TCP/IP for Real-Time Embedded Systems

The Good, The Bad, The Ugly

MLE Mission: "From Software to Silicon!"

High-Performance (Embedded) Compute & Connected Systems-of-Systems need "Offload Engines" for better performance, lower and deterministic latency and improved energy efficiency.

Focus on standards such as PCIe, NVMe, Ethernet, TCP/UDP/IP, TSN.

AMD Adaptive Computing Partner PREMIER

intel partner Gold

FPGA MICROCHIP DESIGN PARTNER

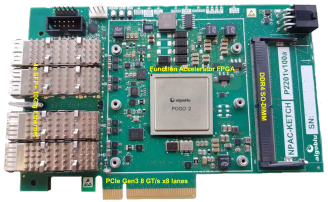
trenz electronic GmbH

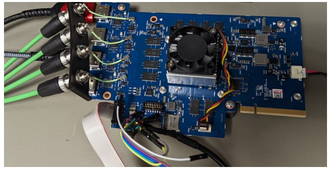
PCI SIG

Fraunhofer Heinrich-Hertz-Institut

Fraunhofer IPMS

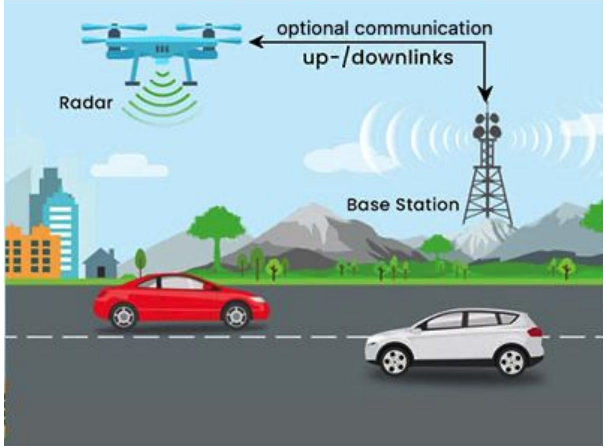
ELEMASTER



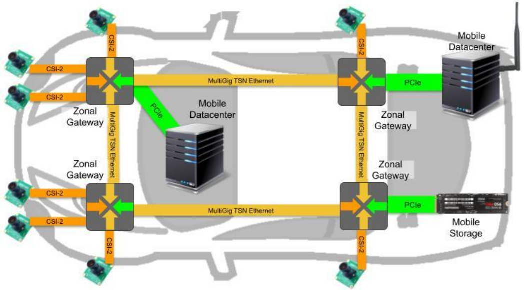


Multi-Gigabit Real-Time Networking Market & Technology Forces

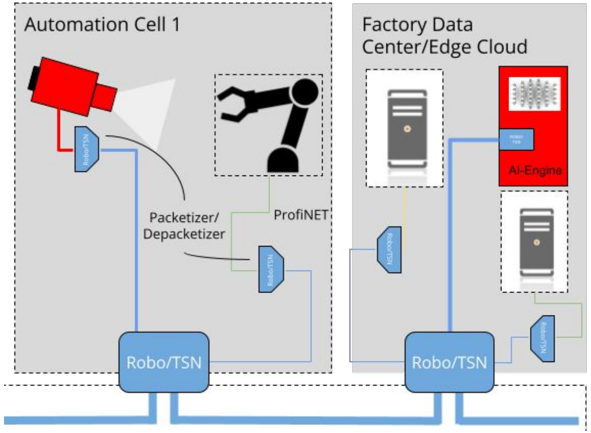
6G Radio Integrated Communication and Sensing (ICAS)



Zone-Based In-Vehicle Networking (Auto/TSN)



100G Real-time Backbone for Virtualized PLC (Robo/TSN)



Work Motivation

Systems-of-systems

- AI inference using high-data-rate sensors (Camera, Radar, Lidar)
- Tightly-coupled: i.e. distributed processing with microservices
- Loosely-connected via networks (which continuously are the bottleneck)

