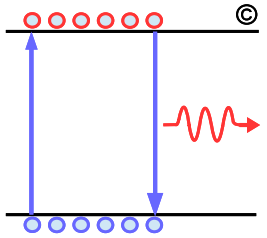


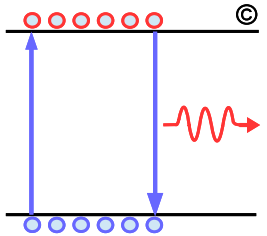
Applications of 100GBASE-SR

Ali Ghiasi
Ghiasi Quantum LLC

IEEE Meeting
Geneva

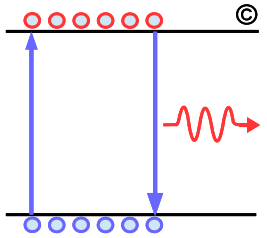
Jan 21, 2020

Overview



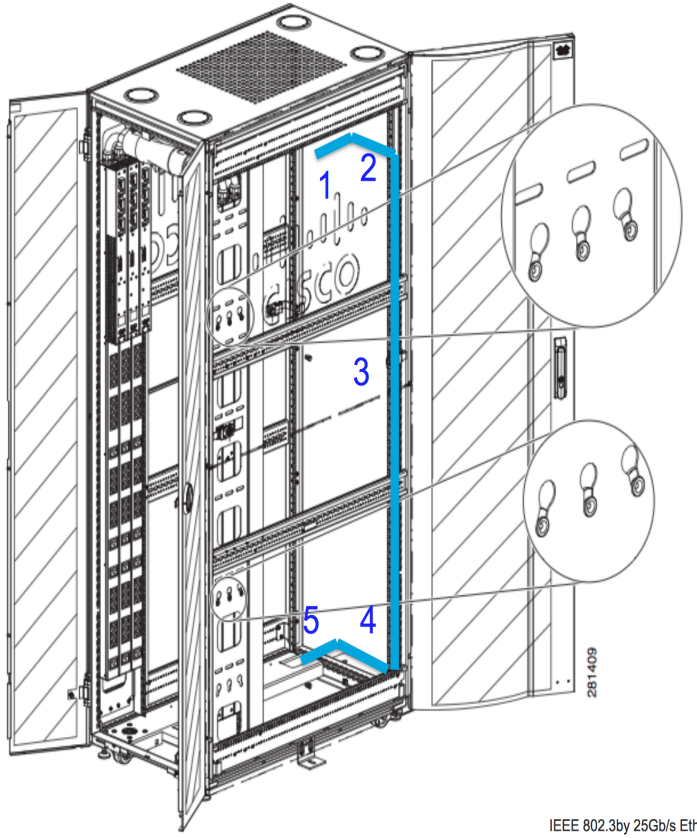
- Evolution of Cu cabling
- Cu cabling no longer addresses TORs
- Server/NIC evolution
- Switch evolution
- DCN evolution
- Sources of dispersions
- Recommendations.

802.3/SFF have defined several Cu DAC PMDs



- **10GSFP+DAC**
  - SFF-8431 defines 8 m reach Cu DAC
- **40GBASE-CR4**
  - Defined in the 802.3ba CL-85 with a reach of 5 m
- **100GBASE-CR4**
  - Defined in the 802.3bj CL-92 with a reach of 3 m
- **25GBASE-CR**
  - Defined in the 802.3by CL110 with a reach of 3 m
- **50GBASE-CR**
  - Defined in 802.3cd CL136 with a reach of 3 m
- **goergen_3by_02a_0715 analysis shows that real life Cu cable needs to be at least 2.69 m which can't be met with 100GBASE-CR with 2 m max reach!**

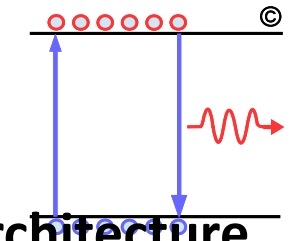
Cabling Installation – Top to Bottom



IEEE 802.3by 25Gb/s Ethernet

- Consider this common strategy
  - 1 – 152mm
  - 2 – 304mm
  - 3 – 1778mm
  - 4 – 304mm
  - 5 – 152mm
- This real life case is 2690mm.

