STORAGE DEVELOPER CONFERENCE
SDC 22
Fremont, CA
September 12-15, 2022
BY Developers FOR Developers
A SNIA Event

Converging PCIe and TSN Ethernet for Composable Infrastructure in High-Performance In-Vehicle Embedded Systems

Endric Schubert, PhD, CTO Missing Link Electronics
Marcus Pietzsch, Research Lead Fraunhofer IPMS

MLE: “If It Is Packets, We Make It Go Faster!”

MLE is an Integrator and Turnkey Solutions / Systems Provider for
High-Performance (Embedded) Compute & Connected Systems-of-Systems

- PCIe (CXL, ISB, NVMe)
- Ethernet (TCP/IP, TSN)
- Audio/Video (HDMI, SDI)

mle missing link electronics

PCI SIG

XILINX Premier Partner

Fraunhofer Heinrich-Hertz-Institut

Fraunhofer IPMS

aws

intel FPGA Design Solutions Network Gold

FPGA Microchip DESIGN PARTNER

OpenPOWER™

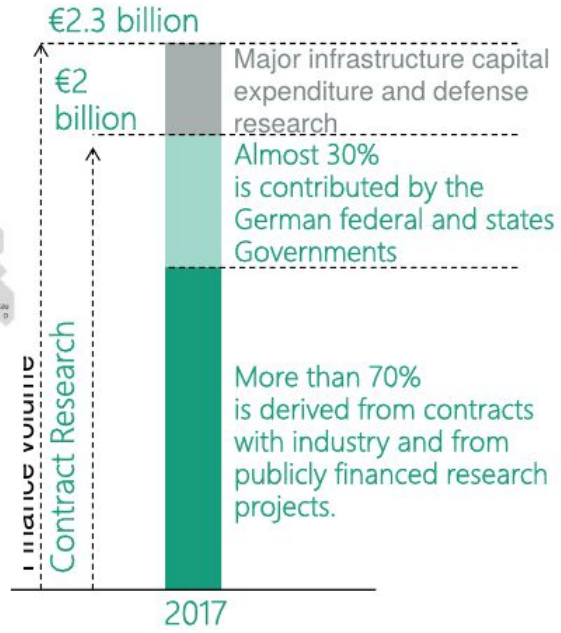
MLE Technology & Manufacturing Ecosystem

Fraunhofer

The Fraunhofer-Gesellschaft undertakes applied research of direct utility to private and public enterprise and of wide benefit to society.

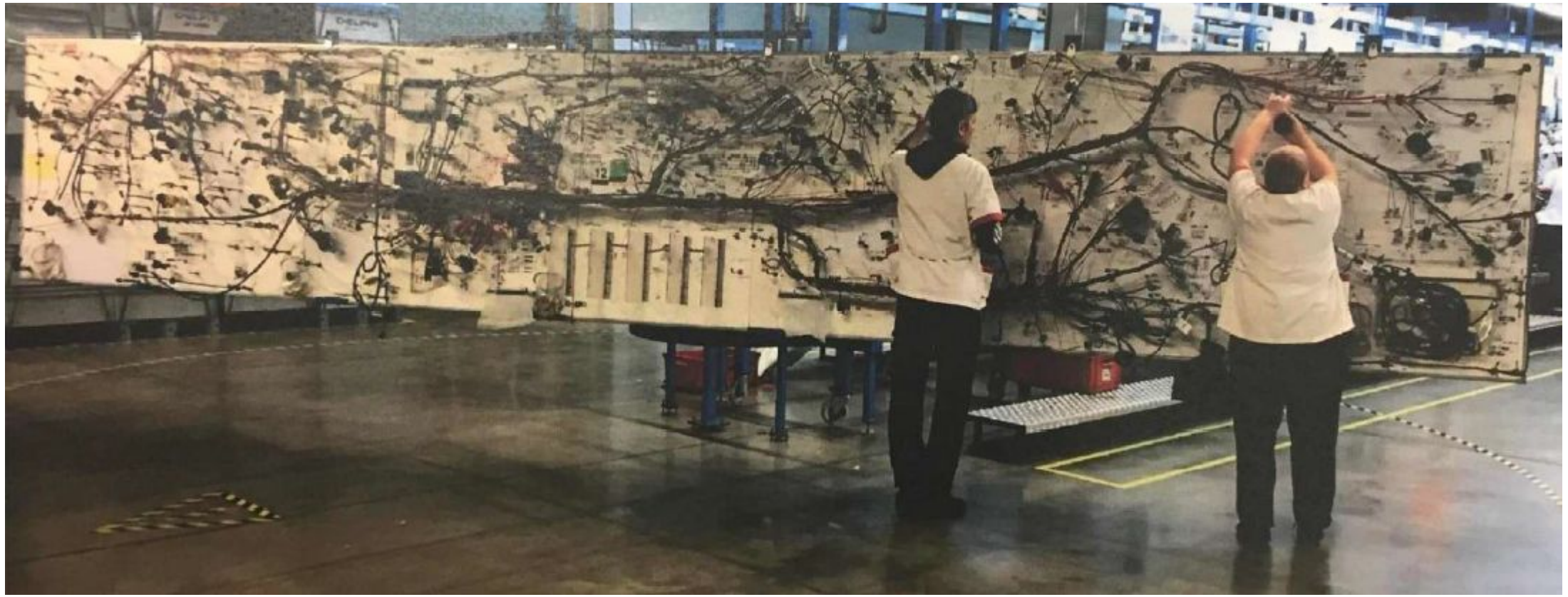
25,327 staff

72 institutes and research units



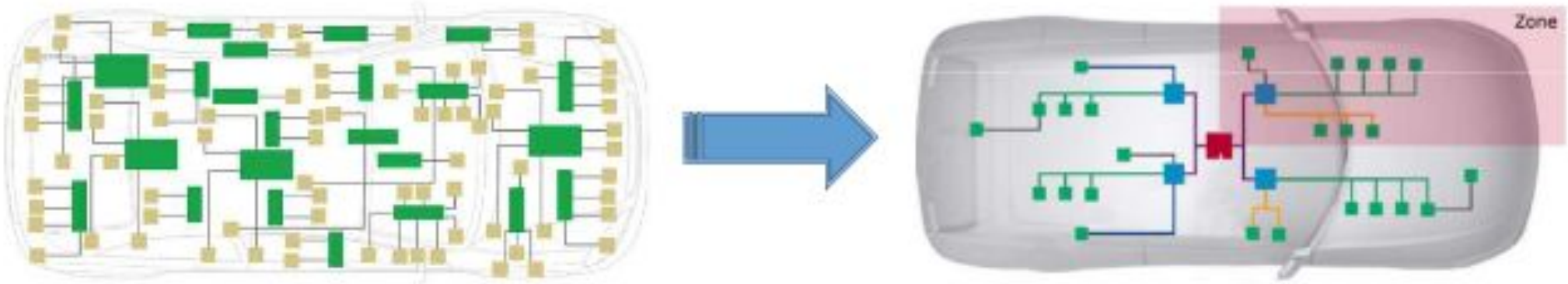
Elemaster

Automotive’s Expensive Wiring Nightmare



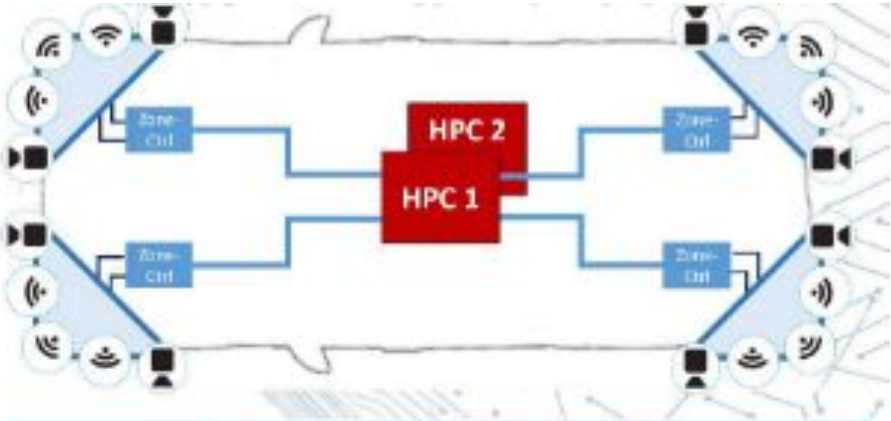
Next: Zone-Based Architectures

- Network connectivity is not based on functional domains, but on *physical location and proximity* inside the vehicle, i.e. “Zones”
- Data aggregation and preprocessing in Zone Controllers:
- High-bandwidth connectivity towards a central “High Performance Computer” HPC



Application Example: “Smart Corner”

- Smart Corner:
  integrates all data sources and sinks located at one corner of a vehicle
- Smart Corner Node contains e.g.:
  - 2 Lidars
  - 2 Radars
  - 2 Cameras
  - 2 Ultrasonics
  - 1 Lighting Unit



Traffic Type	Characteristics
Scheduled Traffic	High-priority real-time traffic transmitted according to a time schedule (time-driven), no interference from other traffic
Stream Reservation	Periodic, guaranteed
Event-driven Traffic	Aperiodic bursts, generated by sporadic events, with real-time constraints
Best-effort Traffic	No guarantees; statistical performance

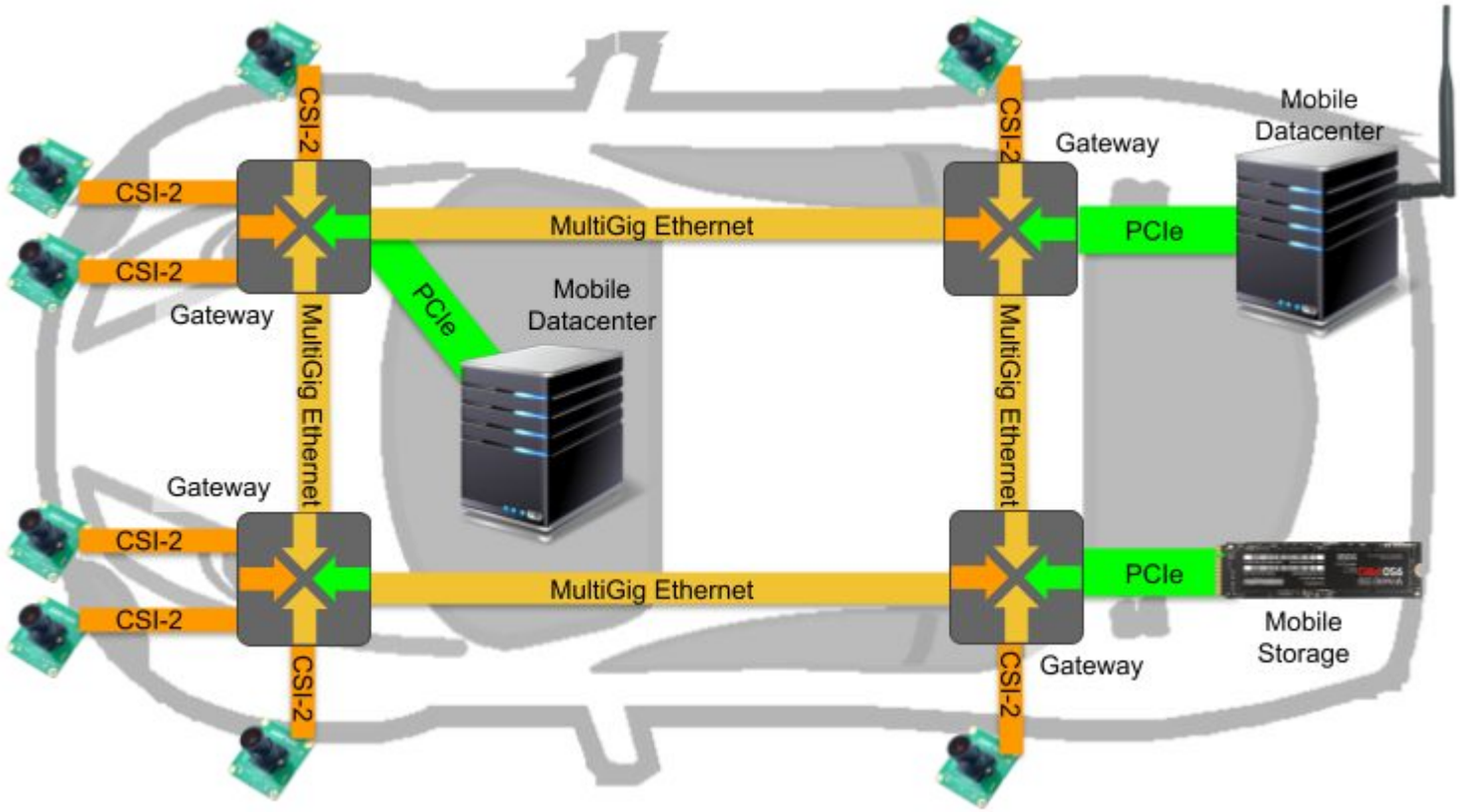
Zone-Based Automotive Networks Need PCIe/NVMe

