

Missing Link Electronics:

“Microsecond Connectivity Backbones for
High-Volume Data – Fast, Secure, Reliable!”



SNIA DEVELOPER CONFERENCE



BY Developers FOR Developers

www.sniadeveloper.org

Real-Time Networking with Stanford's HOMA Protocol

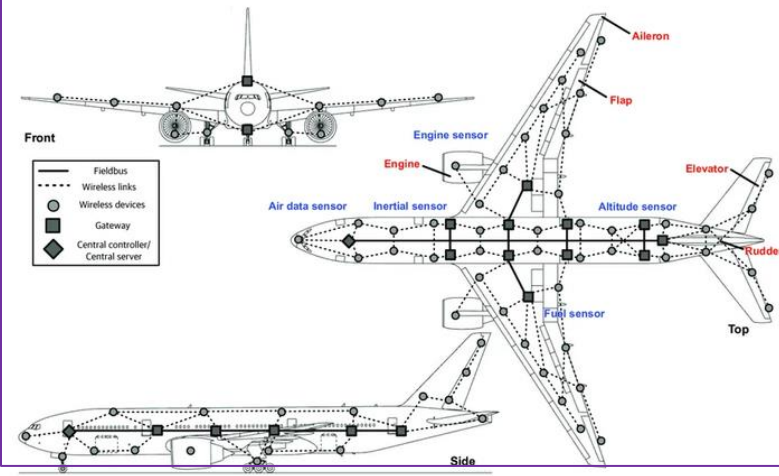
Endric Schubert, Missing Link Electronics
Ulrich Langenbach, Missing Link Electronics



Networking in Distributed Systems

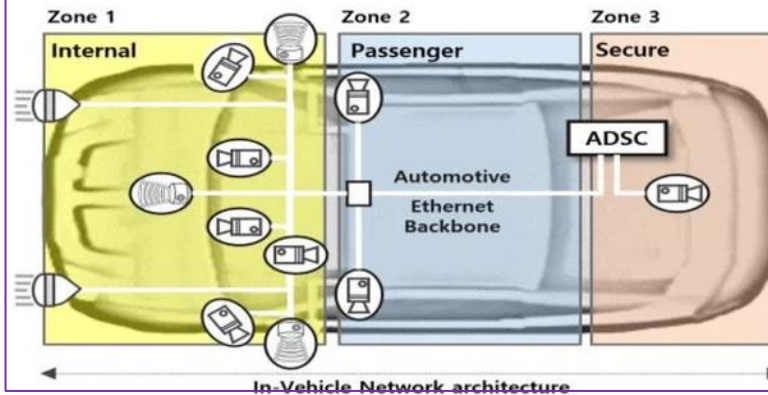
Aerospace

- Data driven (Move to 10s of Gbps)
- Low volume, High Reliability



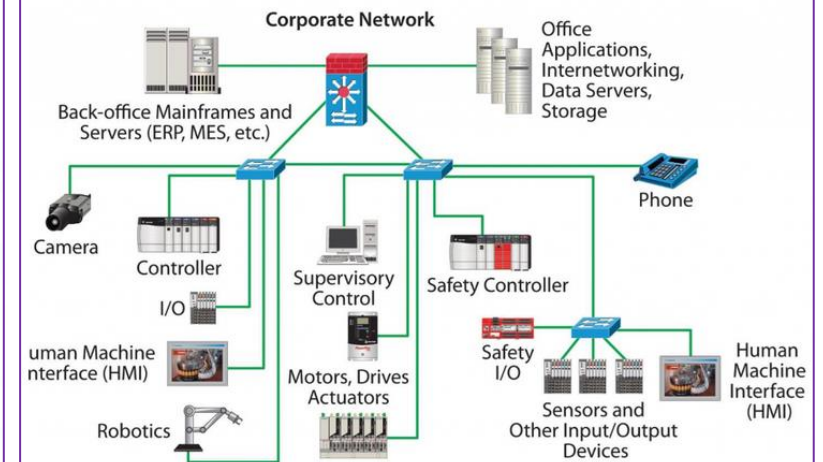
Automotive

- Cost Driven
- High Reliability
- Still Waits for AV



Industrial Automation

- Legacy Protocols
- Highly Distributed
- Op. Equip. Efficiency



Humanoids: A Datacenter on Legs Running on a Small Battery

Unique Challenges for the “Spinal Cord”:

- Physical AI requires significant compute power
- Motion Control requires low, deterministic latencies
- Must meet SWaPC

**A-Priori Knowledge:
Network is Designed to
Meet Specs!**

Sensor Plane:

- 6D Torque Sensors
- Radar / Lidar Sensors
- Camera / ToF Sensors
- Microphone Arrays
- Tactile Sensors
- MEMS / IMU
- Environment Sensors

Actuator Plane:

- Brushless DC Motors
- Harmonic Drives
- Hydraulics
- Structural Parts



Decision Plane:

- Physical AI
- RFM / SLM
- Machine Vision
- ROS2 / NoPLC
- Functional Safety
- CPU / GPU / NPU
- FPGA / SoC-FPGA

Power Plane:

- Batteries and BCM
- Charging
- Cooling
- Drivers

Not Again!!!!

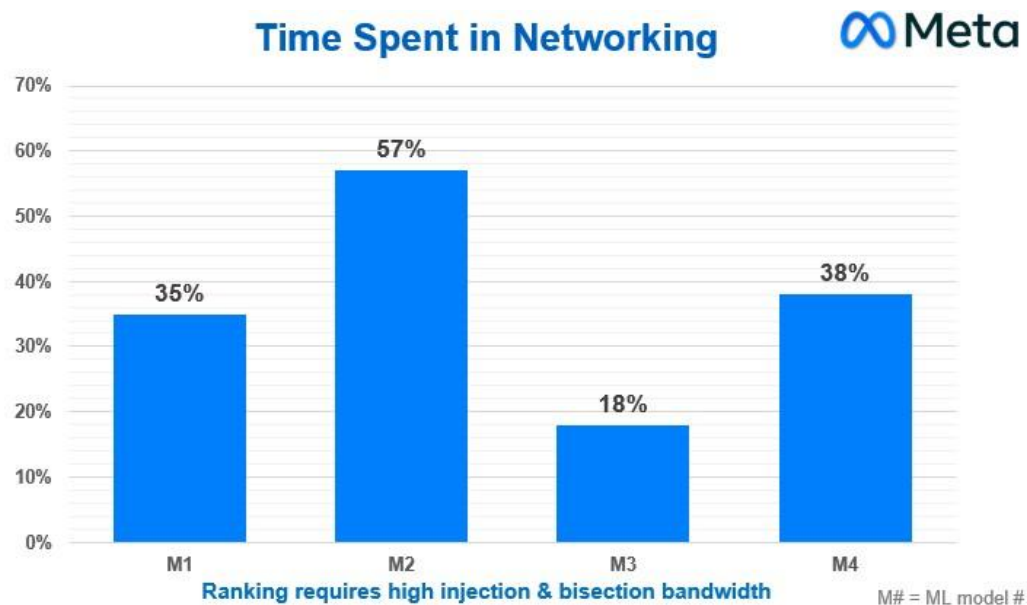


Effects of Latency

Datacenter:



- Utilization of expensive GPUs



Humanoids:



- Regulation on Machinery

29.6.2023	EN	Official Journal of the European Union	L 165/1
REGULATION (EU) 2023/1230 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL			
of 14 June 2023			
on machinery and repealing Directive 2006/42/EC of the European Parliament and of the Council and Council Directive 73/361/EEC			
(Text with EEA relevance)			
THE EUROPEAN PARLIAMENT AND THE COUNCIL OF THE EUROPEAN UNION,			
Having regard to the Treaty on the Functioning of the European Union, and in particular Article 114 thereof,			
Having regard to the proposal from the European Commission,			
After transmission of the draft legislative act to the national parliaments,			
Having regard to the opinion of the European Economic and Social Committee ⁽¹⁾ ,			
Acting in accordance with the ordinary legislative procedure ⁽²⁾ ,			
Whereas:			
(1) Directive 2006/42/EC of the European Parliament and of the Council ⁽³⁾ was adopted in the context of establishing the internal market, in order to harmonise health and safety requirements for machinery in all Member States and to remove obstacles to trade in machinery between Member States.			
(2) The machinery sector is an important part of the engineering industry and is one of the industrial mainstays of the Union economy. The social cost of the large number of accidents caused directly by the use of machinery can be reduced by inherently safe design and construction of machinery and by proper installation and maintenance.			

Part 1: Stanford's HOMA Protocol

Homa - Started by John Ousterhout et al.

It's Time to Replace TCP in the Datacenter

It's Time to Replace TCP in the Datacenter

John Ousterhout
Stanford University

January 18, 2023

*This position paper has been updated since its original publication in October of 2022 in order to correct errors and add clarification. Updates are in *italics*; none of the original text has been modified. The paper has triggered discussion and dissent; for pointers to comments on the paper, see the Homa Wiki: <https://homa-transport.atlassian.net/wiki/spaces/HOMA/overview#replaceTcp>.*

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 expectation 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.

1 Introduction

The TCP transport protocol [9] has proven to be phenomenally successful and adaptable. At the time of TCP's design in the late 1970's, there were only about 100 hosts attached to the existing ARPANET, and network links had speeds of tens of kilobits/second. Over the decades since then, the Internet has grown to billions of hosts and link speeds of 100 Gbit/second or more are commonplace, yet TCP continues to serve as the workhorse transport protocol for almost all applications. It is an extraordinary engineering achievement to have designed a mechanism that could survive such radical changes in underlying technology.

However, datacenter computing creates unprecedented challenges for TCP. The datacenter environment, with millions of cores in close proximity and individual applications harnessing thousands of machines that interact on microsecond timescales, could not have been envisioned by the designers of TCP, and TCP does not perform well in this environment. TCP is still the protocol of choice for most datacenter applications, but it introduces overheads on many levels, which limit application-level performance. For example, it is well-known that TCP suffers from high tail latency for short messages under mixed workloads [2]. TCP is a major contributor to the "datacenter tax" [3, 12], a collection of low-level overheads that consume a significant fraction of all processor cycles in datacenters.

This position paper argues that TCP's challenges in the datacenter are insurmountable. Section 3 discusses each of the major design decisions in TCP and demonstrates that every one of them is wrong for the datacenter, with significant negative consequences. Some of these problems have been discussed in the past, but it is instructive to see them all together in one place. TCP's problems impact systems at multiple levels, including the network, kernel software, and applications. One example is load balancing, which is essential in datacenters in order to process high loads concurrently. 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 [19, 21] 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 congestion in the network core fabric, are eliminated entirely. Other problems, such as congestion control and load balancing, become much easier to address. Homa demonstrates that it is possible to solve all of TCP's problems.

Complete replacement of TCP is unlikely anytime soon, due to its deeply entrenched status, but TCP can be displaced for many applications by integrating Homa into a small number of existing RPC frameworks such as gRPC [6]. With this approach, Homa's incompatible API will be visible only to framework developers and applications should be able to switch to Homa relatively easily.

2 Requirements

Before discussing the problems with TCP, let us first review the challenges that must be addressed by any transport protocol for datacenters.

Reliable delivery. The protocol must deliver data reliably from one host to another, in spite of transient failures in the

A Linux Kernel Implementation of the Homa Transport Protocol

A Linux Kernel Implementation of the Homa Transport Protocol

John Ousterhout, *Stanford University*

<https://www.usenix.org/conference/atc21/presentation/ousterhout>

This paper is included in the Proceedings of the
2021 USENIX Annual Technical Conference.

July 14–16, 2021

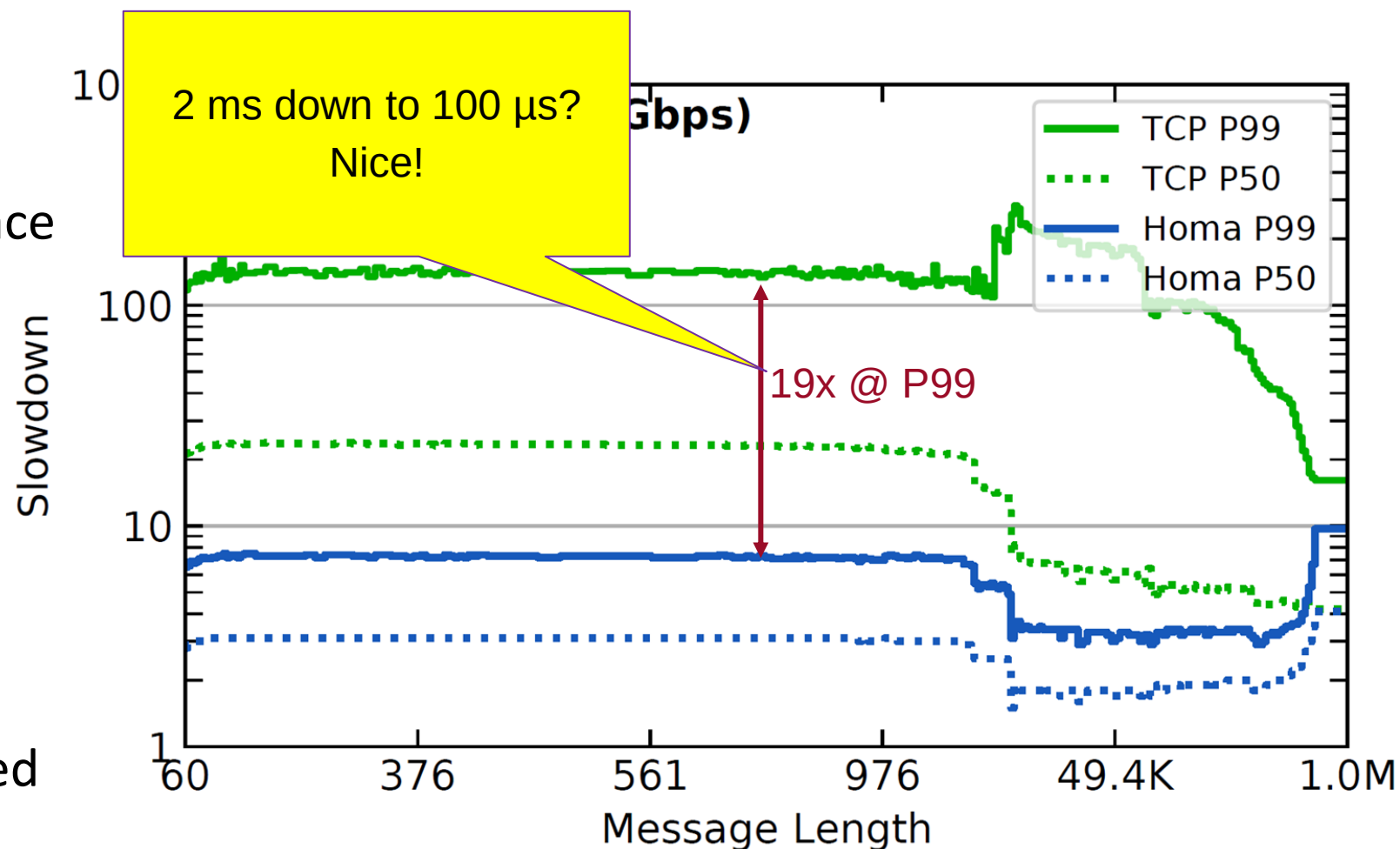
978-1-939133-23-6

Open access to the Proceedings of the
2021 USENIX Annual Technical Conference
is sponsored by USENIX.