Need to optimize

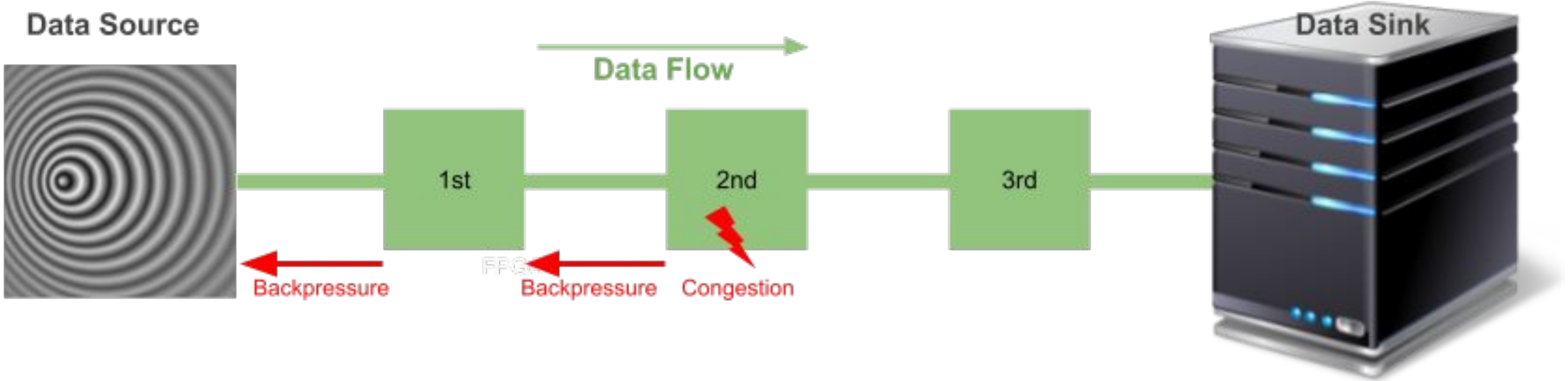
- for power / energy efficiency
- for throughput
- for (deterministic) latency and real-time delivery

Domain-Specific Architectures:

- “Offload” (protocol) processing but yet adhere to (defacto) standards and APIs
- Make networks more deterministic and Time-Sensitive

TCP - The Good

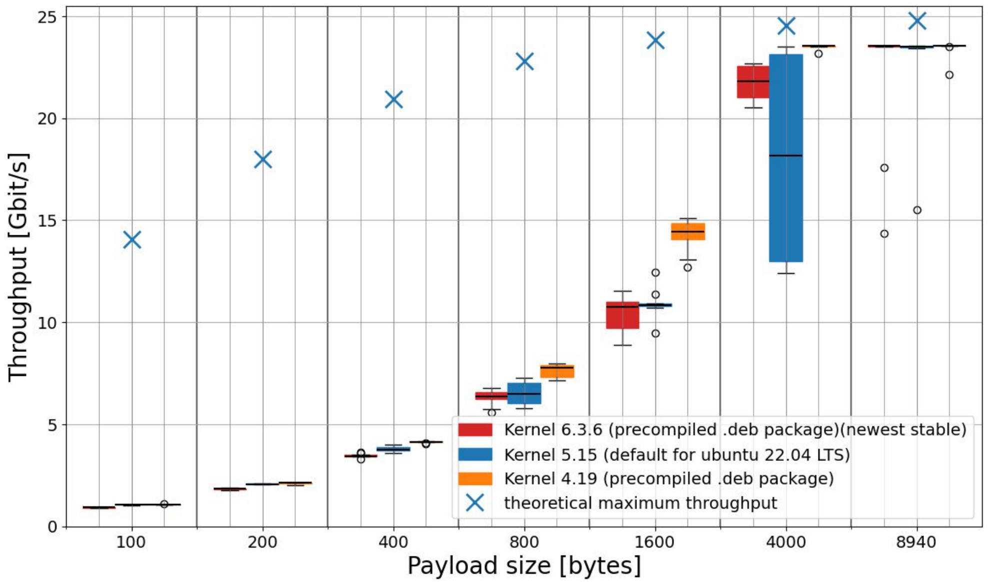
- Very well known protocol, mature (50 year anniversary)
- Reliable, scalable, packet transport
- Ubiquitous use in networking: LAN, WAN, WiFi, 3GPP mobile
- Stream-based with backpressure
  ⇒ Allows to implement self-synchronized dataflow processing systems

