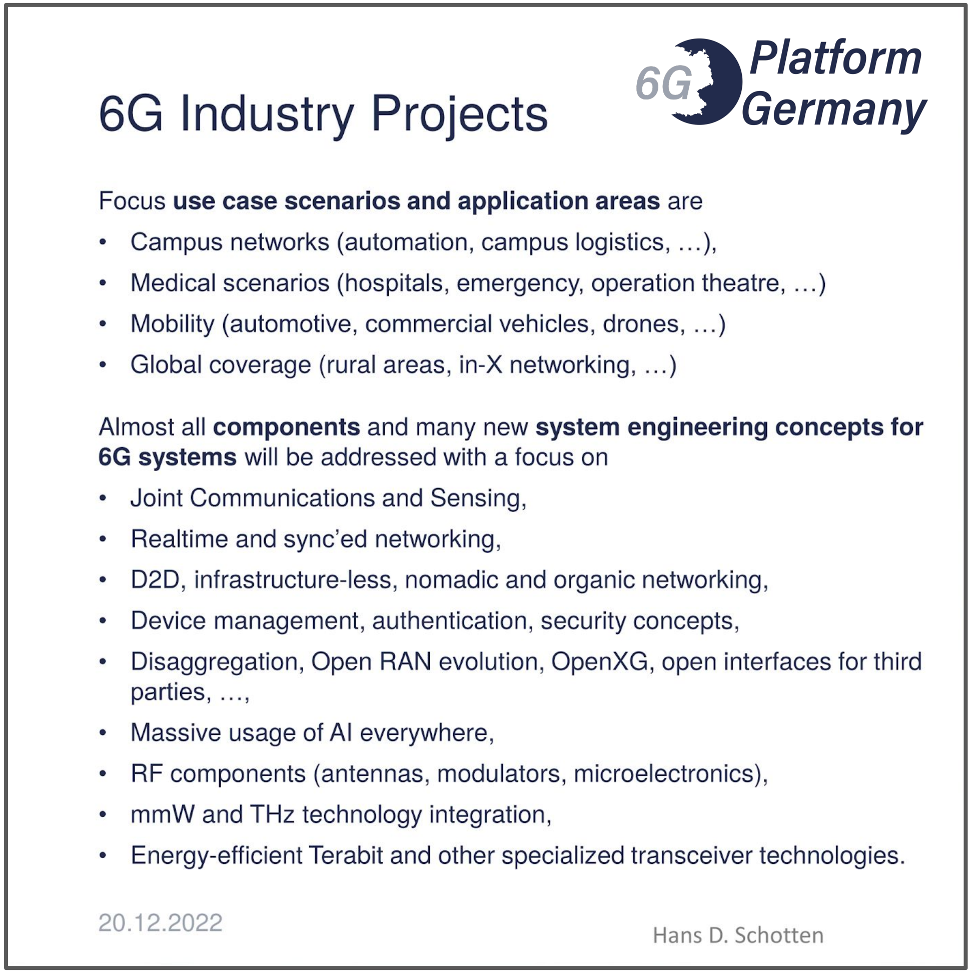
embeddedworld2023
Exhibition&Conference
... it´s a smarter world

Composable Edge Cloud Systems With NVMe over 5G URLLC

Frederik Pfautsch
Endric Schubert



6G Industry Projects

6G Platform Germany

Focus **use case scenarios and application areas** are

- Campus networks (automation, campus logistics, …),
- Medical scenarios (hospitals, emergency, operation theatre, …)
- Mobility (automotive, commercial vehicles, drones, …)
- Global coverage (rural areas, in-X networking, …)

Almost all **components** and many new **system engineering concepts for 6G systems** will be addressed with a focus on

- Joint Communications and Sensing,
- Realtime and sync'ed networking,
- D2D, infrastructure-less, nomadic and organic networking,
- Device management, authentication, security concepts,
- Disaggregation, Open RAN evolution, OpenXG, open interfaces for third parties, …,
- Massive usage of AI everywhere,
- RF components (antennas, modulators, microelectronics),
- mmW and THz technology integration,
- Energy-efficient Terabit and other specialized transceiver technologies.

20.12.2022 Hans D. Schotten

Background

Frederik Pfautsch

frederik.pfautsch@missinglinkelectronics.com

- MSc Computer Engineering
  @TU Berlin

- Master’s thesis in cooperation with
  Fraunhofer HHI, MLE and Uni Ulm

- “5G Berlin” campus network (Release 15)

- MLE
  R&D Lead Engineer for 5G/6G Radio
  Sidelink Comm & Precision Time Synch

Endric Schubert

endric.schubert@missinglinkelectronics.com

- Dipl.-Ing. ET Univ. Karlsruhe
- PhD CS Univ. Tuebingen
- Honorary Professor Univ. Ulm
- Ambassador Startup Sued

- Background in Semiconductors, EDA,
  Domain Specific Architectures w/ FPGA

- 60+ technical publications
- 20+ patents

MLE - Experts for Domain-Specific Compute Architectures

Our Mission:

- Deliver HW and SW for **High-Performance (Embedded) Compute** Systems & Solutions
- Offering pre-validated subsystems with FPGA IP blocks and open-source software
- Support customer projects with deep expertise and hands-on design services

Head-quartered in Silicon Valley with Design Offices in Germany

- Founded 2010, employee owned
- 18+ Certified FPGA Designers
- Customers include technology leaders, US and European government agencies, Fortune 500 companies
- Partners to:

XILINX Premier Partner

PCI SIG

intel FPGA Design Solutions Network Gold

Fraunhofer Heinrich-Hertz-Institut

Fraunhofer IPMS

ulm university universität uulm

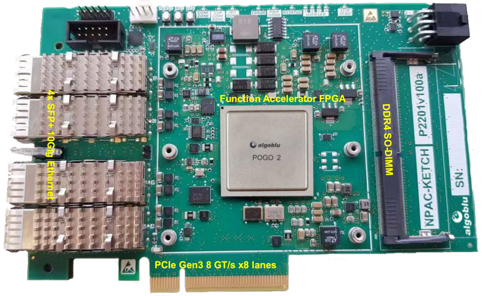
THU Technische Hochschule Ulm

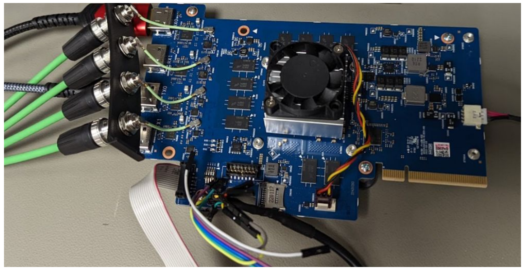
FPGA MICROCHIP DESIGN PARTNER

aws

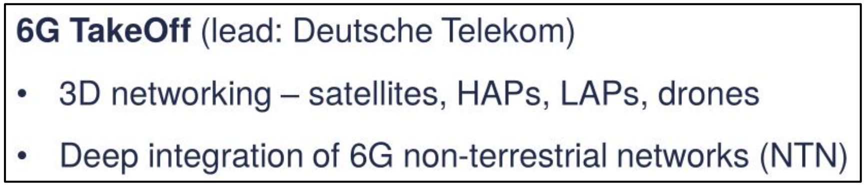
OpenPOWER

ELEMASTER



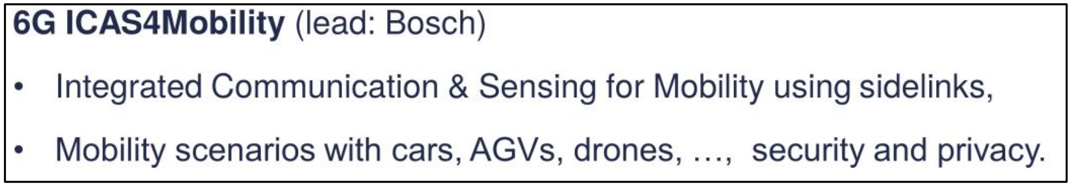


Why?



6G TakeOff (lead: Deutsche Telekom)

- 3D networking – satellites, HAPs, LAPs, drones
- Deep integration of 6G non-terrestrial networks (NTN)



6G ICAS4Mobility (lead: Bosch)

- Integrated Communication & Sensing for Mobility using sidelinks,
- Mobility scenarios with cars, AGVs, drones, …, security and privacy.

6G Platform Germany

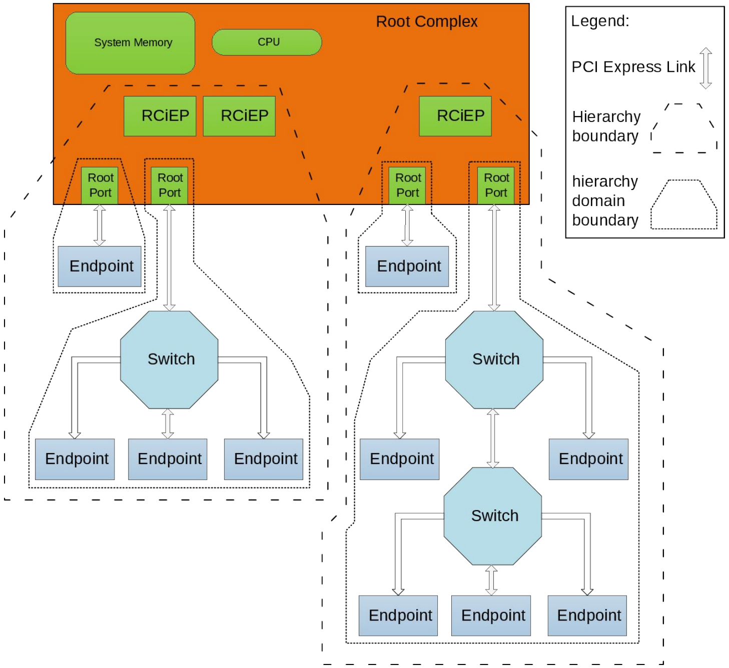
⇒ 5G/6G edge cloud devices with limited storage capacity

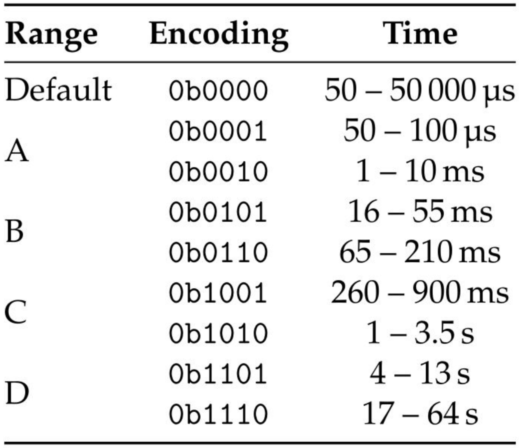
- NVMe over 5G? It should be possible, 5G has the low latency and bandwidth promises!