arXiv:2210.00714v2 [cs.NI] 19 Jan 2023

Homa Reduces Tail Latencies in Loaded Networks

- Experimental results
- 25 GigE Network
- Compares Linux kernel space implementation of
 - TCP/IPv4
 - Homa/IPv4
- X-axis is distribution of message lengths in workload
- Y-axis is Slowdown $\text{RTT}_{\text{loaded}} / \text{RTT}_{\text{unloaded}}$



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)**

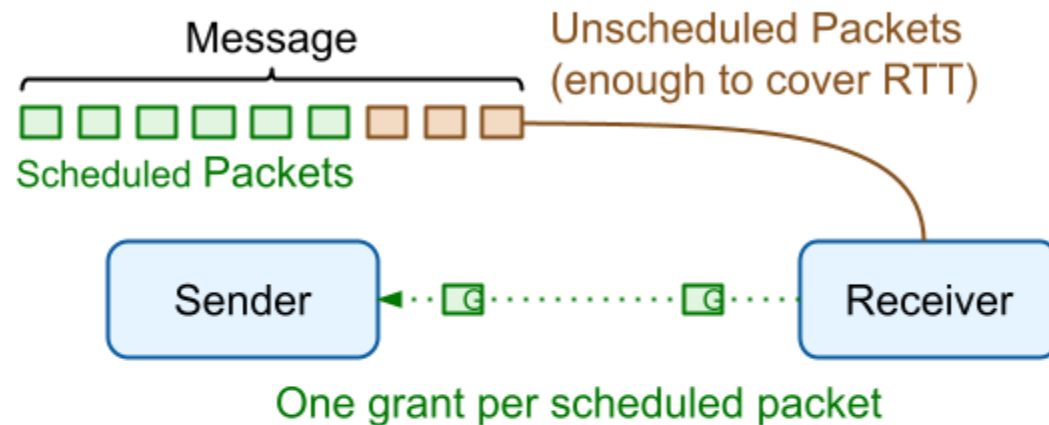
Courtesy of John Ousterhout, Stanford University

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

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

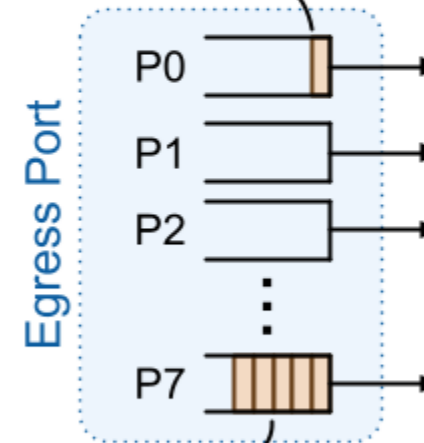
Courtesy of John Ousterhout, Stanford University

Homa Uses Priority Queues

- **Modern switches: 8–16 priority queues per egress port**
- **Homa receivers select priorities for SRPT:**
 - Favor shorter messages
- **Achieve both high throughput and low latency**
 - Need buffering to maintain throughput (e.g. if sender doesn't respond to grant)
 - But buffers can result in delays
 - Solution: **overcommitment**:
 - Grant to multiple messages
 - Different priority for each message

Overcommitment

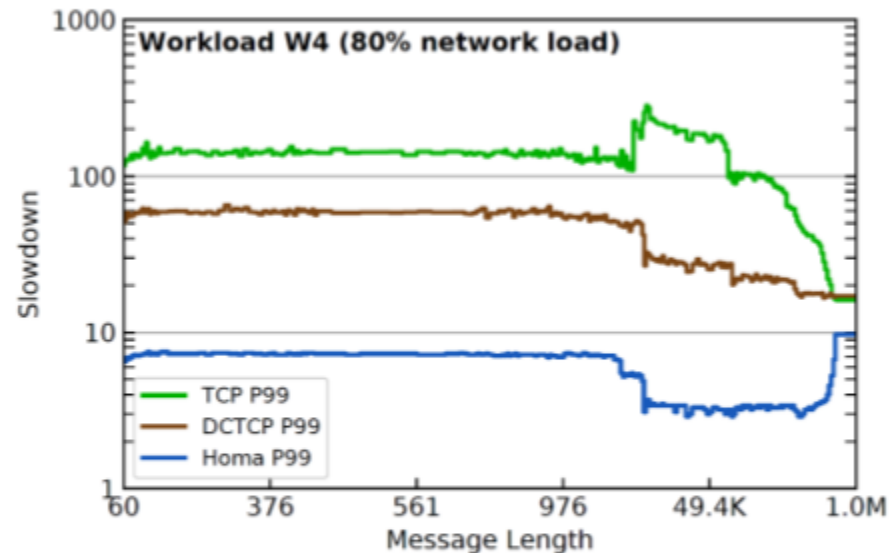
Short messages use high priority queues (low latency)



Buffers accumulate in low-priority queues (ensure throughput)

4. Homa: SRPT

- Combination of grants, priorities
- Run-to-completion improves performance for every message length!
- Starvation risk for longest messages?
 - Use 5-10% of bandwidth for oldest message