Driven by Cost/Performance, i.e. Centralized Compute & Storage:

- PCIe (for Embedded CPUs, GPUs, FPGAs and SoCs)
- NVMe (for SSDs)

Driven by Compliance

- FuSa ISO 26262
- Security ISO/SAE 21434
- SOTIF ISO 21448
- etc



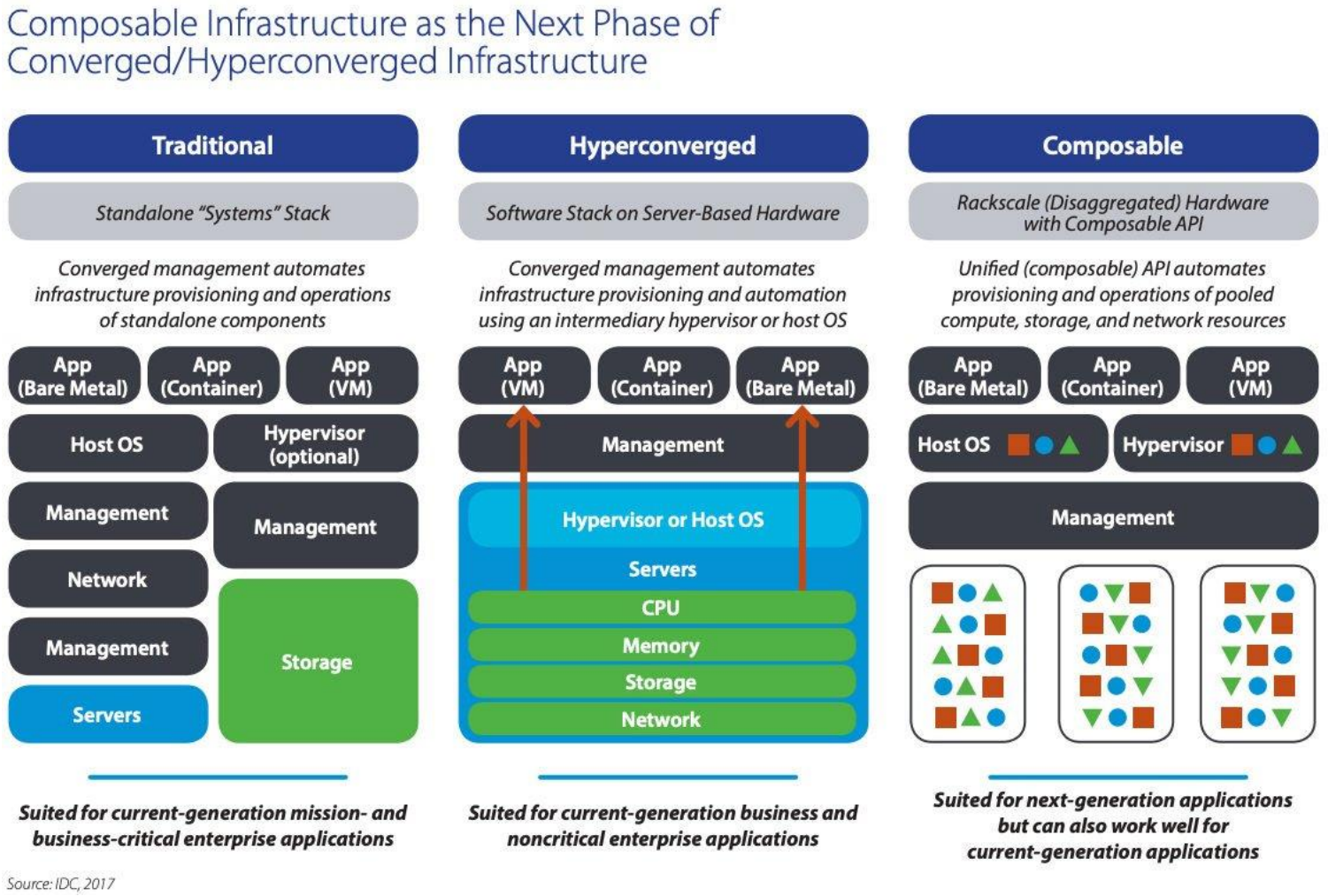
“Borrowing” from Datacenter Infrastructure (... and Giving Back)

Aspect	Datacenter Infrastructure	In-Vehicle Networks
Longevity	~ 15 years	~ 15 years
Proper functioning	High-Availability via SLA	Functional Safety as in ISO 26262
Security	Encryption in flight and at rest ISO 27001 etc	Soon: Encryption in flight and at rest ISO 21434
Network timing behavior	Low tail-end transport latency Avoid congestion and HoL blocking	Deterministic low latency Real-time
Environmental	Low power and high energy efficiency	Low power and high energy efficiency Resistant to shock, vibration, temp cycles
Number of nodes within system	100s of thousands	< 10
Flexibility needs	ability to deal with many different work loads, screwdriver-less add/change HW and SW	ability to field-upgrade functionality and security, screwdriver OK

Composable Datacenter Infrastructure

Automotive

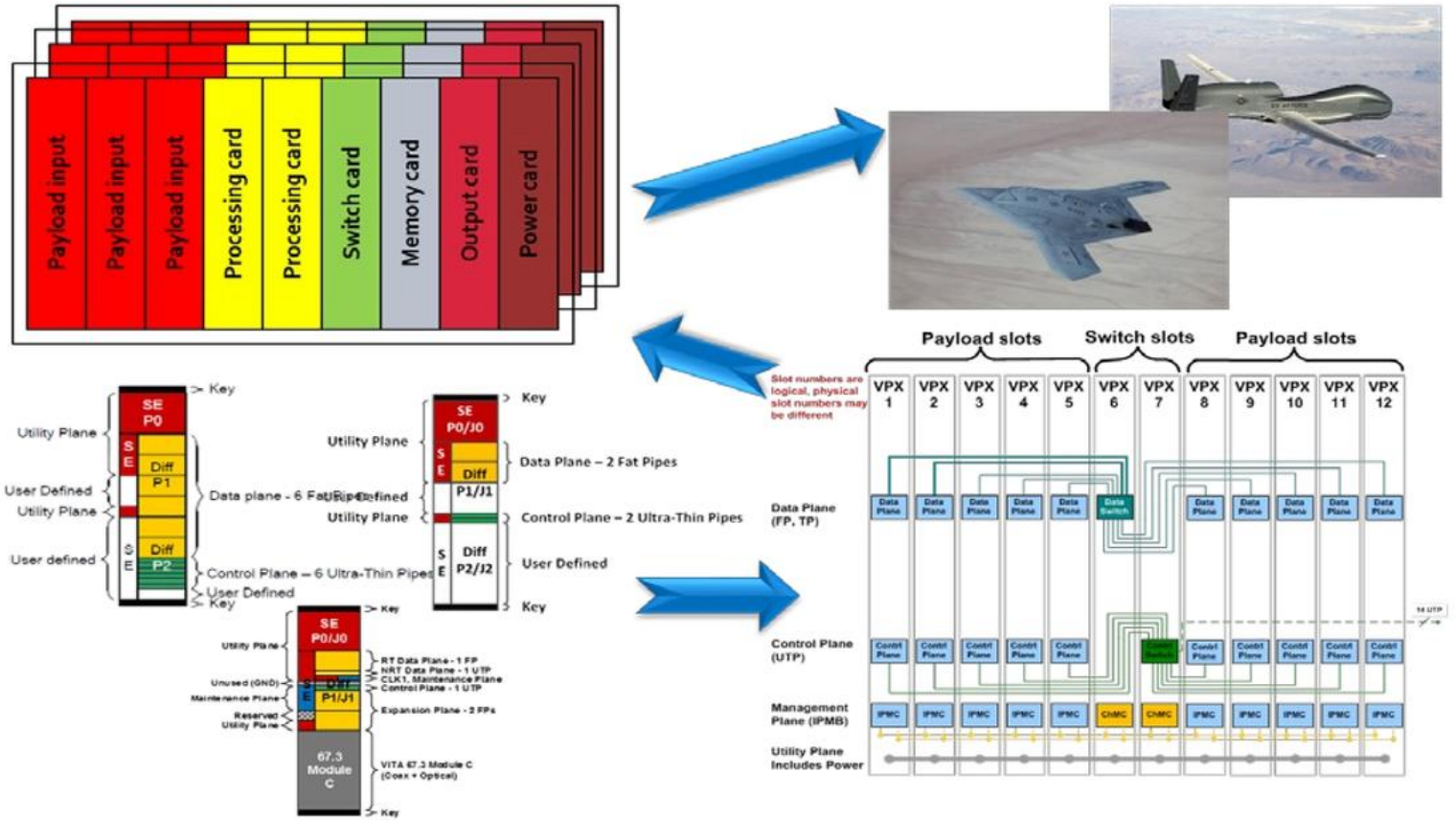
- Today: Traditional (~100 ECUs)
- Next: Leap-frog towards Composable i/f



Sensor Open Systems Architecture (SOSA)

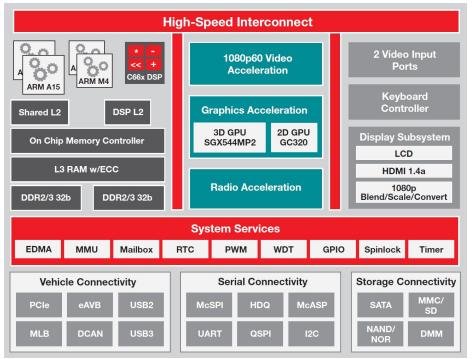
Other Embedded

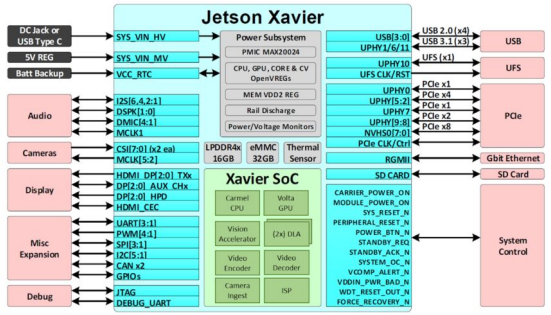
- Similar life-cycle challenges
- Need for field-upgrade and in-field repair
- Connectivity based on
  - PCIe/NVMe
  - Ethernet

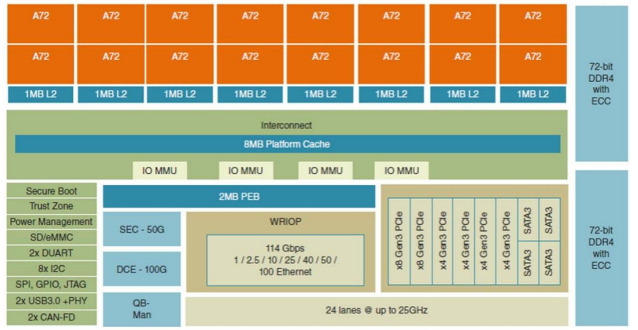


Why PCI Express?

- Future-proof road-map, driven by PCI-SIG
- NVM Express cost/performance/power optimized storage
- PC, Cloud Computing, Embedded Systems drive this roadmap
- Best-in-class price ($) per performance (Gbps) ratio
- Common to modern SoCs, ability to commoditize silicon







Why Ethernet?

