CERN openlab

High-Performance Networking in Distributed Quasi Real-time Systems

Summer Student Lightning Talks

Daniel Lupu

Supervisor: Fabrice Le Goff

15/09/2022

Introduction

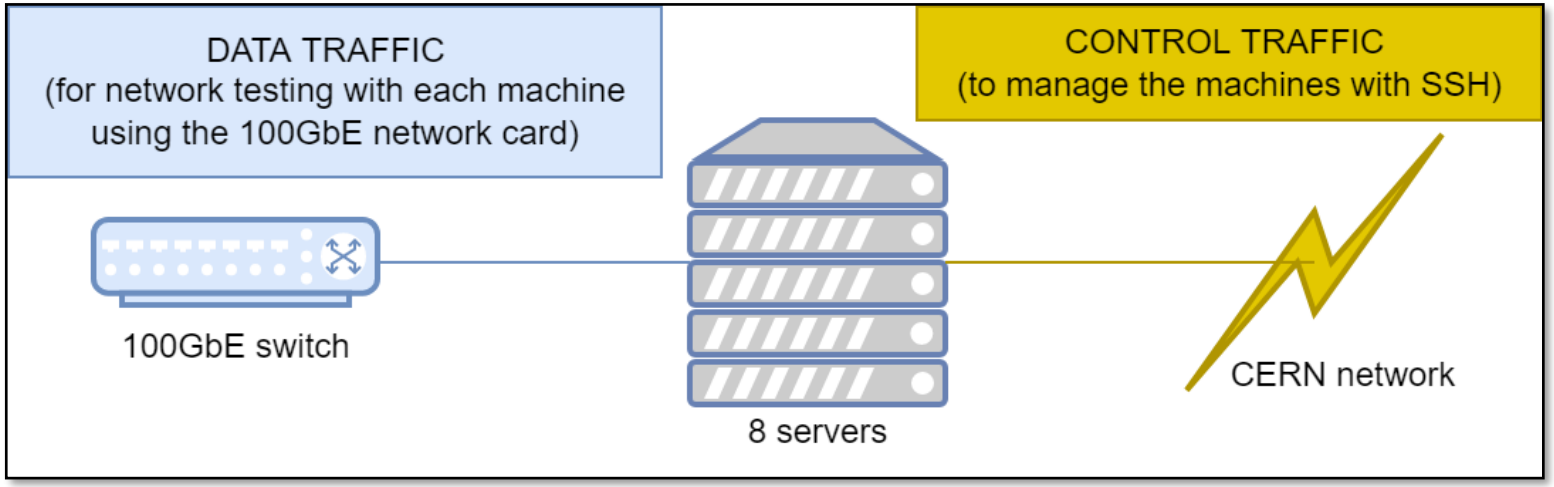
- ATLAS TDAQ will undergo an upgrade to take advantage of HL – LHC
- Dataflow sub-system is evaluating TCP/IP for network communication
- Goal: characterize performance of a 100GbE test bench

Network Topology

- 8 servers running CentOS
- Switch: QFX5120-32C
- NIC: Mellanox ConnectX-5
- Jumbo frames: 9000 bytes

Receiving machine:

