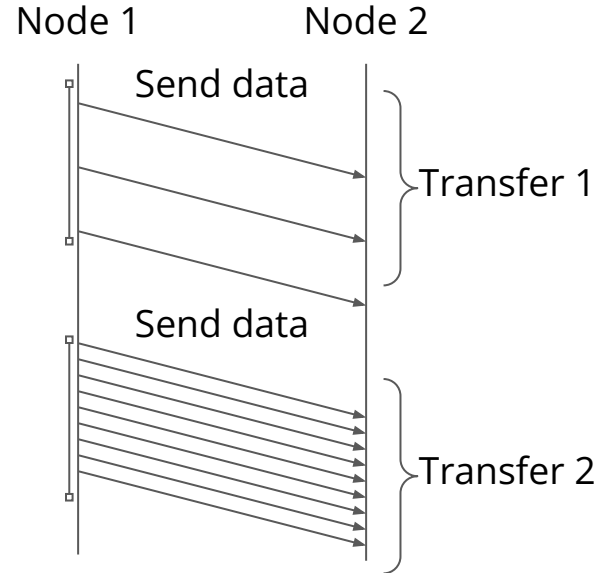
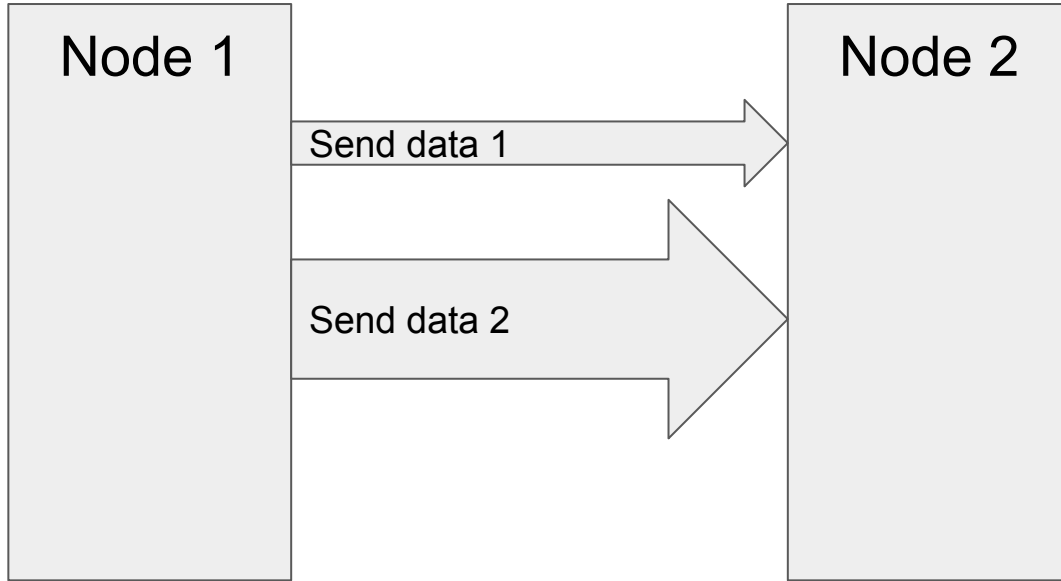
	network throughput	BL = 10	7,21	Gbps
		BL = 100	8,97	Gbps
		BL = 1k	9,23	Gbps
BDP_{avg}	Average bandwidth-delay product	BL = 1	44	kbit
		BL = 10	101	kbit
		BL = 100	255	kbit
		BL = 1k	246	kbit

¹ Burst length BL=1 (=8 kByte)

² Measurement setup = HHI Stack-to-HHI Stack

³ The values are valid for XILINX Virtex 5XC5VFX130T, speed grade - 2 FPGA. The reference test hardware is the HHI 10G EthEval board.

Bandwidth



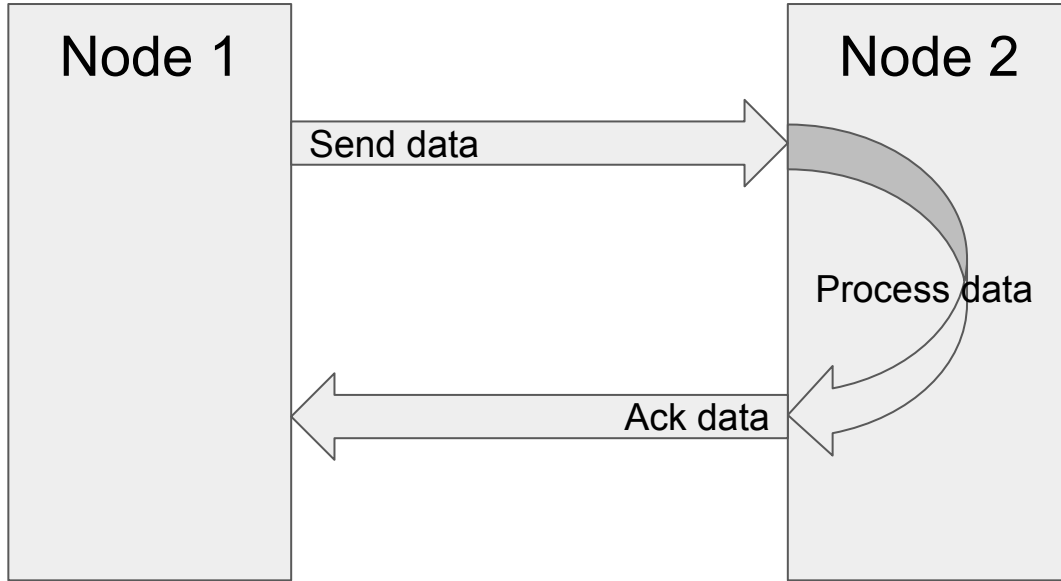
(Equally sized packets = 1 unit [Bit])

$$\text{Transfer 1: } B_1 = 3u / T$$

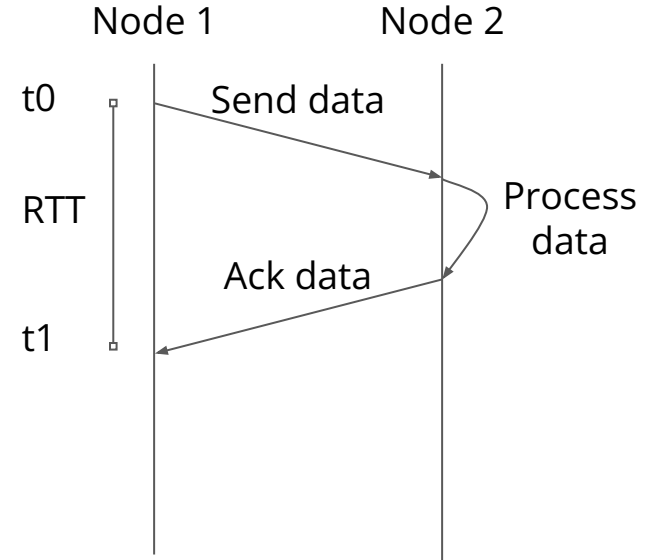
$$\text{Transfer 2: } B_2 = 9u / T = 3 * B_1$$