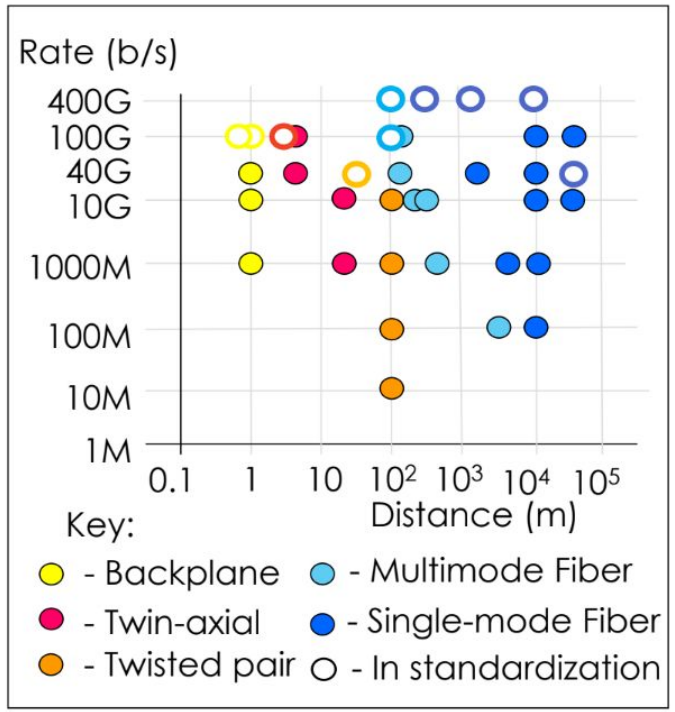
- Future-proof road-map
- PC, Cloud Computing, Embedded Systems drive this roadmap
- Best-in-class price ($) per performance (Gbps) per length (meters) ratio

Distance vs Speed

ethernet alliance

Ethernet operates at different speeds over different distances depending on the media :

- backplanes up to 1m
- Twinax to 15m
- Twisted pair to 100m
- Multimode fiber to 5km
- Single-mode fiber to 40km



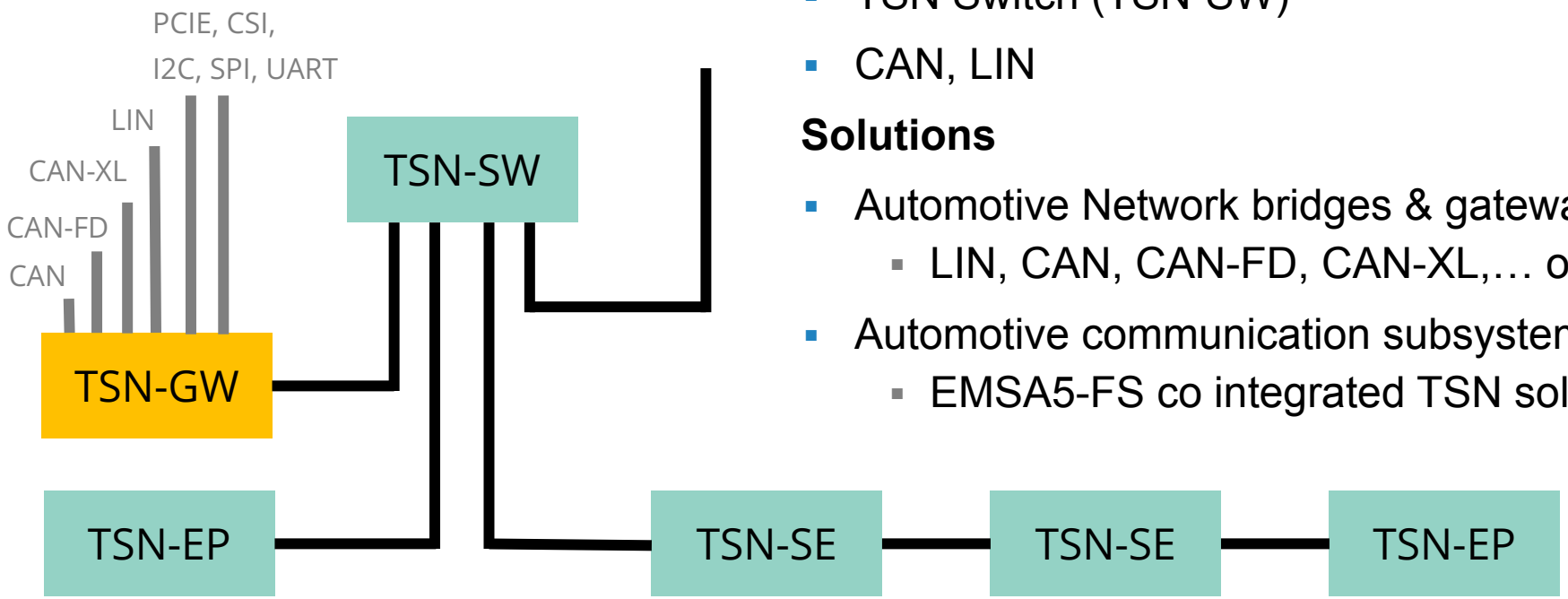
Why TSN?

- **T**ime **S**ensitive **N**etworking
- The history
  - AVB Task Group for latency free delivery of audio/video data
  - 2012 the TSN Task Group evolved from the AVB Task Group (IEEE 802.1)
- TSN is not a single standard
- It‘s a collection of sub standards and extensions
  - Network wide time synchronization
  - Determinism
  - Low latency
  - Low jitter
- Scalable speed

TSN Profiles & Standards

- Standards
  - Time Synchronization (802.1AS & it‘s profile IEEE 1588, 802.1AS/-2020)
  - Bounded low latency (802.1Qav, 802.1Qbv, 802.3br & 802.1Qbu, 802.1Qch, P802.1Qcr, P802.1Qcr, P802.1DC)
  - High availability/reliability (802.1CB, 802.1Qci, 802.1Qca)
  - Resources and API (802.1Qat, 802.1Qcc, 802.1Qcp, P802.1Qcx, P802.1ABcu, P802.1Qcw, 802.1CBcv, P802.1CS, P802.1Qdd, P802.1CBdb..)
- Profiles
  - Audio Video Bridging (802.1BA)
  - Fronthaul (802.1CM)
  - Industrial Automation(IEC/IEEE P60802)
  - Automotive In-Vehicle (P802.1DG)
  - Service Provider (P802.1DF)
  - TSN for Aerospace Onboard Ethernet (P802.1DP)
  - TSN for Avionics (SAE AS-1A2*)

TSN Technology from Fraunhofer IPMS



Pre-Certified & cut through IP Cores

- TSN Endpoint (TSN-EP)
- TSN Switched Endpoint (TSN-SE)
- TSN Switch (TSN-SW)
- CAN, LIN

Solutions

- Automotive Network bridges & gateways
  - LIN, CAN, CAN-FD, CAN-XL,… over TSN
- Automotive communication subsystems
  - EMSA5-FS co integrated TSN solutions

Challenges for TSN at Multi-Gigabit Speeds

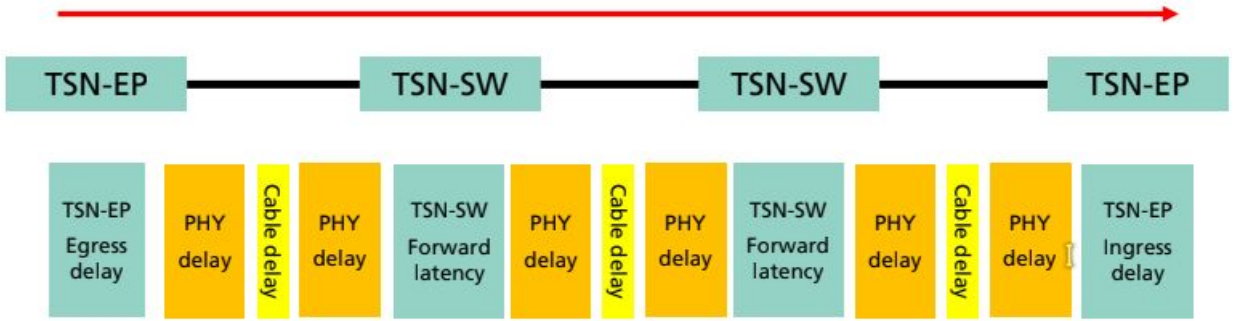
TSN <= 1 Gbps

- Used in industrial networks
- Limited amount of data, mostly for control
- SW-rich systems running RTOS
- CPUs fast enough for data processing
  - DMA i/f to SW
  - Little or no offloading needed

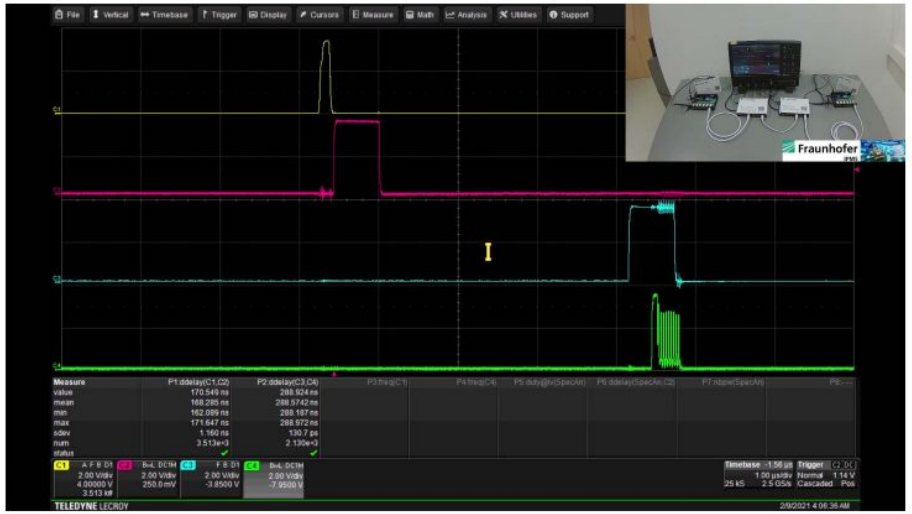
TSN >= 10 Gbps

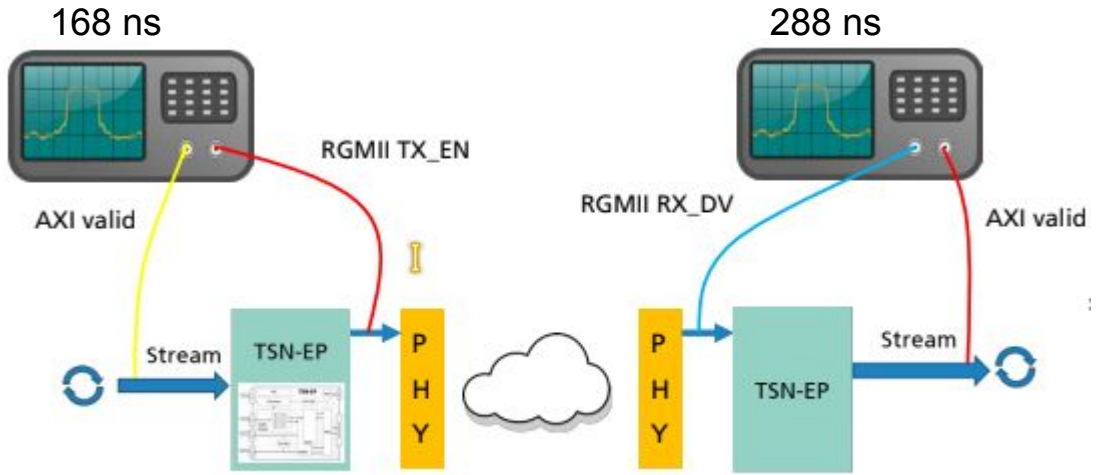
- High-performance distributed systems in vehicles and robots
- Large amounts of data from sensors
  - Radar, Lidar, Cameras
- CPUs too slow for data processing mostly
  - Onchip stream i/f
  - Fixed and programmable function accelerators