- Side effect: Measure capabilities of 5G Release 15 thoroughly

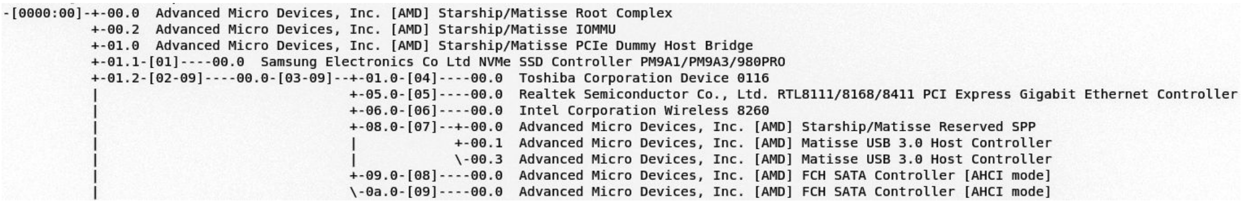
Interlude – PCIe

- De-facto standard for general purpose peripheral connectivity within x86 PCs and servers
- Easy extension of CotS-computers with almost any type of peripheral
- Every new PCIe gen approx. doubles the available bandwidth
  - PCIe Gen 4: 31.5 GByte/s (x16)
  - PCIe Gen 5: 63.015 GByte/s (x16)
- Packet-based, layered protocol (TLPs)





Range	Encoding	Time
Default	0b0000	50 – 50 000 µs
A	0b0001	50 – 100 µs
	0b0010	1 – 10 ms
B	0b0101	16 – 55 ms
	0b0110	65 – 210 ms
C	0b1001	260 – 900 ms
	0b1010	1 – 3.5 s
D	0b1101	4 – 13 s
	0b1110	17 – 64 s



```
-[0000:00]-+-00.0  Advanced Micro Devices, Inc. [AMD] Starship/Matisse Root Complex
           +-00.2  Advanced Micro Devices, Inc. [AMD] Starship/Matisse IOMMU
           +-01.0  Advanced Micro Devices, Inc. [AMD] Starship/Matisse PCIe Dummy Host Bridge
           +-01.1-[01]----00.0  Samsung Electronics Co Ltd NVMe SSD Controller PM9A1/PM9A3/980PRO
           +-01.2-[02-09]----00.0-[03-09]--+-01.0-[04]----00.0  Toshiba Corporation Device 0116
           |                               +-05.0-[05]----00.0  Realtek Semiconductor Co., Ltd. RTL8111/8168/8411 PCI Express Gigabit Ethernet Controller
           |                               +-06.0-[06]----00.0  Intel Corporation Wireless 8260
           |                               +-08.0-[07]--+-00.0  Advanced Micro Devices, Inc. [AMD] Starship/Matisse Reserved SPP
           |                               |            +-00.1  Advanced Micro Devices, Inc. [AMD] Matisse USB 3.0 Host Controller
           |                               |            \-00.3  Advanced Micro Devices, Inc. [AMD] Matisse USB 3.0 Host Controller
           |                               +-09.0-[08]----00.0  Advanced Micro Devices, Inc. [AMD] FCH SATA Controller [AHCI mode]
           |                               \-0a.0-[09]----00.0  Advanced Micro Devices, Inc. [AMD] FCH SATA Controller [AHCI mode]
```

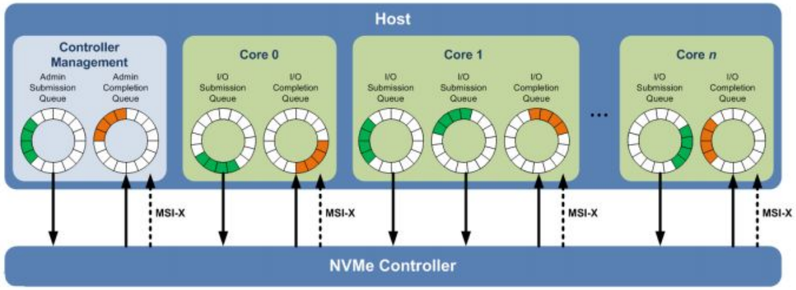
Interlude – NVMe

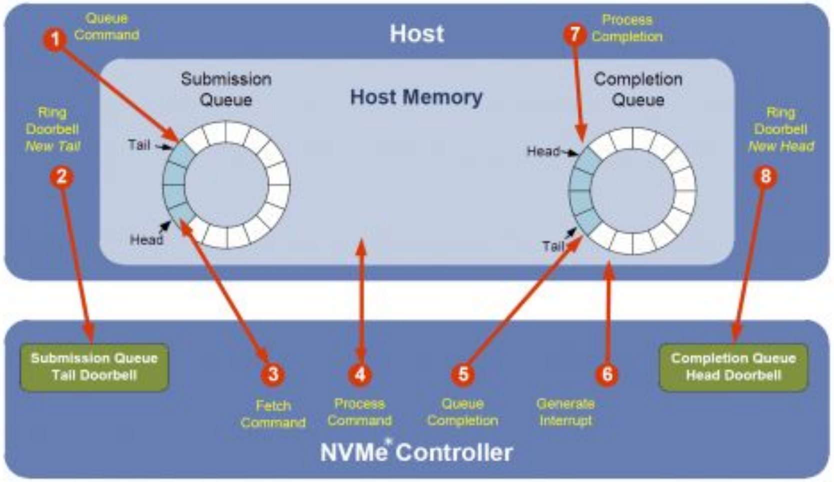
NVMe is an example of a modern, fast, PCIe based communication protocol.

- Avoid software reads to device registers
- Hardware device implementation can issue multiple reads in parallel, masking the round trip time
- Also software can only transfer 64 bits per access

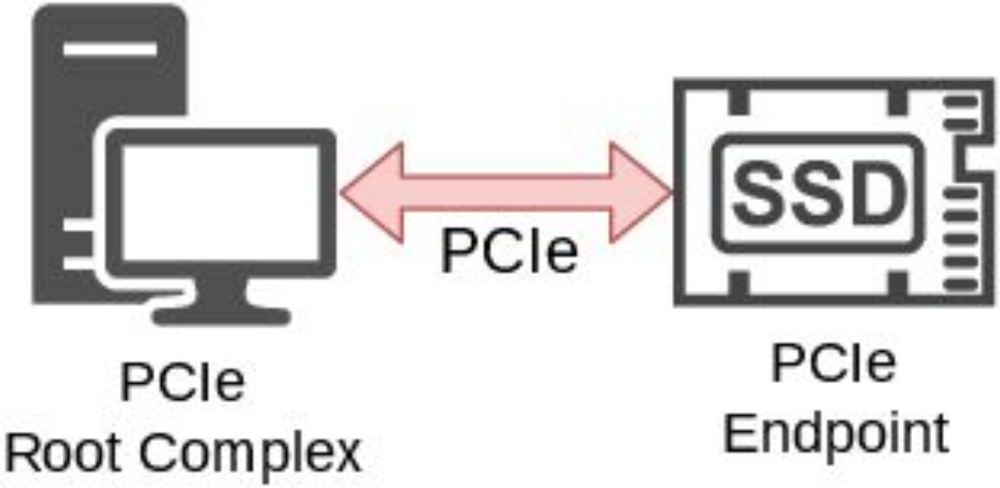
- Pipeline processing for example by allowing for lazy pointer updates of queues

- Scale with the number of CPU cores by having independent queues/ringbuffers and MSI-X interrupts

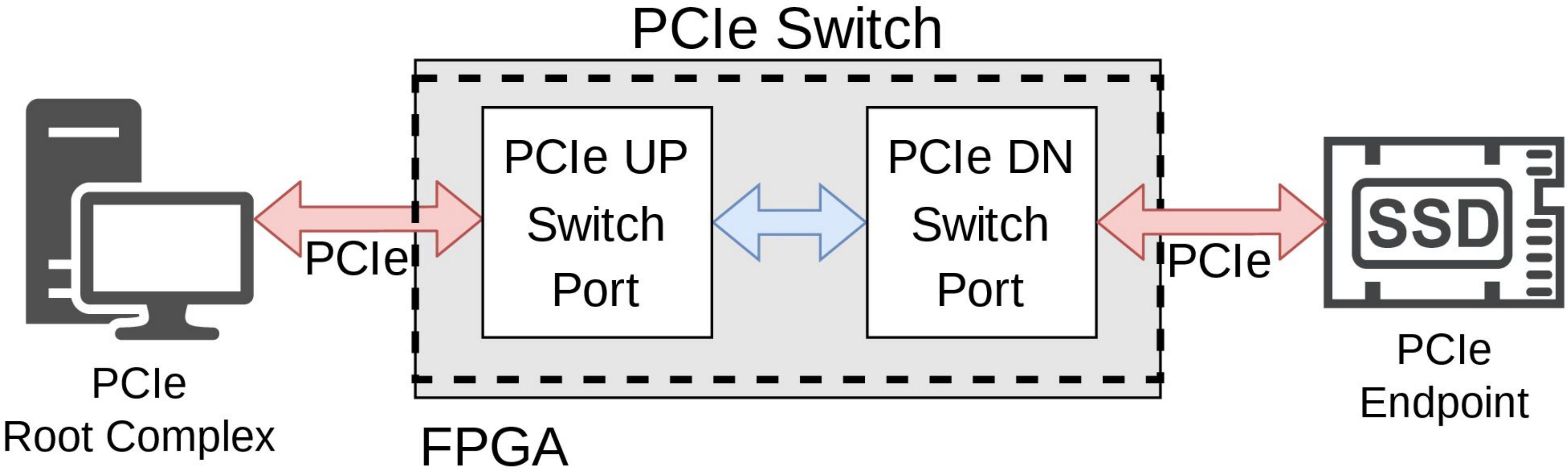




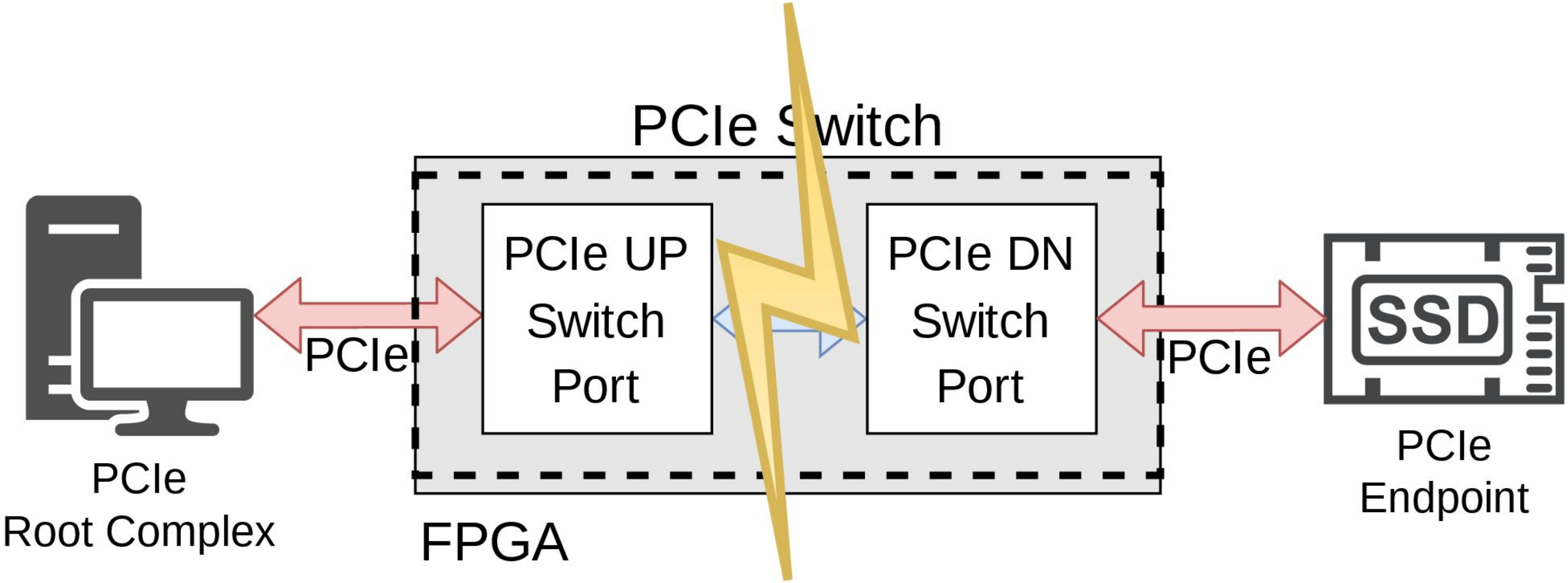
Why?