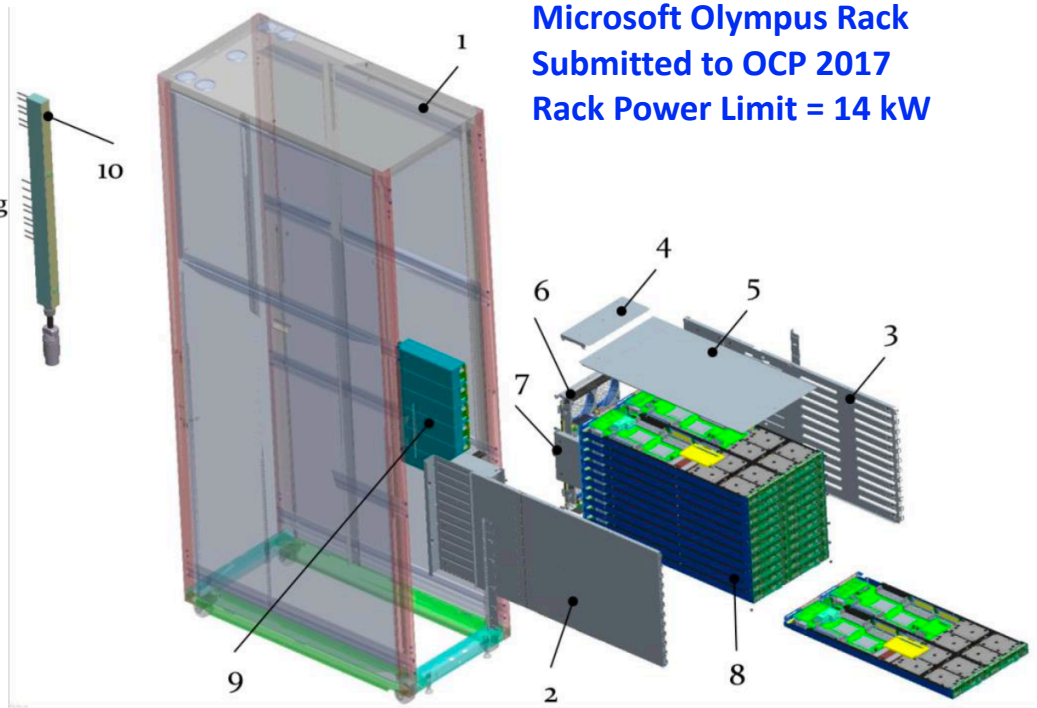
Number of Server per Rack Are Decreasing



- **A decade ago 48-96 servers were common combined with SFP+ started the TOR architecture**
  - Today common server rack only have about 24 servers with increased CPU cores (16-48) the PD per CPU has increased in excess of 300 W
  - A high end 4 socket server in 2RU form factor may have 1.8 kW PD!

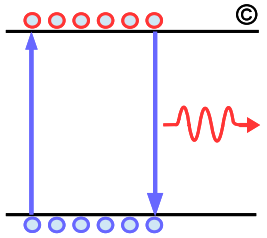
Microsoft Olympus Rack
Submitted to OCP 2017
Rack Power Limit = 14 kW

1. Rack
2. Left Side Wall (including Power Dist Board)
3. Right Side Wall
4. Rear Covers
5. Top Cover
6. Fan Door
7. Chassis Manager
8. Trays
9. Power Supplies
10. Power Distribution Unit

