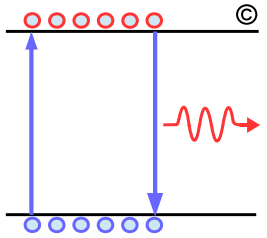


# The Need for 100Gb/s/lane MMF PMDs

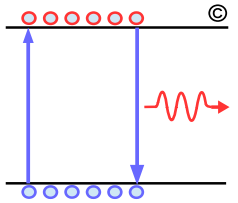
**Ali Ghiasi**
**Ghiasi Quantum LLC**

**Next-generation 200 Gb/s and 400 Gb/s MMF PHYs Study Group**

**Geneva**

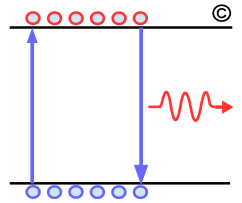
**January 22, 2018**

# List of Supporters



- **Brad Booth – Microsoft**
- **Rich Baca – Microsoft**

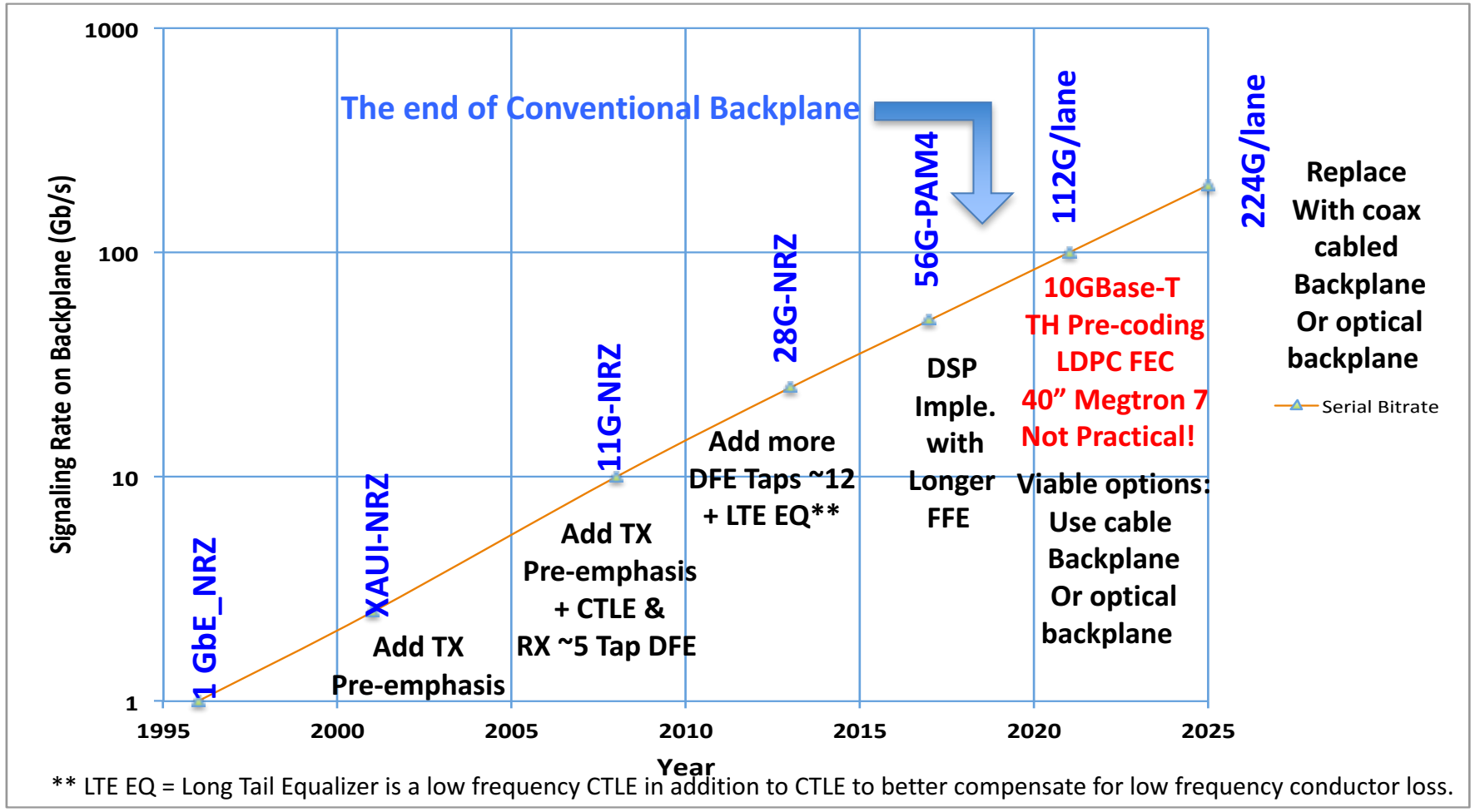
# Overview



- **Next-generation 200 Gb/s and 400 Gb/s MMF PHYs Study Group need to consider key market trend:**
  - Introduction of 100Gb/s/lane switches in about 2 years
  - Switch radix increase from 128 to 256 may eliminate TOR switches and 2 m Cu “Apple Pie” cable
  - A large radix switch may be placed in one rack but will connect several adjacent racks
  - Or the large radix switch may be placed middle of row “MOR” or in some case end of row “EOR”
  - With switch radix increasing, where single switch must connect several racks or a ~1 m Cu passive cable no longer interesting at 100G/s/lane
- **With Cu passive cable reach either too short or no longer practical at 100 Gb/s/lane what are the options**
  - AOC is one option but the OSFP/QSFP-dd plugs are very bulky
  - Active Cu is an option potentially for ~5 m but cable and plug are both bulky
  - Short reach ultra-low cost optical PMD with ~ 30 m could replace Cu cables and
- **With conventional backplane not longer viable what are the potential alternatives**
  - Cabled backplane – complex
  - Small chassis with short ~12” backplane - low port count and limits Clos radix
  - AOC/Optical PMD multi-chassis system – enable larger radix Clos switch
- **The MMF study group should consider 2 emerging tectonic trends:**
  - Unless MMF PMDs operate at 100Gb/s/lane in the next 3 years their use will be limited to laggard networks
  - The greatest opportunity for MMF could emerge with introduction of 100Gb/s/lane host replacing Cu cables and Cu backplanes.

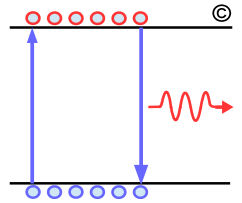
# Evolution of Signaling Rate

- **Mass volume deployment of 112G/lane are expected by 2021**
  - Product introduction 2019/2020