TSN Technology from Fraunhofer IPMS



TSN-EP ingress/egress – Stream in

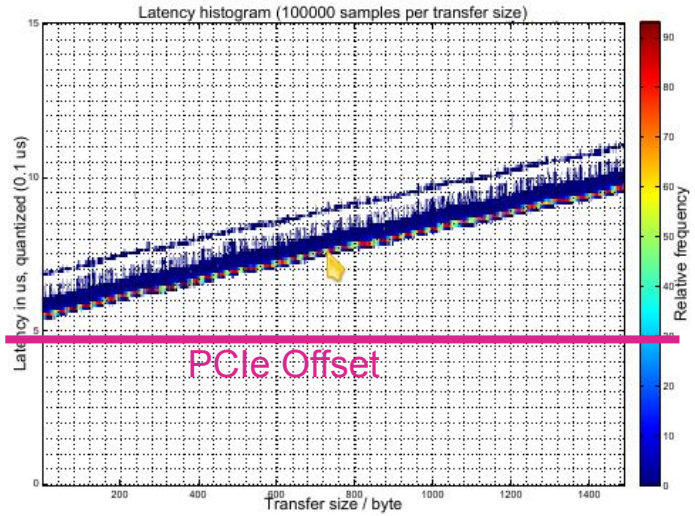


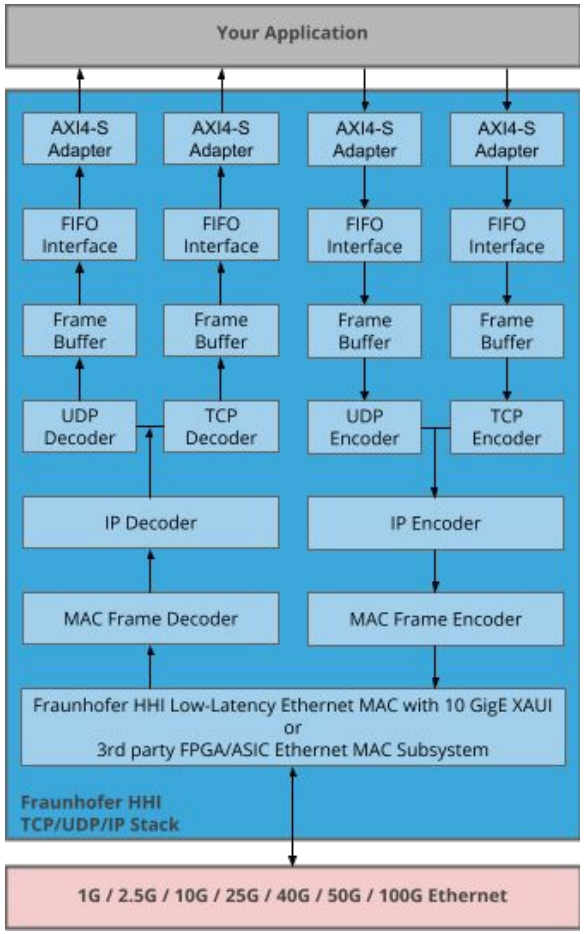


Why TCP?

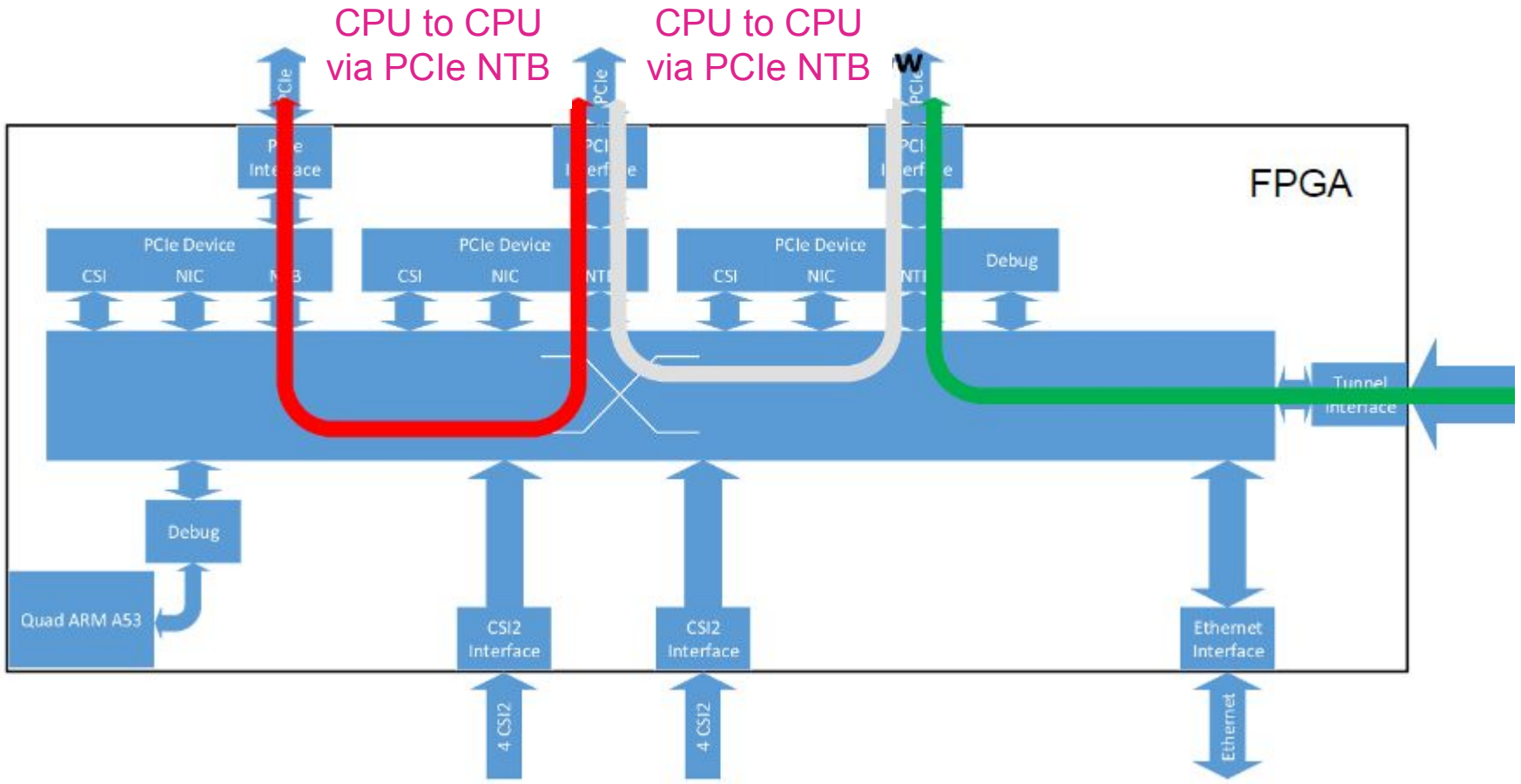
Benefits

- PCIe Transport Layer required reliability
- Long established and well understood SW API
- On-chip Full Accelerator from Fraunhofer HHI
  - Industry-proven
  - Resource efficient
  - 128 bit wide for up to 100 Gbps linerate processing
  - Low and deterministic latency (700 ns RTT for 100B)





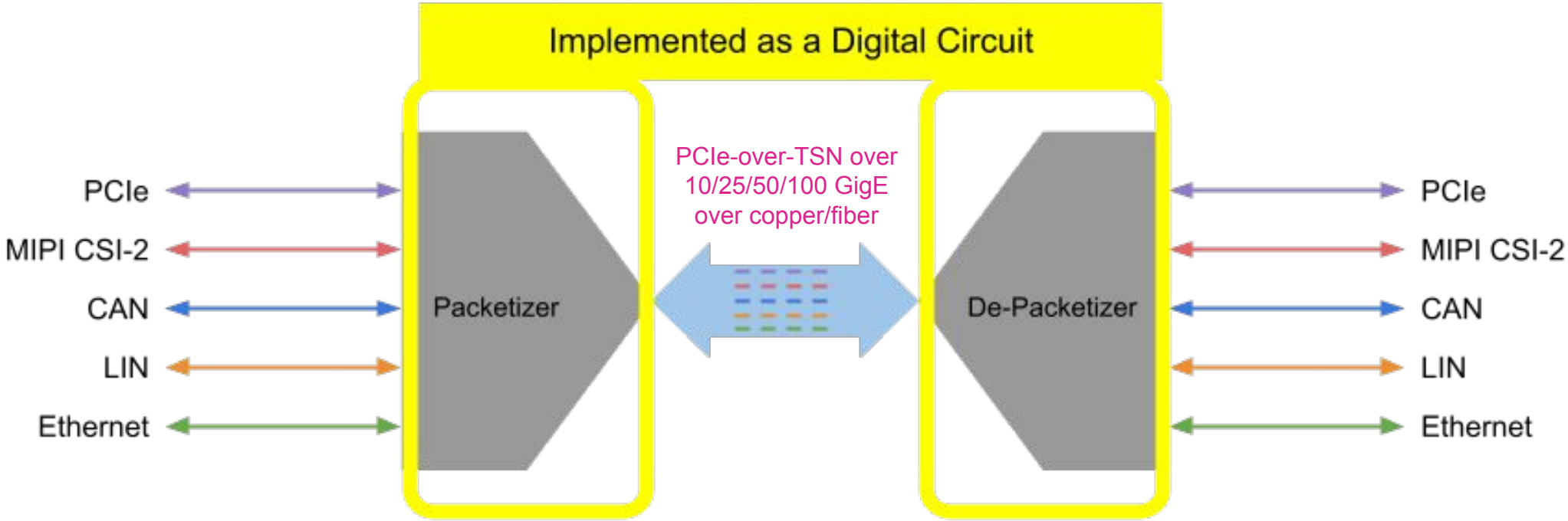
Converging PCIe/NVMe and TSN - Automotive Connectivity



PCIe-over-TSN
NVMe-over-TSN
over 10/25/50/100 GigE
over copper/fiber

PCIe-over-TSN / NVMe-over-TSN Hardware Portion

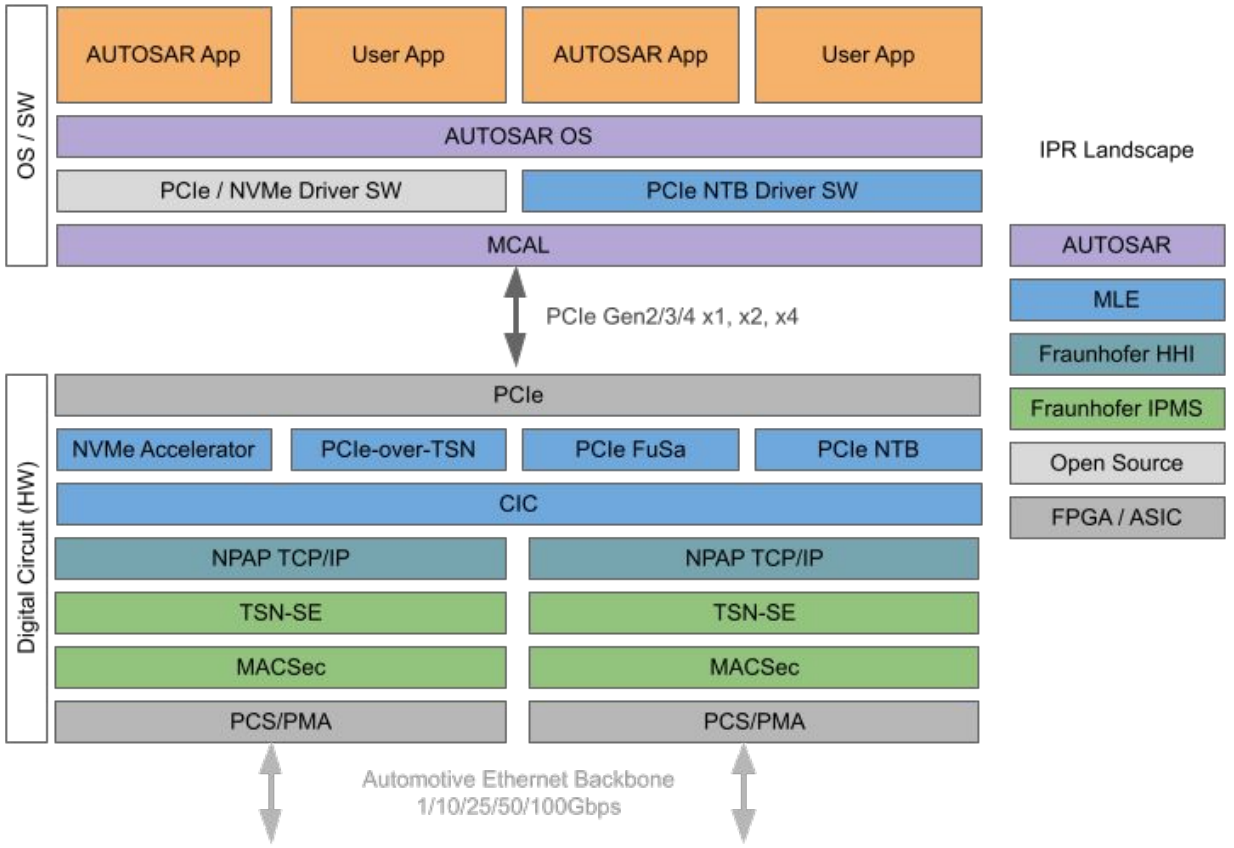
- PCIe from PCI-SIG, TSN from IEEE
- Symmetric for CPU-to-CPU (e.g. PCIe NTB) or Asymmetric Sensor-to-CPU
- US Patents 10,140,049 10,708,199 10,848,442 11,356,388