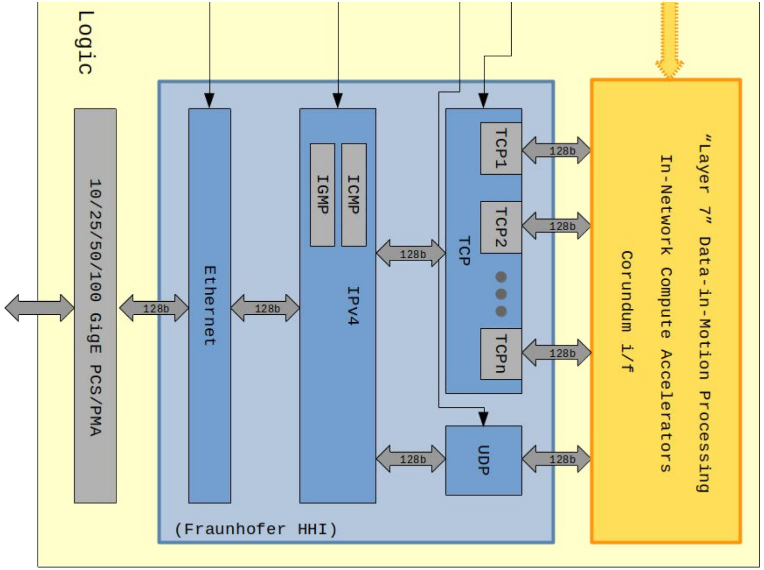


TCP - The Bad

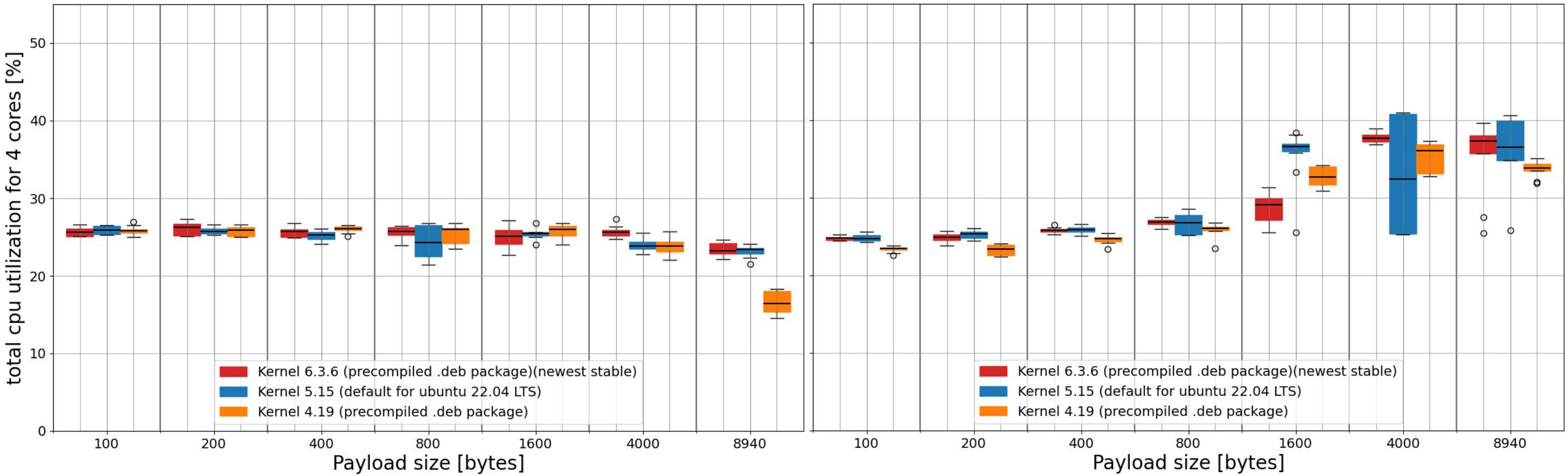
- Large compute burden for TCP in SW
- Different behavior across different implementations