Bandwidth B = data quantity per time interval
= data quantity / (t1 - t0) [Bit/s]

Delay a.k.a. RTT

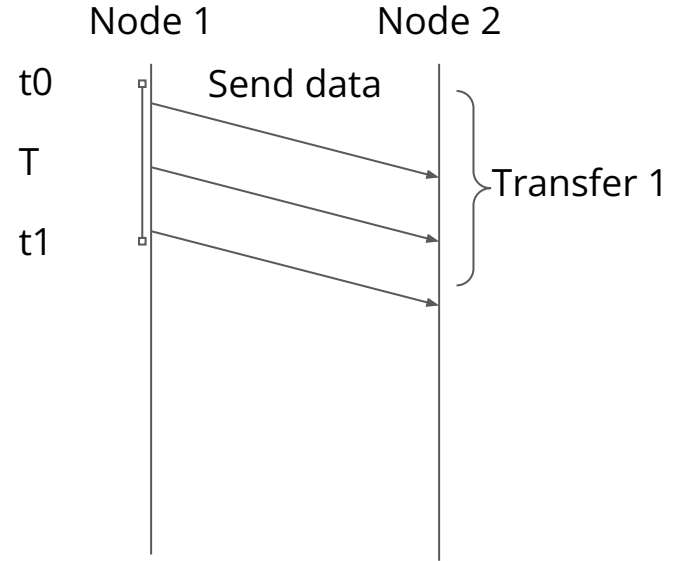
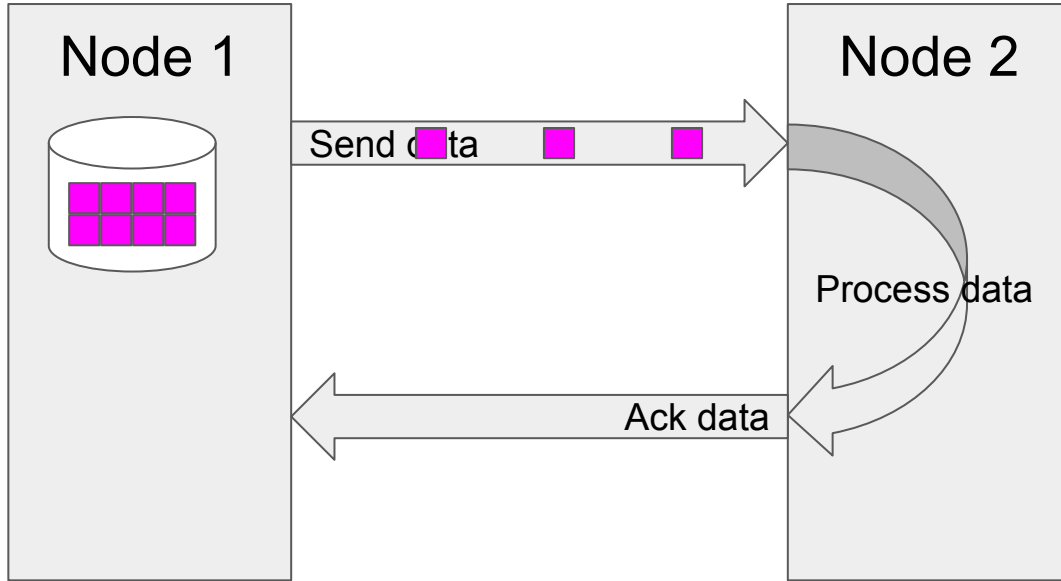


$RTT = 2 * \text{Latency}$ (for symmetric systems)



$RTT = t_1 - t_0$ [s]

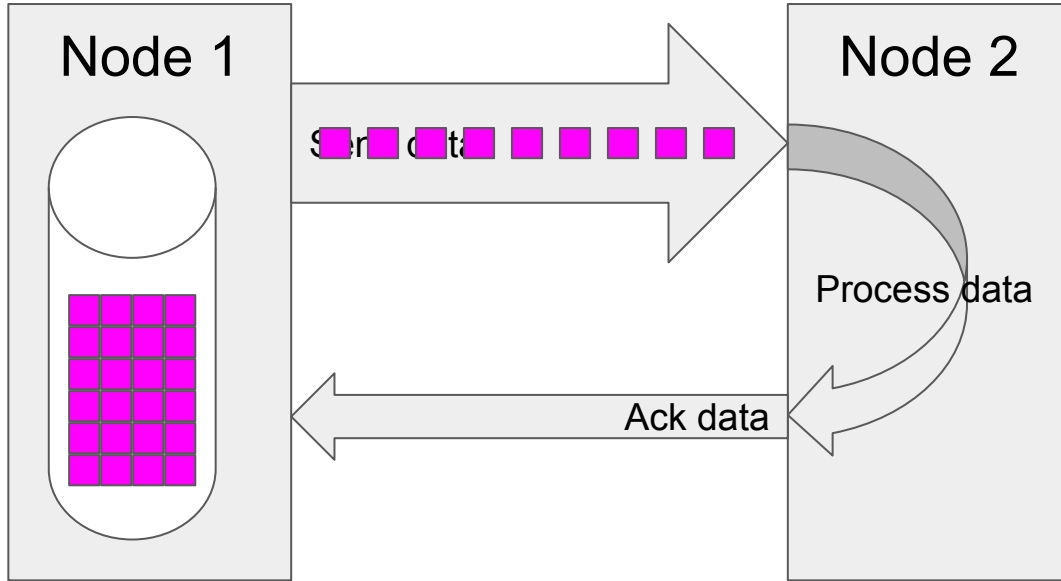
Bandwidth-Delay-Product - Low Bandwidth



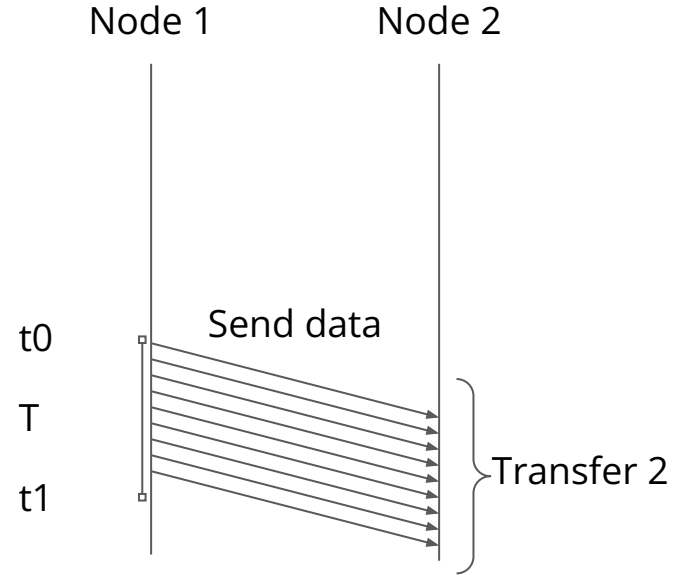
$$\begin{aligned}\text{Bandwidth-Delay-Product} &= \text{Bandwidth} * \text{Delay} \\ &= B \text{ [Bit/s]} * T \text{ [s]} \\ &= \text{BDP [Bit]}\end{aligned}$$