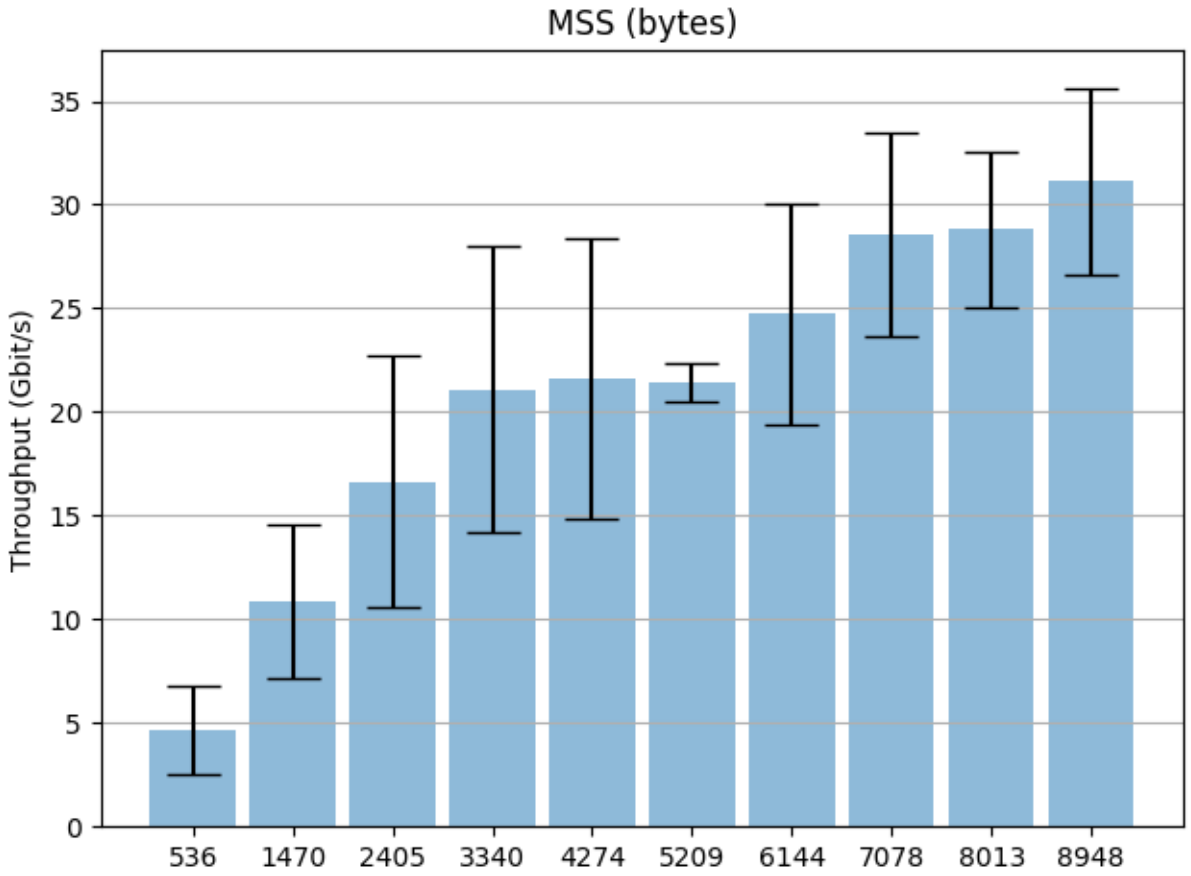
- CentOS Stream 8
- 1× AMD EPYC 7302P
- NIC firmware and driver updated to the latest version



Tools: *iperf* command and *Python* in a GNU/Linux environment

Single Connection – MSS

- The Maximum Segment Size (MSS) is the largest amount of data in a TCP segment, bounded by the frame size
- Almost linear increase in throughput
- Receiver CPU core saturated for all values
- In power-saving mode throughput decreased by 35 %

