ToR Switch Architectures and Implications for 100G Electrical Lane Interfaces

Rob Stone

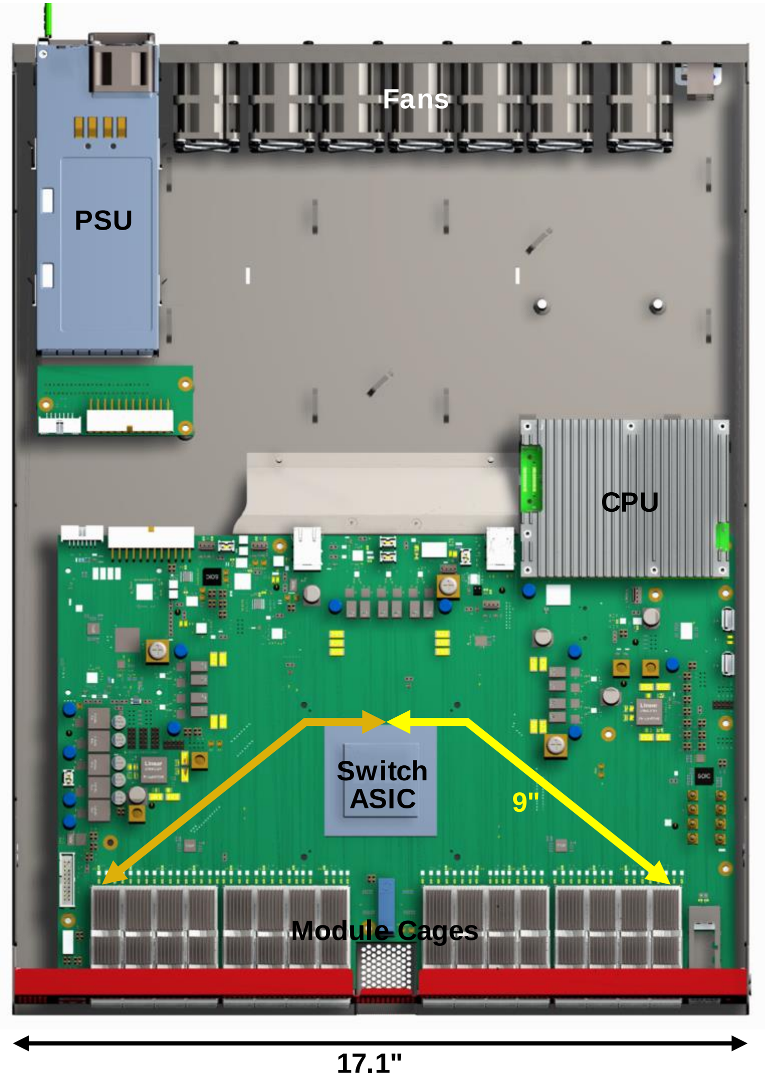
BROADCOM

100GEL Study Group, Chicago 2018

Recap

- Historically the ToR switch serves several purposes:
  - Aggregation of server IO using low cost DAC cables
  - Enable oversubscription (i.e. more server bandwidth than optical uplink bandwidth) to minimize optics spend
- Highly cost sensitive
  - More ToRs than other class of switch in a datacenter
  - DAC has been the favored server attach media as it has provided the lowest cost per bit
- ToR design is often multi-purpose
  - ToR box can be used as an all optical switch (aggregation or spine)
- ToR bandwidth needs to be right-sized to total server rack bandwidth + uplink bandwidth
  - End users don’t want to pay for unused bandwidth

Typical ToR



10, 25 and 50G / lane generation ToRs have the following characteristics:

- Generally a single switch ASIC per box, 1 RU
- Every port is universal
  - DAC, MMF, SMF optics - ***compatible host loss budgets***
- Power and cost optimized
  - No additional components (gearboxes, retimers)
- ~ 9" longest trace to most distant module
  - Historically OK to do this without a retimer for both DAC and VSR channels at 10, 25 and 50G / lane
- Switch lane speed is matched to server lane speed
  - Eliminates any gearboxing required to match server IO (drives cost and power)

How does 100G DAC fit within ToR application space?

- Useful ToR bandwidth is set by oversubscription ratio (Uplink : Downlink BW), server speed, and number of servers per rack
- Downlink bandwidth is equal to server bandwidth
- Number of servers may be rack power limited
  - high-end servers are ~ 365 W each[1], rack power limit ~ 15 kW

Total ToR Bandwidth (Tb/s), 1:3 OSR						
Server Bandwidth (Gb/s)	Servers per Rack					
	18	24	32	48	64	128
25	0.6	0.8	1.1	1.6	2.1	4.3
50	1.2	1.6	2.1	3.2	4.3	8.5
100	2.4	3.2	4.3	6.4	8.5	17.0
200	4.8	6.4	8.5	12.8	17.0	34.0
400	9.6	12.8	17.0	25.5	34.0	68.1