(Equally sized packets = 1 unit [Bit/s])

Bandwidth-Delay-Product - High Bandwidth

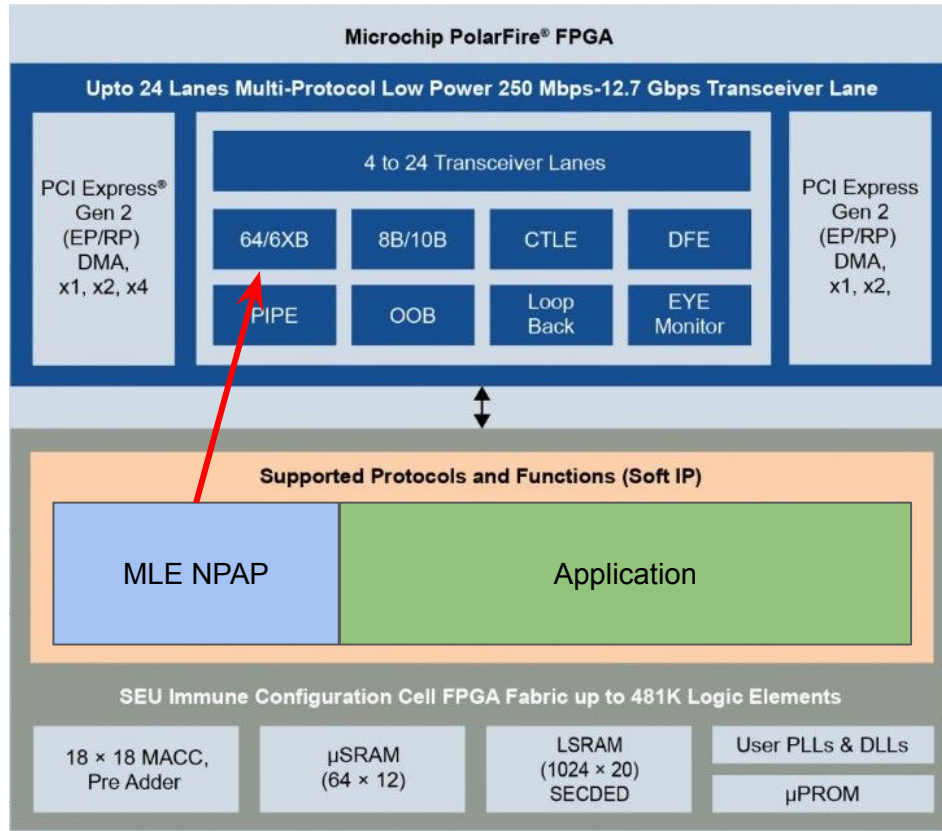


More data in flight during the RTT -> larger buffer required to cover for potentially missed packets (re-transmission buffer)



(Equally sized packets = 1 unit)

NPAP on PolarFire



Resource Utilisation

The following table shows resources synthesized for MicroSemi PolarFire MPF300TS-1FCG1152I using Libero 2021.1 - instantiating the following design features:

- Ethernet
- IPv4
- UDP
- 3 instances of TCP

Instance	Fabric 4LUT	Fabric DFF	Interface 4LUT	Interface DFF	uSRAM 1K	LSRAM 18K	Math 18x18	Chip Global
npap_tcp_udp_wrapper_u0	60339	31116	6792	6792	122	146	2	12
npap_tcp_udp_top_u0	60339	31116	6792	6792	122	146	2	12
Primitives	13	1	0	0	0	0	0	0
interface_adapter (all)	57	1	0	0	0	0	0	0
wrapper_ll_ip_tcp_u0	60269	31115	6792	6792	122	146	2	12
Primitives	349	195	0	0	0	0	0	0
i_netLayerConv	249	206	0	0	0	0	0	0
i_tcpTxMux	123	2	0	0	0	0	0	0
gen_WithUdp.i_udp	5598	3997	1548	0	24	35	0	0
gen_tcpConnections[0].i	15560	7037	1740	1740	31	37	0	3
gen_tcpConnections[1].i	13287	6649	1068	1068	26	19	2	3
gen_tcpConnections[2].i	15707	7112	1752	1752	35	37	0	3
iBusScheduler8	106	34	0	0	0	0	0	0
i_internetLayer	3717	2995	216	216	6	4	0	1
i_networkLayer	5306	2665	504	504	0	14	0	0

Challenges migrating to Microchip Polarfire FPGAs

FPGAs of other vendors

- Registers are initialised during FPGA SRAM cell initialisation (bitstream load) to a specific value
 - a. An initial value is chosen by the tool
 - b. An initial value is provided by the developer
- Power-on-Reset is nice-to-have, but a good design practice

Microchip Polarfire FPGA

- Registers cannot be initialised during SRAM cell initialisation (bitstream load)
- To provide a defined POR state a Power-on-Reset is a must on this platform!

Challenges migrating to Microchip Polarfire FPGAs

- RTL descriptions written for other vendors' devices / families must not necessarily perform similar on Microchip Polarfire devices
- Code must be carefully reviewed and re-written to implement
 - POR for all registers that define the circuit state,
 - including re-writing potentially present initial values into a POR

NPAP Applications

Distributed PCIe NTB

- NTB: Non-transparent Bridge
- Prototypical implementation based on Xilinx ZU+ devices