- However, TOEs (TCP Offload Engines) and TCP Full Accelerators exist





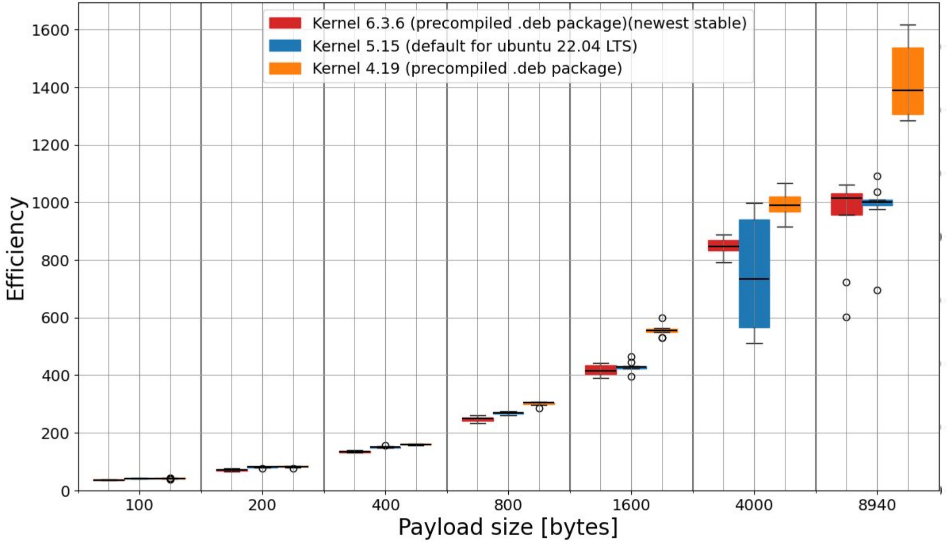
Netperf TCP_STREAM Results - CPU Load



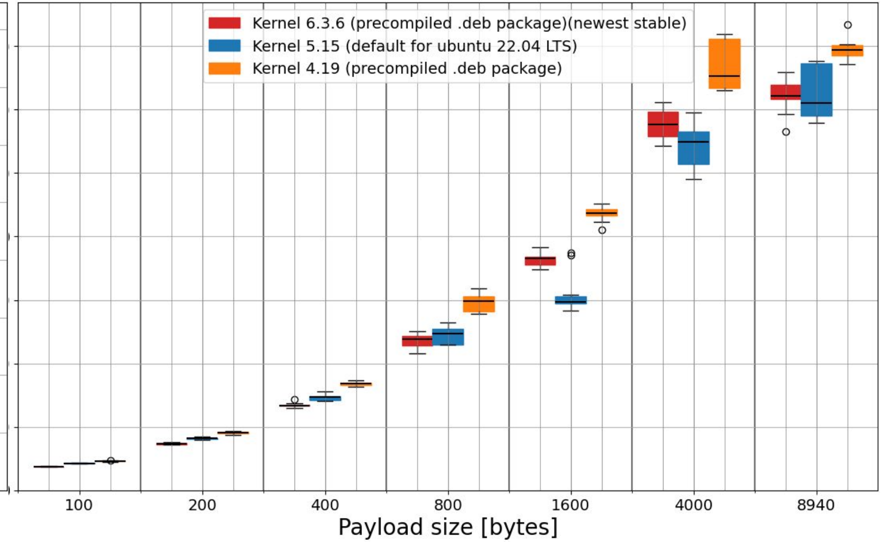
Tx Side

Rx Side

Netperf TCP_STREAM Results: Efficiency



Tx Side

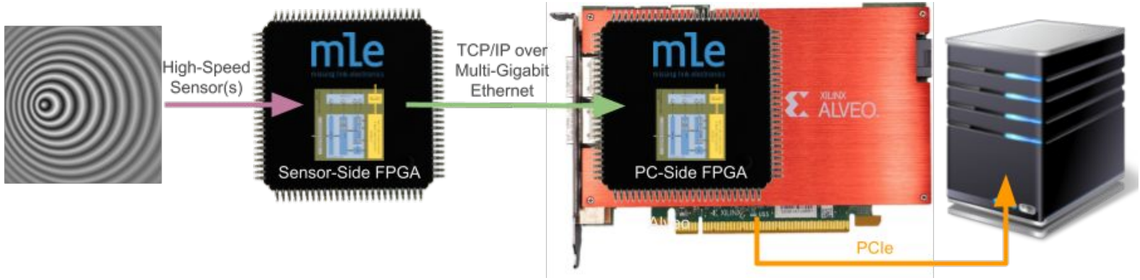


Rx Side

Benefits of PC-Side TCP Full Accelerators

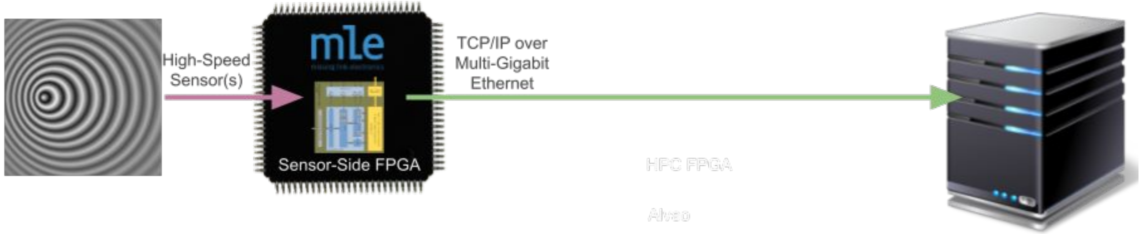
With PC-side TCP Full Accelerators:

- All PC CPU cores can fully be used for sensor data processing
- Round-trip time (RTT) is minimal, so only small Tx buffers are needed in the sensor-side FPGA
- All data streams between sensor and server can be prioritized using QoS, scheduling, traffic shaping etc
- Optional in-network processing with IEEE 1588-2019 HA PTP