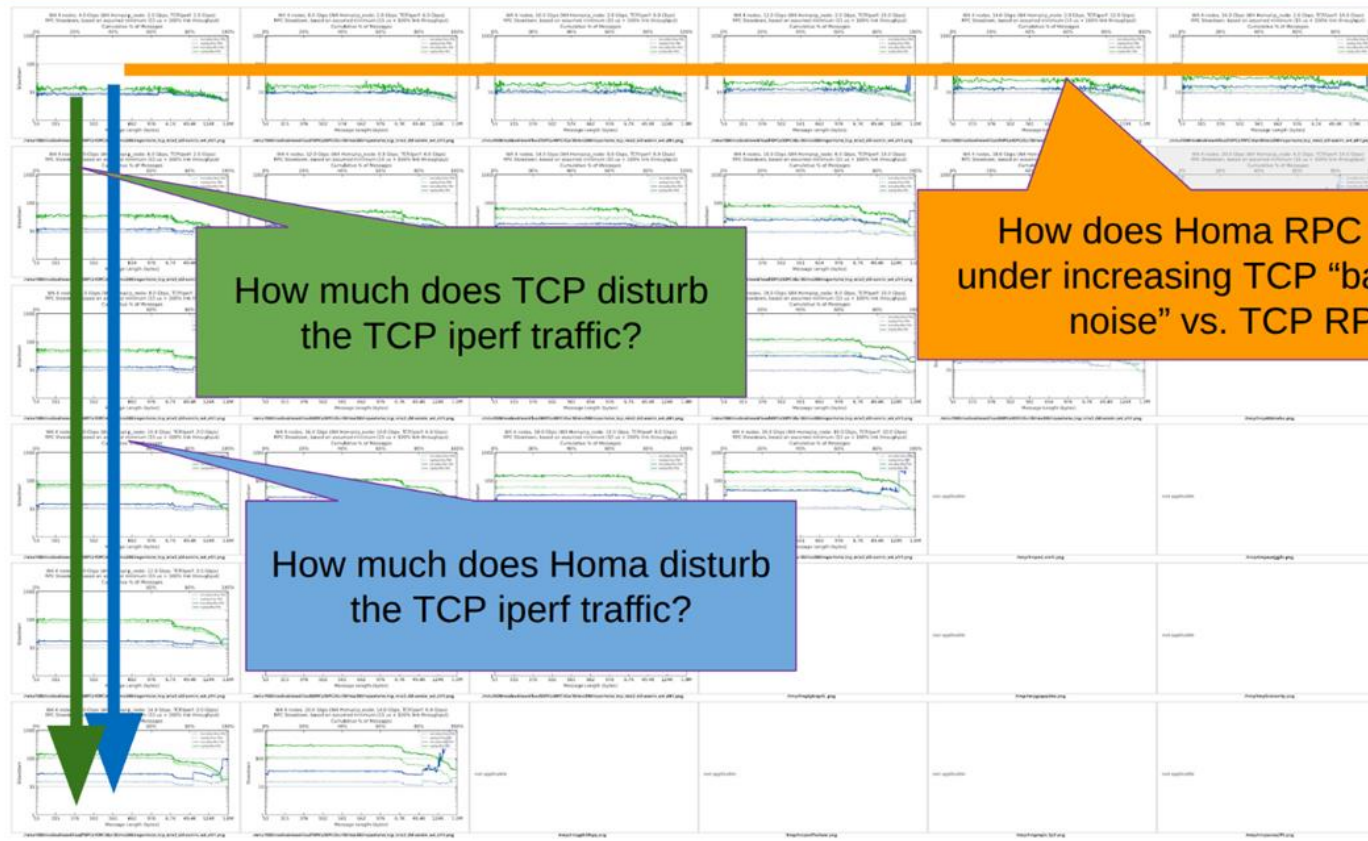
Courtesy of John Ousterhout, Stanford University

5. Homa: No Order Requirement

- **Can use packet spraying in datacenter networks**
 - Hypothesis: will eliminate core congestion (unless core fabric systemically overloaded)
- **Better load balancing across CPU cores**

SDC24: Can HOMA Co-exist With TCP? YES!

Slowdown Results Workload W4



18 | ©2024 SNIA. All Rights Reserved.

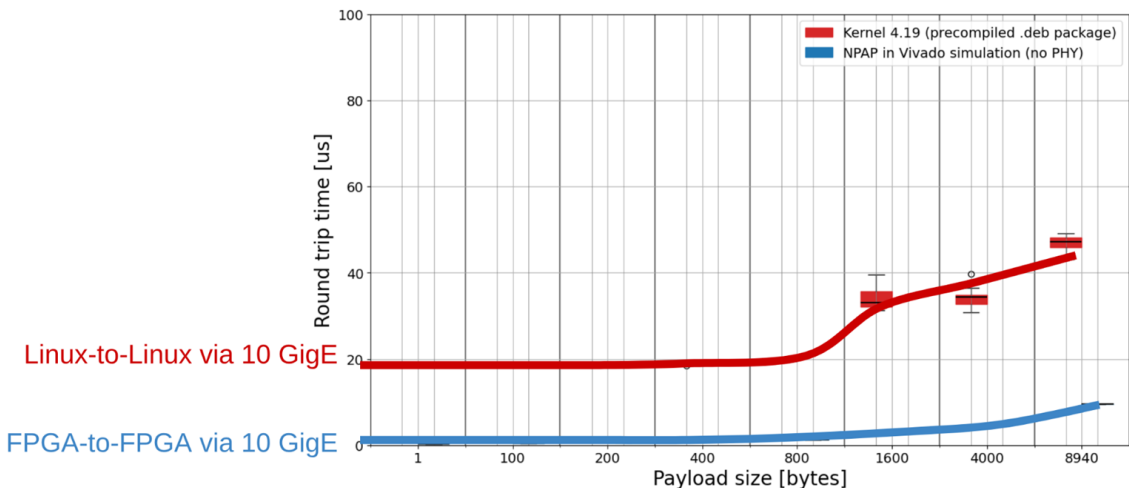


HOMA, Accelerated: RRRRP

(Rapid, Reliable Request-Response Protocol)

Option 1: Full Acceleration

MLE NPAP - Network Protocol Accelerator Platform for ASIC & FPGA



Option 2: Offload Engines

Open Source Multi-Queue
FPGA NIC corundum.io

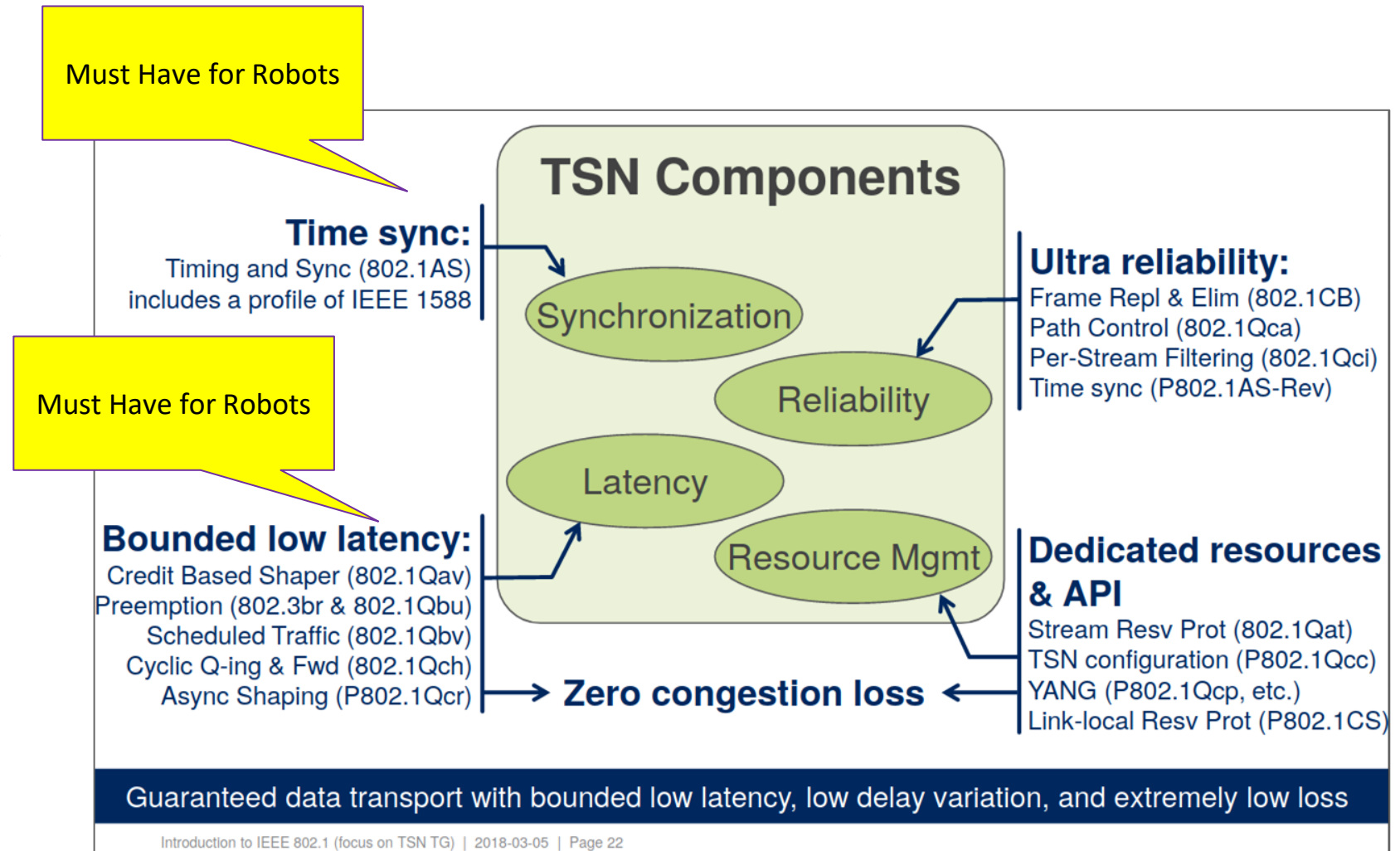
TCP (MLX 5)	HomaModule	RRRRP
Checksum Offload	N/A	N/A
Receive Side Scaling	Receive Flow Steering	Receive Side Scaling
Large Segmentation Offload	Generic Segmentation Offload	Large Segmentation Offload
Large Receive Offload	Generic Receive Offload	Large Receive Offload