PCIe-over-TSN / NVMe-over-TSN System Stack

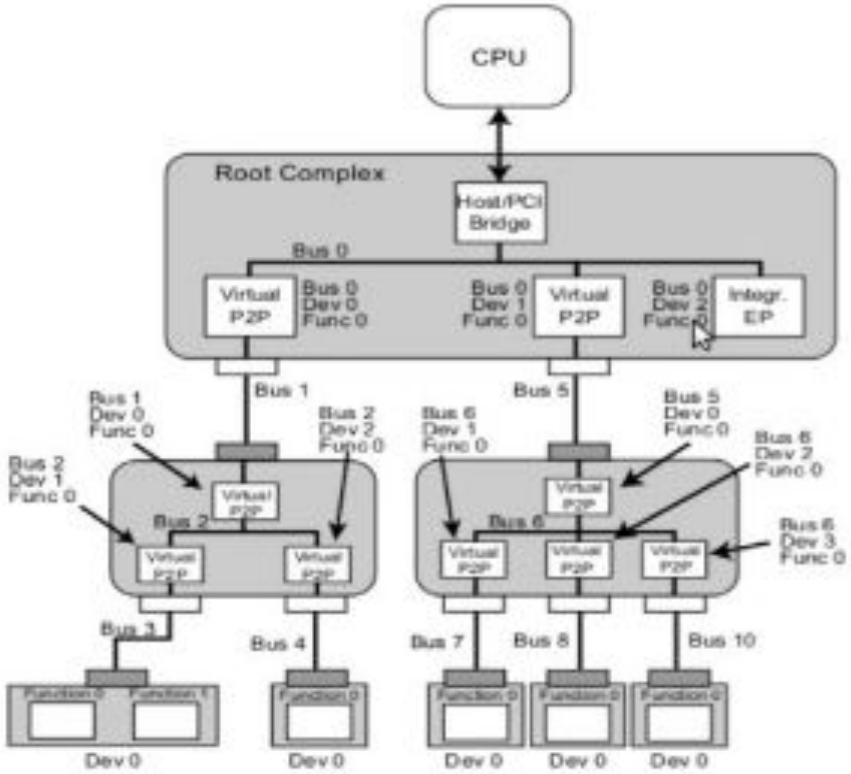
System Stack is
- Hardware (Digital Circuit)
- Software (Drivers)

Features
- PCIe Endpoint and Root-Port in FPGA/ASIC
- PCIe Switch in FPGA/ASIC
- PCIe NTB in FPGA/ASIC
- TCP/UDP/IP over TSN in FPGA/ASIC
- netdev Linux Device Drivers



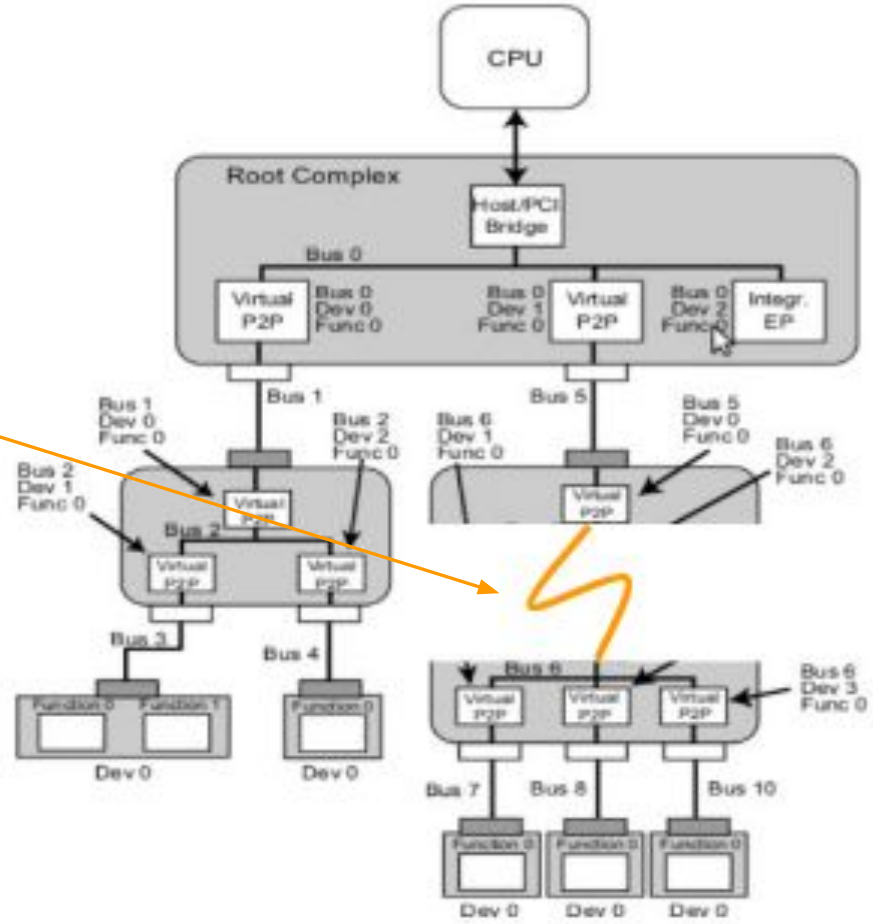
PCIe-over-TSN Concept: A Distributed PCIe Switch

PCIe Hierarchy with PCIe Switches



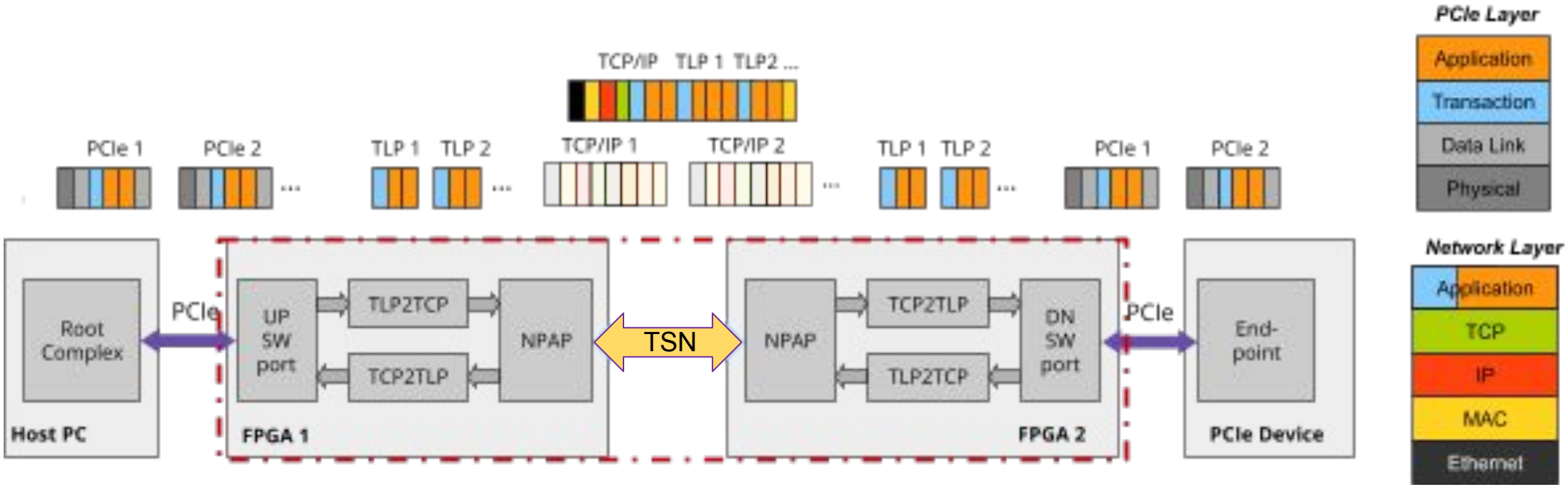
TSN

PCIe Long Range Tunnel "cuts open" PCIe Switch



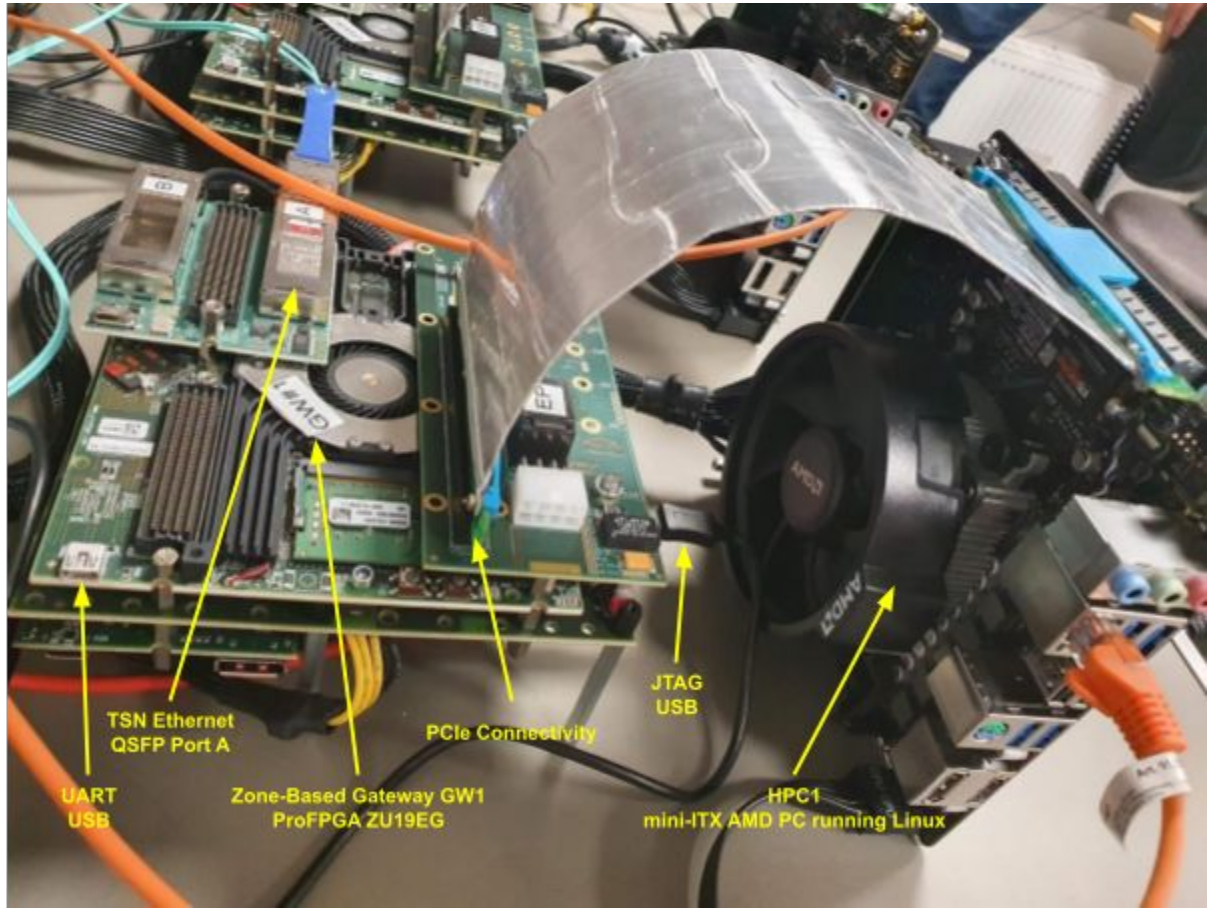
PCIe-over-TSN Concept: Distributed PCIe Switch

Encapsulate and Decapsulate PCIe TLPs. PCIe demands reliability, therefore we transport TLPs over TCP/IP over TSN over Ethernet.