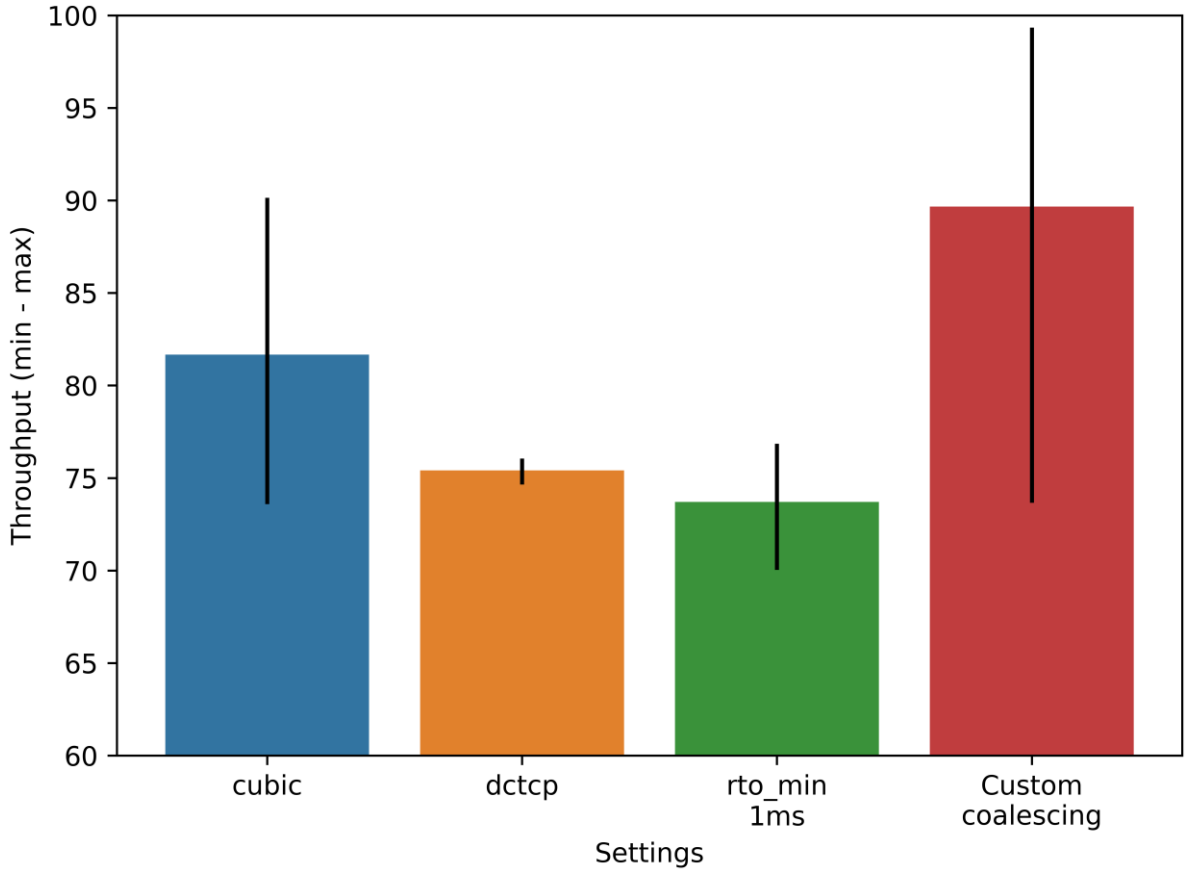


Multiple Connections

Many-to-one traffic pattern

At least three connections are needed to achieve a throughput more than 90 Gbit/s

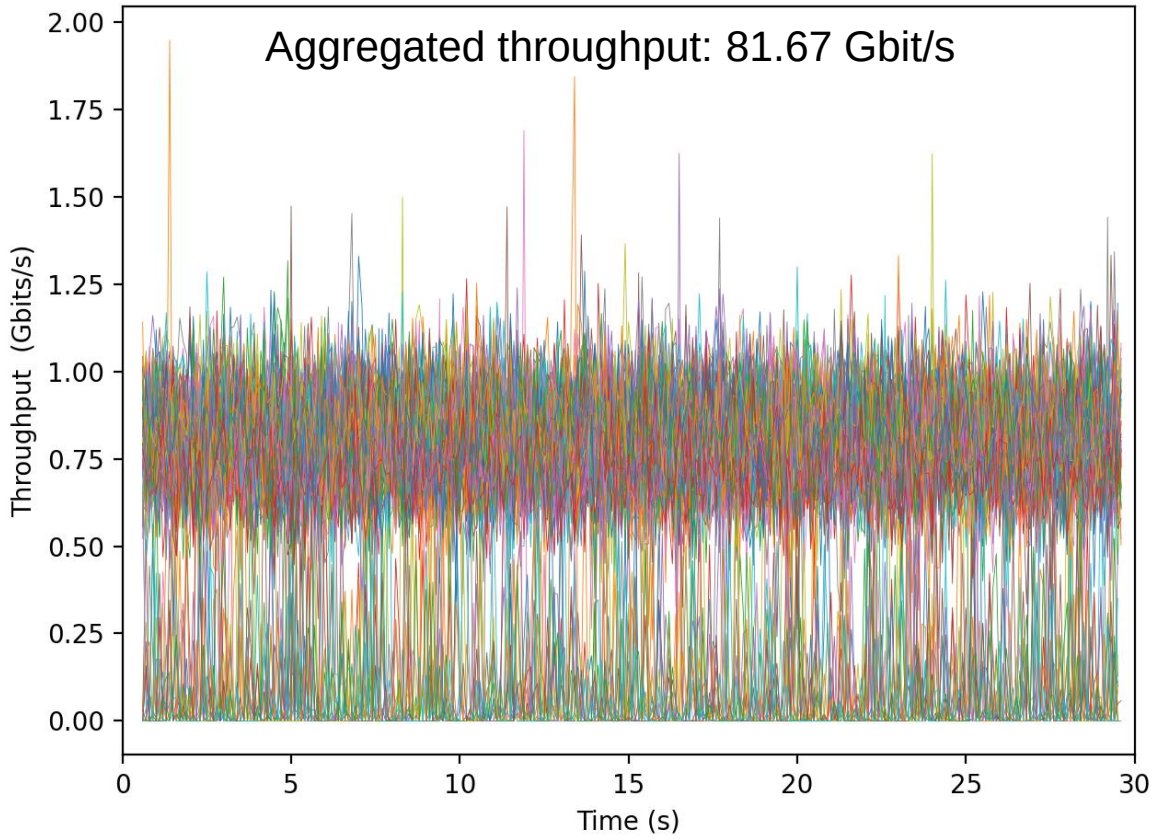
For a more realistic scenario to a DAQ system, 15 connections per machine were started, for a total of 105 senders to a single receiver