Part 2:

Time Sensitive Networking (TSN)

Time-Sensitive Networking (TSN)

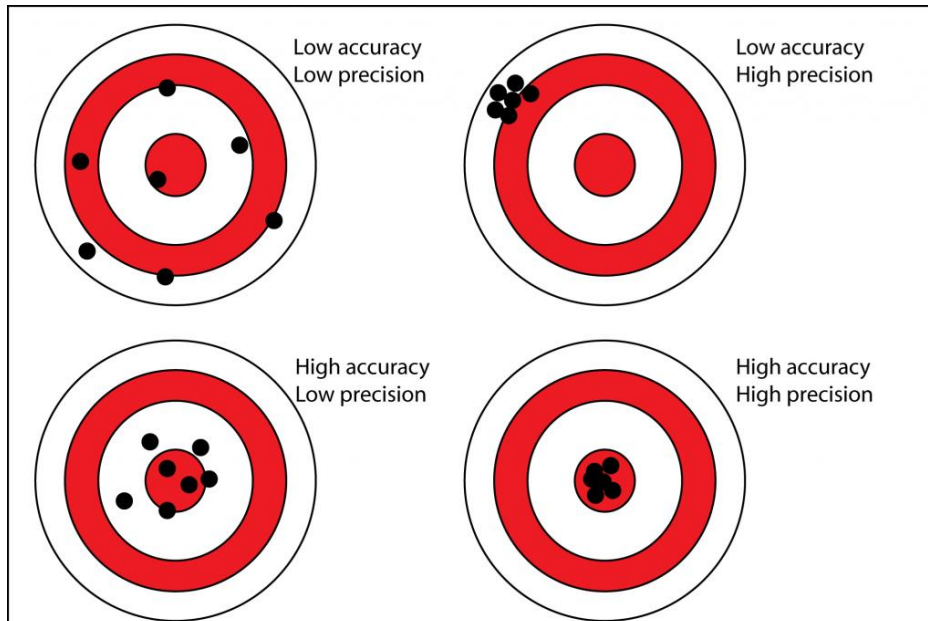
- 2002: Audio Video Broadcasting (AVB)
- 2012: Became TSN, a set of IEEE standards
- Slow Adoption:
 - Manufacturing
 - Automotive
 - Datacenter



Time Synchronization

Accuracy and Precision

- Needed for Sensor Coherency



White Rabbit from CERN

- Cost Optimized: Light Rabbit

Light Rabbit on AMD ZCU102

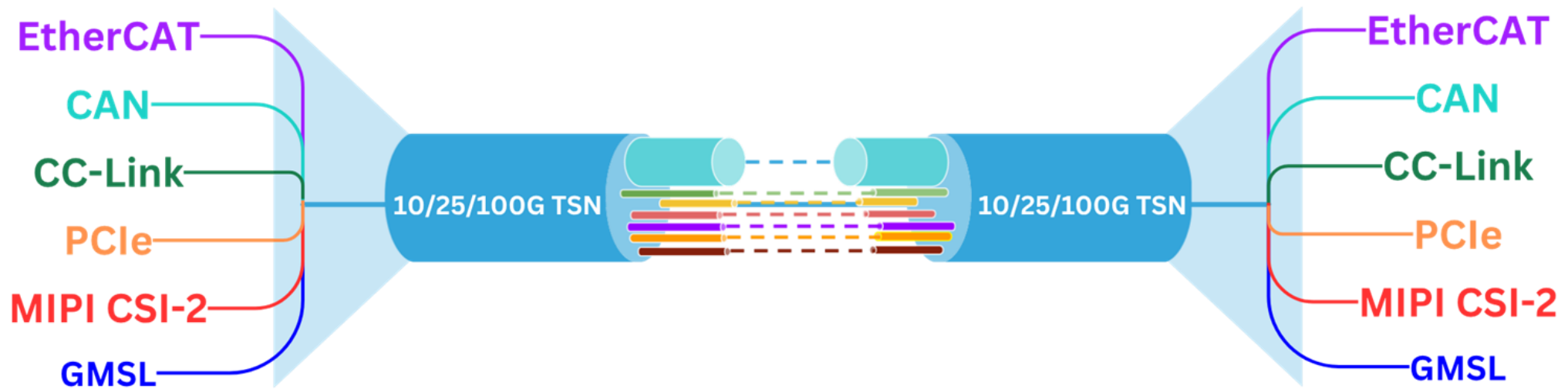


PPS stddev	< 24 ps
PPS Range	< 170 ps
PPS mean	< 1.4 ns
Measurements	7304

Uncalibrated Device

Traffic Shaping for a “Virtualized Spinal Cord”

- Implement a safe and cost-efficient means for legacy interconnect
- Separate **Motion Control Messages** from **Sensor Traffic Streams**