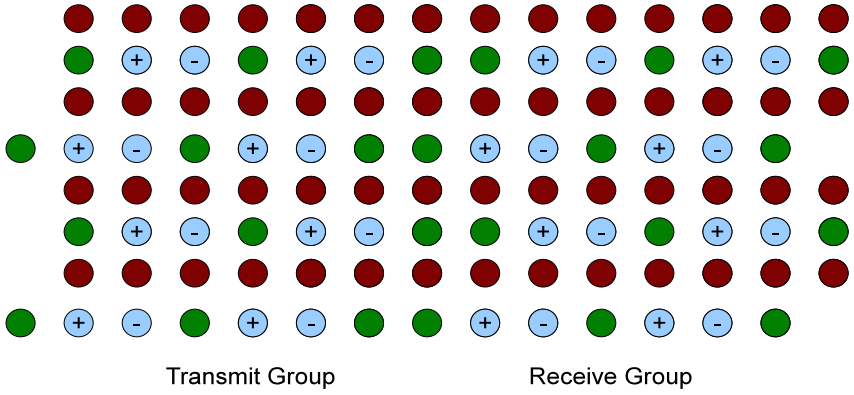


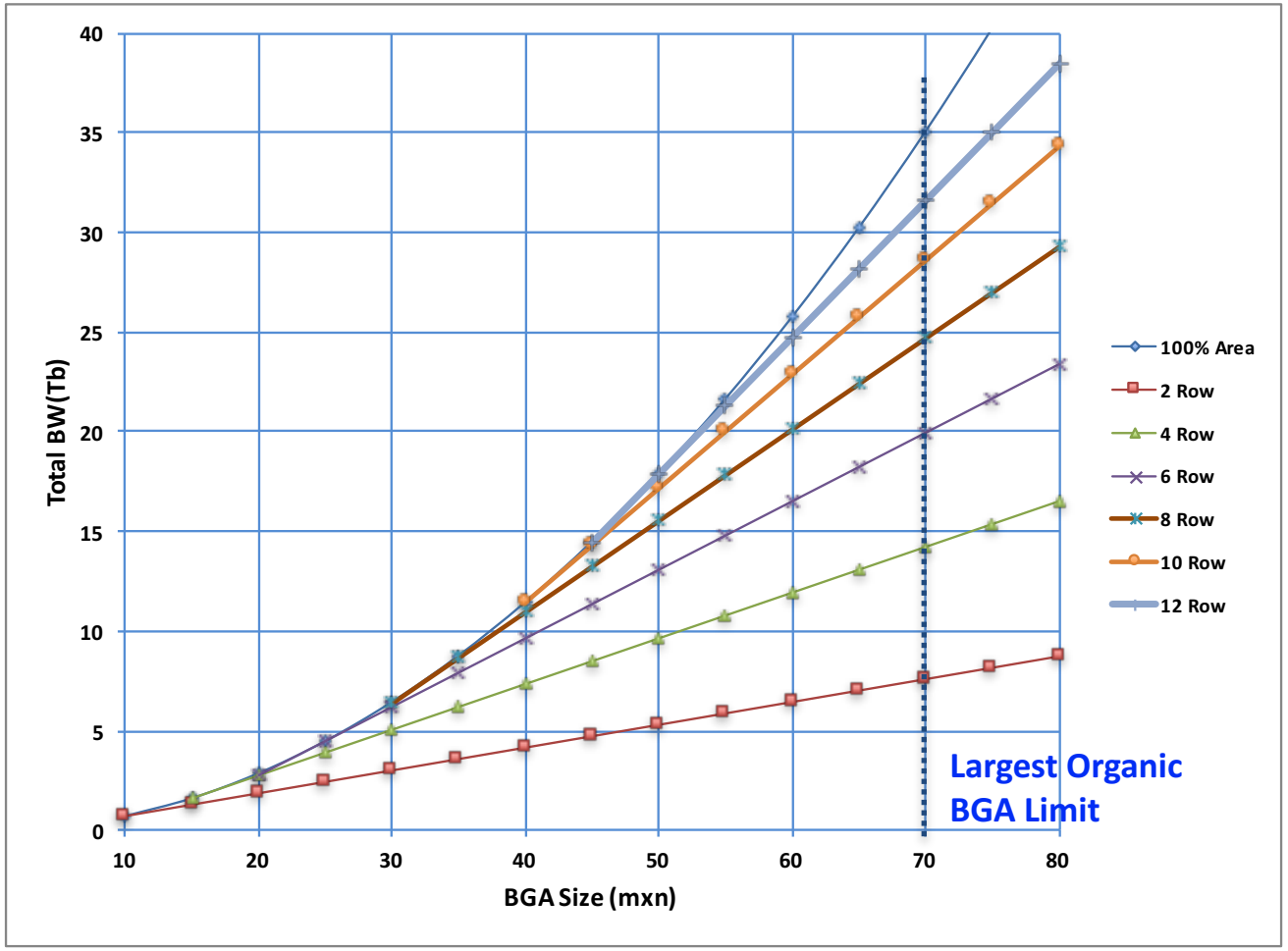

** LTE EQ = Long Tail Equalizer is a low frequency CTLE in addition to CTLE to better compensate for low frequency conductor loss.

# Switch ASIC BW Assuming 100G IO

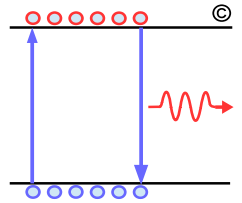


- **Updating Ghiasi invited talk at IEEE Photonics Group IV 2012 to 100 Gb/s**
  - A 70x70 mm package can support 12.8 Tb assuming 50G IO but 25.6 Tb switch generation would need 100G IO
  - Assume BGA ball map for calculation of the BW