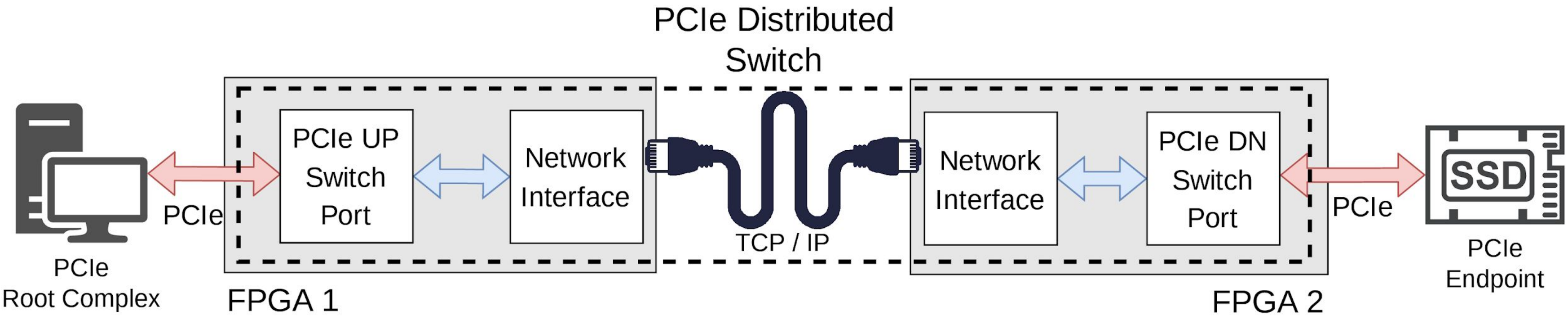
Why?


PCIe Switch
PCIe
Root Complex
PCIe
PCIe UP
Switch
Port
PCIe DN
Switch
Port
PCIe
SSD
PCIe
Endpoint
FPGA

Why?

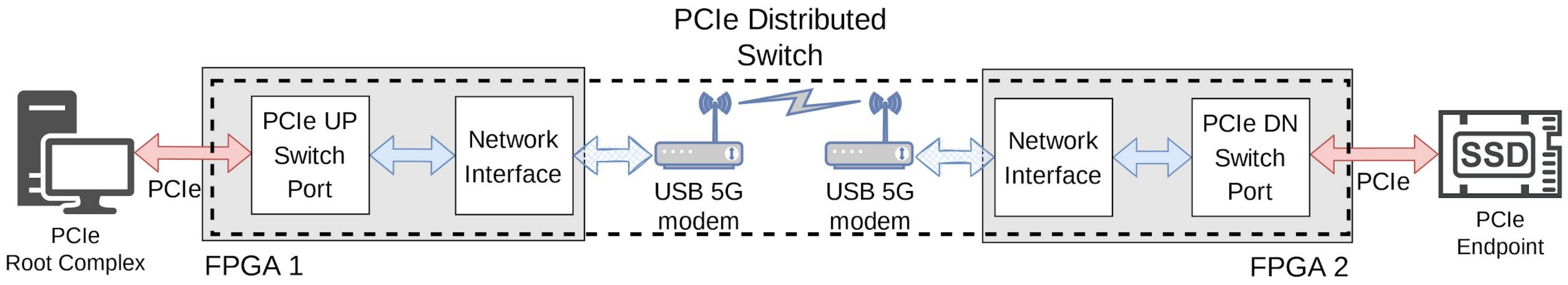

PCIe Switch
PCIe UP Switch Port
PCIe DN Switch Port
PCIe
PCIe
SSD
PCIe Root Complex
FPGA
PCIe Endpoint

Why?



See: Schubert, Braun and Langenbach: “PCI Express over IP - Accelerated” Embedded World Conference, 2016

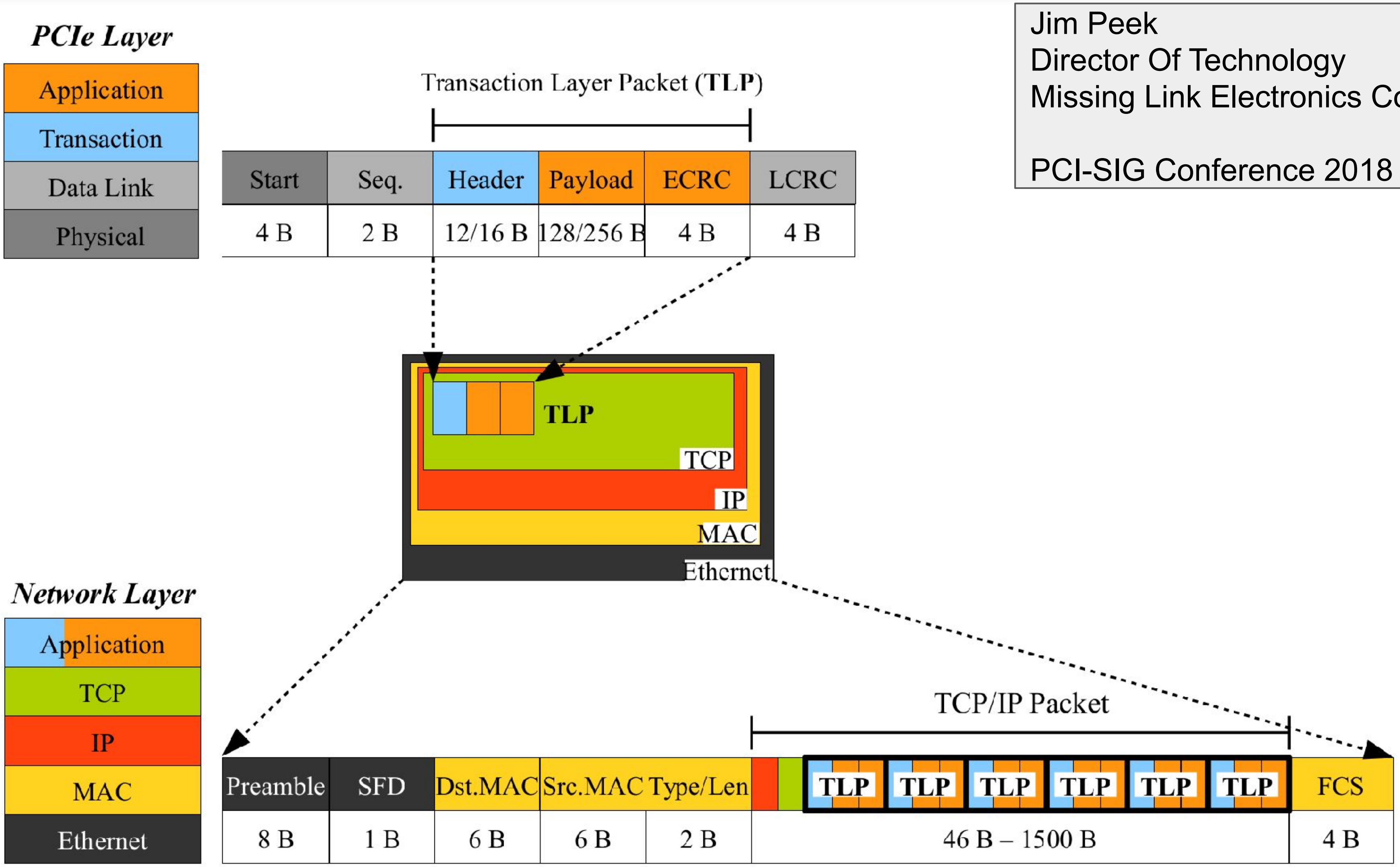
Why?


PCIe Distributed
Switch
PCIe
Root Complex
PCIe
PCIe UP
Switch
Port
Network
Interface
FPGA 1
USB 5G
modem
USB 5G
modem
Network
Interface
PCIe DN
Switch
Port
FPGA 2
PCIe
SSD
PCIe
Endpoint