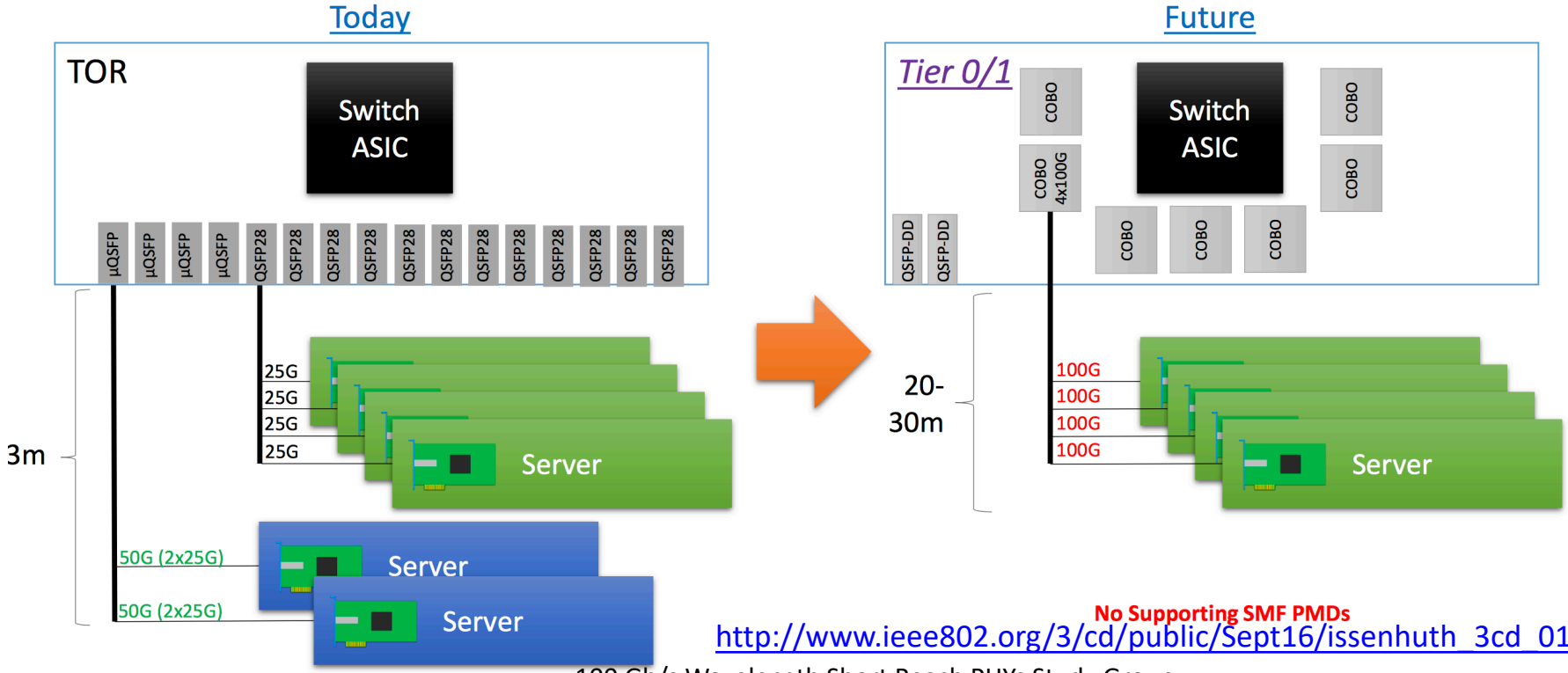


Google Barreleye Rack
Submitted to OCP 2018
Rack Power Limit = 30 kW

Emerging Trend: Server Connecting to MOR Switch



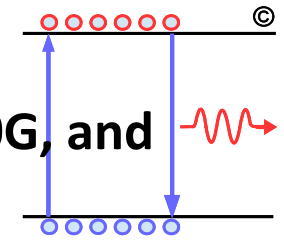
- **Microsoft evolution showing server directly connecting to MOR/Tier 0/1 switches as result of switch radix increase from 128 to 256 and fewer servers in a rack**
  - Passive Cu cable with reach limited ≤2 m at 100 Gb/s/lane has limited usefulness
  - Passive Cu cable require adding retimer on ~50% of the ports or any trace > ~5”
  - 100GBASE-SR with C2M interface with 11 dB loss budget can operate without need for retimers.