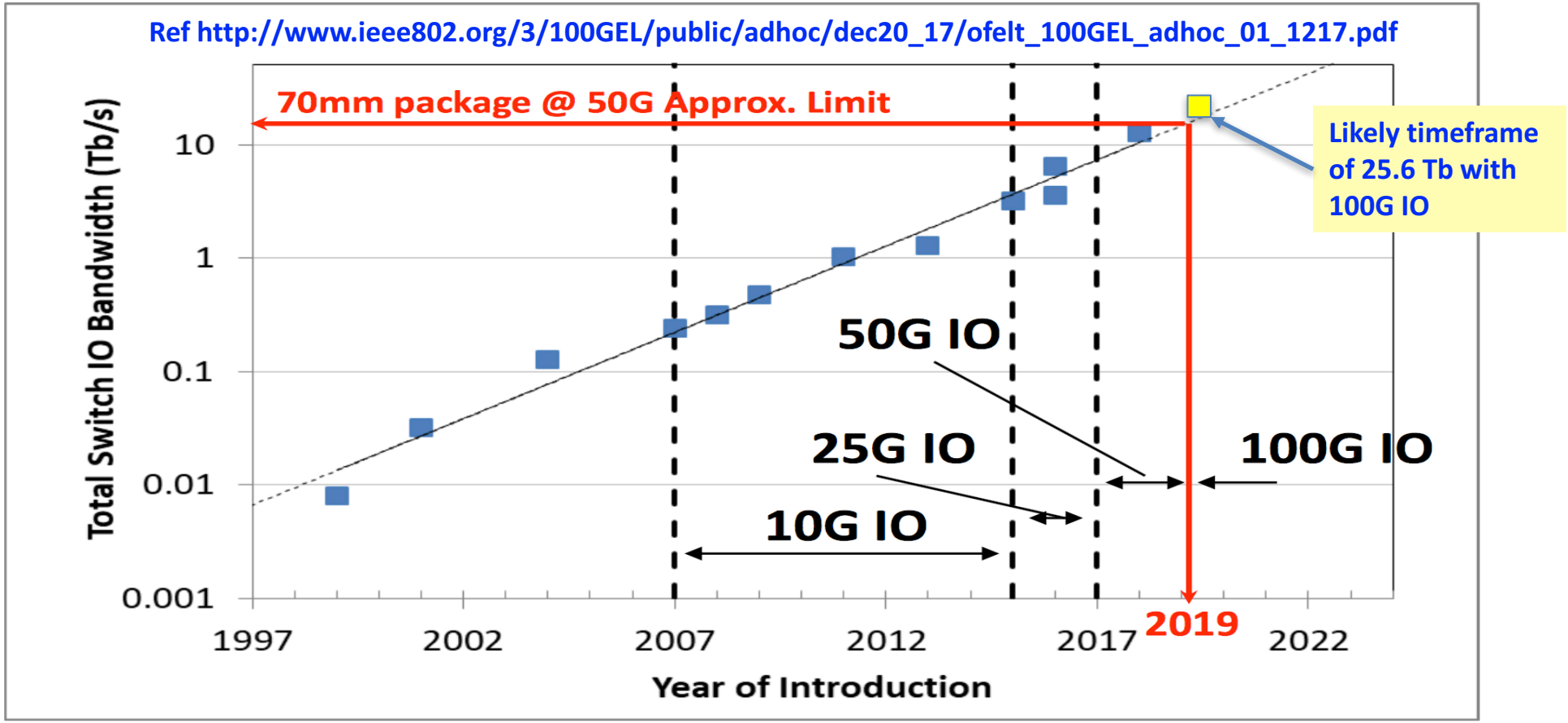


# Why 112G IO Necessary



- **112G is necessary to double current 256x50G (12.8 Tb) ASIC BW to 25.6 Tb given current limitation of organic packages to 70x70 mm**
  - Unless VCSEL MMF can support 100G/s/lane MMF use will be limited to laggard networks!