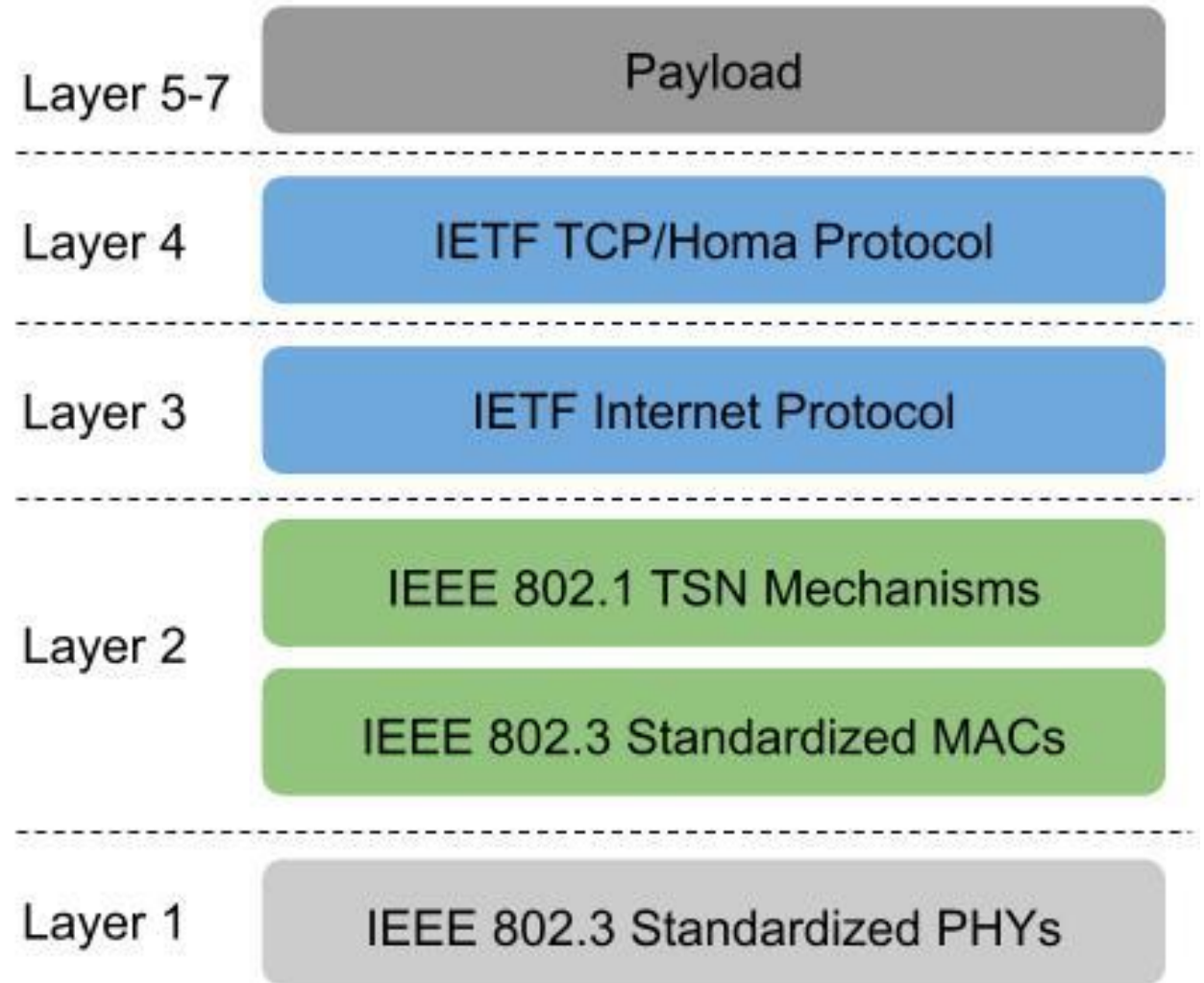


Part 3: HOMA and TSN

Running HOMA over TSN

A-Priori Knowledge: Network is Designed to Meet Specs!

- TSN Traffic Shaping is key to meet the hard real time constraints!
- But: Traffic Shaping has negative effects on TCP and on HOMA!
- HOMA as a reliable UDP with receiver-driven congestion control



BUT ...

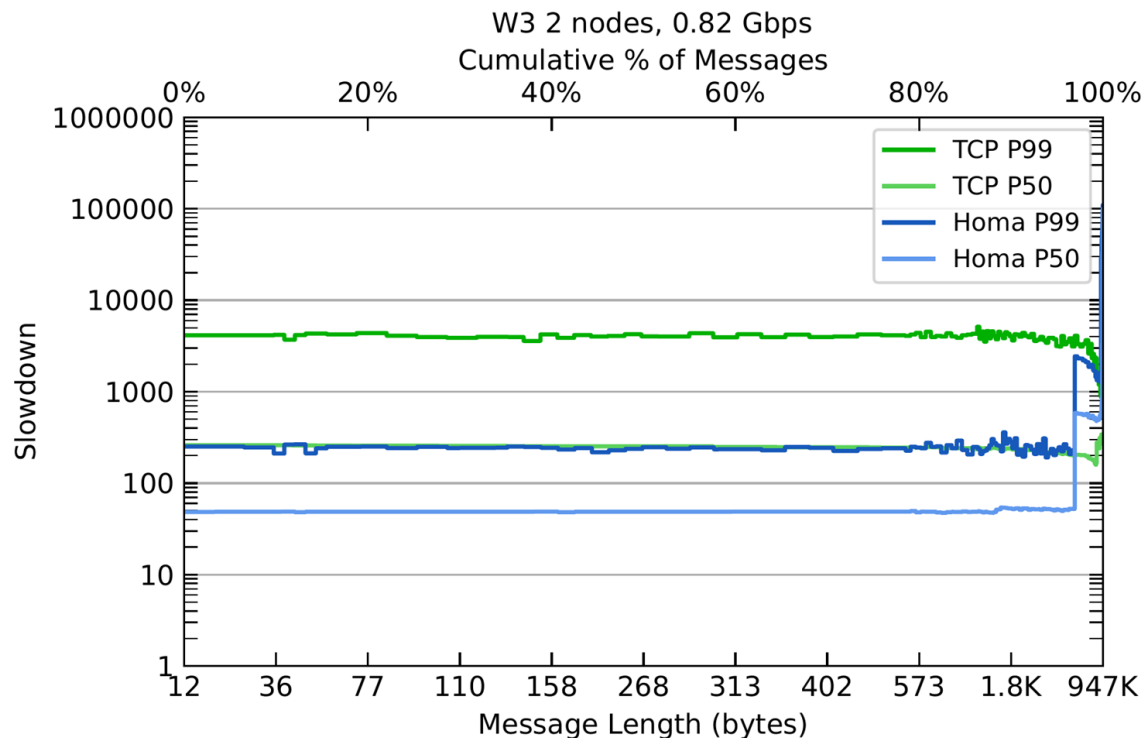
Mechanism	Control Paradigm	Latency Performance	HOMA Suitability with TSN
Strict Priority Queuing (SP) - 802.1Q	“Greedy” - Send Immediately	Low average latency, high P99 jitter	Good for Bulk (Low Priority) because of 100% bandwidth Utilization
Credit-Based Shaping (CBS) - 802.1Qav	Rate-Limiting/Token	High average delay	Does not align with HOMA’s Run-to-Completion for bulky LP
Time-Aware Shaper (TAS) - 802.1Qbv	Time-Triggered (TDMA)	Bounded latency, but with gate-miss delay	Static schedule fights dynamic priorities (SRPT)
Asynchronous Traffic Shaping (ATS) - 802.1Qcr	Event Driven	Predictable latency, flexible response to burst	Good for High Priority unscheduled homa traffic but would not align with HOMA’s Run-to-completion for bulky LP Traffic

Experimental Results with IEEE 802.1Qcr (ATS)