PCIe over TCP/IP Tunneling

Jim Peek
Director Of Technology
Missing Link Electronics Corp

PCI-SIG Conference 2018



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

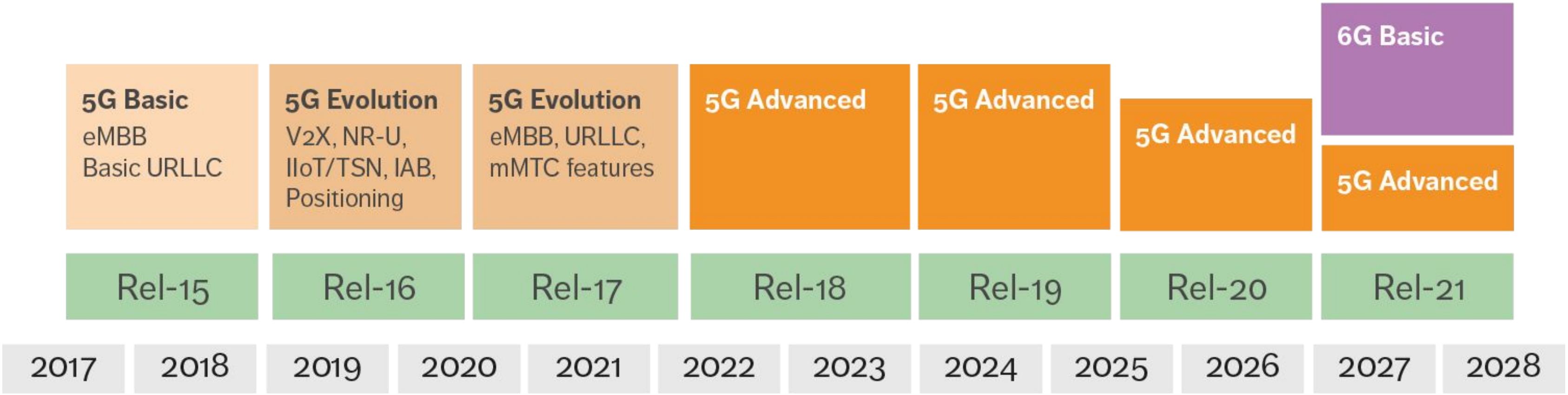
TCP/IP Packet

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

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**

5G



5G Basic
eMBB
Basic URLLC

5G Evolution
V2X, NR-U,
IIoT/TSN, IAB,
Positioning

5G Evolution
eMBB, URLLC,
mMTC features

5G Advanced

5G Advanced

5G Advanced

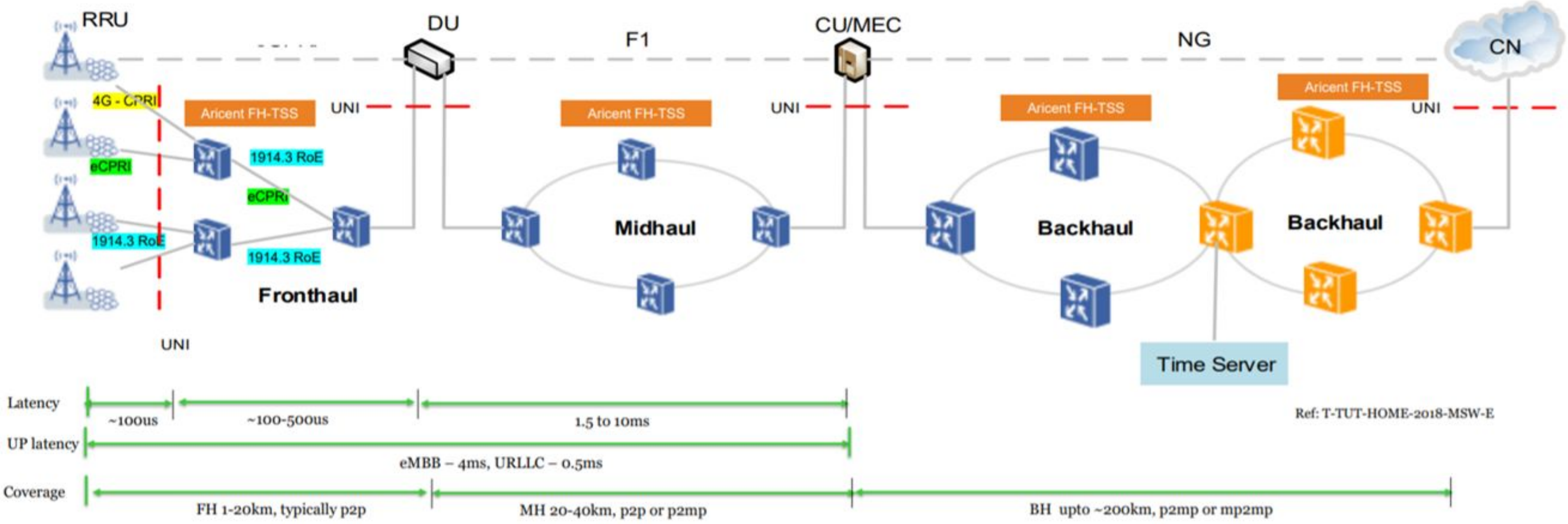
6G Basic

5G Advanced

Rel-15 Rel-16 Rel-17 Rel-18 Rel-19 Rel-20 Rel-21

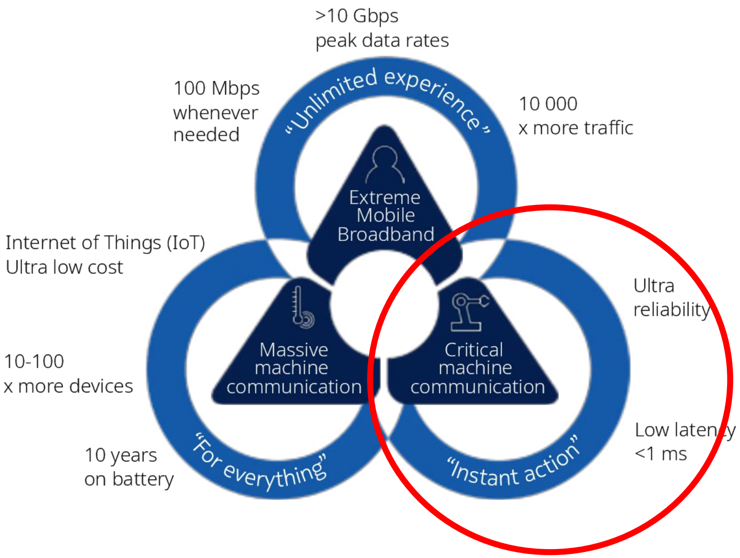
2017 2018 2019 2020 2021 2022 2023 2024 2025 2026 2027 2028

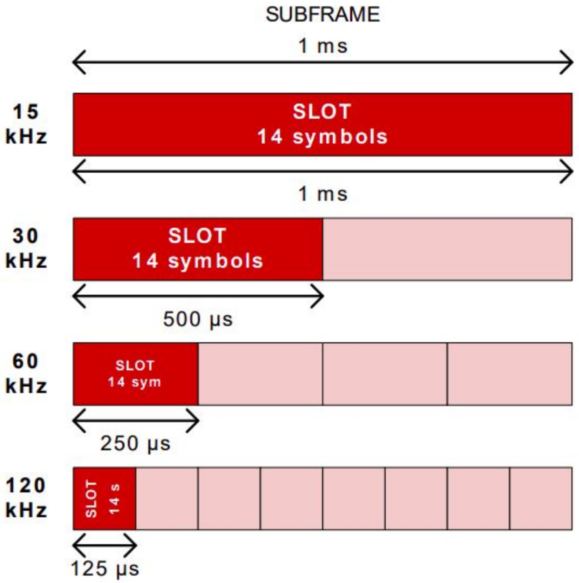
5G

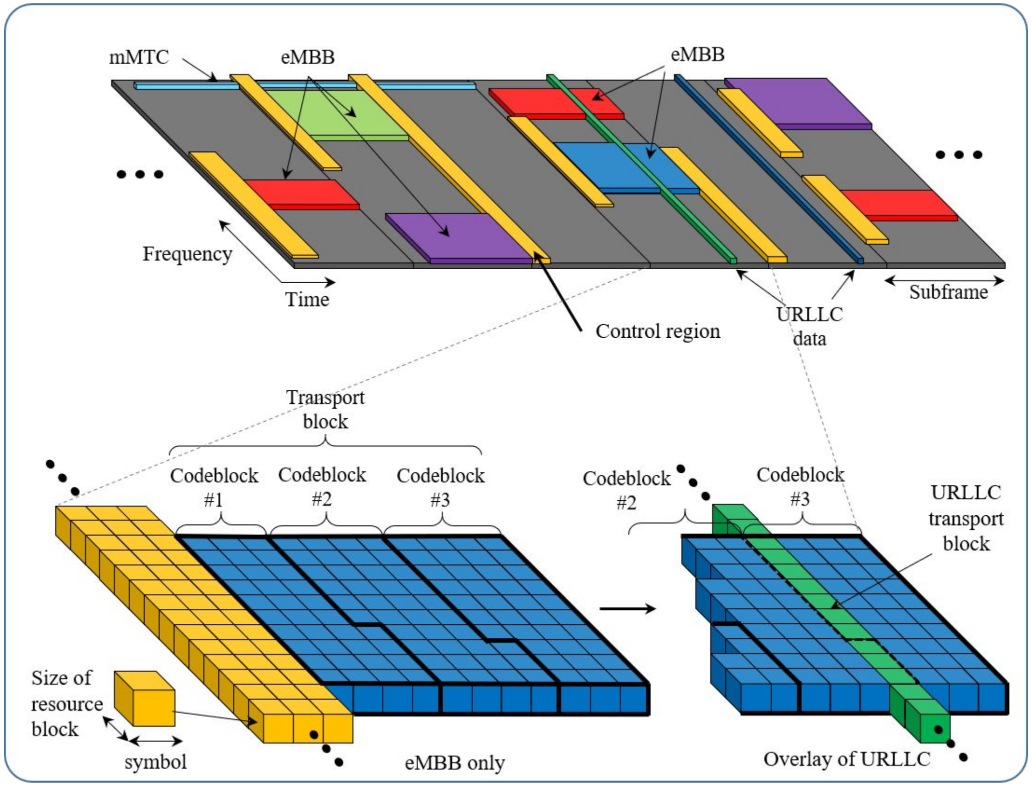


- ✓ Fronthaul – Network between RRU/RU (Remote Unit) and DU (Distributed Unit) – can be CPRI or eCPRI or IEEE 1914.3
- ✓ Midhaul – Network between DU and CU (Centralized Unit) – “F interface”
- ✓ Backhaul – Network between CU and 5G NGC (and EPC)

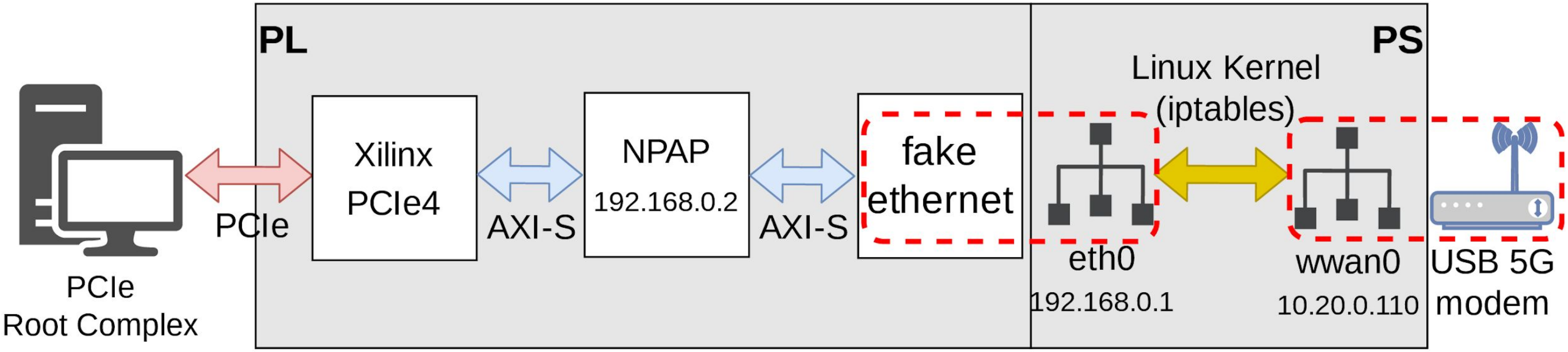
5G – URLLC

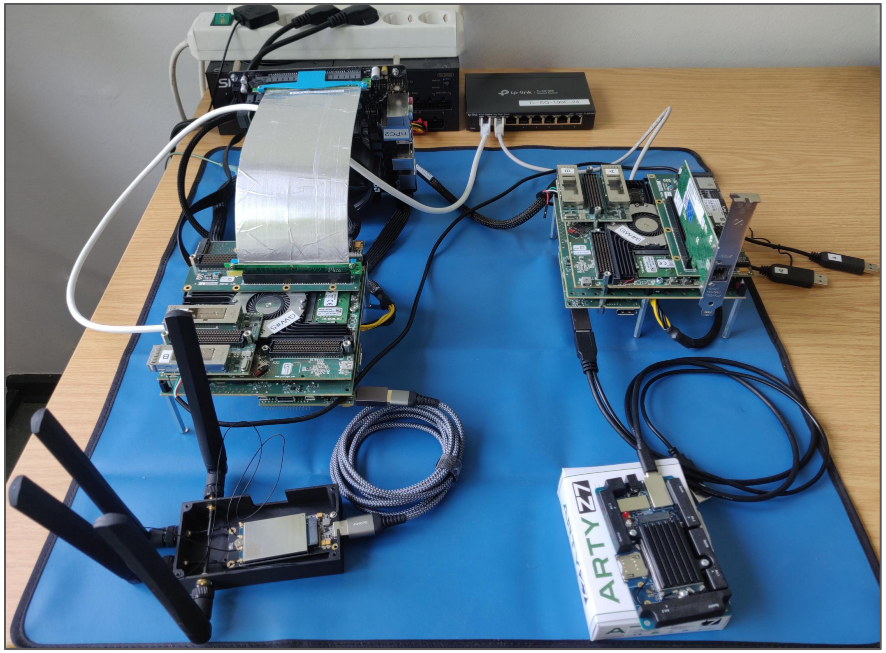

>10 Gbps
peak data rates
100 Mbps
whenever
needed
10 000
x more traffic
"Unlimited experience"
Extreme
Mobile
Broadband
Internet of Things (IoT)
Ultra low cost
Ultra
reliability
10-100
x more devices
Massive
machine
communication
Critical
machine
communication
Low latency
<1 ms
10 years
on battery
"For everything"
"Instant action"