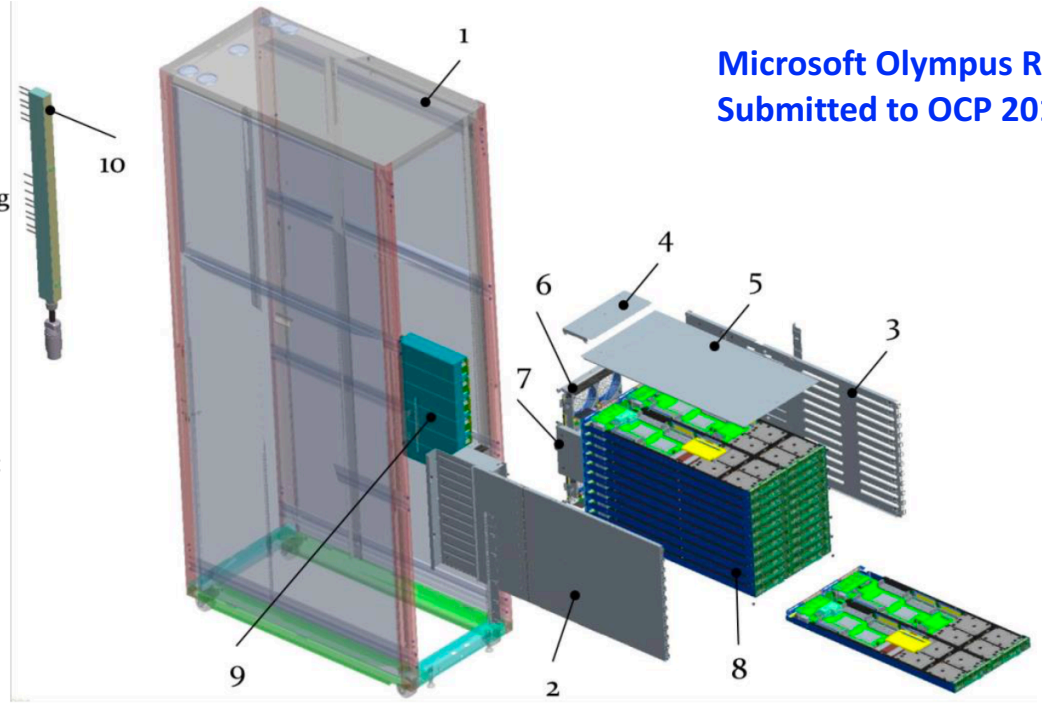



> 2019 ASIC requirements are expected to exceed BW delivered by a conventional BGA with 50G IO

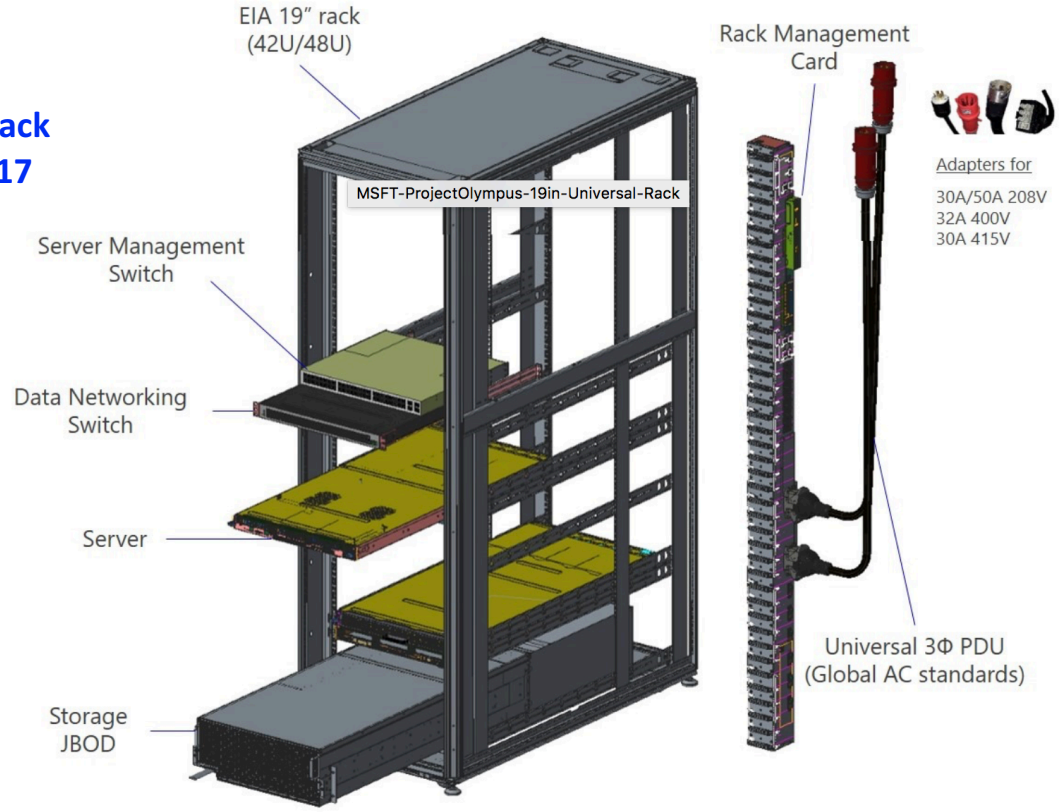
# Example Server Rack and TOR

- **A decade ago half-width servers with 96 servers in a rack were common**
- **Today common server rack implementation only have 24-48 servers as result of**
  - Larger CPUs with more cores/memory and racks having JBOD, JBOF, and GPU.

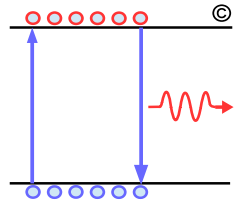
1. Rack
2. Left Side Wall (including Power Dist Board)
3. Right Side Wall
4. Rear Covers
5. Top Cover
6. Fan Door
7. Chassis Manager
8. Trays
9. Power Supplies
10. Power Distribution Unit