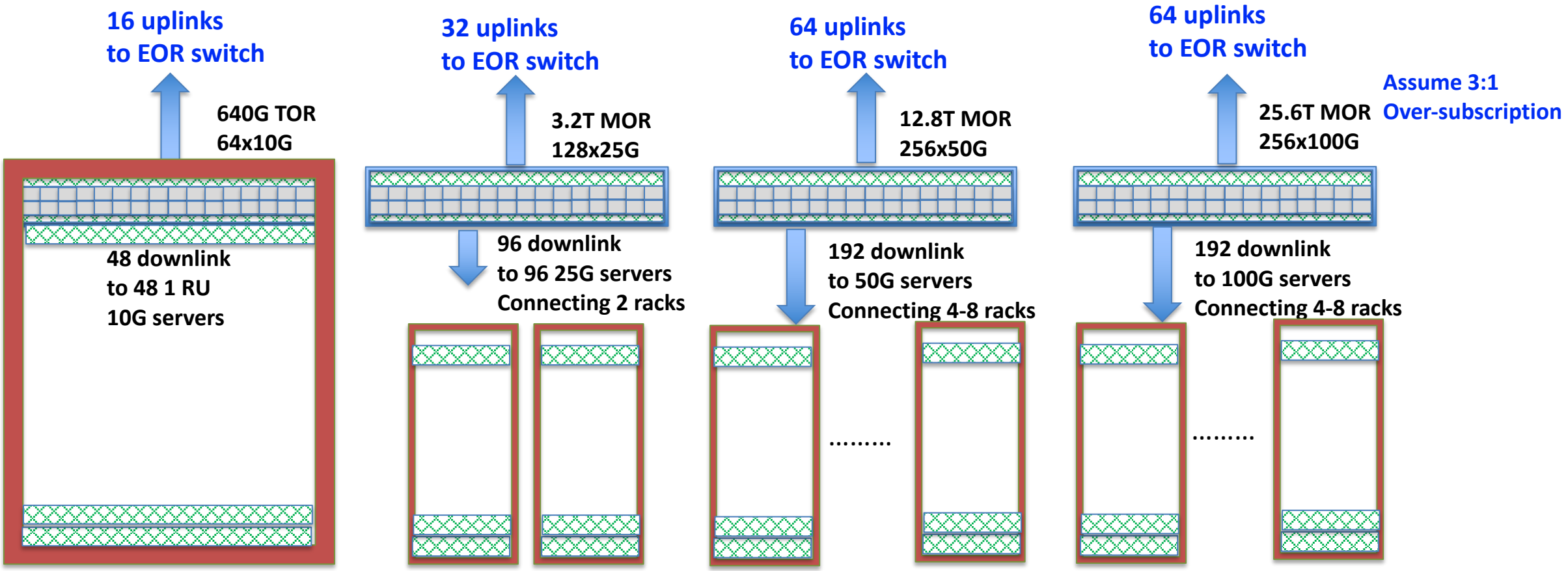
No Supporting SMF PMDs

http://www.ieee802.org/3/cd/public/Sept16/issenhuth_3cd_01a_0916.pdf

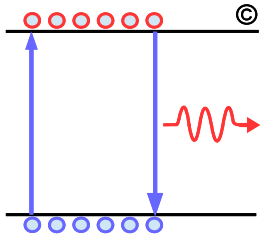
Datacenter Trends



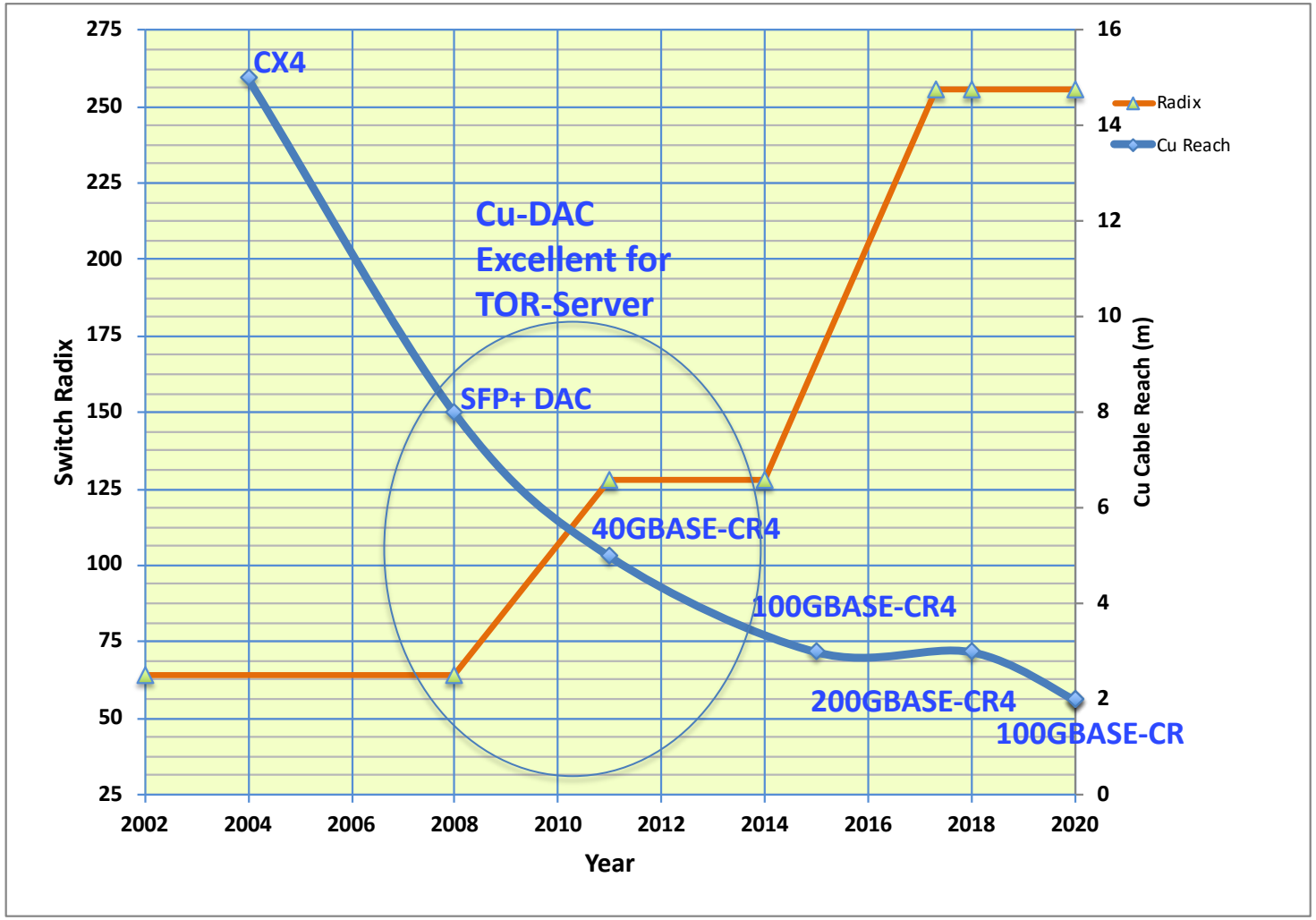
- **Switch radix over the last 12 years has increased from 64x10G, 128x25G, now to 256x50G, and likely to 256x100G by 2020**
  - With this trend a 2 m Cu DAC no-longer will be ubiquitous server-TOR solution
  - A low cost-low power 100G/lane SR PMD with 15 m reach without mid-span connector would be ideal to connect 4-10 racks.



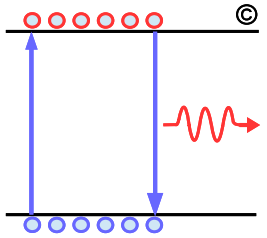
Why we need low cost 100G-SR



- **SFP+ DAC with 8 m cable reach not only supported TOR but could connect up to 5 racks!**
- **With introduction of 128 radix switches single switch became too large for one rack of servers**
  - Over the last 10 years the number of servers per rack have decreased from ~48 to ~24 while the DAC cable reach decreased from 8 m to 3 m
- **With introduction of 256x100G switches Cu DAC with 2 m reach no longer is a viable servers to 1st layer switch!**



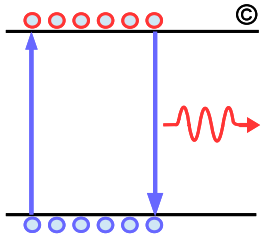
Datacenters are Evolving and they are Getting Smaller