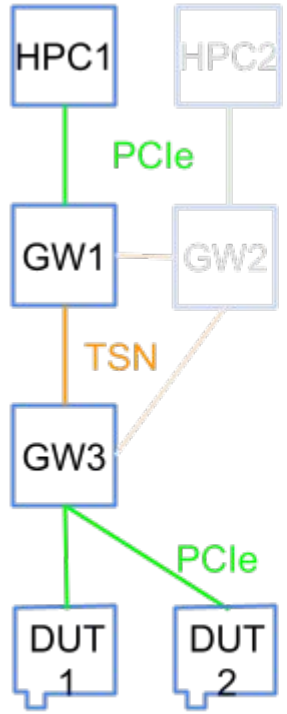


PCIe-over-TSN / NVMe-over-TSN Lab Car

Labcar Setup w/ PCIe Connect to HPC



Labcar Setup for Experiments PCIe/NVMe

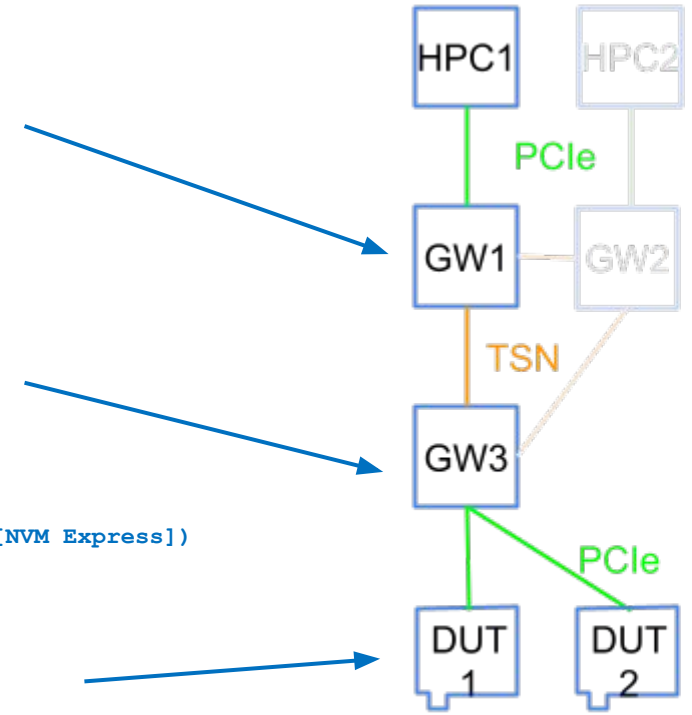


PCIe = 8GT/s x4 TSN = 1GE and 10GE

NVMe-over-TSN Results With SSD

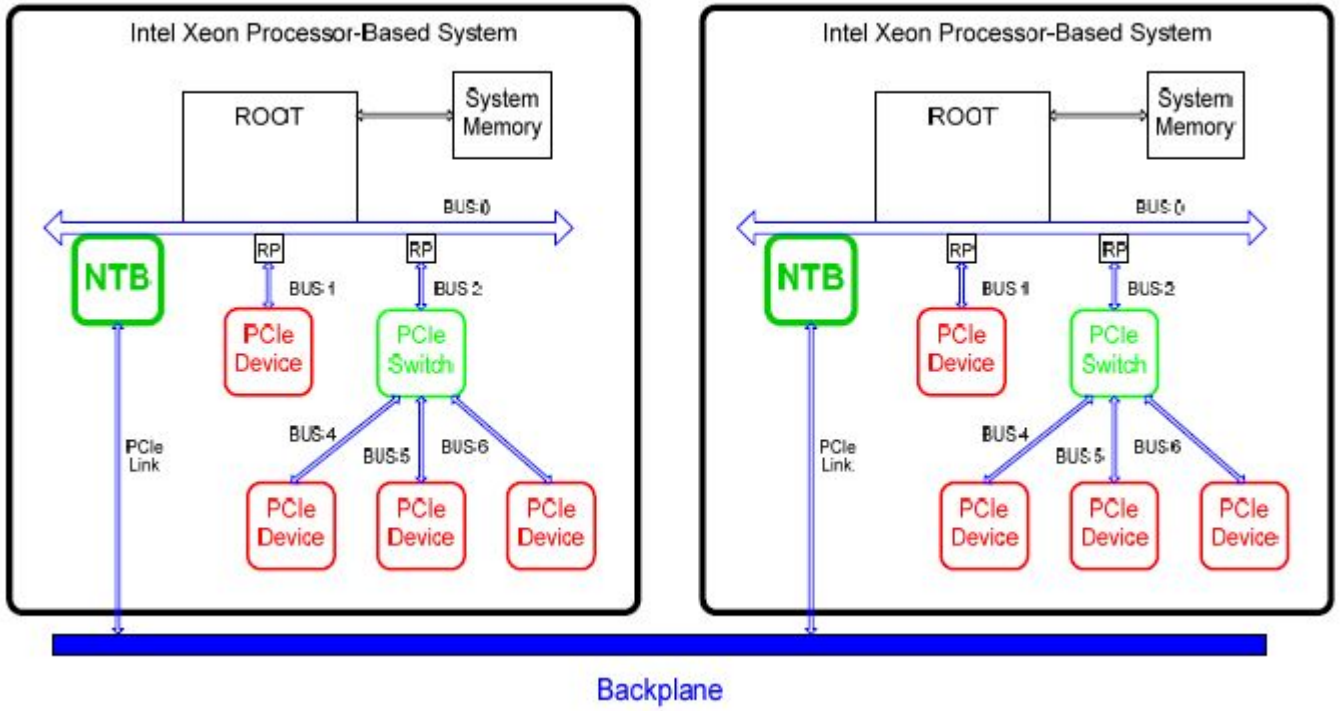
Linux lspci

```
00:00.0 Host bridge [0600]: Advanced Micro Devices, Inc. [AMD] Device [1022:15d0]
00:01.0 Host bridge [0600]: Advanced Micro Devices, Inc. [AMD] Device [1022:1452]

01:00.0 PCI bridge [0604]: Xilinx Corporation Device [10ee:9034] (prog-if 00 [Normal decode])
Control: I/O+ Mem+ BusMaster+ SpecCycle- MemWINV- VGASnoop- ParErr- Stepping- SERR- FastB2B- DisINTx-
Status: Cap+ 66MHz- UDF- FastB2B- ParErr- DEVSEL=fast >TAbort- <TAbort- <MAbort- >SERR- <PERR- INTx-
Latency: 0, Cache Line Size: 64 bytes
Bus: primary=01, secondary=02, subordinate=03, sec-latency=0
Capabilities: [70] Express (v2) Upstream Port, MSI 00
Capabilities: [1c0 v1] #19
Kernel driver in use: pcieport
Kernel modules: shpchp

02:00.0 PCI bridge [0604]: Xilinx Corporation Device [10ee:9134] (prog-if 00 [Normal decode])
Control: I/O+ Mem+ BusMaster+ SpecCycle- MemWINV- VGASnoop- ParErr- Stepping- SERR- FastB2B- DisINTx-
Status: Cap+ 66MHz- UDF- FastB2B- ParErr- DEVSEL=fast >TAbort- <TAbort- <MAbort- >SERR- <PERR- INTx-
Latency: 0, Cache Line Size: 64 bytes
Bus: primary=02, secondary=03, subordinate=03, sec-latency=0
Capabilities: [70] Express (v2) Downstream Port (Slot+), MSI 00
Capabilities: [1c0 v1] #19
Kernel driver in use: pcieport
Kernel modules: shpchp

03:00.0 Non-Volatile memory controller [0108]: Samsung Electronics Co Ltd Device [144d:a808] (prog-if 02 [NVM Express])
Subsystem: Samsung Electronics Co Ltd Device [144d:a801]
Control: I/O- Mem+ BusMaster+ SpecCycle- MemWINV- VGASnoop- ParErr- Stepping- SERR- FastB2B- DisINTx+
Status: Cap+ 66MHz- UDF- FastB2B- ParErr- DEVSEL=fast >TAbort- <TAbort- <MAbort- >SERR- <PERR- INTx-
Latency: 0, Cache Line Size: 64 bytes
Interrupt: pin A routed to IRQ 30
NUMA node: 0
Region 0: Memory at fcf00000 (64-bit, non-prefetchable) [size=16K]
Kernel driver in use: nvme
Kernel modules: nvme
```



PCIe Non-Transparent Bridge

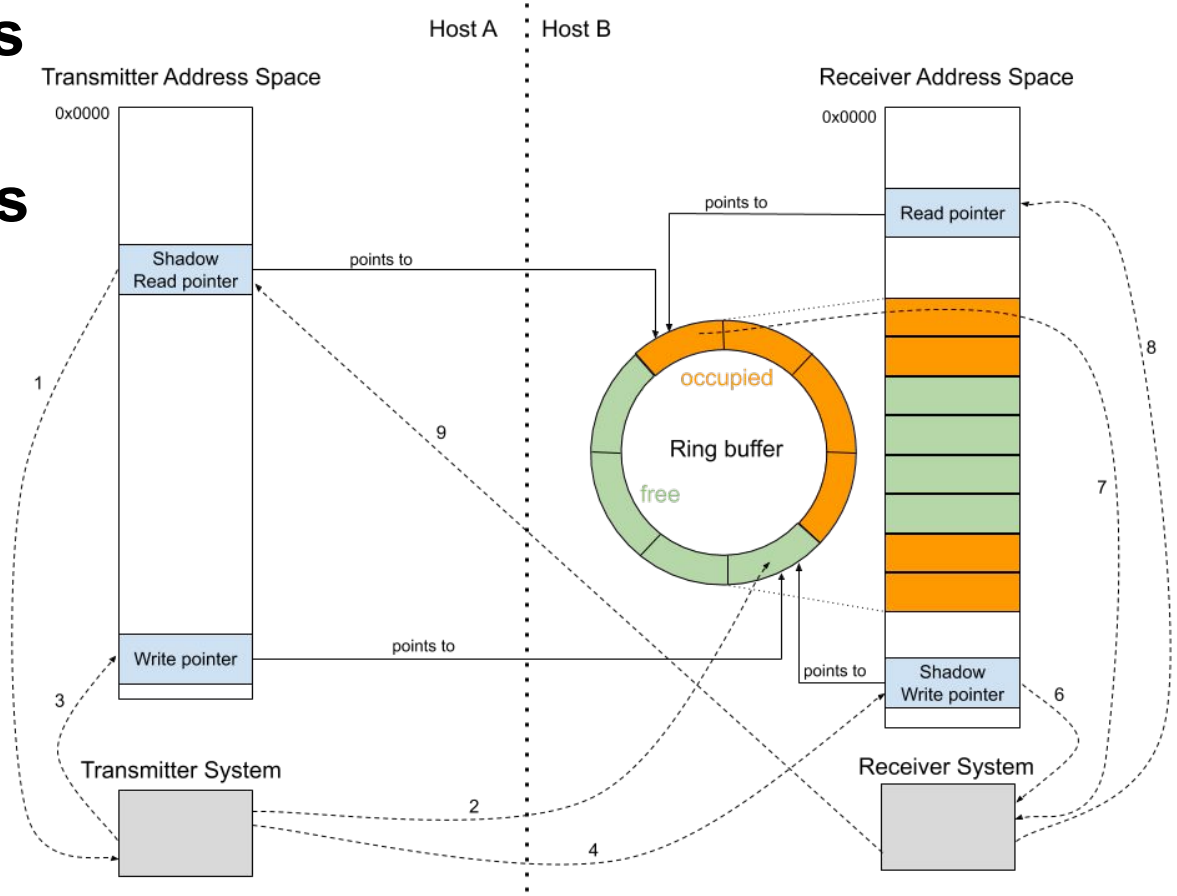
- **Non-Transparent Bridge (NTB) connects multiple Root Ports**
- **Example of NTB Back-2-Back**
  (Example from Intel Xeon C5500)