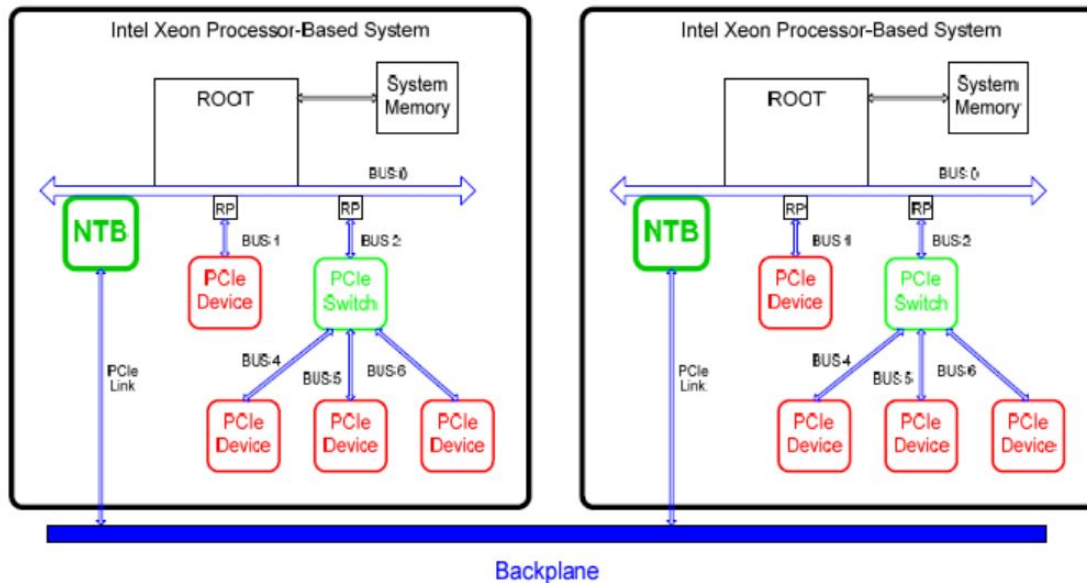
**Sensor Fusion and Data-in-Motion Processing
for Autonomous Vehicles**

**Endric Schubert, Ph. D.
CTO
Missing Link Electronics (MLE)**

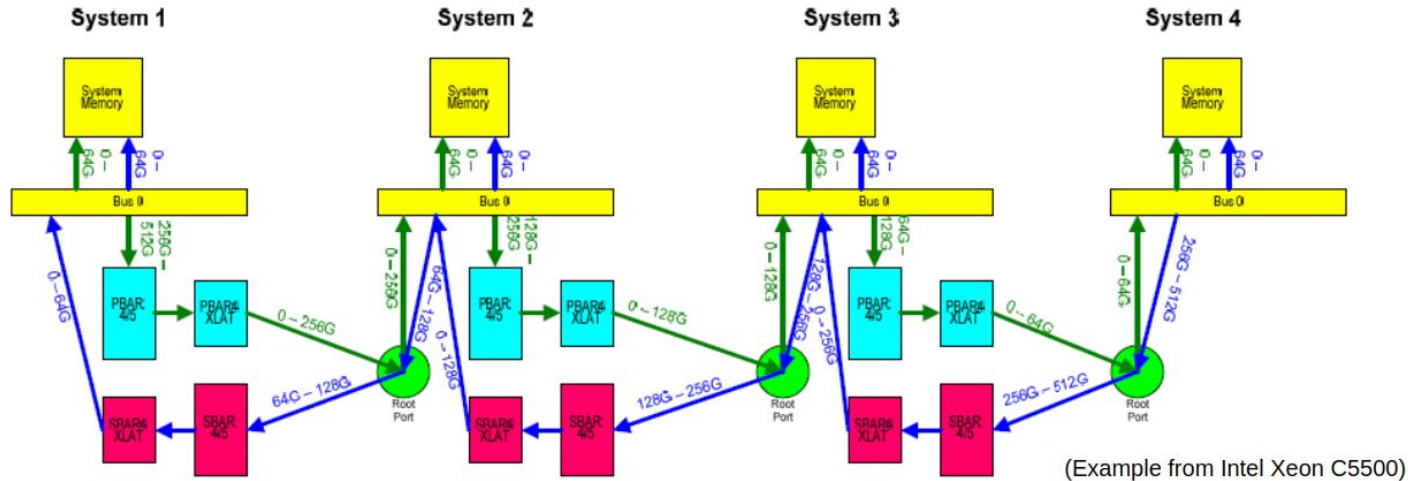
Copyright © 2019 PCI-SIG® - All Rights Reserved

PCIe Non-Transparent Bridge

- o **Non-Transparent Bridge (NTB) connects multiple Root Ports**
- o **Example of NTB Back-2-Back**
(Example from Intel Xeon C5500)



NTB: Multi-CPU Interconnect via a Daisy-Chain

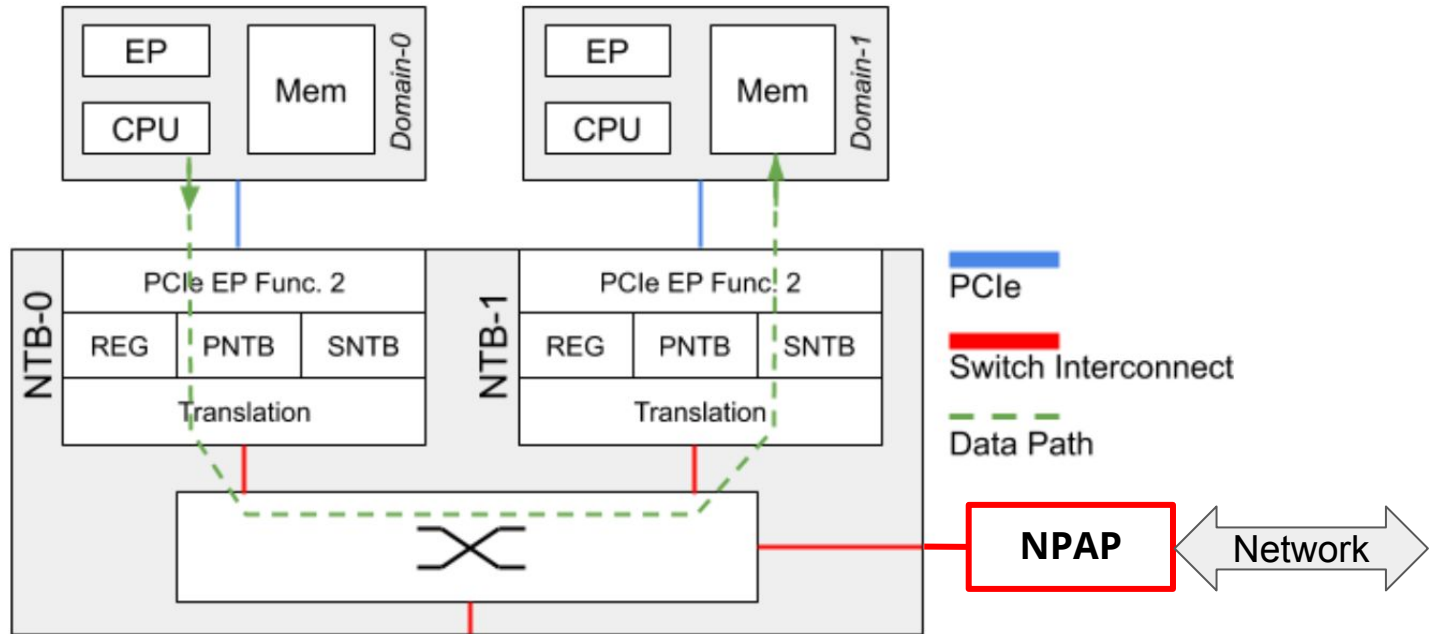


o Not optimal for Automotive ECU

- Shared Bandwidth
- Not resilient to HW failures
- Added Latency for ID translation