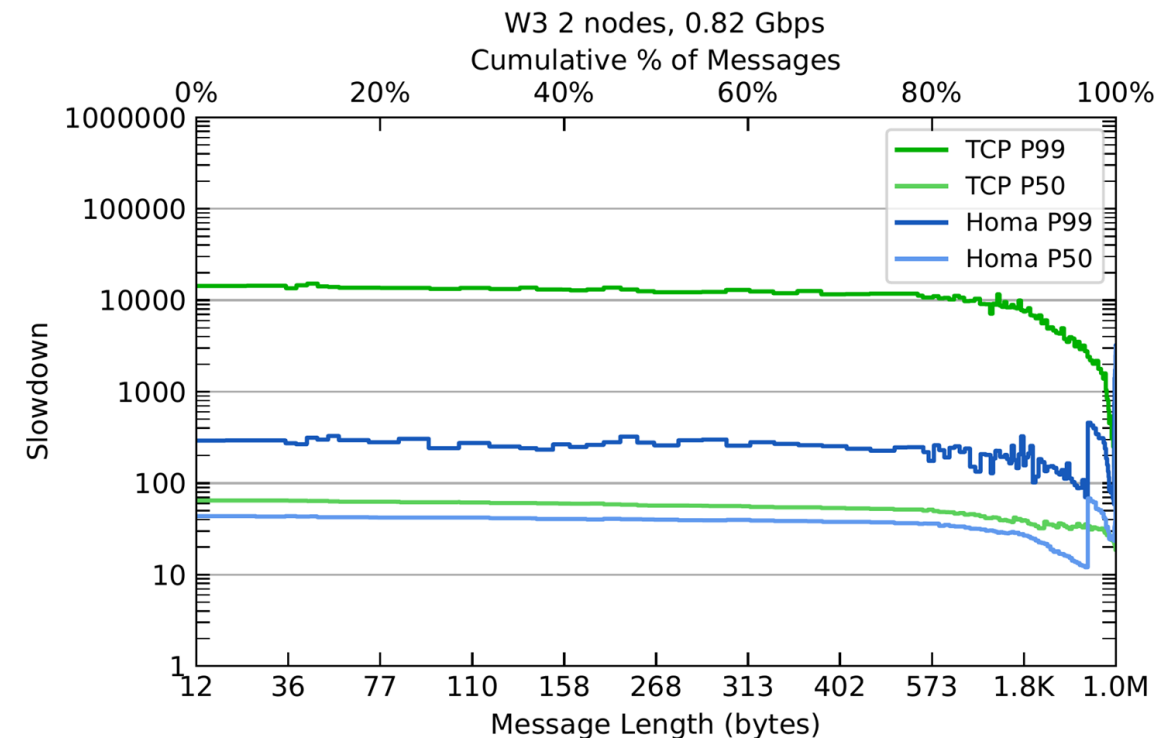
Test DD 9/17

Workload W3 over 1 GigE, CIR 60 Mbps

w/o Traffic Shaping



With Traffic Shaping

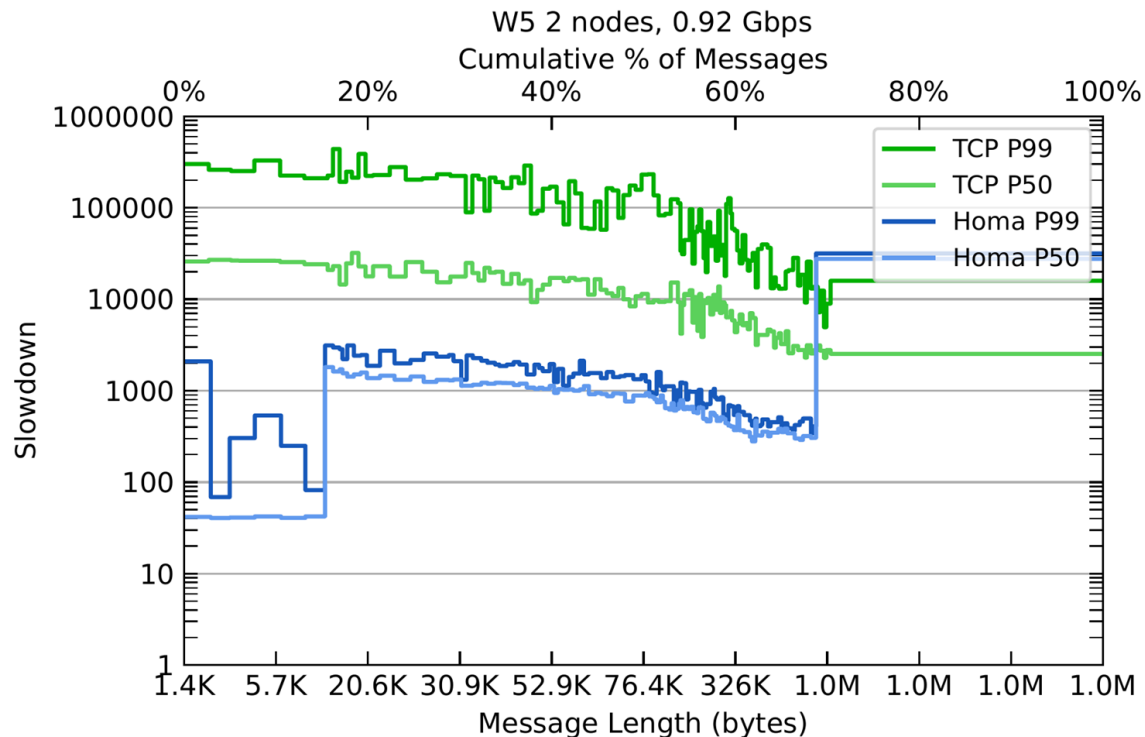


Experimental Results with IEEE 802.1Qcr (ATS)

