Fiber to Machine

Zuowei Shen
Google System Infrastructure

IEEE P802.3db 100 Gb/s, 200 Gb/s, and 400 Gb/s Short Reach Fiber Task Force
Interim Teleconference, November 5, 2020

Overview

Objective: 400G SR4 over 50m OM4 MMF to connect machines to stage 1 switches.

- Passive copper cable reach is <2m at 100Gb/s per lane.
- Optics enables networking disaggregation from machine racks: remote TOR.
- 50m is sufficient for stage 1 switch to machine, with flexibility in rack placement.
- Lower cost, lower power consumption and lower latency can be achieved by limiting fiber reach to 50m.

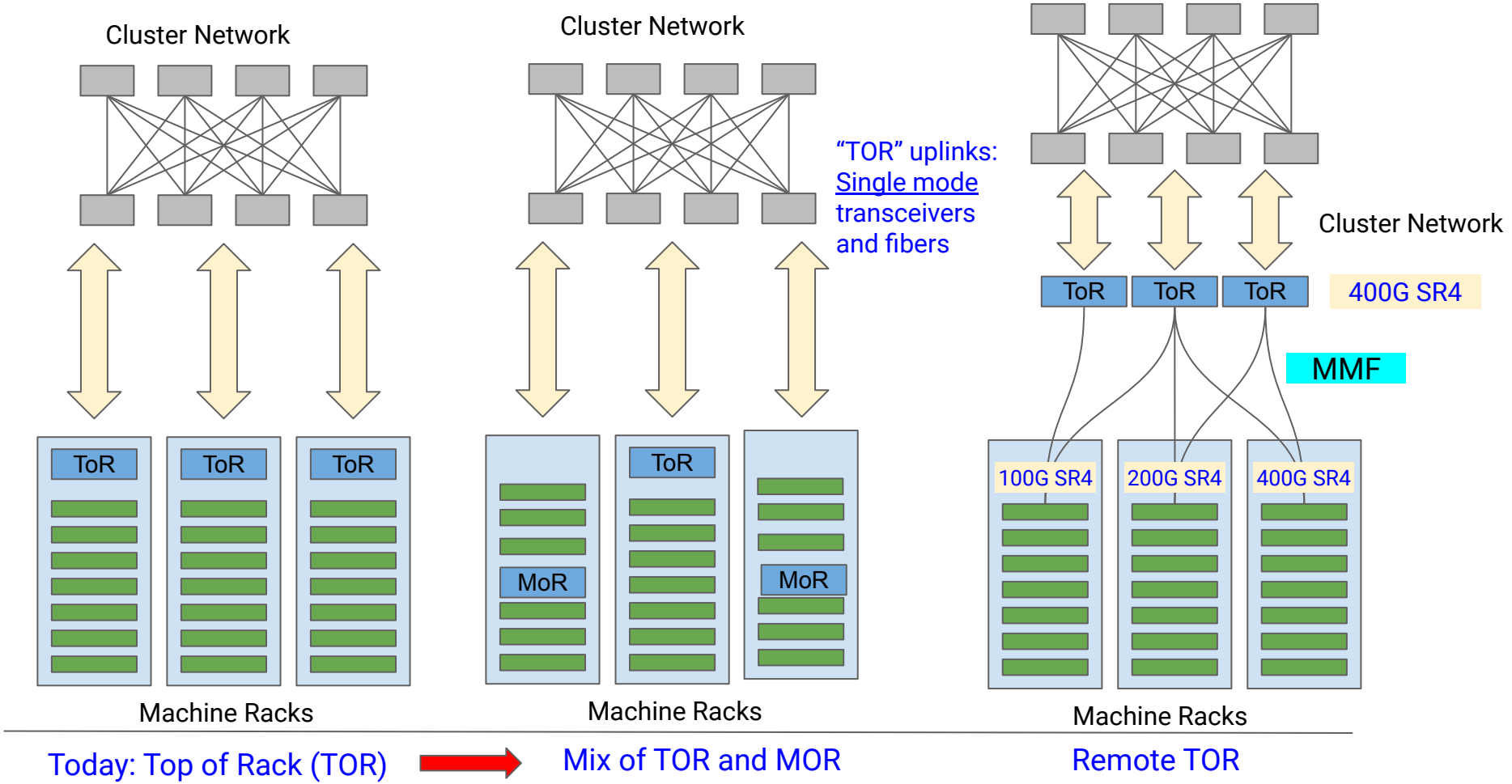
Review of Passive Copper Cable Reach

- At 100Gb/s, passive copper reach is reduced to 2m.
- Moving TOR to the middle of rack (MOR) creates stranding and deployment complexity.
- Active copper/optical interconnect is required when S1 switch is placed as TOR or remotely

NIC	SerDes Data Rate	Modulation	Passive Cu Cable Reach (IEEE)
100G	4x25Gbps	NRZ	5m
200G	4x53Gbps	PAM4	3m
400G	4x106Gbps	PAM4	2m

Why Fiber to Machines

- DAC reach can’t support current rack design at 100Gb/s per lane
  - 2m reach limit
  - Cable management becomes more challenging with thicker copper cables.
- Optical interconnect enables network disaggregation and reduces BW stranding.
  - Provides flexible BW to compute/storage/ML racks.
  - Could Improve deployment velocity
- BW demand per machine rack increases at a slower rate than switch BW
  - Fewer number of servers per racks due to increased power consumption of machines.
  - Machine BW is a function of CPU generations, spindles, SSD, rack space & power.


Cluster Network
Cluster Network
"TOR" uplinks:
Single mode
transceivers
and fibers
Cluster Network
ToR
ToR
ToR
400G SR4
MMF
ToR
ToR
ToR
MoR
ToR
MoR
100G SR4
200G SR4
400G SR4
Machine Racks
Machine Racks