NTB: Multi-CPU Interconnect via a Daisy-Chain

Network-on-Chip for Any-2-Any Connectivity between PCIe Roots



PCIe Range Extension via TCP/IP

- Presented at PCI-SIG Developers Conference 2018
- Results of a prototypical implementation based on Xilinx Z7000
- Since then a new generation of prototypes is available based on Xilinx ZU+ devices



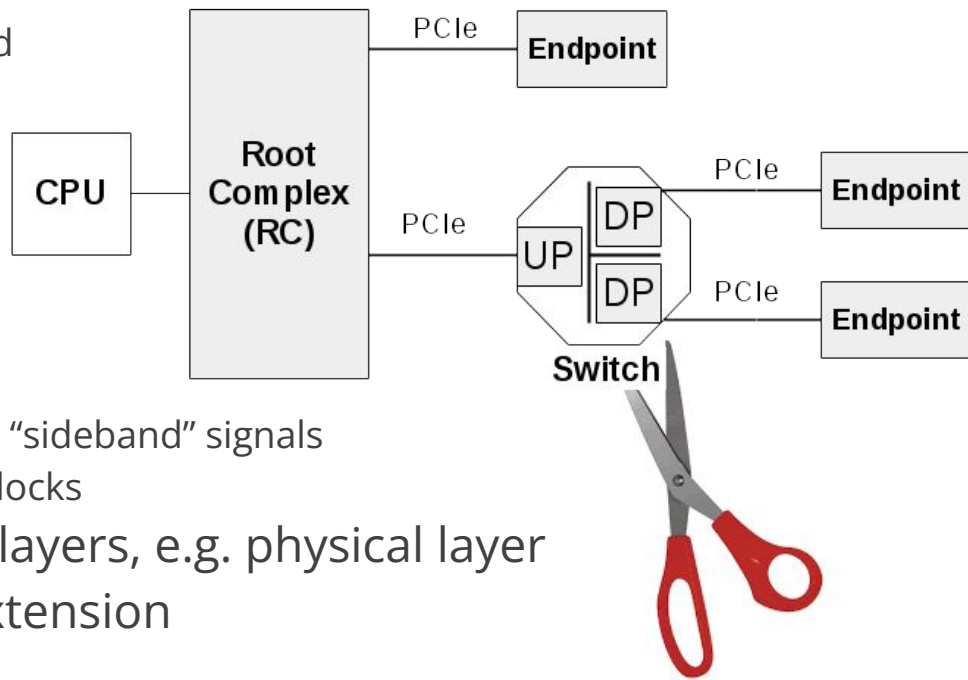
PCIe Range Extension via Robust, Long Reach Protocol Tunnels

Jim Peek
Director of Technology
Missing Link Electronics

Copyright © 2018 PCI-SIG® - All Rights Reserved

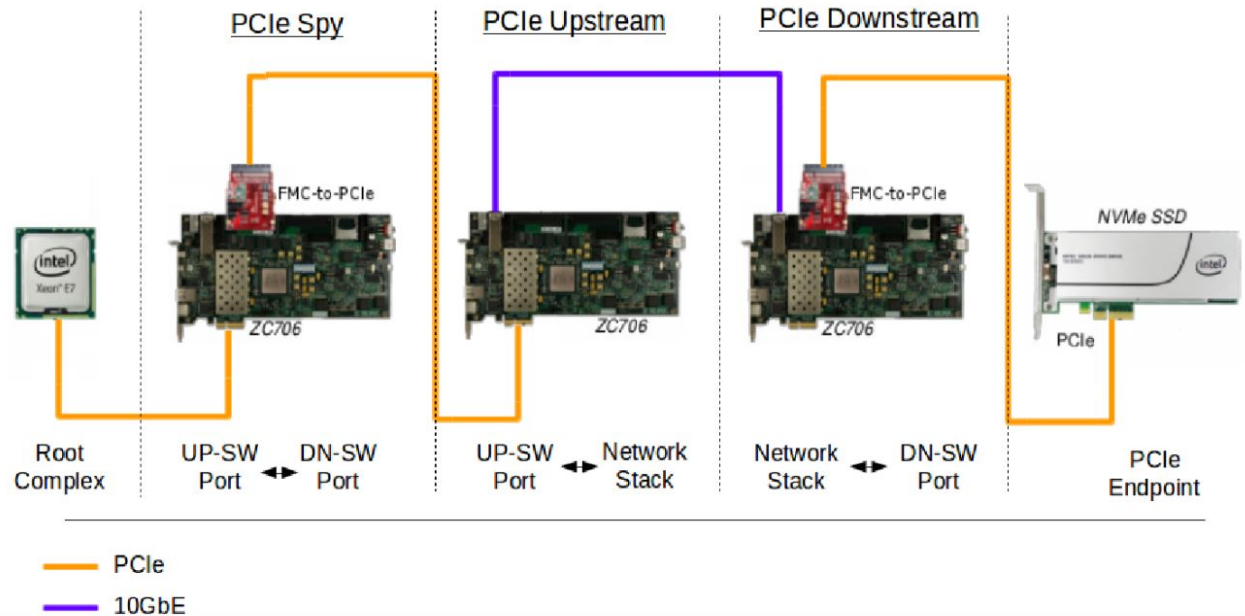
PCIe Transport via TCP/IP

- 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
- A “distributed” PCIe Switch
 - In accordance to PCIe Spec
 - Scalable via TCP session count
 - Supports latency requirements for “sideband” signals
 - Special care needed to avoid deadlocks
- Independent of lower network layers, e.g. physical layer
- Could be used as PCIe range extension



A Prototypical Implementation

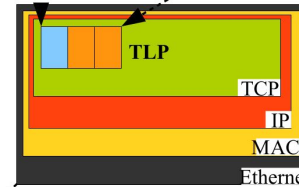
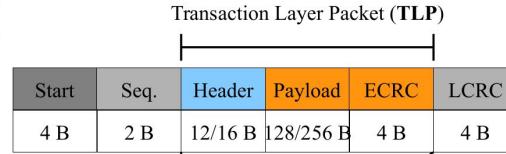
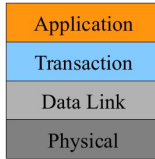
- Adds visibility to the PCIe path
 - Traditional Network tools now applicable to PCIe, e.g. Wireshark
- Using PCIe over TCP/IP also opens PCIe for simple (performance) monitoring via network traces**