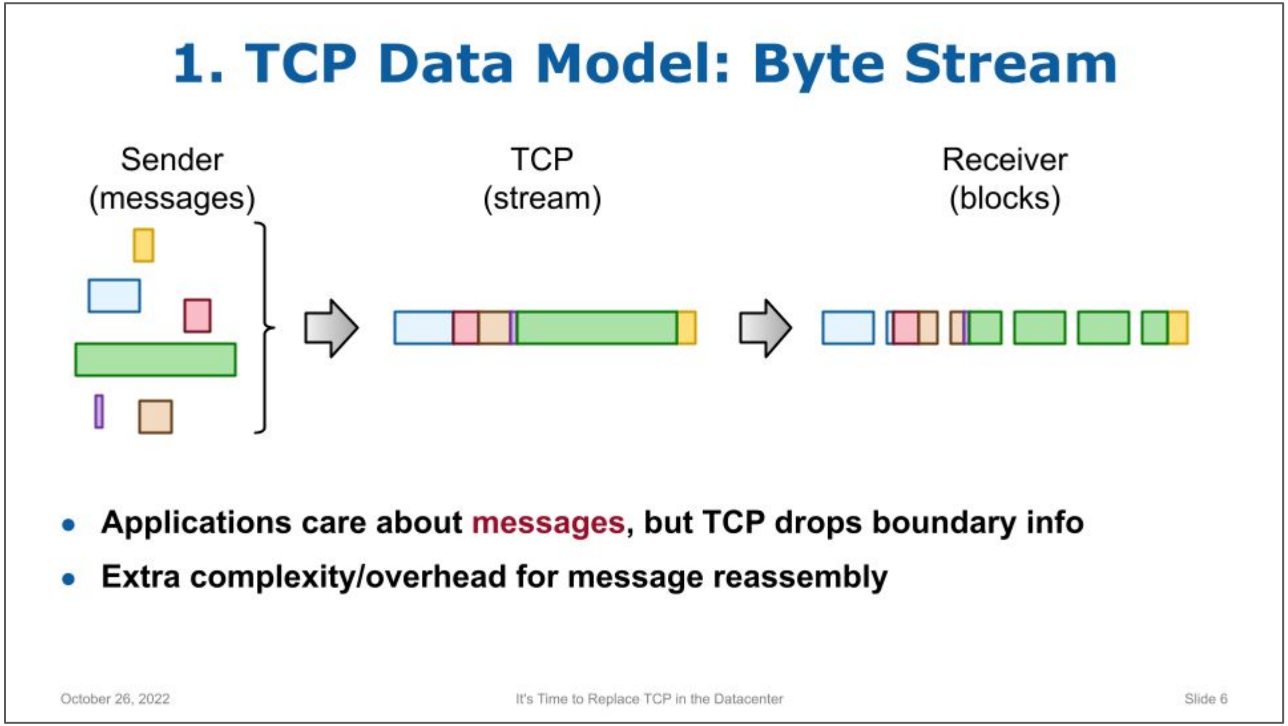
Unaccelerated PC:

- Precious CPU cores are consumed for network protocol processing
- Round-trip time (RTT) is high which eats up BRAM resources in the sensor-side FPGA
- TCP scheduling and congestion control restricted to Operating System capabilities