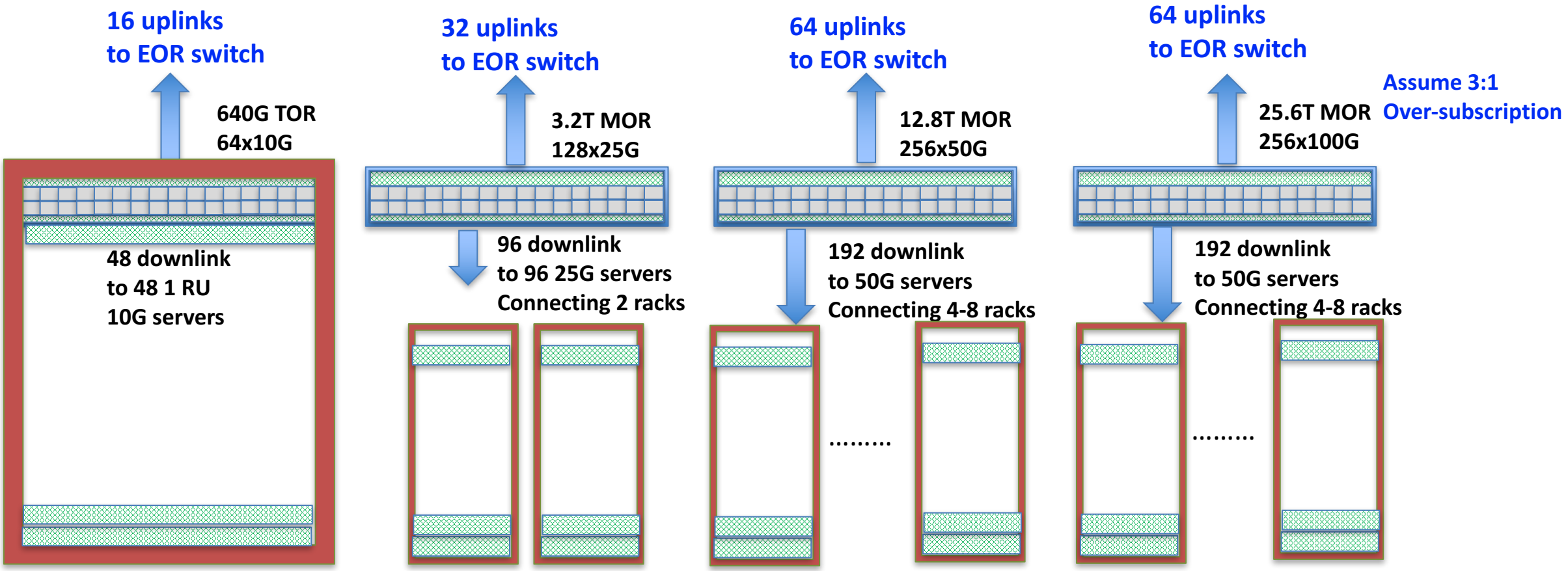

**Microsoft Olympus Rack**
**Submitted to OCP 2017**



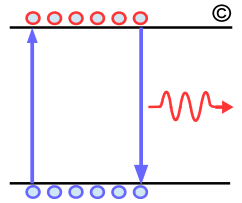
# Datacenter Trends



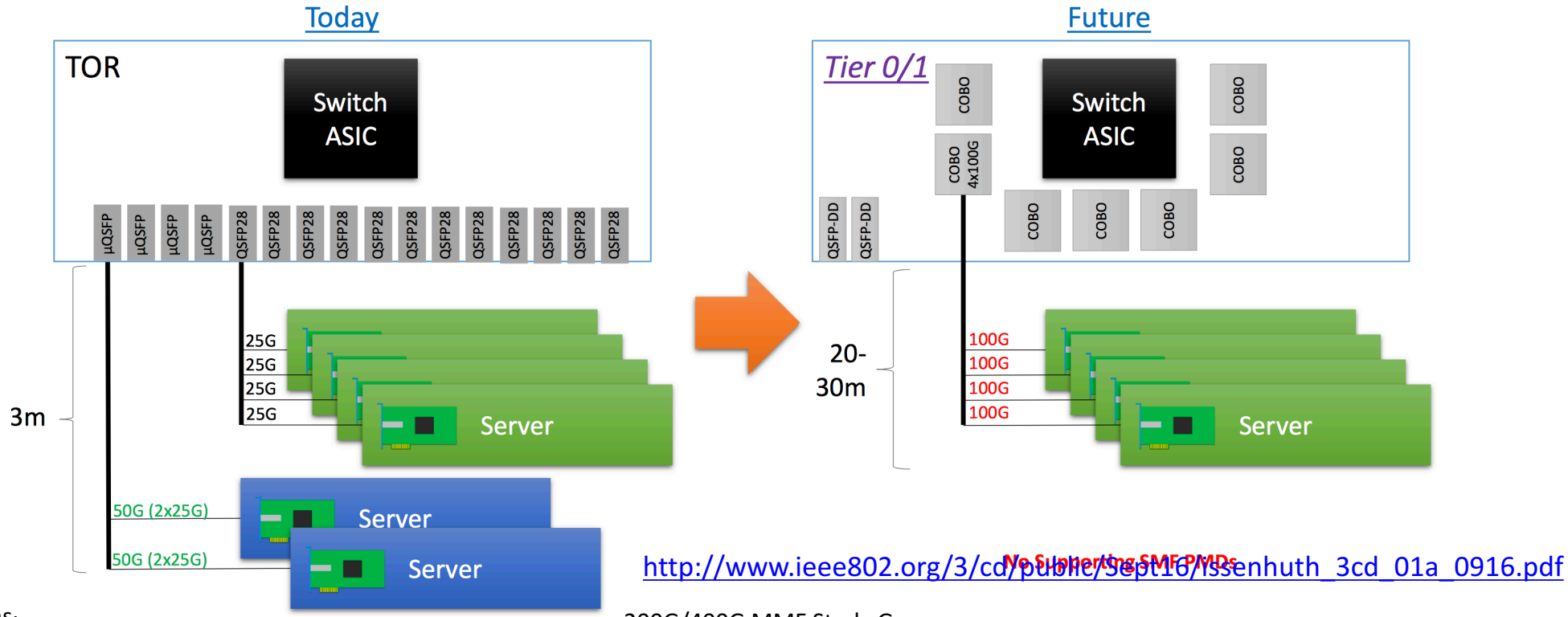
- **Switch radix over the last 9 years has increased from 64x10G, 128x25G, now to 256x50G, and likely to 256x100G by 2019/2020**
  - To mitigate full rack failure dual MOR switches may connect to each rack.



# Emerging Trend: Server Connecting to MOR Switch



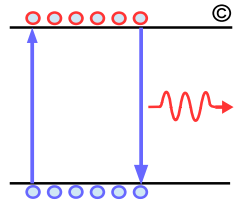
- **Microsoft evolution showing server directly connecting to MOR/Tier 0/1 switches as result of switch radix increase from 128 to 256 and fewer servers in a rack**
  - Passive Cu cable with reach limited ~1 m at 100 Gb/s/lane not very useful
  - AOC is an option but an octopus AOC is difficult to use instead an ultra-low cost 30 m PMD would be preferable.