Total Server Power (kW) / Rack	6.6	8.8	11.7	17.5	23.4	46.7

Addressed with current techology at 50G / lane and below

▨ Rack Power Limited

[1] 2016 United States Data Center Energy Usage Report
http://eta-publications.lbl.gov/sites/default/files/lbnl-1005775_v2.pdf

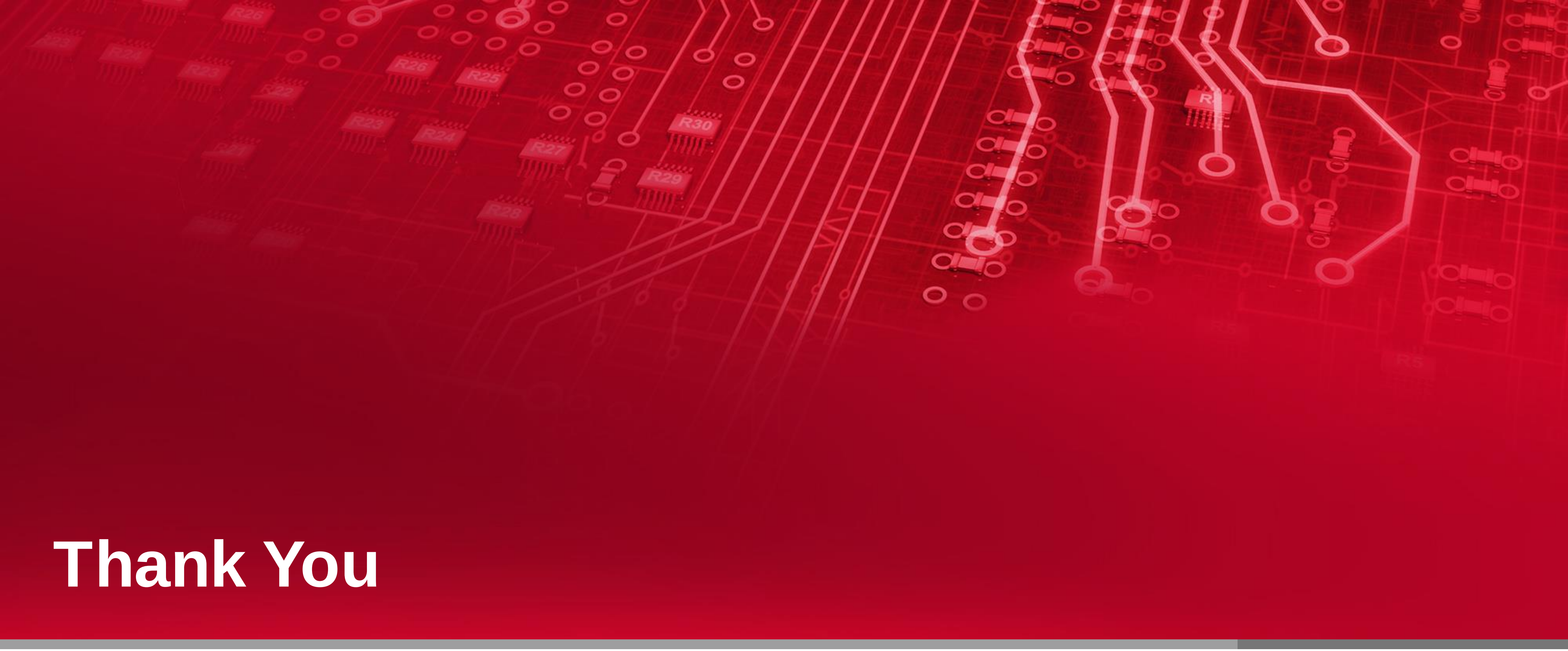
Should we specify DAC to support a shorter host channel?

- Several presentations on this option
  - *lim_100GEL_adhoc_01_022618.pdf, haser_100GEL_adhoc_01_022618.pdf, mellitz_100GEL_adhoc_01_021218.pdf, tracy_100GEL_01_0318*
- Will force use of retimers for many ToR ports
  - No longer DAC in the true sense!
  - Higher cost for these systems will negatively impact BMP, or drive architectural transition to EoR architectures
- Server side doesn't require 100G / lane IO
  - Not IO pin limited like the switch ASIC side of the link
  - Example: Today can support up to 48 x 200GE (4x50G) servers with 1:3 oversubscription on current technology
  - Unlikely servers will move to 100G / lane unless economics are favorable to do so

100G / lane DAC Broad Market Potential

Suggested Requirements for Success

- Requires 100G / lane Servers
- Support a “Universal Switch Port”
  - Requires 9” host PCB traces
  - Don’t sacrifice C2M budget for optical modules to support DAC budget – doing so increases the power for all optical spine and EoR switches!
- Ensure DAC continues to provide a total low cost solution
  - Fully passive, no gearbox or retimers
- Minimum 2 m reach
  - See *goergen_100GEL_01_0318*



Thank You