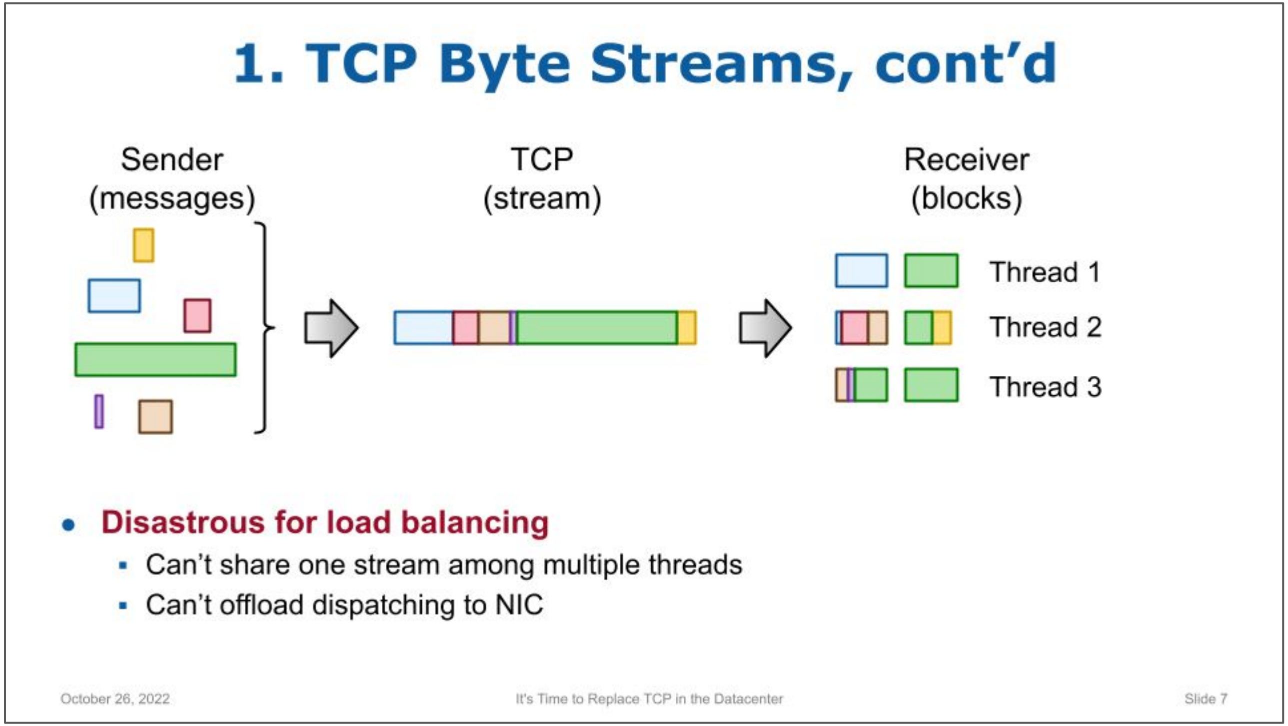
TCP - The Ugly

Courtesy of John Ousterhout, Stanford University



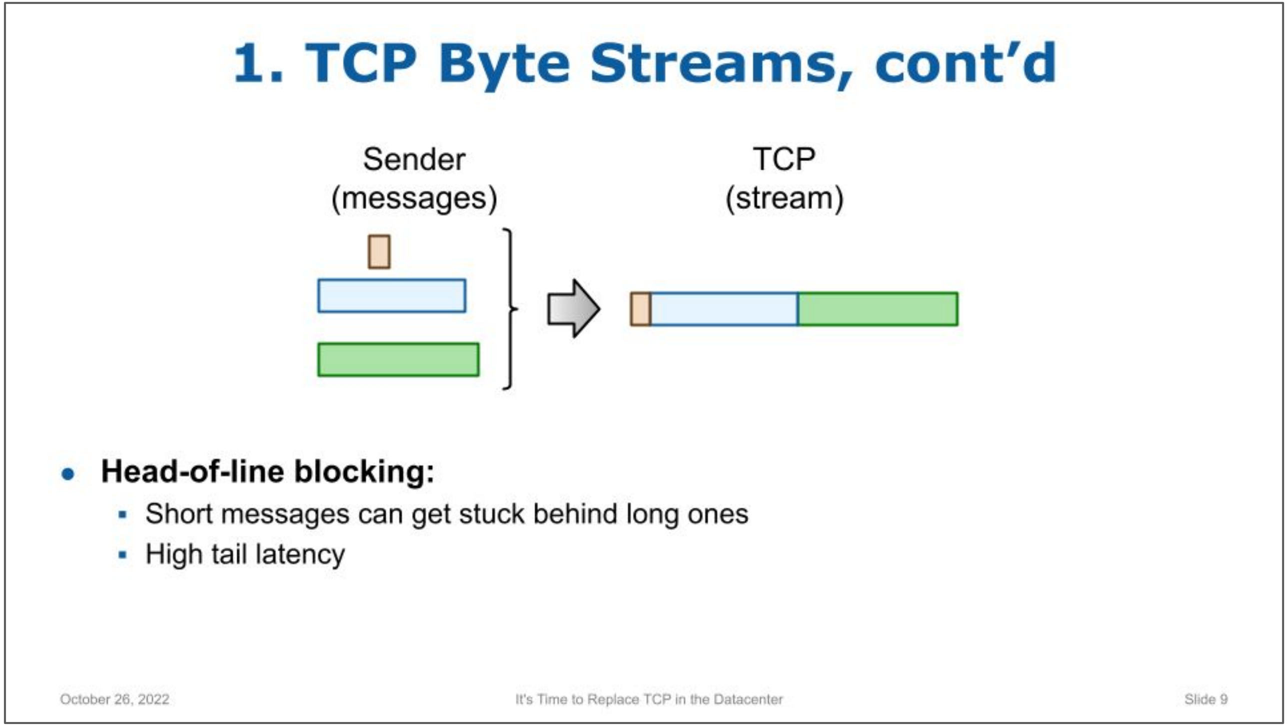
TCP - The Ugly

Courtesy of John Ousterhout, Stanford University



TCP - The Ugly

Courtesy of John Ousterhout, Stanford University



TCP - The Ugly

Courtesy of John Ousterhout, Stanford University

2. TCP is Connection-Oriented

- **Requires long-lived state for each stream**
  - ~2000 bytes per connection in Linux, not including packet buffers
  - Individual datacenter apps can have thousands of connections
  - Mitigate with connection pooling/proxies (e.g. Facebook)? Adds overhead
  - Challenging for NIC offloading (e.g. Infiniband): thrashing in connection caches
- **Before sending any data, must pay round-trip for connection setup**
  - Problematic in serverless environments: can't amortize setup cost
- **Motivation for connections:**
  - Enable reliable delivery, flow control, congestion control
  - But, all these can be achieved without connections

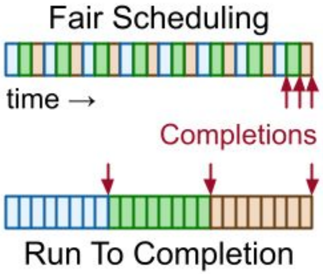
October 26, 2022 It's Time to Replace TCP in the Datacenter Slide 11