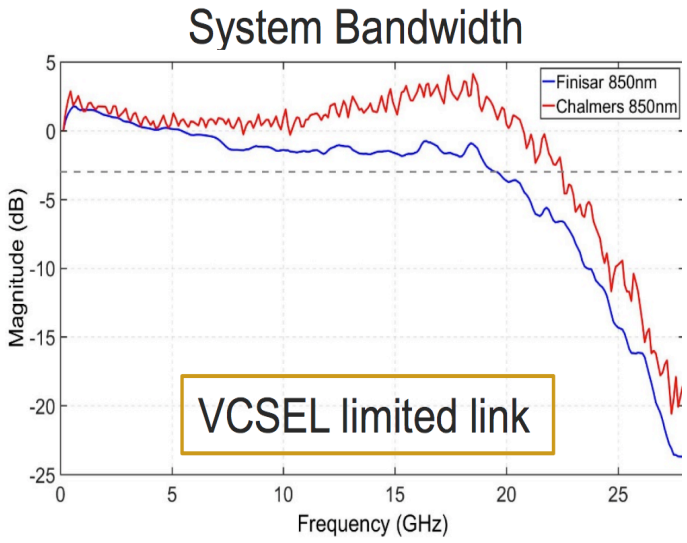

http://www.ieee802.org/3/cd/public/Sept16/issenhuth_3cd_01a_0916.pdf

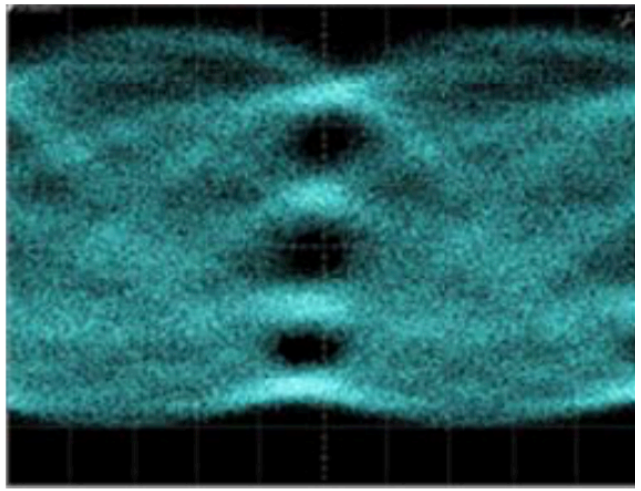
No Supporting SMF PMDs

# Early Proof of 100Gb/s VCSEL Technical Feasibility

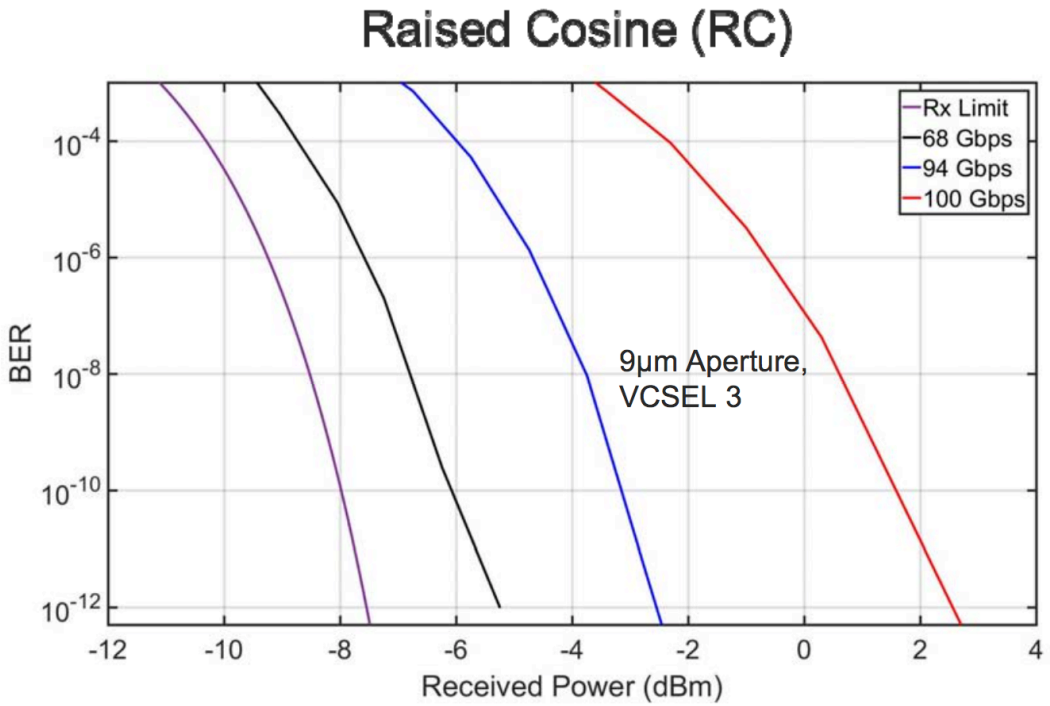


- **Used raised cosine pulse shaping with no receive equalization for 100 Gb/s VCSEL transmission, Stephen Ralph, ECOC 2017**
  - Results to the right are for Chalmers VCSEL
  - Arguably Finisar VCSEL with some compensation might be as good if not better!

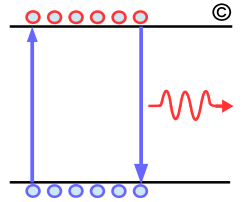


100Gbps RC



# Summary



- **Given that at 100 Gb/s/lane switches likely will be introduced in the next two years**
  - All MMF PMDs based on 50Gb/s/lane optics addressing leading edge data center would require a high cost inverse Mux/De-mux chip – defeating the purpose of using MMF
  - VCSEL/MMF PMDs based on 50G/lane would only address laggard datacenter networks and enterprise
- **With potential elimination of TOR the volume for optical links/AOC from servers to the 1st switch likely would as large total data center optics volume!**
- **Next-generation 200 Gb/s and 400 Gb/s MMF study group should consider defining MMF PMDs based on 100 Gb/s/lane to address 100 Gb/s/lane switch introduction**
  - A shorter reach 30 m objective could be within time frame of study group by using existing ~23 GHz VCSELs to address MOR-servers and cluster/spines applications
  - A longer reach ~70 m maybe feasible but would require faster VCSELs that don't exist yet but likely outside time frame of current study group
  - Not having 100G/lane MMF PMDs in 2020 means MMF deployment will further decline in favor of DR/DR4 and AOCs in sub-100 m applications
- **100 GbE is necessary for next generation servers but currently not in the scope of the project**
  - One option is to at least add an informative annex how to repurposes 400GBASE-SR4 for 100GBASE-SR applications or do a CFI to expand the scope.