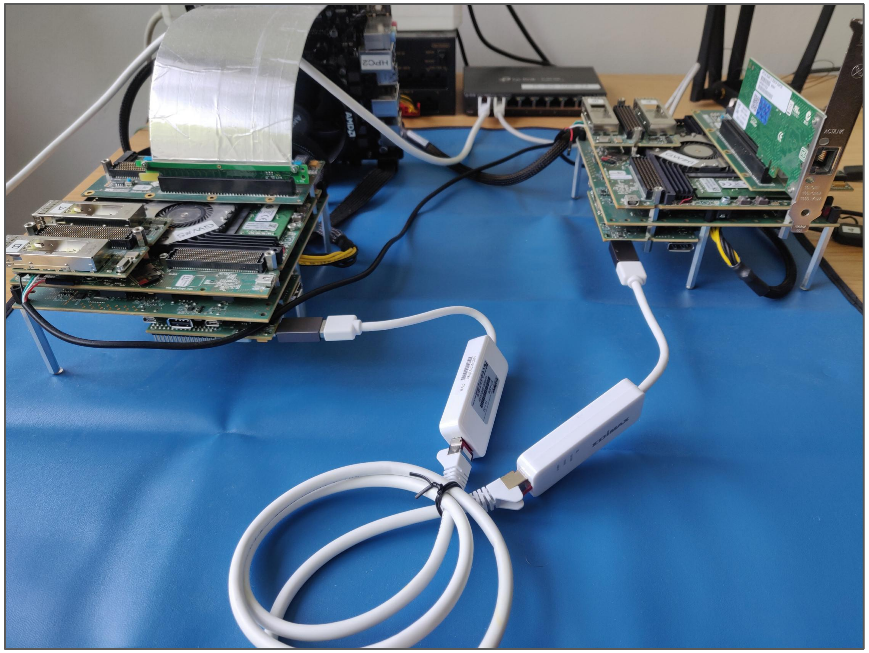

SUBFRAME
1 ms
15
kHz
SLOT
14 symbols
1 ms
30
kHz
SLOT
14 symbols
500 µs
60
kHz
SLOT
14 sym
250 µs
120
kHz
SLOT
14 s
125 µs


mMTC
eMBB
eMBB
Frequency
Time
Control region
URLLC
data
Subframe
Transport
block
Codeblock
#1
Codeblock
#2
Codeblock
#3
Codeblock
#2
Codeblock
#3
URLLC
transport
block
Size of
resource
block
symbol
eMBB only
Overlay of URLLC

Setup

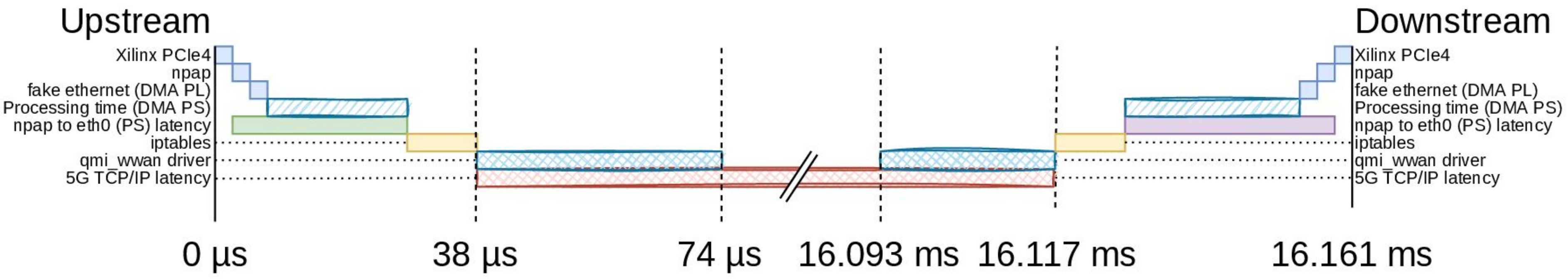

PL
PS
PCIe
Root Complex
PCIe
Xilinx
PCIe4
AXI-S
NPAP
192.168.0.2
AXI-S
fake
ethernet
Linux Kernel
(iptables)
eth0
192.168.0.1
wwan0
10.20.0.110
USB 5G
modem





Latency Chain

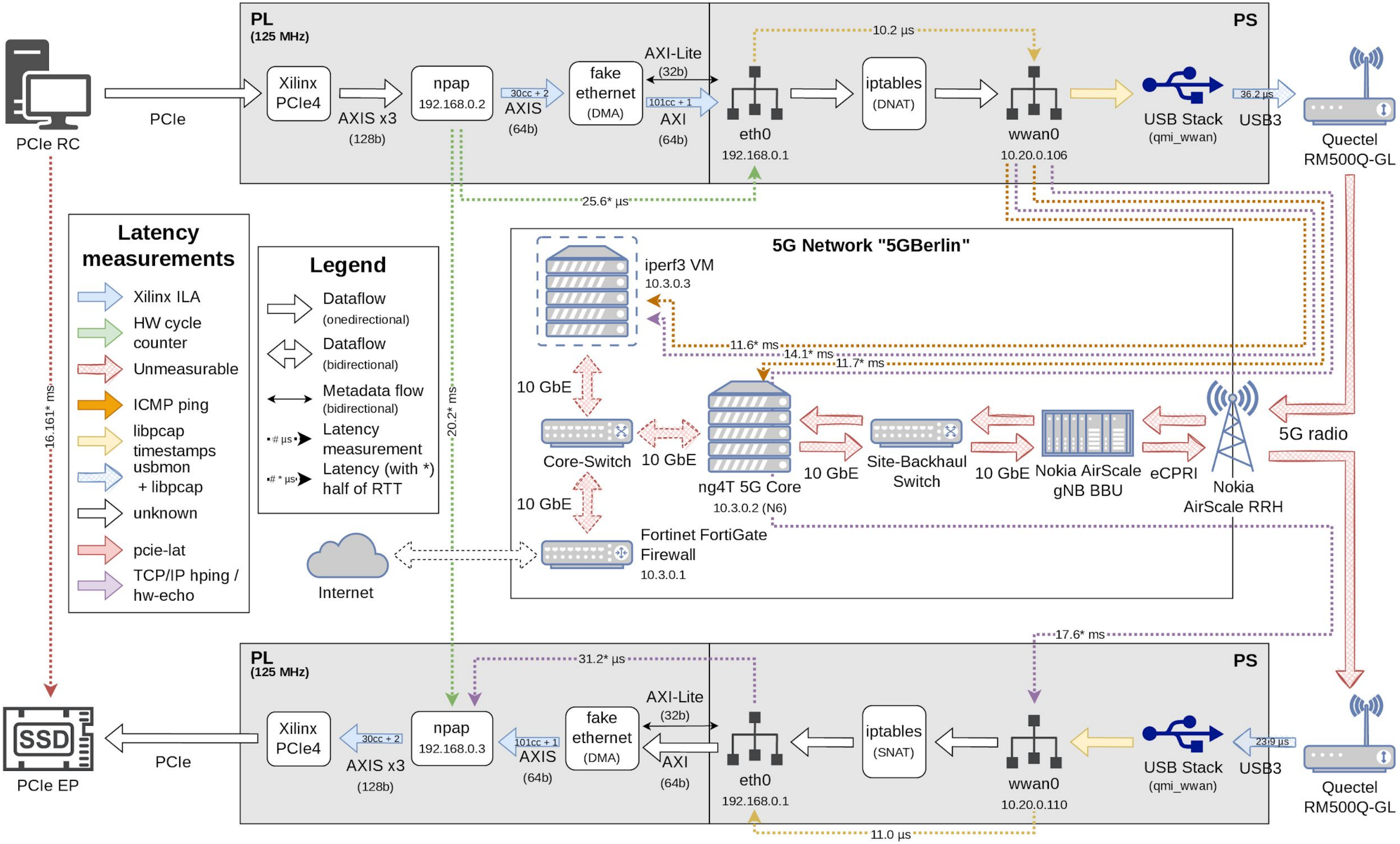
ø 16.2 ms !



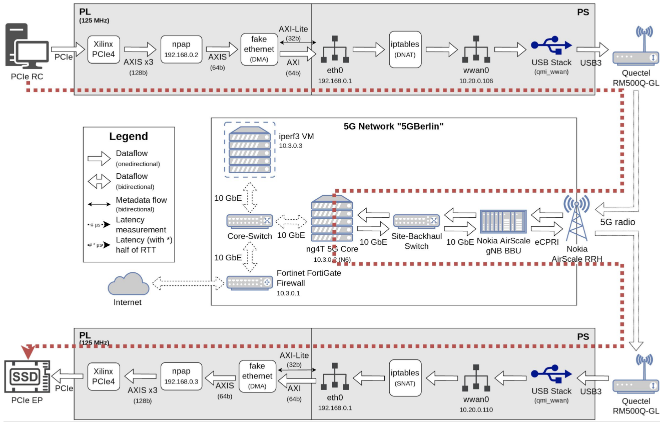
PCIe over 5G is Feasible!

Default PCIe Completion Timeout: 50 µs to 50 ms

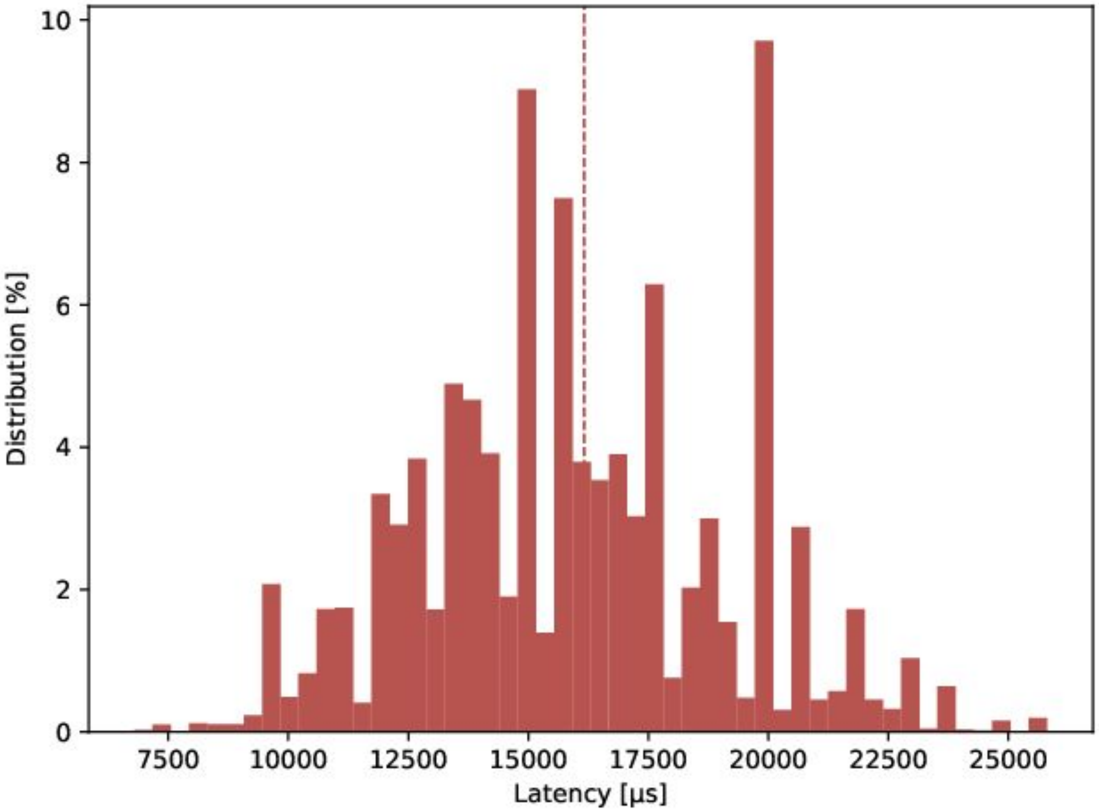
Latency Map


PL
(125 MHz)
PS
PCIe RC
PCIe
Xilinx PCIe4
AXIS x3
(128b)
npap
192.168.0.2
30cc + 2
AXIS
(64b)
fake ethernet
(DMA)
AXI-Lite
(32b)
101cc + 1
AXI
(64b)
eth0
192.168.0.1
10.2 µs
iptables
(DNAT)
wwan0
10.20.0.106
USB Stack
(qmi_wwan)
36.2 µs
USB3
Quectel RM500Q-GL
25.6* µs
16.161* ms
20.2* ms
Latency measurements
Xilinx ILA
HW cycle counter
Unmeasurable
ICMP ping
libpcap timestamps
usbmon + libpcap
unknown
pcie-lat
TCP/IP hping / hw-echo
Legend
Dataflow
(onedirectional)
Dataflow
(bidirectional)
Metadata flow
(bidirectional)
# µs
Latency measurement
# * µs
Latency (with *) half of RTT
5G Network "5GBerlin"
iperf3 VM
10.3.0.3
11.6* ms
14.1* ms
11.7* ms
10 GbE
Core-Switch
10 GbE
ng4T 5G Core
10.3.0.2 (N6)
10 GbE
Site-Backhaul Switch
10 GbE
Nokia AirScale gNB BBU
eCPRI
Nokia AirScale RRH
5G radio
10 GbE
Fortinet FortiGate Firewall
10.3.0.1
Internet
17.6* ms
PL
(125 MHz)
PS
31.2* µs
PCIe EP
PCIe
Xilinx PCIe4
30cc + 2
AXIS x3
(128b)
npap
192.168.0.3
101cc + 1
AXIS
(64b)
fake ethernet
(DMA)
AXI-Lite
(32b)
AXI
(64b)
eth0
192.168.0.1
iptables
(SNAT)
wwan0
10.20.0.110
USB Stack
(qmi_wwan)
23.9 µs
USB3
Quectel RM500Q-GL
11.0 µs