Machine Racks
Today: Top of Rack (TOR)
Mix of TOR and MOR
Remote TOR
100Gb/s Serdes

SMF vs MMF Comparison for Ultra Short Reach

- MMF over 50m OM4 is the preferred interconnect solution for fiber to machines.
- At 100m reach, 400G SR4+OM4 loses cost advantages comparing to 400G PSM4+SM fiber.
- Fiber to machine doesn't need reach >50m. Latency in fiber increases 5ns/m.

	SMF (<500m)	**MMF (Recommendation)** 50m OM4 reach	MMF 100m OM4 reach
Transceiver Cost	Higher	Lower	Higher
Fiber Cost	Lower	Medium	Higher
Power consumption	Medium	Lowest	Highest with stronger EQ in serdes
Backward compatibility	No Lack of 200G PSM4	Yes Interop with 200G&100G SR4	Yes Interop with 200G&100G SR4

Success Metrics

- Reach: 30m OM3, 50m OM4 for fiber to machines
- Cost effective and multiple VCSEL sources
  - 100m reach adds more stringent requirement on VCSEL BW, spectral width, RIN
- Low power consumption required by thermal requirement in various server/storage/accelerator trays
  - Avoid overdesign of Rx serdes and Tx nonlinear compensation.
- Low latency FEC: nice to have.
- Connector:
  - MPO8_APC
  - SN connector: friendly to breakout applications.

Summary

- Fiber to machine is a new application with broad market demand.
- Host interconnect requires cost effective and power efficient.
- Serdes shall be optimized for 50m OM4 only to optimize serdes power.
- Shorter reach is required to optimize fiber latency. Fiber latency over 100m is 250ns more than over 50m.
- Multimode solution may not have cost advantage for >50m TOR uplinks at 100Gb/s compared to parallel single mode. SM solution also offers forward compatibility at 200Gb/s and beyond.

Supporters

- Ilya Lyubomirsky, Inphi
- Ryan Latchman, Macom
- Osa Mok, Innolight
- Chongjin Xie, Alibaba
- Ali Ghiasi, Ghiasi Quantum
- Vipul Bhatt, II_VI
- Piers Dawe, Nvdia