TCP - The Ugly

Courtesy of John Ousterhout, Stanford University

3. TCP Uses Fair Scheduling

- **When loaded, share bandwidth equally among active connections**
- **Well-known to perform poorly: everyone finishes slowly**
- **Run-to-completion approaches (e.g. SRPT) are better**
  - But requires message sizes



TCP - The Ugly

Courtesy of John Ousterhout, Stanford University

4. TCP: Sender-Driven Congestion Control

- **Senders responsible for scaling back transmission rates when needed**
  - But, they have no direct knowledge of congestion
- **Congestion signals based on buffer occupancy:**
  - Packets dropped if queues overflow
  - Congestion notifications based on queue length
- **Problems:**
  - Significant buffer occupancy when system is loaded
  - Queuing causes delays, especially for short messages

October 26, 2022 It's Time to Replace TCP in the Datacenter Slide 14

Alternatives for Systems-of-Systems: Homa

Courtesy of John Ousterhout, Stanford University

1. Homa is Message-Based

- **Dispatchable units are explicit in the protocol**
- **Enables efficient load balancing**
  - Multiple threads can safely read from a single socket
  - Future NICs can dispatch messages directly to threads
- **Enables run-to-completion (e.g. SRPT)**

October 26, 2022 It's Time to Replace TCP in the Datacenter Slide 18

Alternatives for Systems-of-Systems: Homa

Courtesy of John Ousterhout, Stanford University

2. Homa is Connectionless

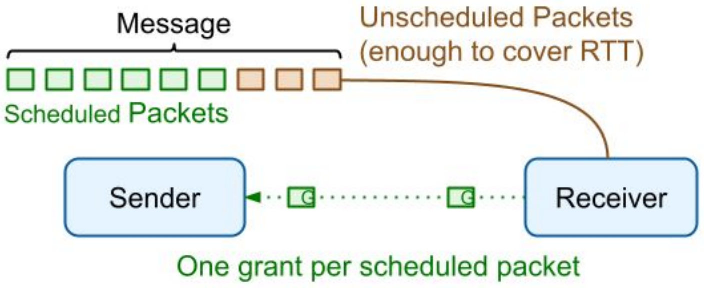
- **Fundamental unit is a remote procedure call (RPC)**
  - Request message
  - Response message
  - RPCs are independent
- **No long-lived connection state**
  - (But there is long-lived per-peer state: ~200 bytes)
- **No connection setup overhead**
  - Use one socket to communicate with many peers
- **Homa ensures end-to-end RPC reliability**
  - No need for application-level timers