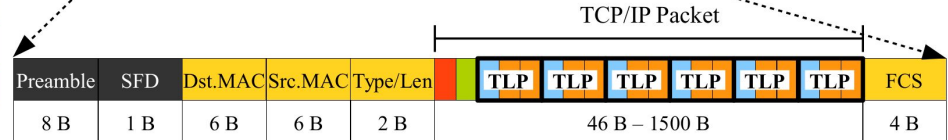
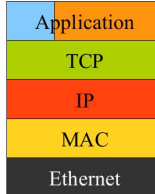
Concept of PCIe-over-TCP (1)

- Network Protocol Stack
- Encapsulation of PCIe Transaction Layer Packets (TLPs) into TCP

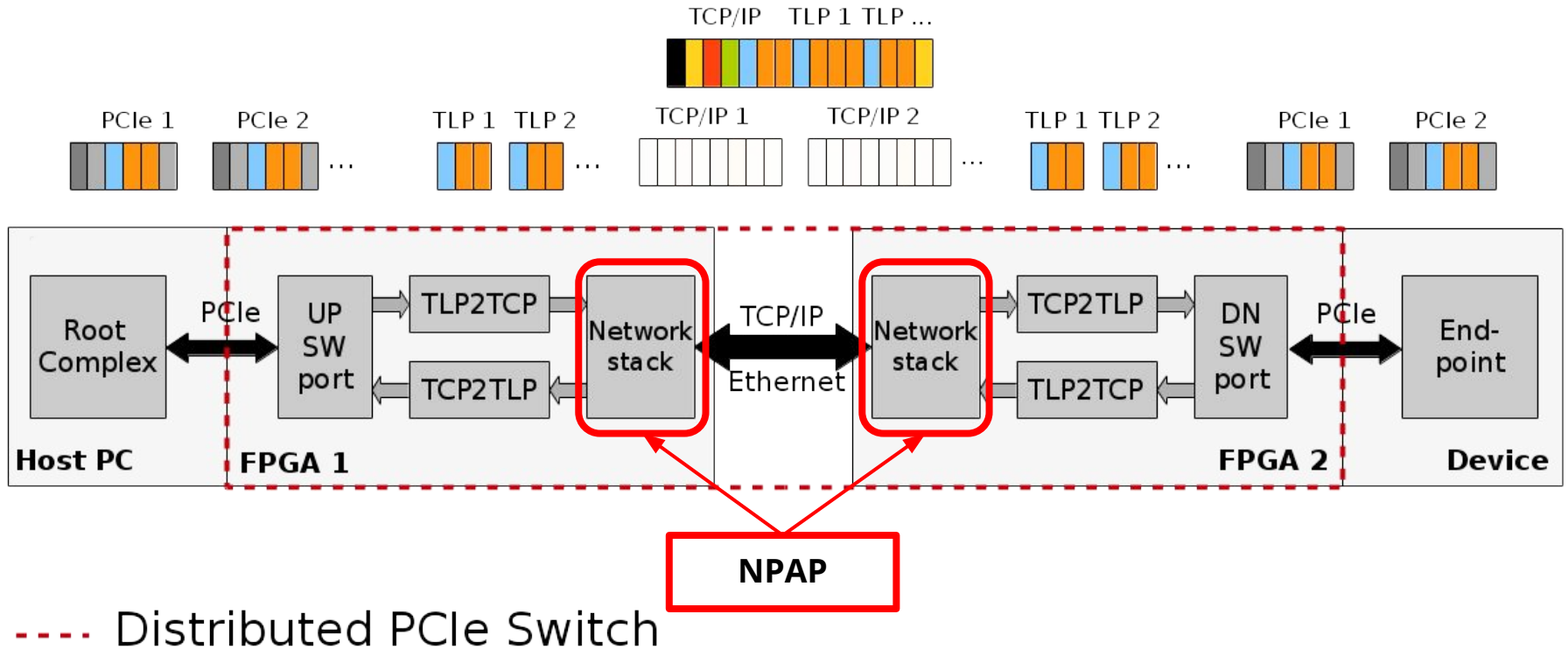
PCIe Layer



Network Layer

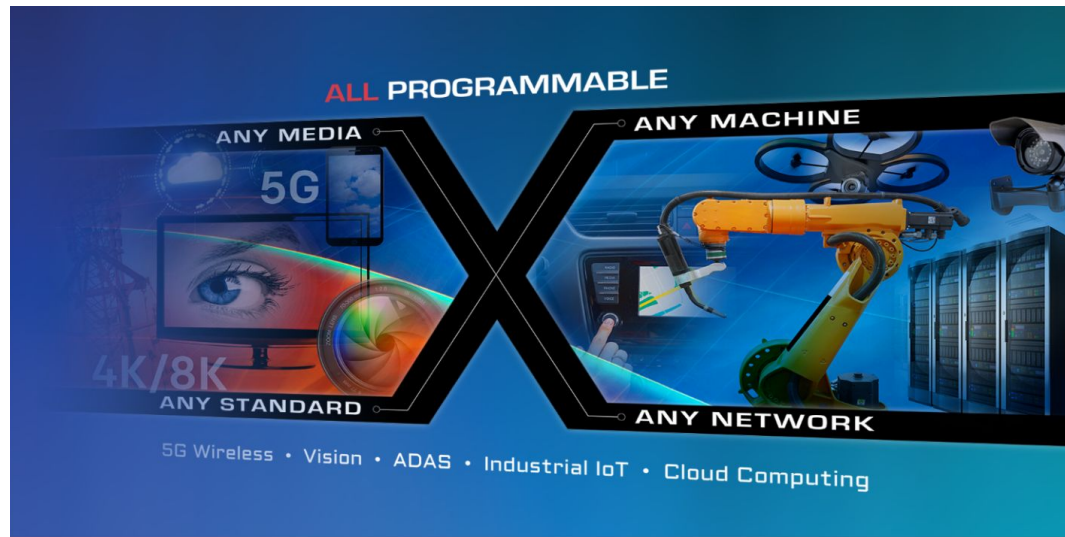


Concept of PCIe-over-TCP (2)



Key-Value-Store Hardware-Acceleration

- Joint work of MLE and Xilinx Research Ireland
- Presented at SNIA SDC 2016 and SNIA SDC 2017
- Single-Chip Solution