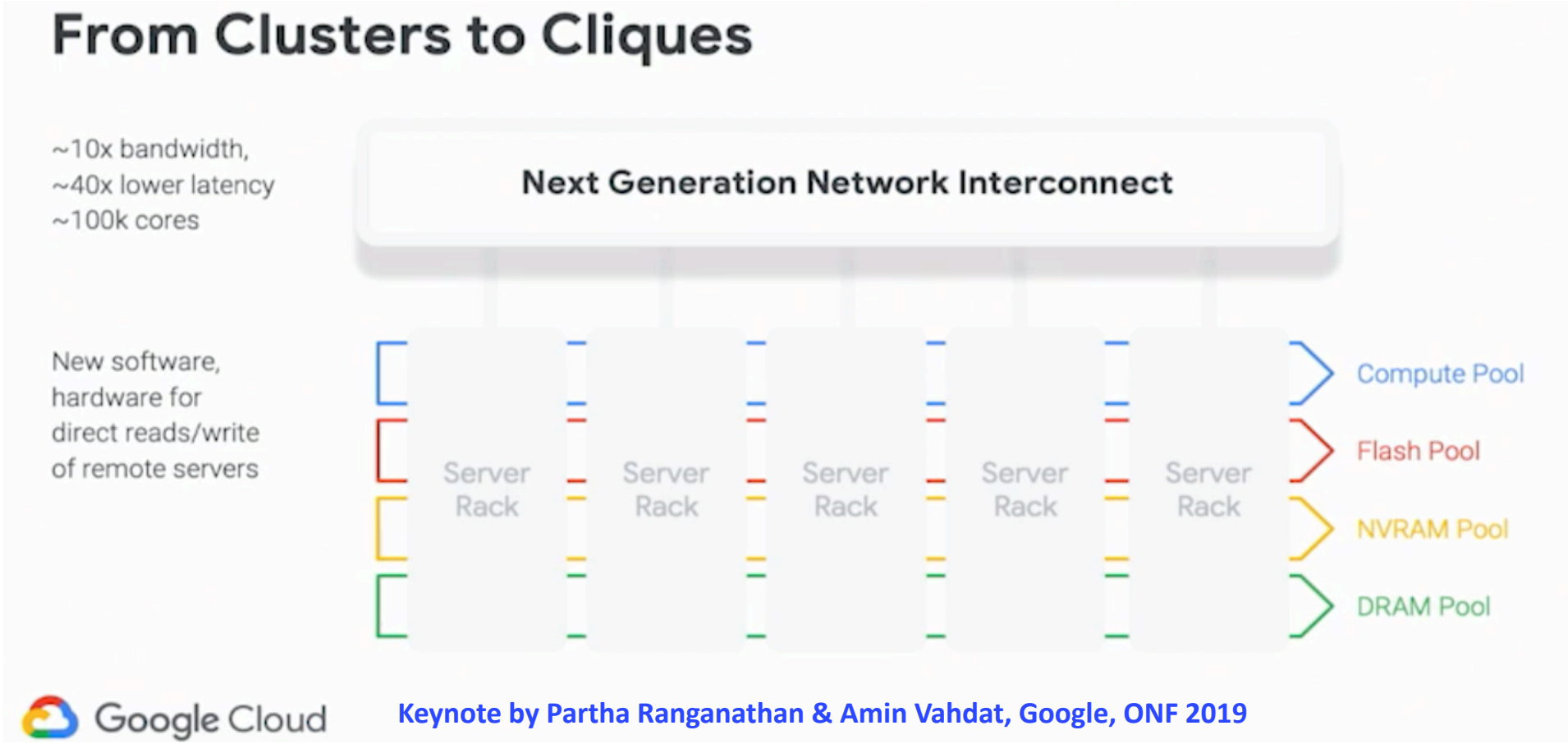
- **With increased CPU core and power dissipation one could hit 100 MW power limit with even less than 100k servers**
- **100 MW power would only be available in remote location near hydro-electric or wind farms**
- **Datacenter operator in recent year have been building smaller 10-20 MW datacenters interconnected with DCI links**
  - These 10-20 MW datacenters may only have 10k-20k servers or about 400-1000 racks only
- **Google is finding that large data center clusters have 50+ µs latency resulting in ~30% CPU cycles lost as the memory and flash have become faster**
  - Google* wants to build HPC like clusters with 1000-3000 servers or about 50-200 racks
- **Given the current trend 100G-SR not only will address TOR/MOR to servers but also can address TOR-EOR or MOR-EOR in the merging smaller DCNs.**