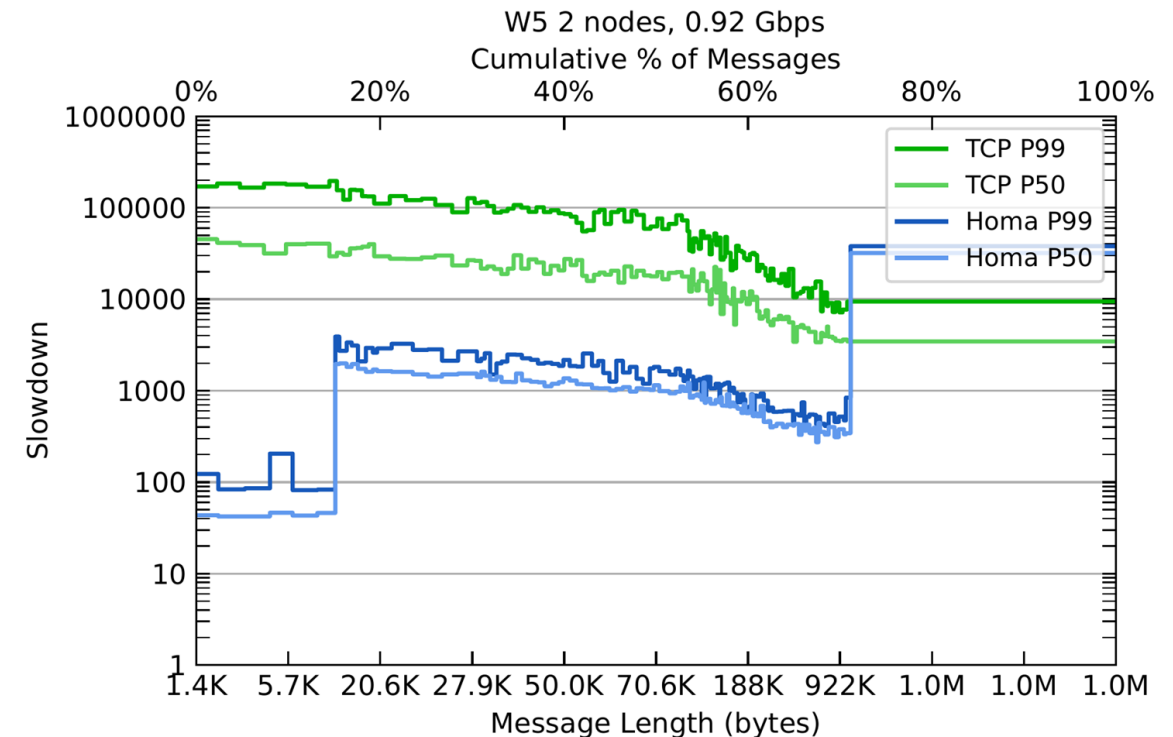
Test DD 9/16

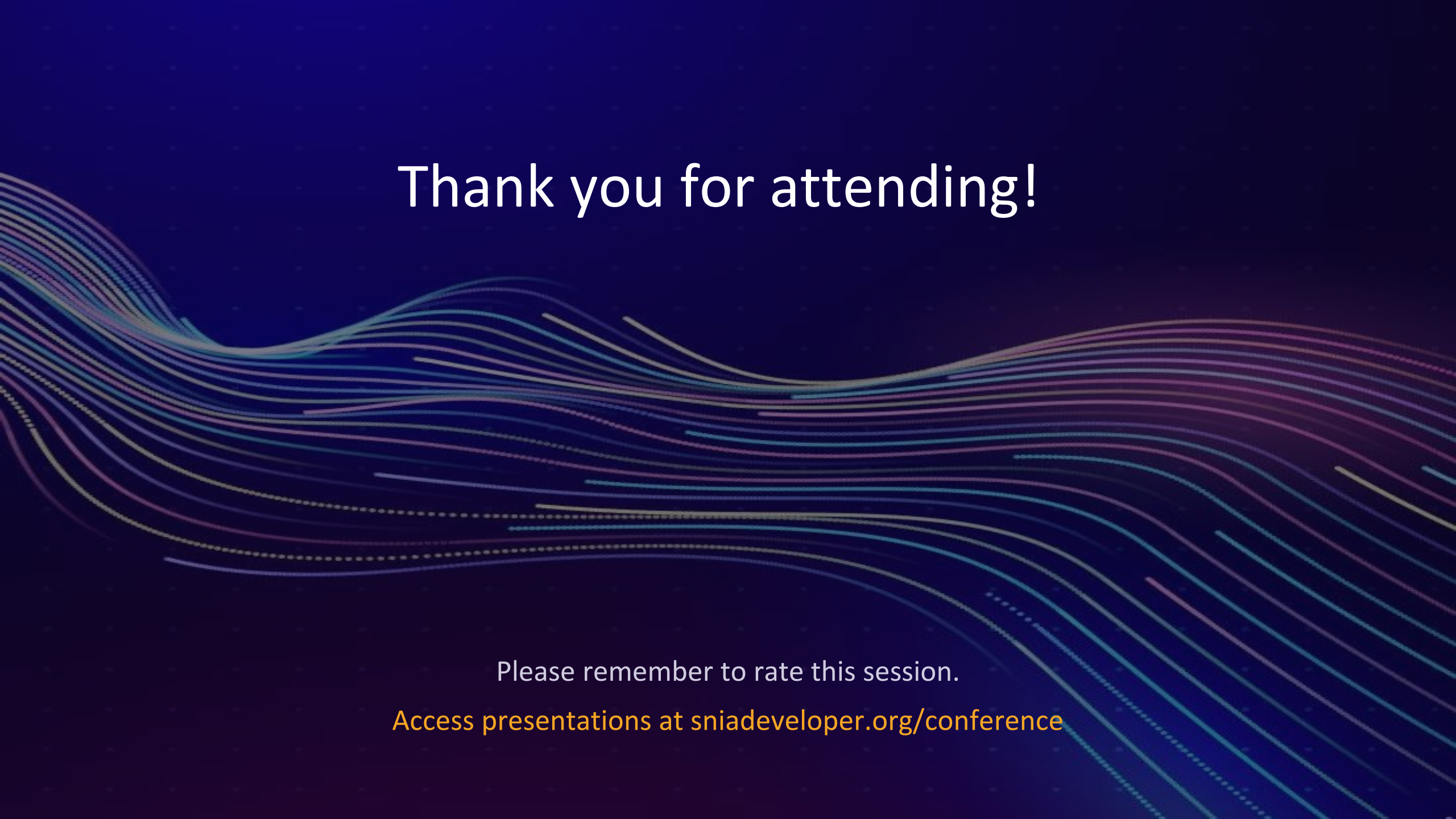
Workload W5 over 1 GigE, CIR 300 Mbps

w/o Traffic Shaping



With Traffic Shaping





Thank you for attending!

Please remember to rate this session.

Access presentations at sniadeveloper.org/conference

Real-Time Networking with Stanford's HOMA Protocol

At SDC 2023 and SDC 2024 we presented Homa, a reliable datacenter transport-layer protocol invented by John Ousterhout at Stanford University. We were able to demonstrate that Homa and TCP can co-exist in the same network “peacefully”, and how Homa can be accelerated by something similar to what Offload Engines do for TCP.

Today, (almost) everybody's eyes are on datacenter infrastructure and AI clusters because of the challenging Terabit line rates and the high volume numbers of network ports. Nevertheless, there are many other applications which can also benefit from Homa:

So-called systems-of-systems such as zone-based automotive networks, telecommunication core networks, converged IT/OT networks in industrial automation, backbones in humanoid robots, or Sensor Open Systems Architectures in aerospace & defense need a reliable transport-layer protocol which minimizes tail latency and optimizes infrastructure efficiency to deliver bandwidths of 10 Gbps, and more.

Many of these systems require hard real-time behavior which - in theory - matches with Homa's attributes of being a) message-based, b) connection-less, c) using receiver-driven congestion control, and d) run-to-completion (SRPT). Time Sensitive Networking (TSN) has become the de-facto choice using IEEE standard Ethernet. TSN comes with time-synchronization, traffic shaping, reliability and resource management and works well with higher-layer protocols such as TCP.

Hence, in our presentation this year we will share insights on how Homa can be run over TSN and, in particular, the effects of different forms of traffic shaping on Homa's core functions. We plan to start with a refresher on Homa and then dive into architecture choices when running TSN and multiple Gbps. This is complemented by presenting “Light Rabbit” which is a cost-optimized variant of CERN's White Rabbit high-accuracy time synchronization which utilizes programmability of modern PLLs for frequency and phase synchronization to achieve nano-second accuracy. We close by presenting first experimental results of running Homa over TSN to achieve the best of both worlds.

Learning Objectives: Tail Latencies in TCP, Homa - A Replacement for TCP From Stanford University, TSN, Traffic Shaping, CERN White Rabbit