Delivering Performance for PCIe NTB

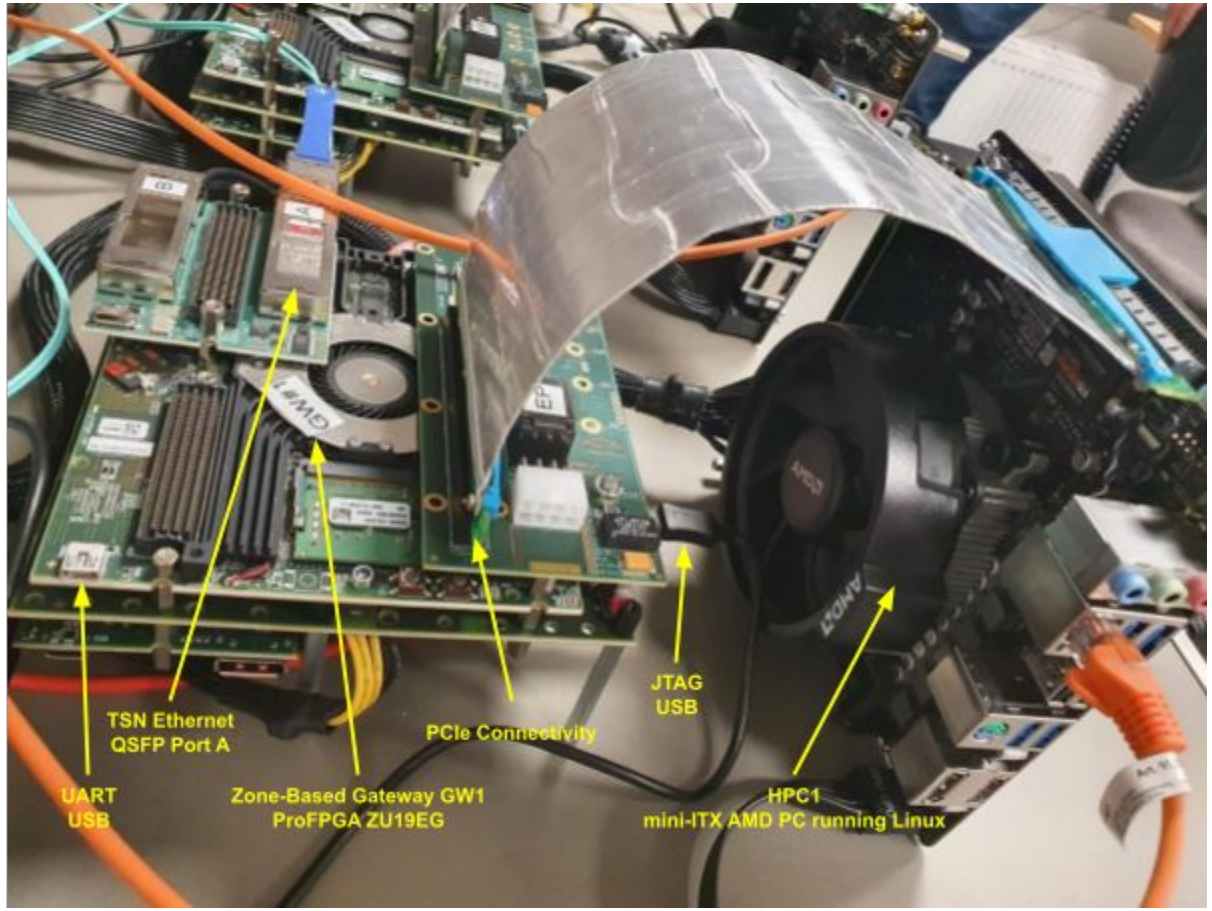
Write-Only Communication via Doorbells - NVMe-style

- Avoids difficulties of multi-device
- Scales to >32 RCs
- Posted Writes



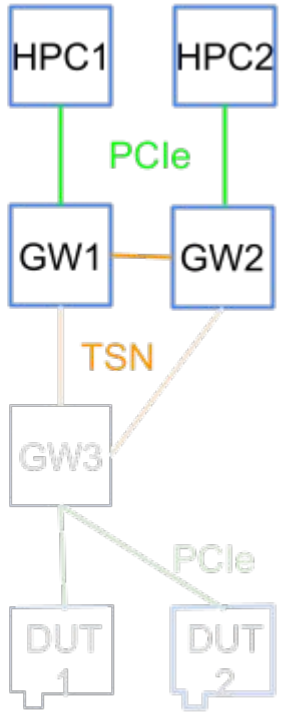
PCIe-over-TSN for NTB Lab Car

Labcar Setup w/ PCIe Connect to HPC



Labcar Setup for Experiments

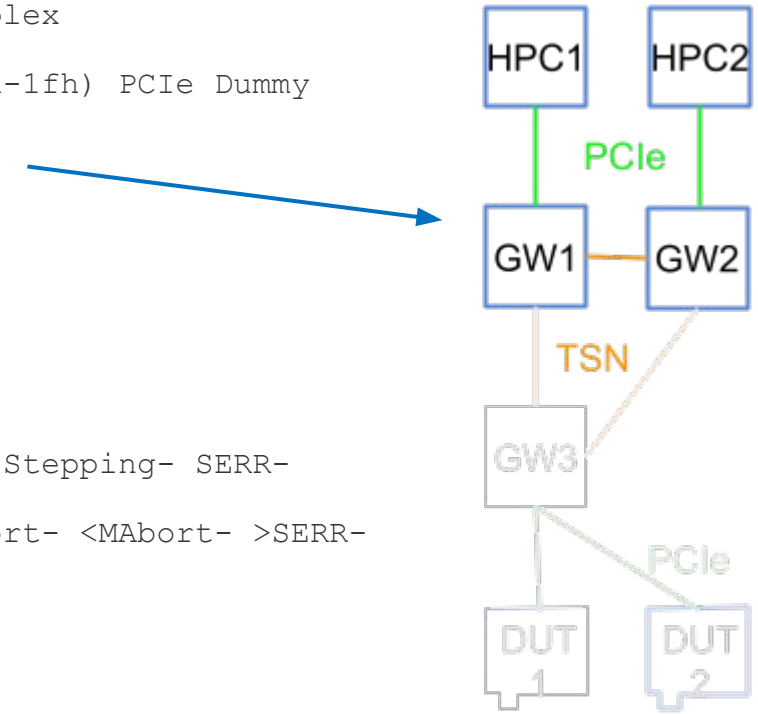
PCIE NTB



PCIe = 8GT/s x4 TSN = 1GE and 10GE

PCIe-over-TSN for Non-Transparent Bridging (NTB)

```
> lspci -vt

-[0000:00]-+-00.0  Advanced Micro Devices, Inc. [AMD] Raven/Raven2 Root Complex
           +-00.2  Advanced Micro Devices, Inc. [AMD] Raven/Raven2 IOMMU
           +-01.0  Advanced Micro Devices, Inc. [AMD] Family 17h (Models 00h-1fh) PCIe Dummy
                   Host Bridge
           +-01.1-[01]----00.0  Missing Link Electronics Device 22fb
...
```



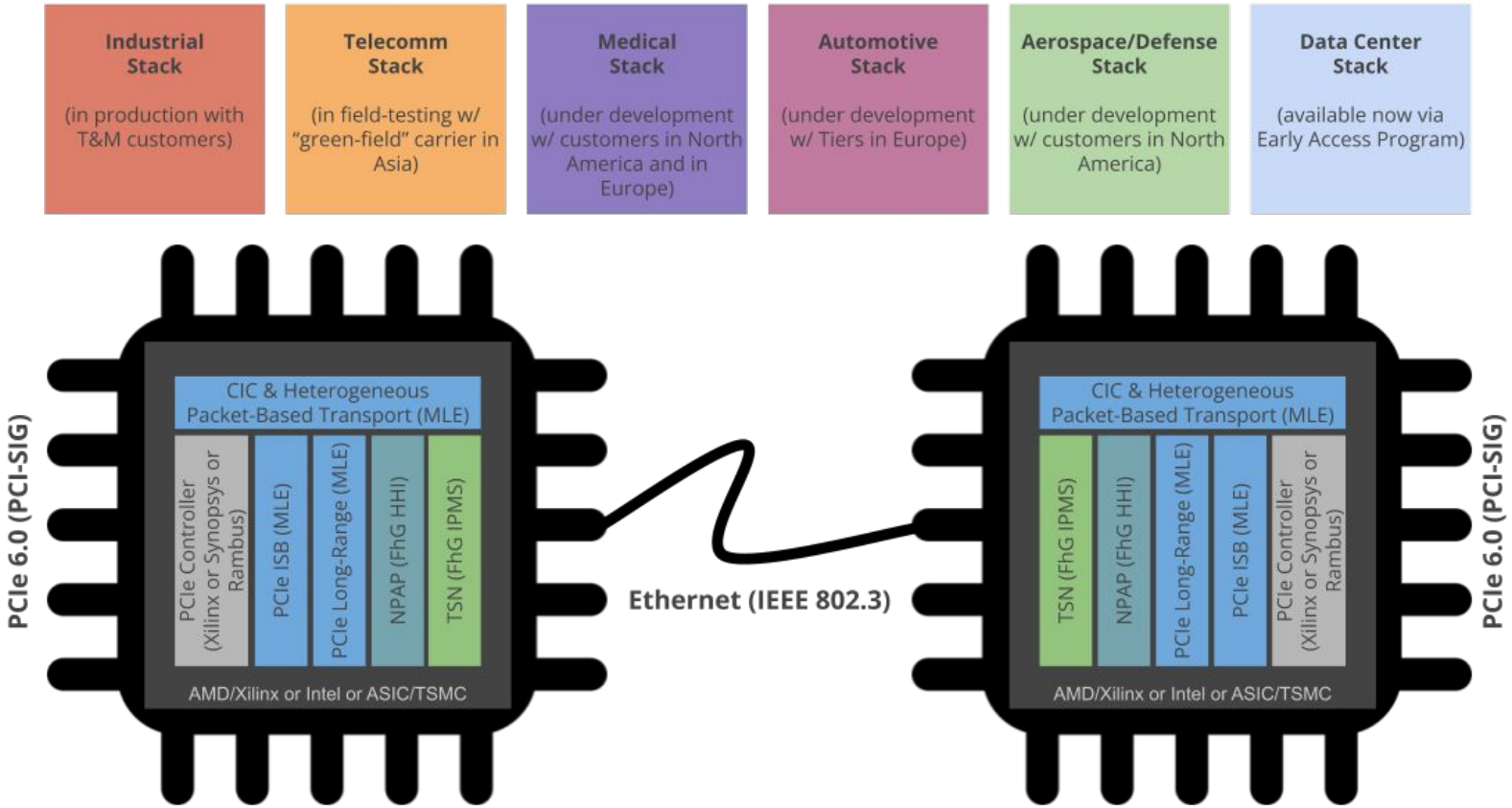

```
> lspci -vv -s 01:00.0

01:00.0 Memory controller: Missing Link Electronics Device 22fb
        Subsystem: Xilinx Corporation Device 0007
        Control: I/O- Mem+ BusMaster+ SpecCycle- MemWINV- VGASnoop- ParErr- Stepping- SERR-
          FastB2B- DisINTx+
        Status: Cap+ 66MHz- UDF- FastB2B- ParErr- DEVSEL=fast >TAbort- <TAbort- <MAbort- >SERR-
          <PERR- INTx-
        Latency: 0, Cache Line Size: 64 bytes
        Region 0: Memory at d0000000 (64-bit, prefetchable) [size=256M]
        Region 2: Memory at c0000000 (64-bit, prefetchable) [size=256M]
        Region 4: Memory at fcf10000 (32-bit, non-prefetchable) [size=64K]
        Region 5: Memory at fcf00000 (32-bit, non-prefetchable) [size=64K]
        Capabilities: [1c0 v1] #19
        Kernel driver in use: ntb_hw_mle
        Kernel modules: ntb_hw_mle
```