Multiple connections - issues

- Connections dropping to 0 Gbit/s

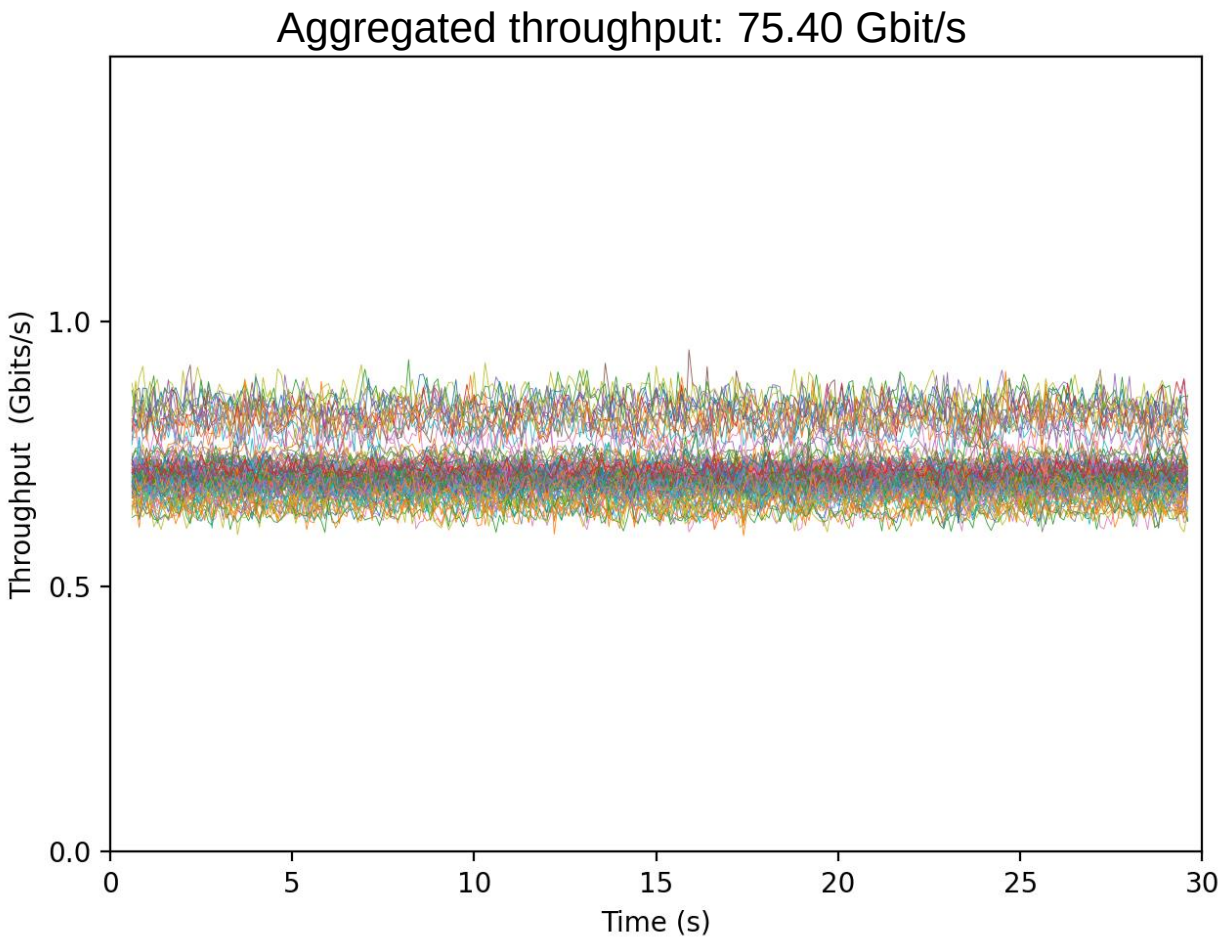
- **rx_discards_phy** counter increased after each test