Heterogeneous Multi-Processing for SW-Defined Multi-Tiered Storage Architectures

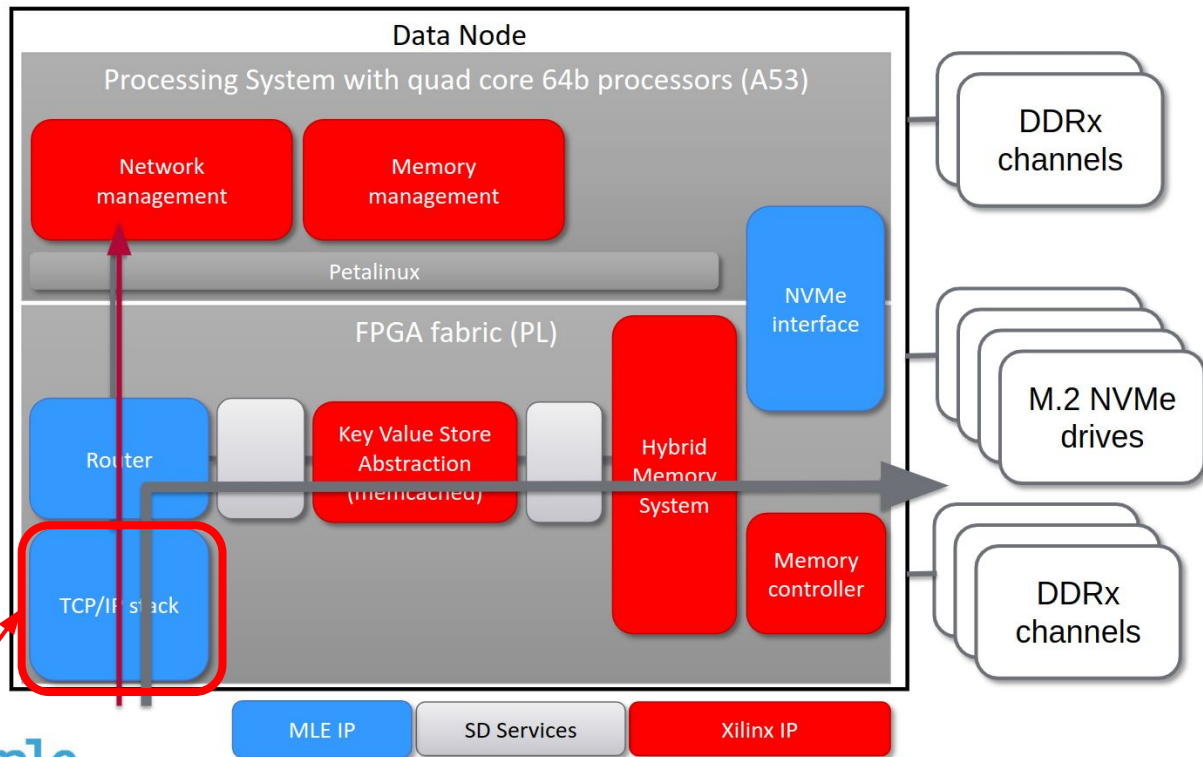
Endric Schubert (MLE)
Ulrich Langenbach (MLE)
Michaela Blott (Xilinx Research)

SDC, 2017

Key-Value-Store Hardware-Acceleration

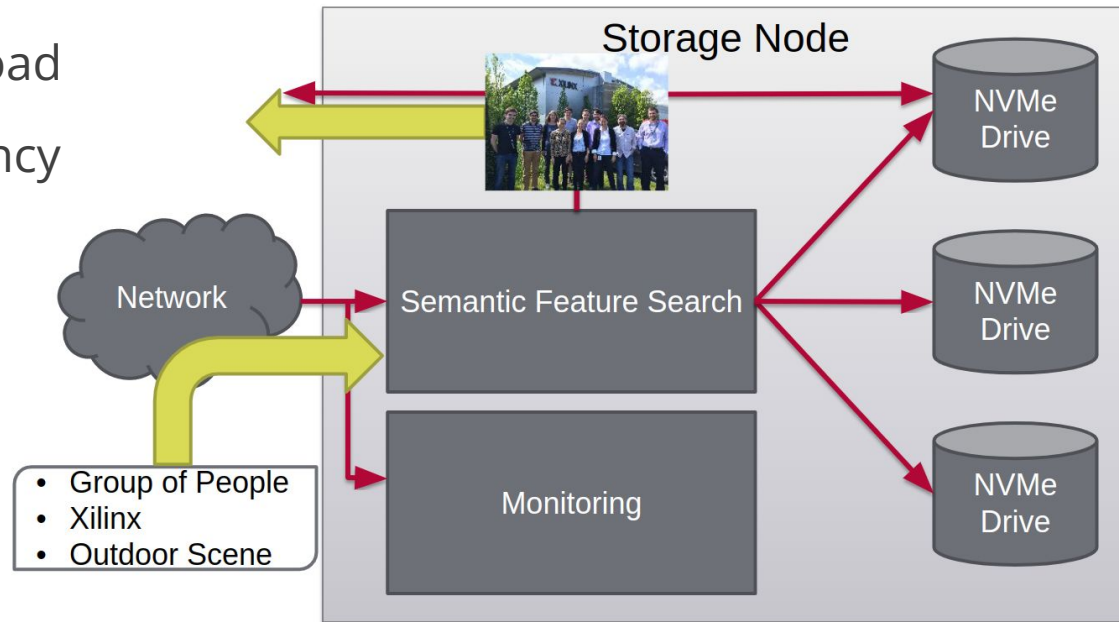
- Ethernet Attached KVS Storage Node
- Fully Pipelined
- Local NVMe memory
- Local DRAM memory
- Multi-hierarchy storage
 - Small and Fast
 - Large and "Slow"