Latency – PCIe



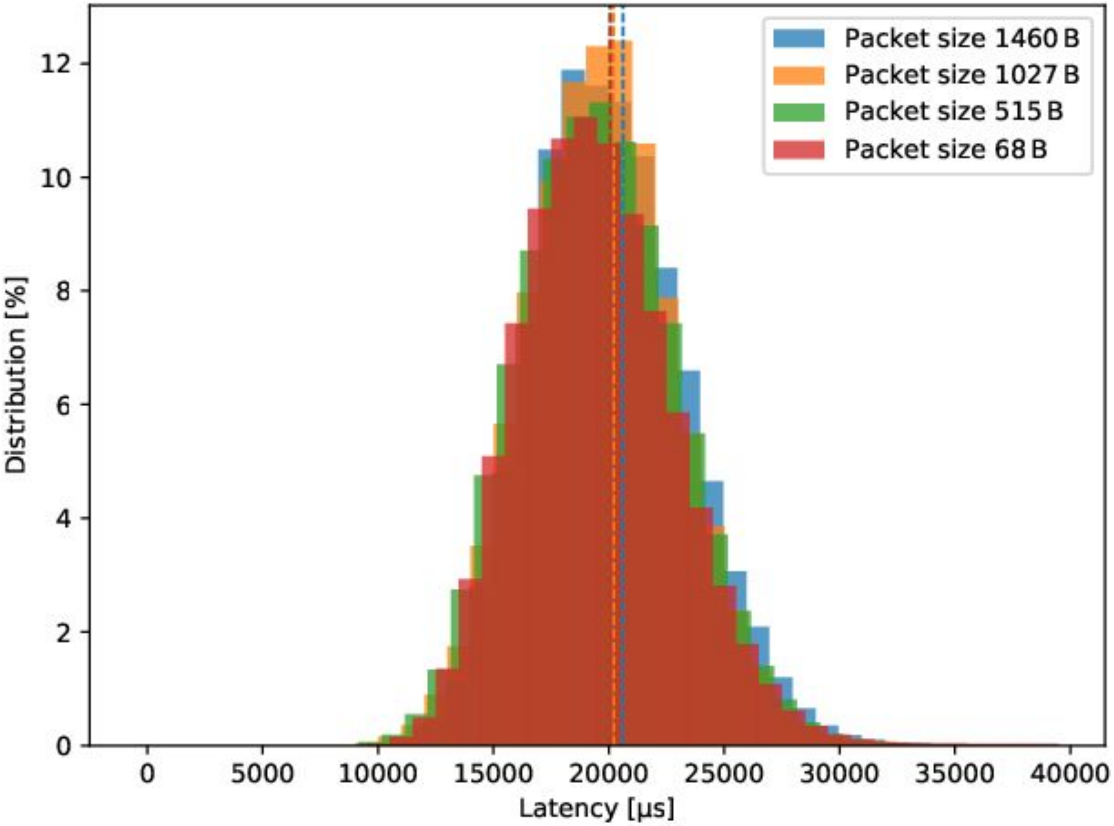
ø 16.2 ms

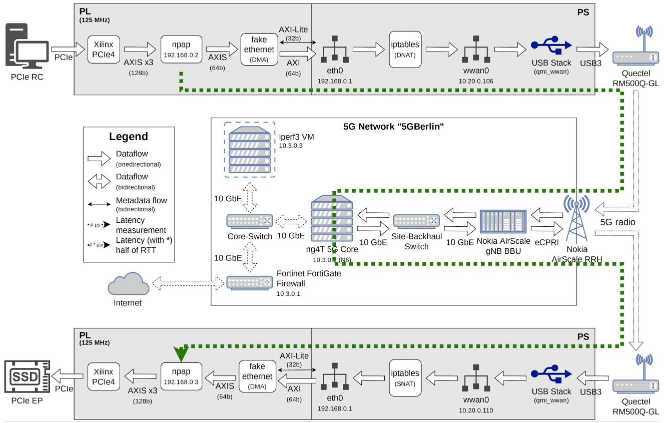


- “PCIelat”
  - Kernel module
  - Ruby script

- Default PCIe Completion Timeout: **50 µs to 50 ms**

Latency – PL2PL

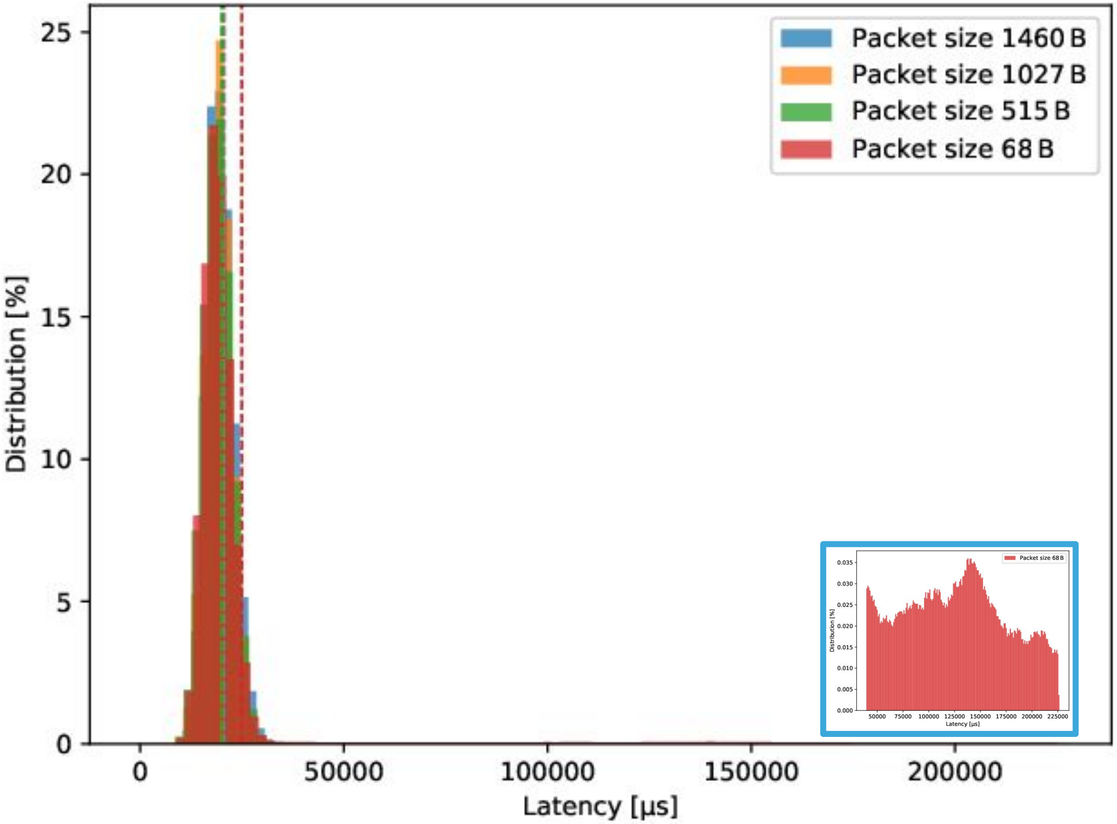


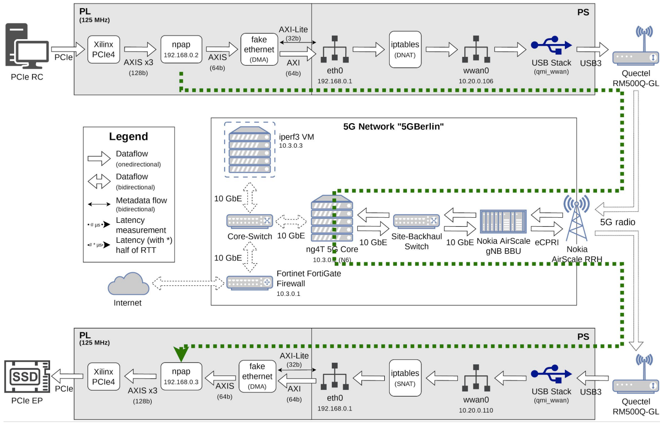


ø 20.2 ms

- VHDL counter
- Replaces PCIe traffic
- Customizable packet size

Latency – PL2PL

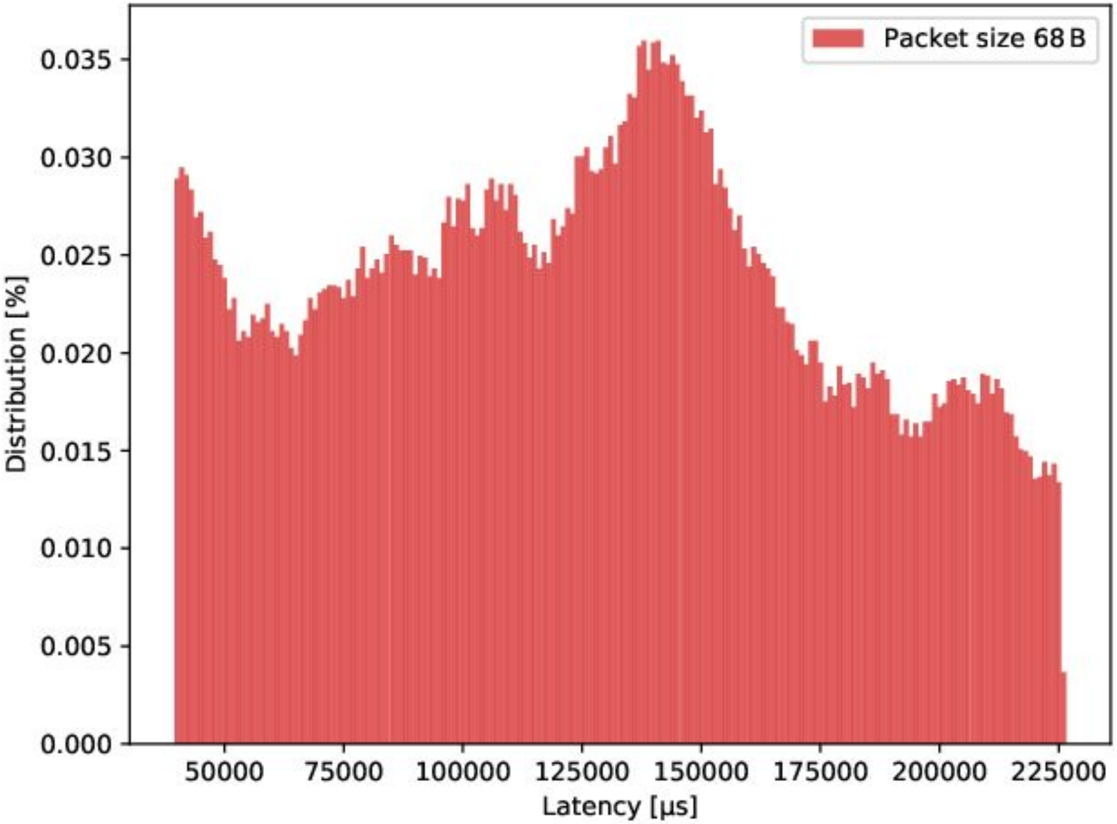


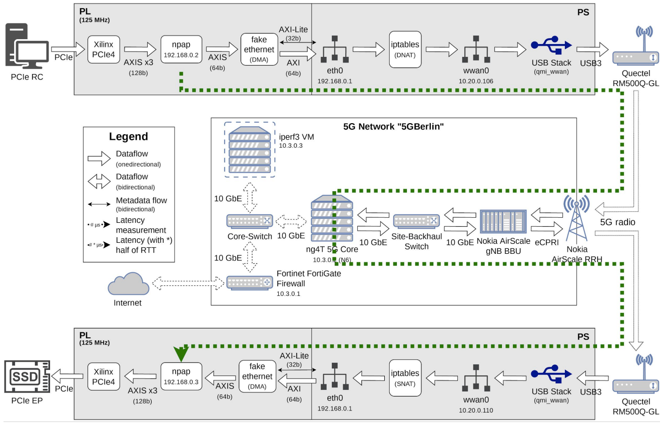


ø 20.2 ms

- VHDL counter
- Replaces PCIe traffic
- Customizable packet size

🔍 Latency – PL2PL





ø 20.2 ms

- VHDL counter
- Replaces PCIe traffic
- Customizable packet size

Summary

- It works :)

- High latency, high variance (tail latency)
  ⇒ Attached PC does not boot, re-enumeration necessary

- Latency measured by reference setup is comparable to other published setups

- 5G Release 15 introduces the majority (>99%) of latency in our setup!

- Latency is mostly independent of packet size (difference vanishes due to the high latency in general)

Outlook URLLC

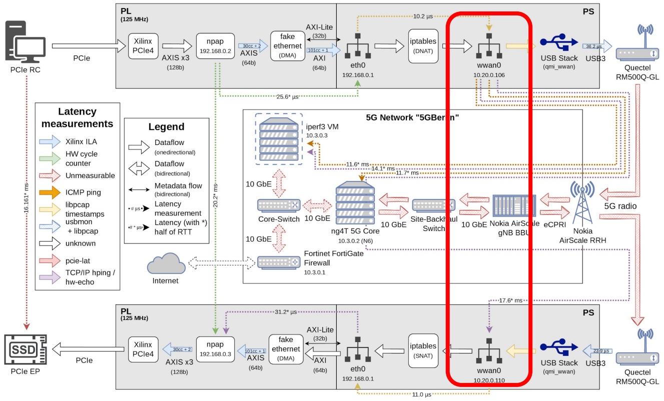