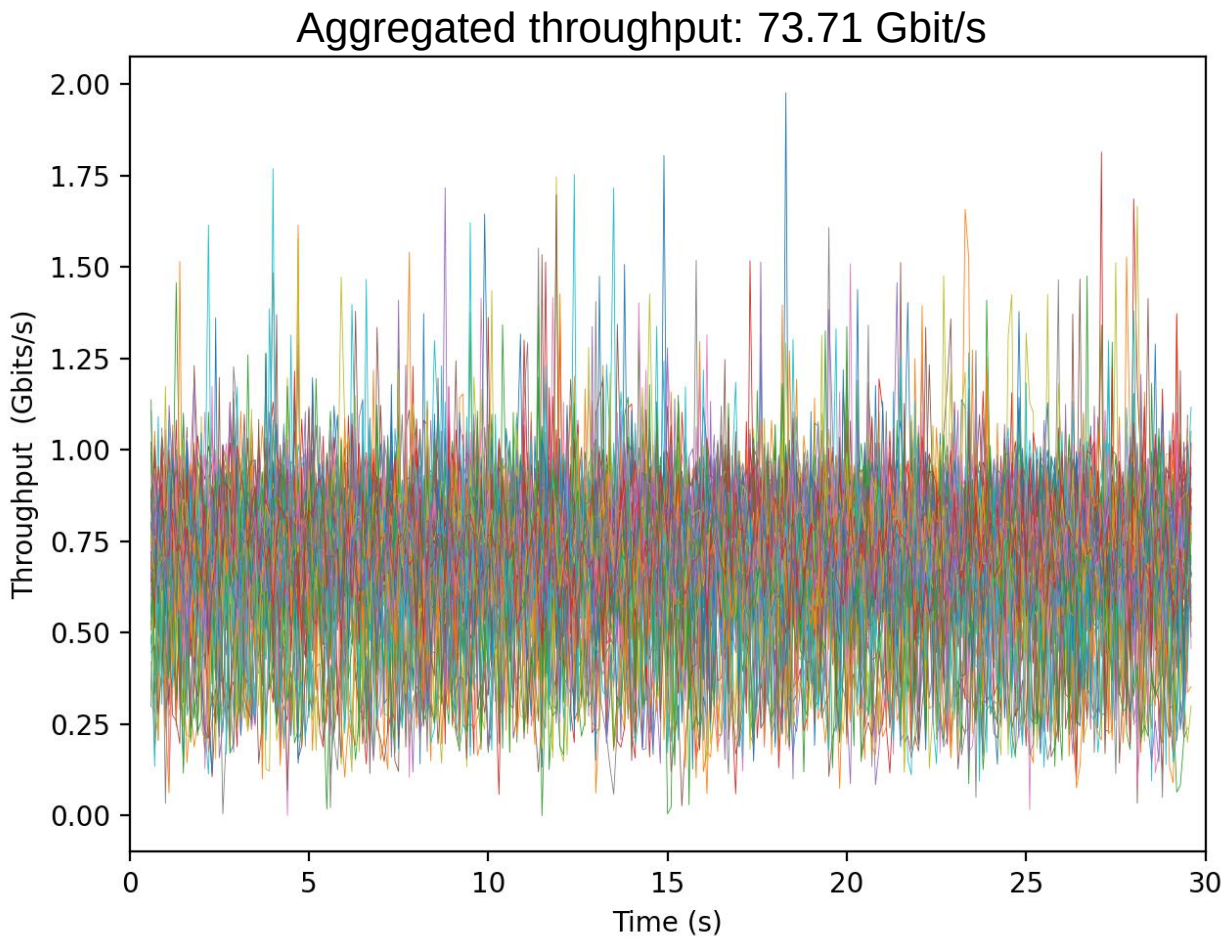
	rx_discards_phy (#packets discarded)	Standard deviation	Discarded data $mss \times \#packets$	Received data
Default	2968204	17368	24,7 GB	306,2 GB
Custom coalescing	162289	6790	1,35 GB	336,3 GB