*** Keynote by Partha Ranganathan & Amin Vahdat, Google, ONF 2019**

From Clusters to Cliques “HPC”

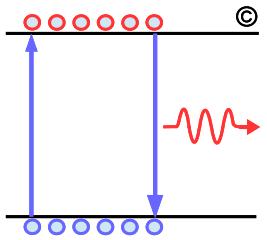


- **Google plan to build HPC clusters with ~100K cores**
  - Assuming 48 core results in ~2100 servers or ~ 100 racks!

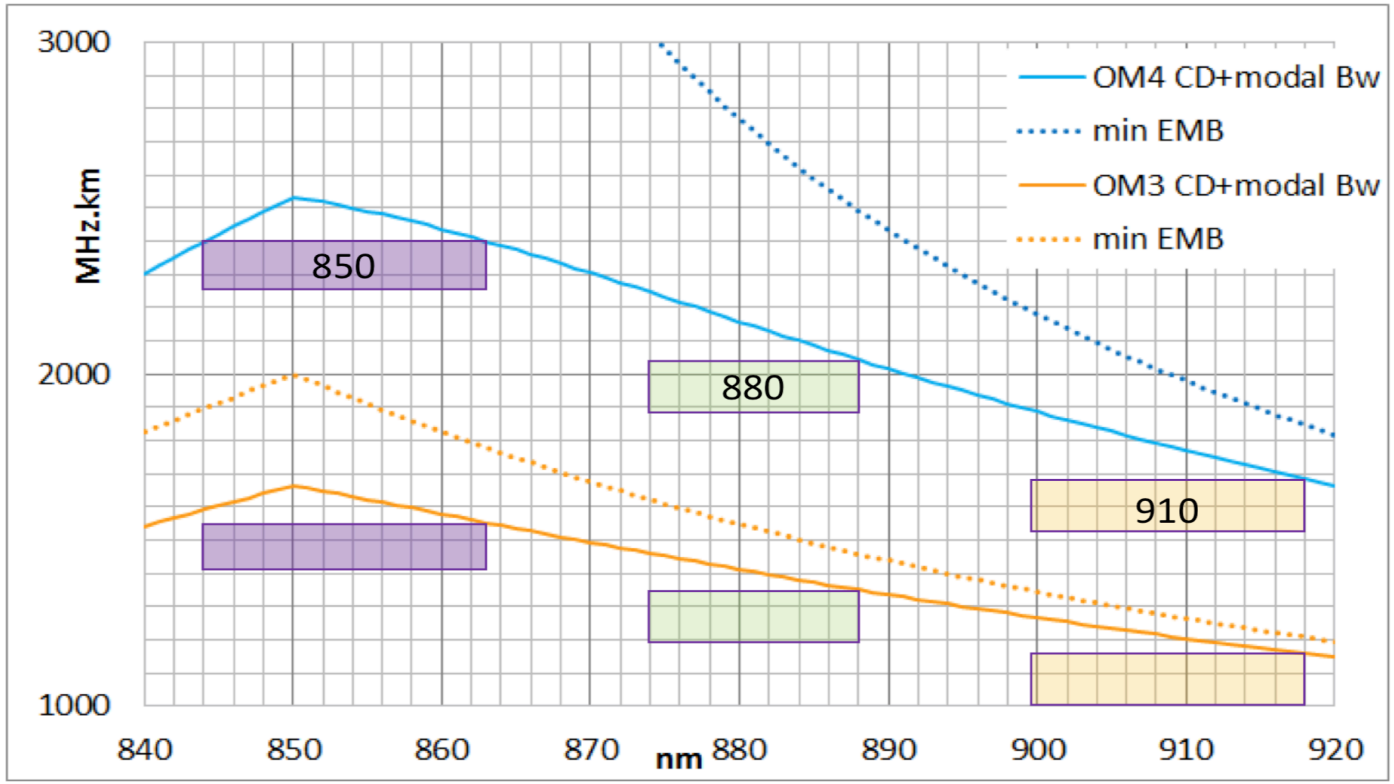


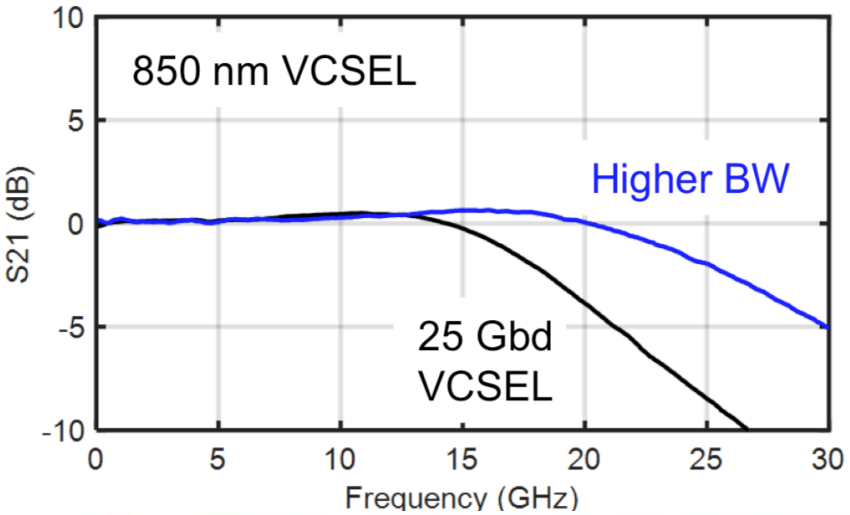
Keynote by Partha Ranganathan & Amin Vahdat, Google, ONF 2019

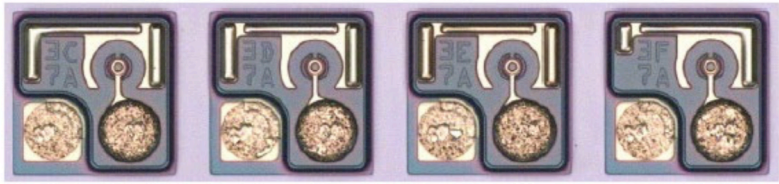
Fiber BW isn't the Dominant Source of Dispersion



- **OM4 fiber BW is ~2000 MHz.km king_3cm_adhoc_01_062818. where A 50 m link will have 40 GHz of BW**
  - The VCSEL and PIN TIA with estimated BW of 25-28 GHz CFI_01_1119 dominants in limiting link and required equalization.

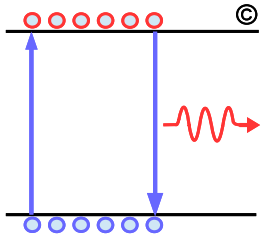






Broadcom 100G VCSEL under development

Recommendation



- **The study group should consider an ultra low cost SR PMD with reach of 15 m addressing TOR to server applications**
  - Single jumper without the need for mid-span connectors just like AOC and Active DAC
  - May want to also consider lower latency FEC such as 1/2 length KP FEC in addition to KP FEC
- **The study group should also consider a 50 m SR PMD addressing TOR-EOR and MOR-EOR applications**
  - Support up to 4 mid-span connectors
  - Based on KP FEC
  - Given that most 100G DSP have significantly more capability than just 5T FIR supporting 50 m reach should not be an issue and given the dominant source of dispersions are from EO/OE devices
- **2 years I suggest defining 100G single λ MMF ghiasi_NGMMF_01_jan18 obviously we are going to miss the initial single λ deployment**
  - But the combination of lower cost, power, and ease of use will enable broad set of applications from HPCs, AI/GPUs, servers, TOR switches, to MOR/EOR switches.