- 5G Release 15 only implements the basic requirements for URLLC, such as “micro slots”

- 5G Release 16 and 17 will begin to support URLLC
  - Does URLLC offer enough bandwidth?

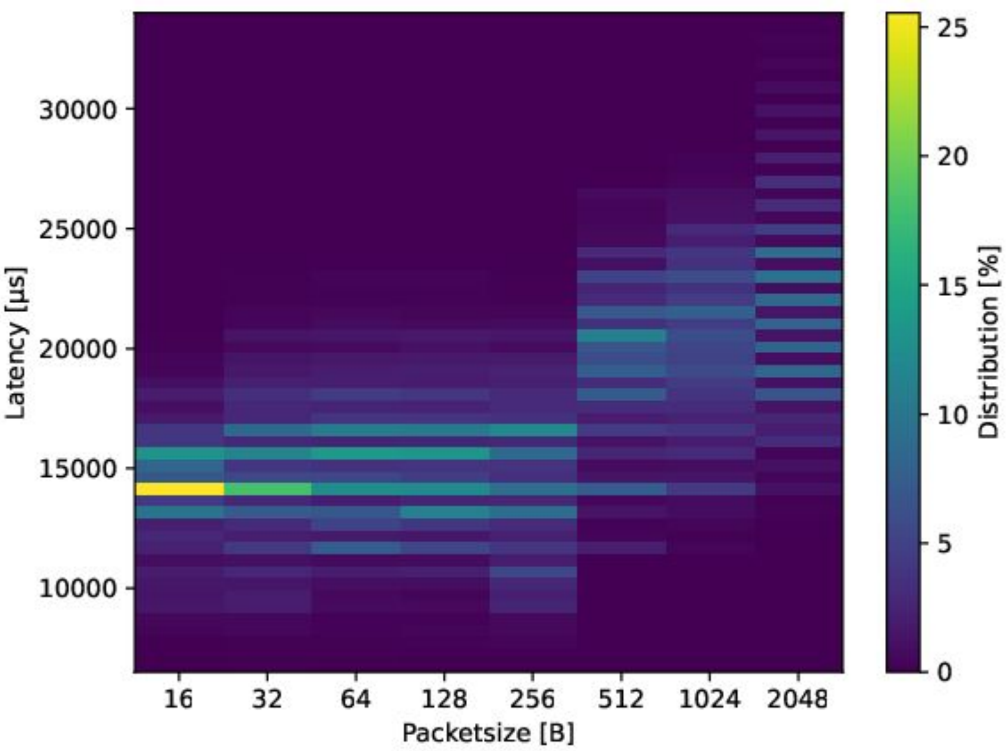
- Hardware improvements during 2022 supporting Release 16
  - Mediatek M80 chip platform released in Q1 2022
  - Qualcomm X65 or X62

Backup Slides

Latency – PS2PS

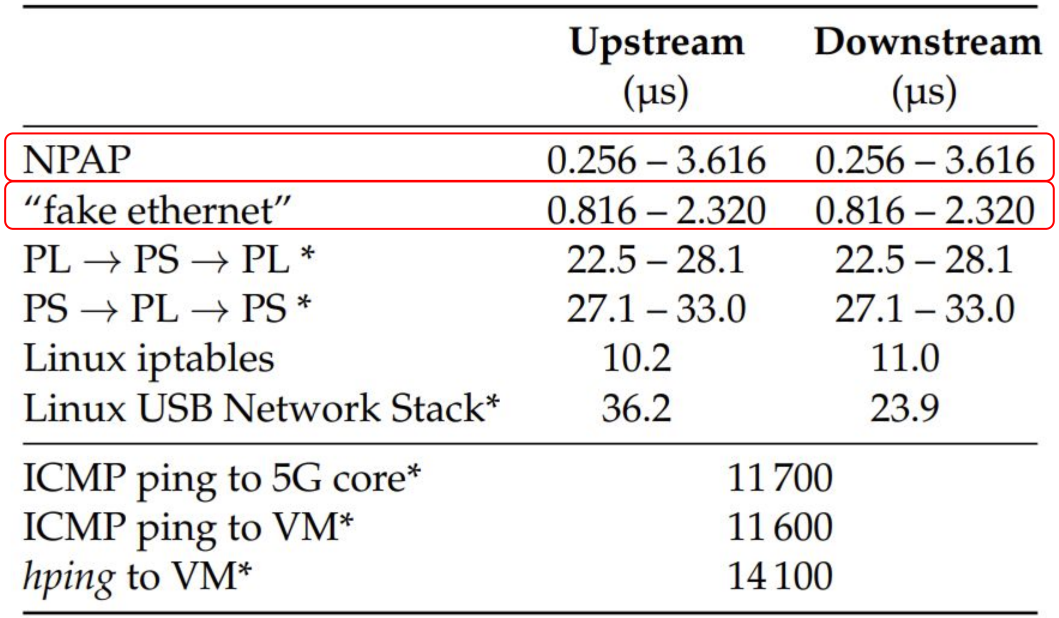


ø 17.7 ms



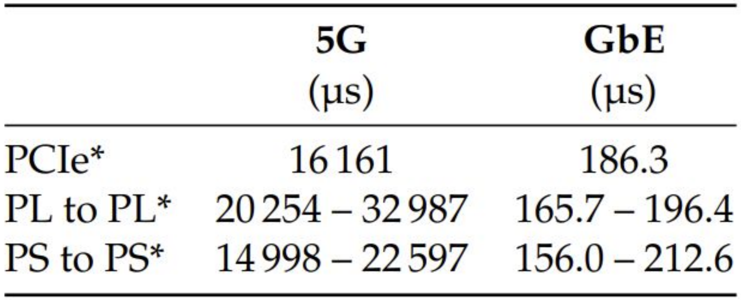
- User space C-program
- Server/Client with `gettimeofday()`
- Customizable packet size

Latency – PL



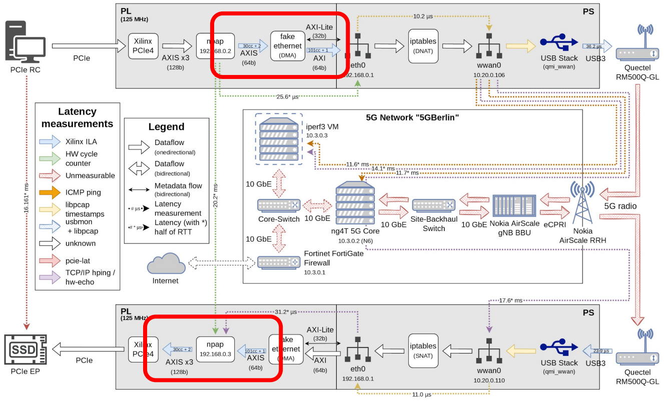
	Upstream (µs)	Downstream (µs)
NPAP	0.256 – 3.616	0.256 – 3.616
“fake ethernet”	0.816 – 2.320	0.816 – 2.320
PL → PS → PL *	22.5 – 28.1	22.5 – 28.1
PS → PL → PS *	27.1 – 33.0	27.1 – 33.0
Linux iptables	10.2	11.0
Linux USB Network Stack*	36.2	23.9
ICMP ping to 5G core*	11 700	
ICMP ping to VM*	11 600	
hping to VM*	14 100	

(a) Individual component latencies



	5G (µs)	GbE (µs)
PCIe*	16 161	186.3
PL to PL*	20 254 – 32 987	165.7 – 196.4
PS to PS*	14 998 – 22 597	156.0 – 212.6