NPAP

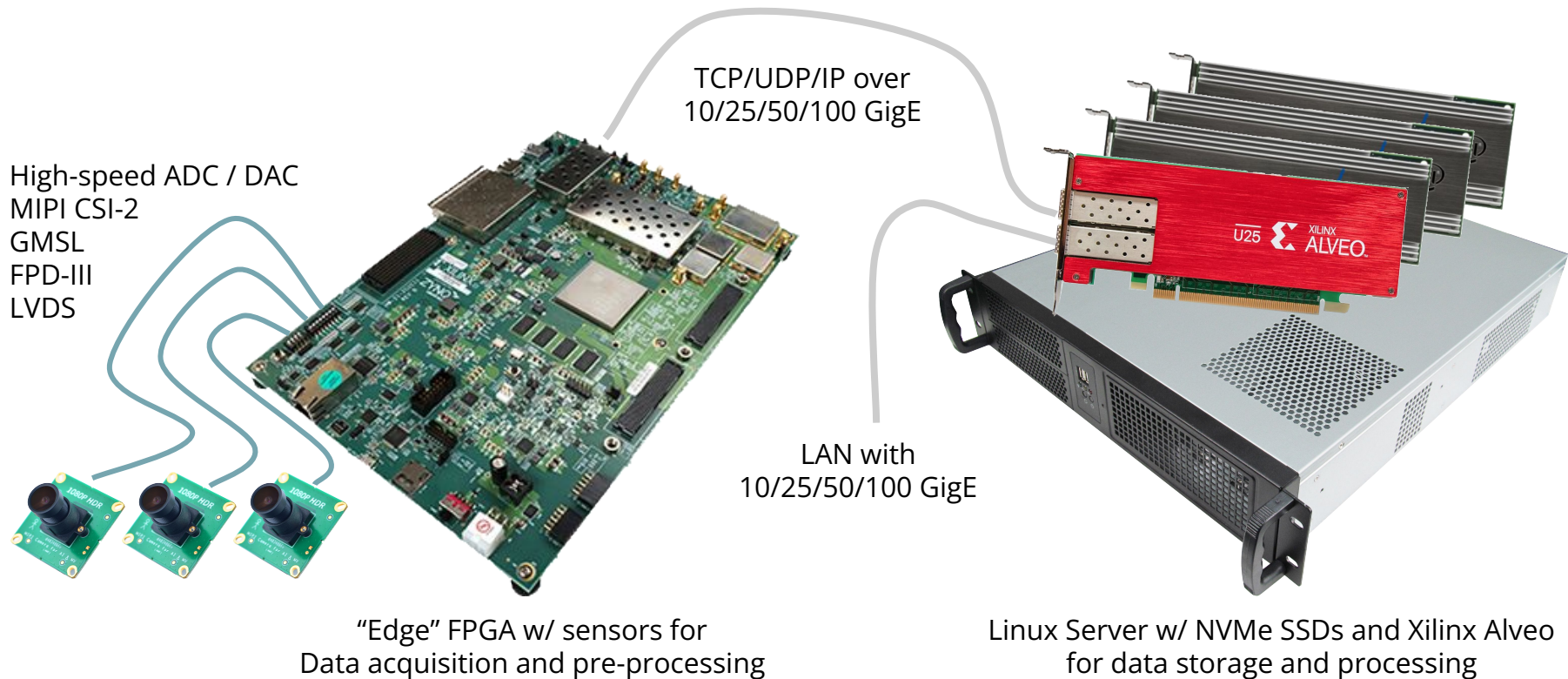


Key-Value-Store Hardware-Acceleration

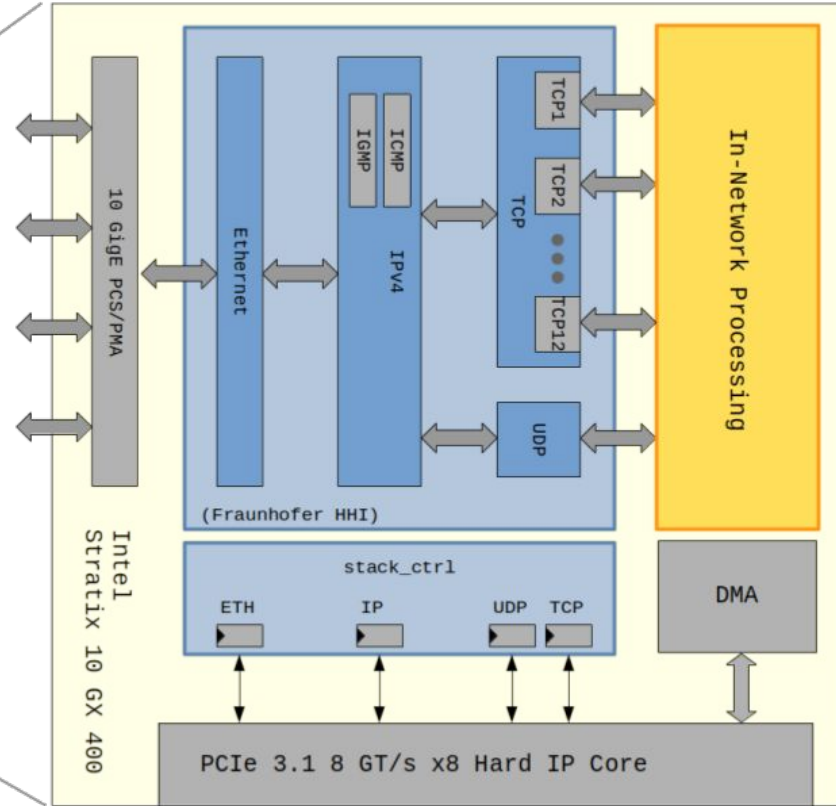
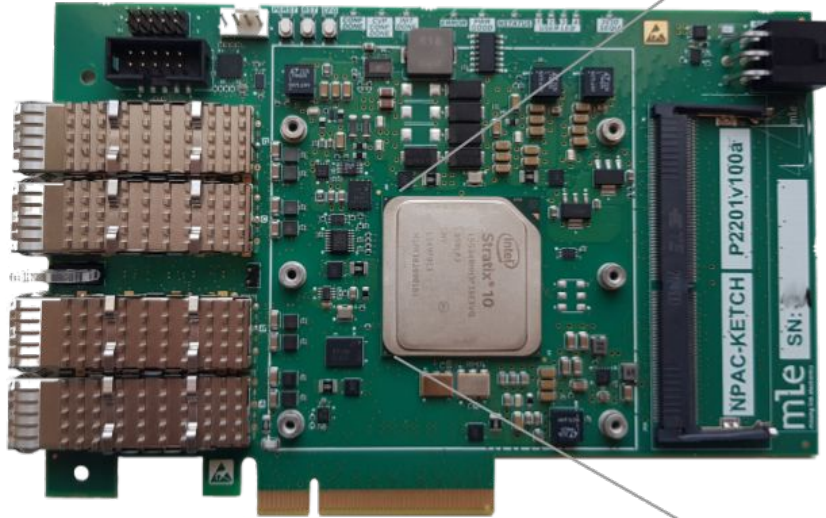
- Data-Flow Processing enables real on-the-fly meta-data extraction
- No additional server load
- Provides very low latency



FPGA-Based "Edge" - Server Connectivity



NPAC - Network Protocol Accelerator Card



NPAC - Features

First MLE NPAC PCIe card, namely NPAC-KETCH, will be available soon:

- Targeted to Intel Stratix 10 GX 400
- Netperf and TCP-/UDP-Loopback example instances
- 4x SFP+ for 4x 10 GigE via Twinax or Fibre
- Supports Quartus design flow with High-Level Synthesis design option
- Runs on MLE NPAC-40G Cost-Optimized SmartNIC

Other device vendors on the Roadmap: Microchip, Xilinx

Contact Information

Email contact: sales-web@missinglinkelectronics.com

Missing Link Electronics, Inc.
+1 (408) 475-1490
2880 Zanker Road, Suite 203
San Jose, CA 95134
United States

Missing Link Electronics GmbH
+49 (731) 141149-0
Industriestraße 10
89231 Neu-Ulm
Germany