Convergence of PCIe/NVMe and TSN - Conclusion

Foundation Technology With Many Applications

Built on top of open standards from PCI-SIG and IEEE

Implemented in ASIC and FPGA



Outlook & Future Work

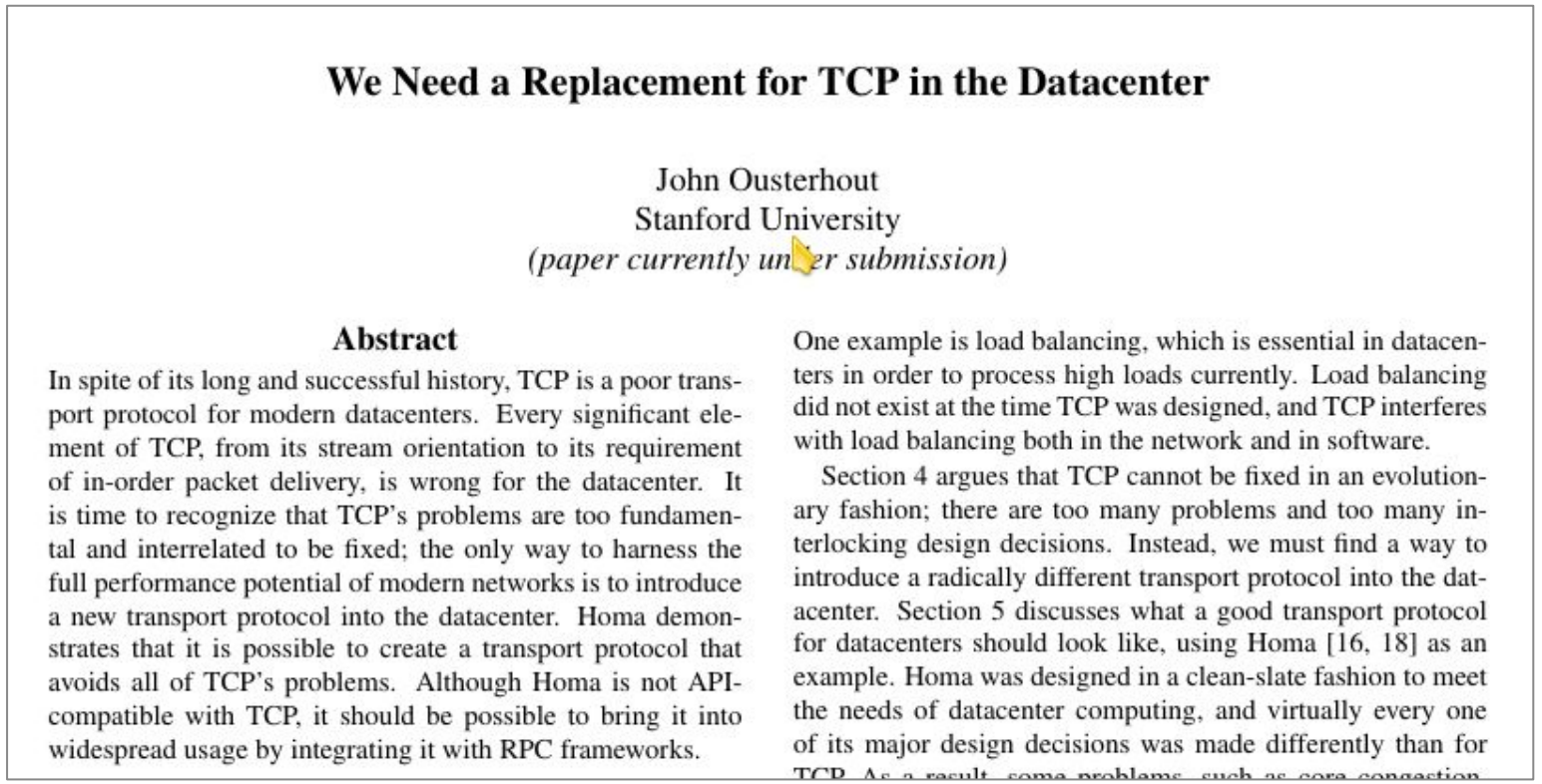
NVMe-over-Homa?

Today TCP:

- PCIe-over-TSN / NVMe-over-TSN

Next Homa:

- Alternatives to TCP for Storage
- Add security



We Need a Replacement for TCP in the Datacenter

John Ousterhout
Stanford University
(paper currently under submission)

Abstract

In spite of its long and successful history, TCP is a poor transport protocol for modern datacenters. Every significant element of TCP, from its stream orientation to its requirement of in-order packet delivery, is wrong for the datacenter. It is time to recognize that TCP's problems are too fundamental and interrelated to be fixed; the only way to harness the full performance potential of modern networks is to introduce a new transport protocol into the datacenter. Homa demonstrates that it is possible to create a transport protocol that avoids all of TCP's problems. Although Homa is not API-compatible with TCP, it should be possible to bring it into widespread usage by integrating it with RPC frameworks.

One example is load balancing, which is essential in datacenters in order to process high loads currently. Load balancing did not exist at the time TCP was designed, and TCP interferes with load balancing both in the network and in software.

Section 4 argues that TCP cannot be fixed in an evolutionary fashion; there are too many problems and too many interlocking design decisions. Instead, we must find a way to introduce a radically different transport protocol into the datacenter. Section 5 discusses what a good transport protocol for datacenters should look like, using Homa [16, 18] as an example. Homa was designed in a clean-slate fashion to meet the needs of datacenter computing, and virtually every one of its major design decisions was made differently than for TCP. As a result, some problems, such as core congestion

Today: PCIe-over-TSN / NVMe-over-TSN

Next Steps: MLE Investigating Alternatives to TCP for Storage:

NVMe-over-Homa

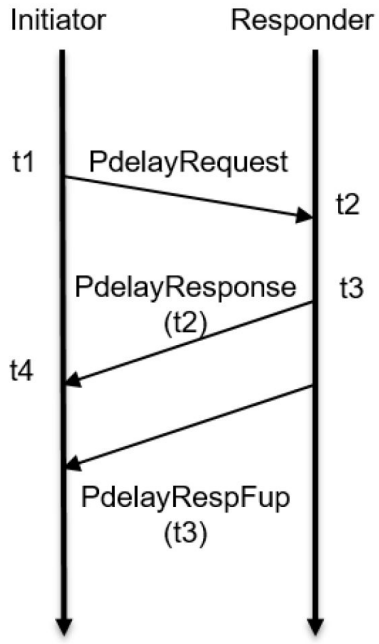
Please take a moment to rate this session.

Your feedback is important to us.

TSN Standards (1)

Time synchronization - IEEE 802.1AS (AS-2020)

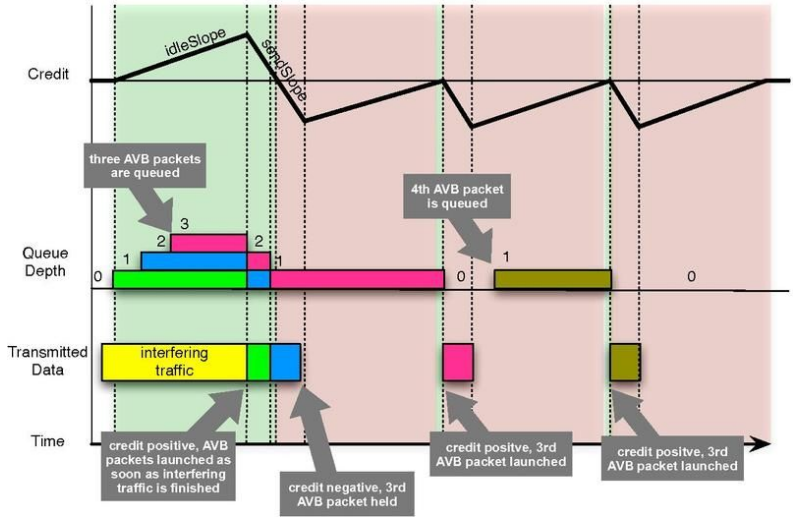
- Network wide operation
- Non 802.1as capable devices break up network
- Periodic announce messages
- Grand Master (GM) is selected for device with the best master clock algorithm (BMCA)
- Periodical Sync + Followup frames
- delay measurement is a two-step peer-to-peer path delay algorithm



TSN Standards (2)

IEEE 802.1Qav - Credit Based Shaper

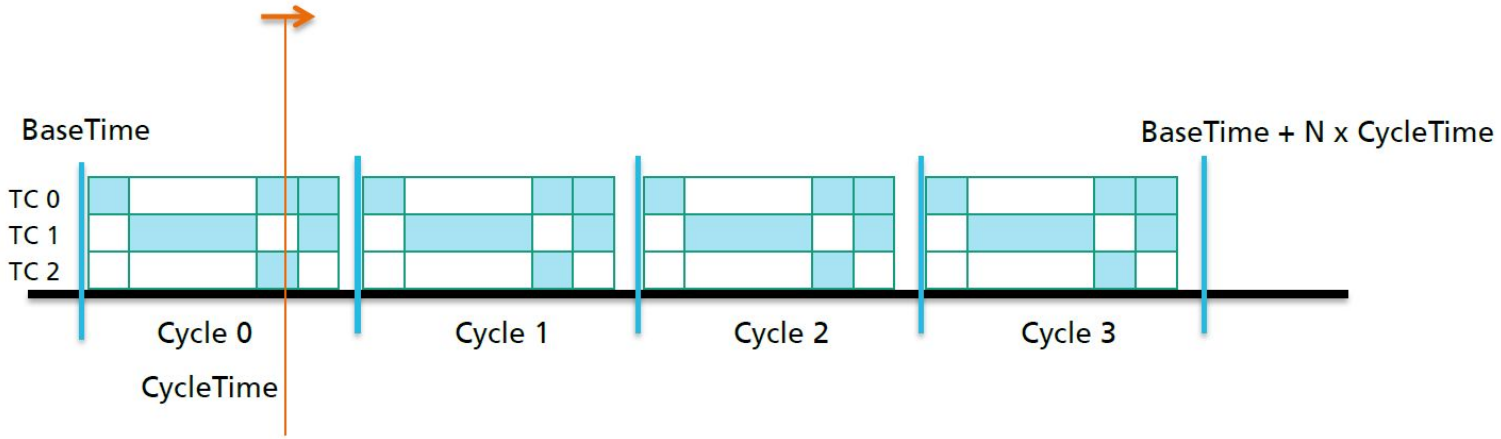
- Forwarding and Queuing Enhancements for Time-Sensitive Streams
- Allready used in AVB
- Credit based scheduling
- Positive credit allowing traffic to be sent
- Negative credit will prevent packets to be send



TSN Standards (3)

IEEE 802.1Qbv - Time Aware Shaper

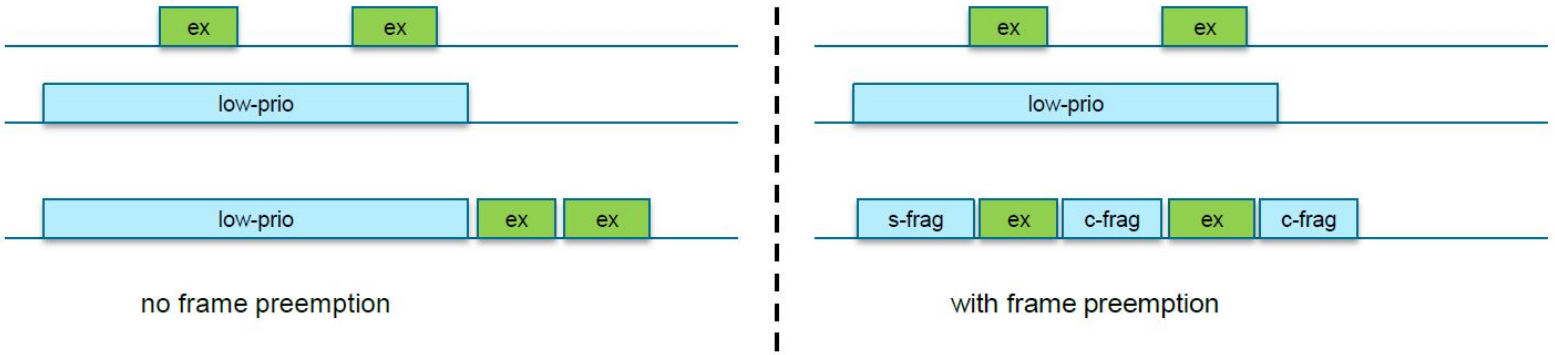
- Cycle based scheduling of frames
- Cycle length
- A number of gate operations
- Guard bands prevent violation of cycle timings



TSN Standards (4)

802.3br & 802.1Qbu – Frame preemption

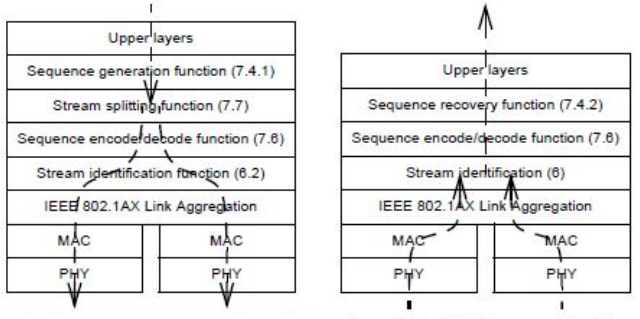
- Extreme low latency for chosen traffic (express traffic)
- Special mPackets (express packet, preemptable packet, fragment of a packet)
- 64bytes of minimal fragment size



TSN Standards (5)

IEEE 802.1CB - Frame replication and elimination for redundancy

- Sequence generation
- Split/Recovery
- Redundancy tag seq encode/decode
- Stream identification
- Link aggregation (802.1AX)



TSN Standards (6)

IEEE 802.1Qci – Per-Stream Filtering and Policing (PSFP)

- Filtering and policing and frame queue decisions made on a per-stream basis for received frames
- Stream gate id □ open/closed