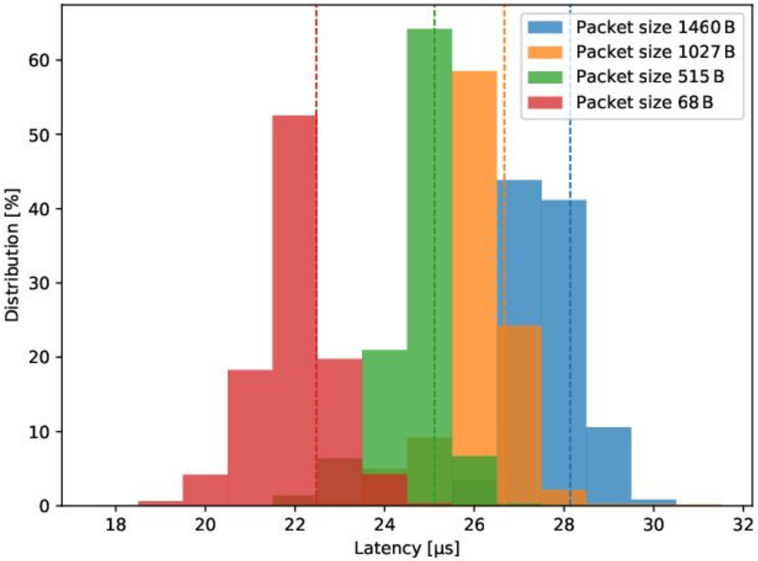
(b) E2E latencies

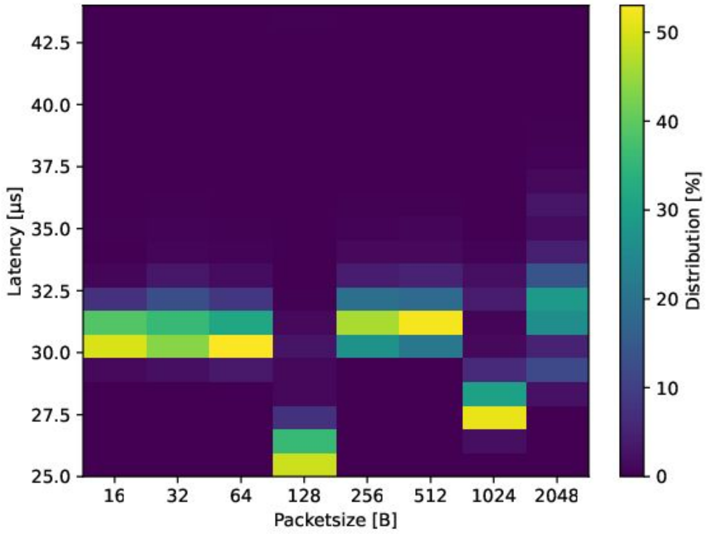


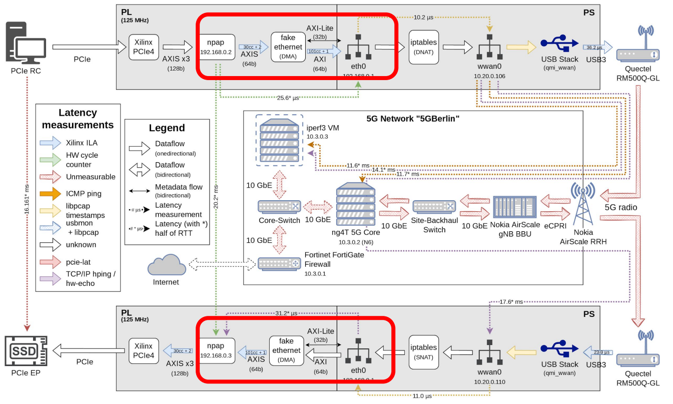
- Xilinx Integrated Logic Analyzer

- Count cycles

🔍 Latency – PS/PL







US: ø 25.6 µs

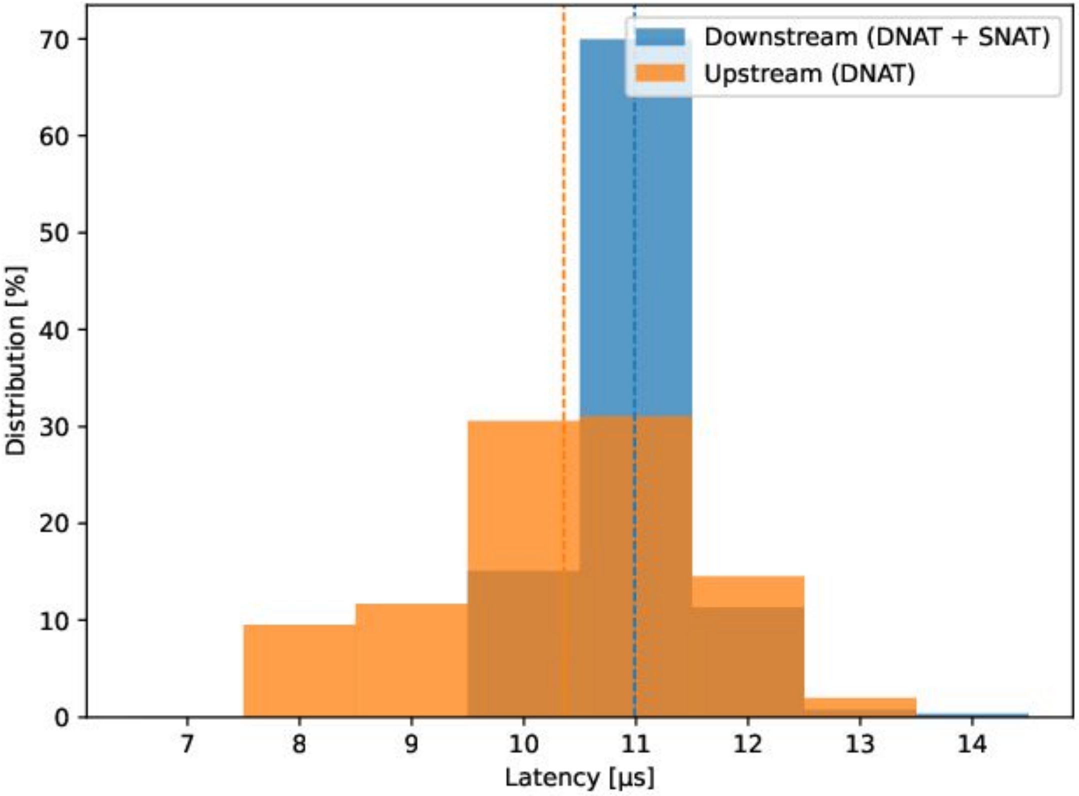
DS: ø 32.2 µs

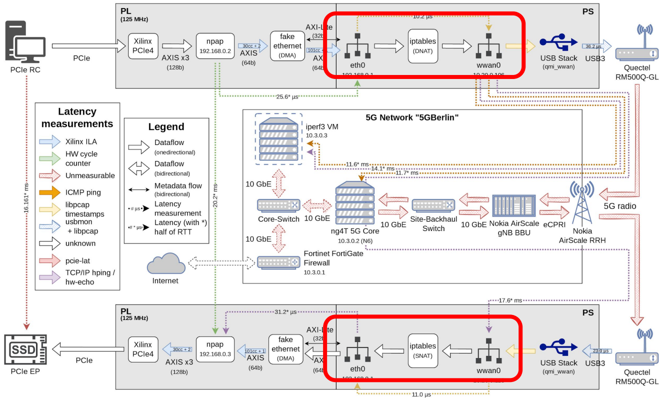
- PL -> PS -> PL:
  VHDL Counter

- PS -> PL -> PS:
  User space C-program

mle missing link electronics

🔍 Latency – iptables



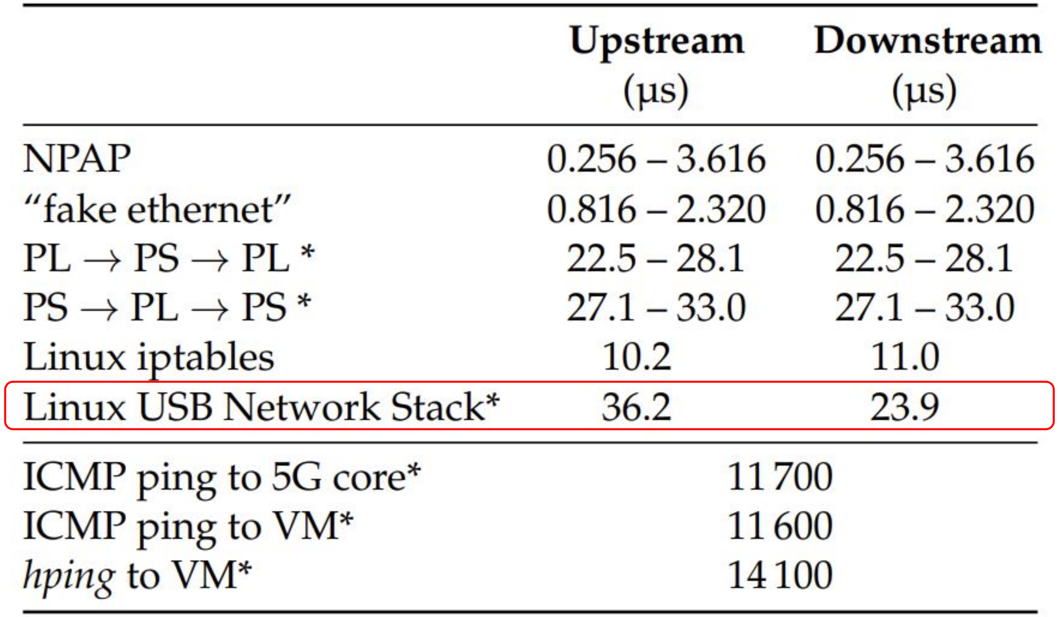


US: ø 10.2 µs

DS: ø 11.0 µs

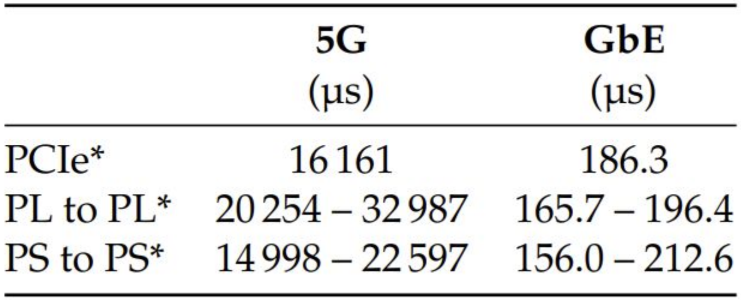
- libpcap timestamps
- tcpdump of both interfaces
- SNAT/DNAT latency

🔍 Latency – Linux USB stack



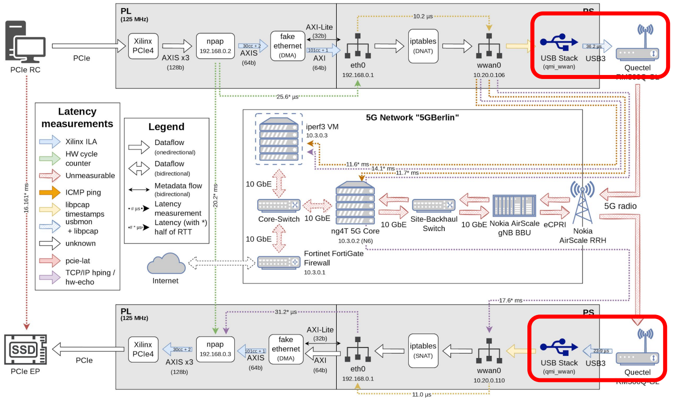
	Upstream (µs)	Downstream (µs)
NPAP	0.256 – 3.616	0.256 – 3.616
"fake ethernet"	0.816 – 2.320	0.816 – 2.320
PL → PS → PL *	22.5 – 28.1	22.5 – 28.1
PS → PL → PS *	27.1 – 33.0	27.1 – 33.0
Linux iptables	10.2	11.0
Linux USB Network Stack*	36.2	23.9
ICMP ping to 5G core*	11 700	
ICMP ping to VM*	11 600	
hping to VM*	14 100	

(a) Individual component latencies



	5G (µs)	GbE (µs)
PCIe*	16 161	186.3
PL to PL*	20 254 – 32 987	165.7 – 196.4
PS to PS*	14 998 – 22 597	156.0 – 212.6

(b) E2E latencies



- tcpdump with usbmon
- Match USB packets to network packets