Multiple connections – TCP parameters

Congestion control algorithm: dctp

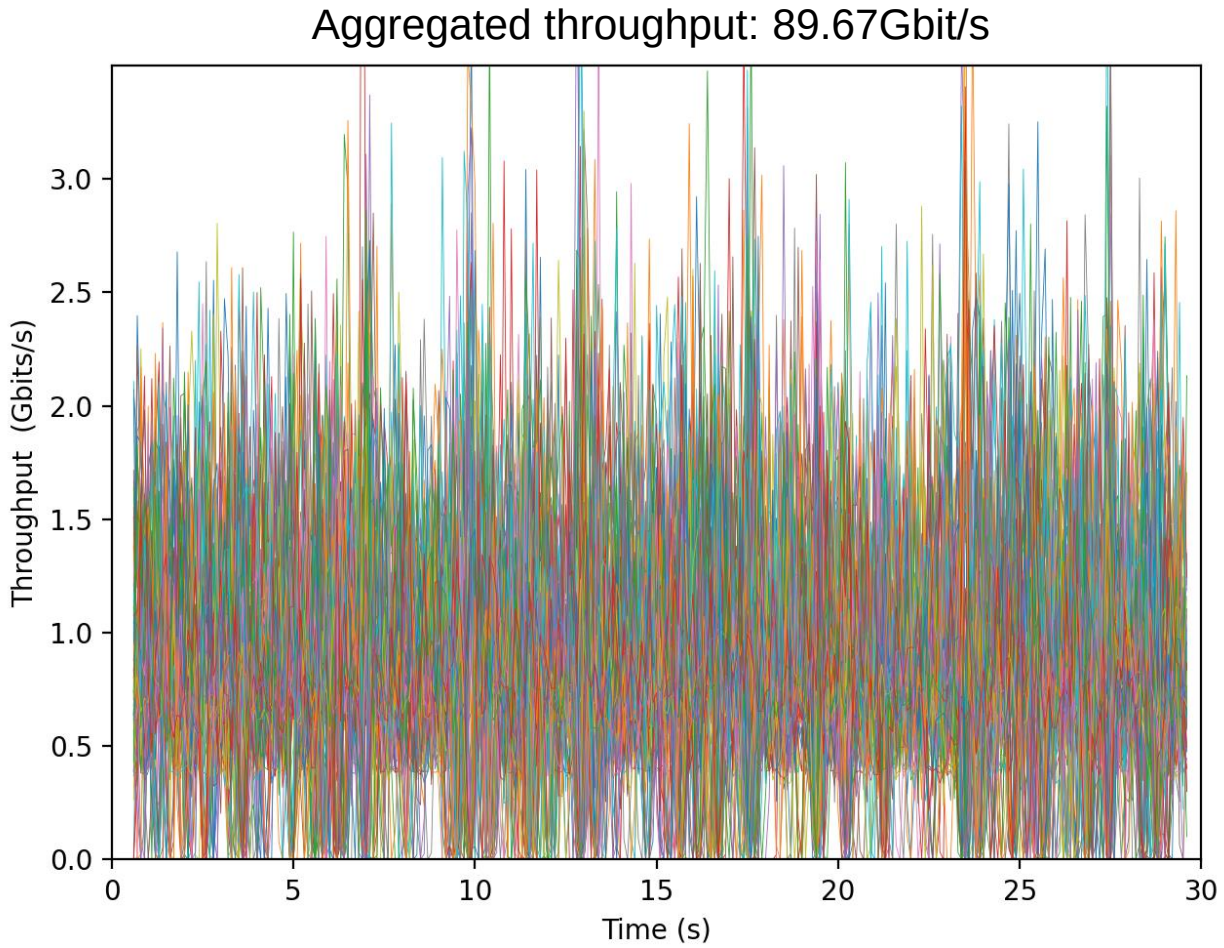


Reducing TCP retransmission timeout



Multiple connections – NIC settings

Interrupt coalescing



Conclusions

CPU plays a major role in a high-performance networking scenario, both for single and multiple connections

Applying data center aware settings can be a sensible choice in a high-performance networking distributed system

QUESTIONS?

daniel.lupu@studenti.unipd.it