October 26, 2022 | It's Time to Replace TCP in the Datacenter | Slide 19

Alternatives for Systems-of-Systems: Homa

Courtesy of John Ousterhout, Stanford University

3. Homa: Receiver-Driven Congestion Control

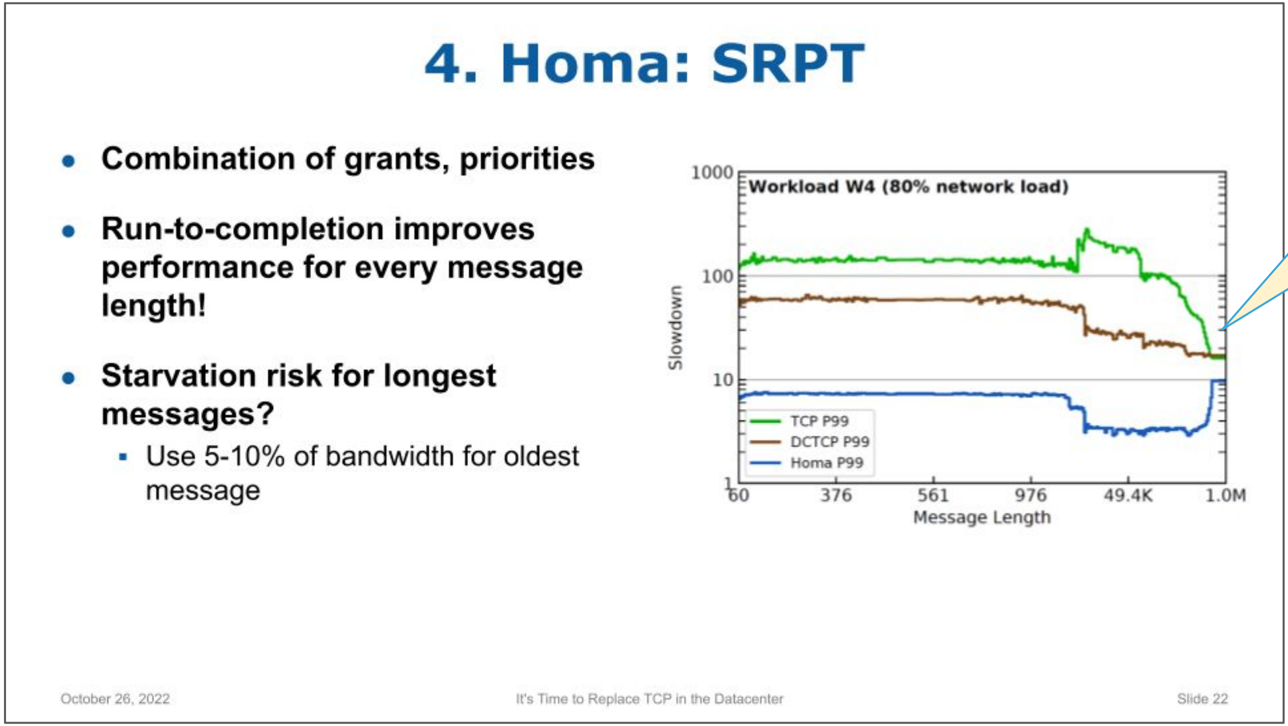


- **Receiver can delay grants to:**
  - Reduce congestion in TOR
  - Prioritize shorter messages
- **Message sizes allow receivers to predict the future:**
  - Faster, more accurate response to congestion

October 26, 2022 It's Time to Replace TCP in the Datacenter Slide 20

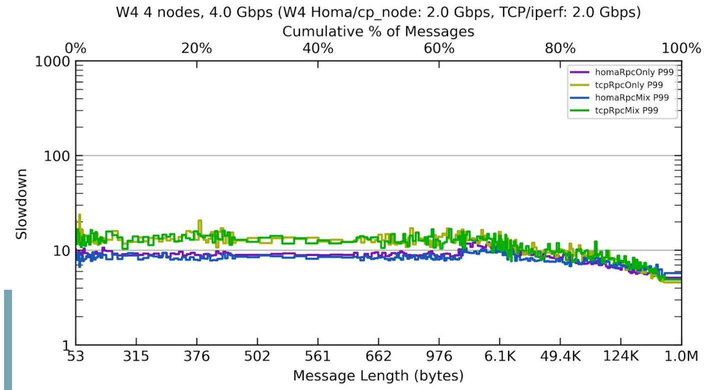
Alternatives for Systems-of-Systems: Homa

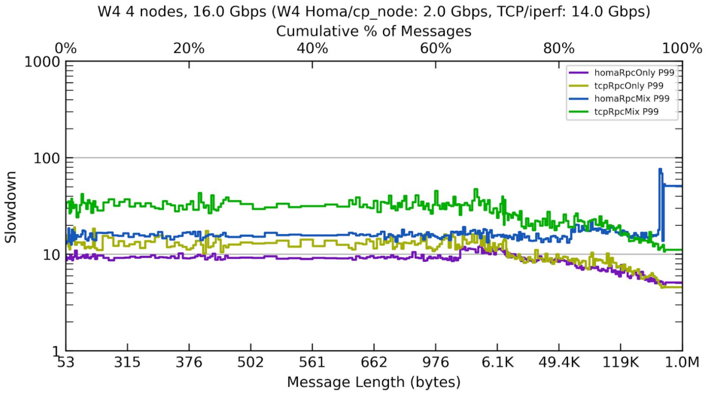
Courtesy of John Ousterhout, Stanford University

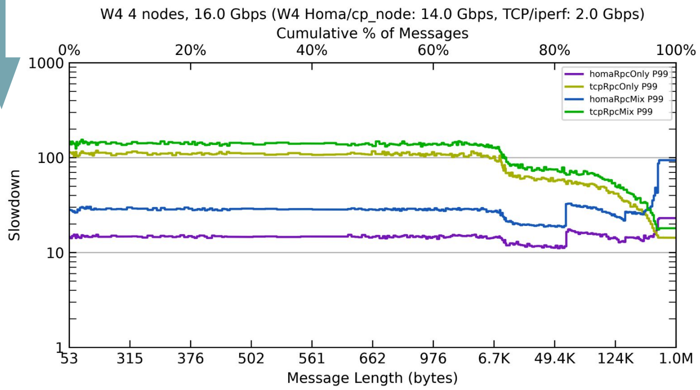


Homa has 50x .. 100x lower tail-end latency over TCP

Alternatives for Systems-of-Systems: Homa







Homa

- "peacefully co-exists with TCP

- behaves better in congested networks

MLE is implementing a Homa Accelerator Rapid, Reliable, Request-Response Protocol (Quad-R P)

Our Contact Information

Missing Link Electronics, Inc.
San Jose, CA 95134, United States

Missing Link Electronics GmbH
89231 Neu-Ulm, Germany

Email contact: sales-web@mlecorp.com