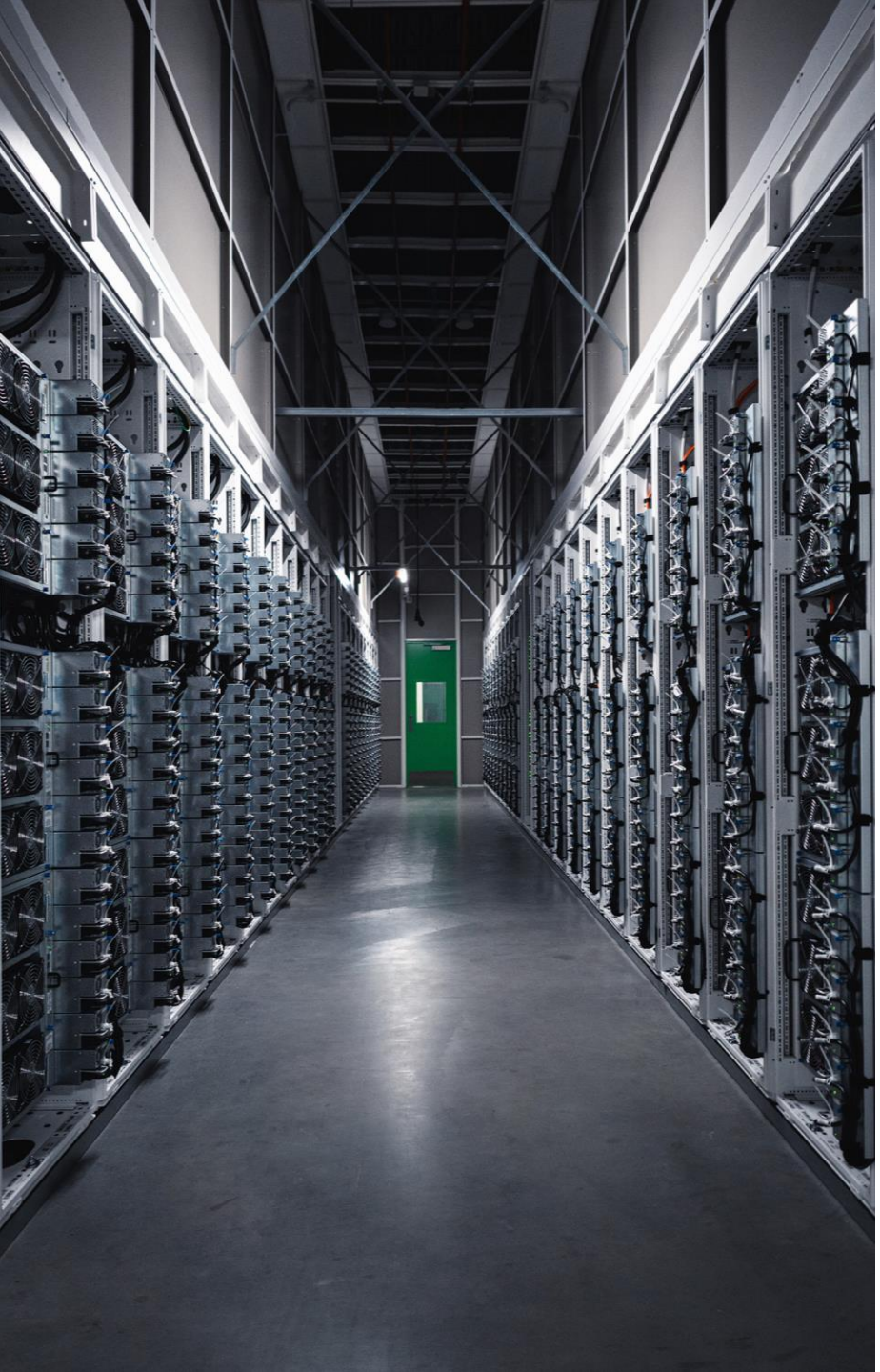
The race to 800G: a reality check

Mark Filer

Principal Engineer, Azure Hardware Architecture (AHA)

Microsoft Azure



Microsoft Azure

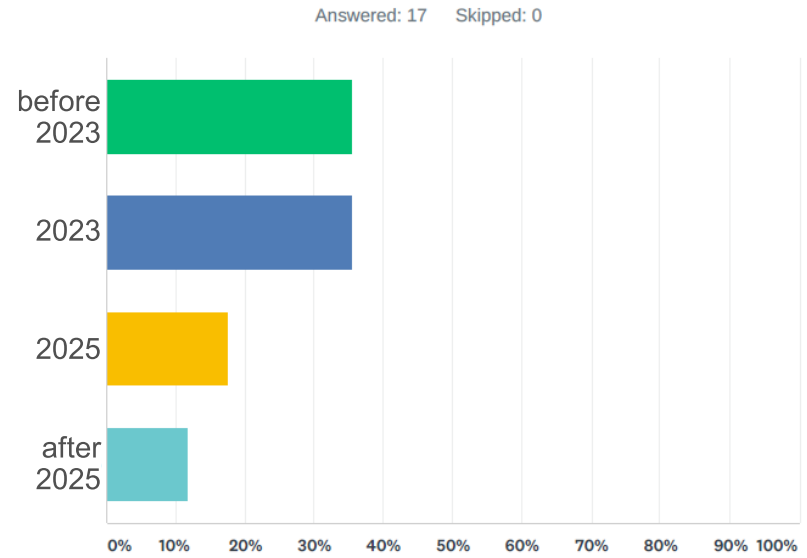
Disclaimer – Statements of Future State

Material does not necessarily represent opinions of Microsoft and certainly cannot be construed as any form of commitment by Microsoft towards pursuing concepts described herein

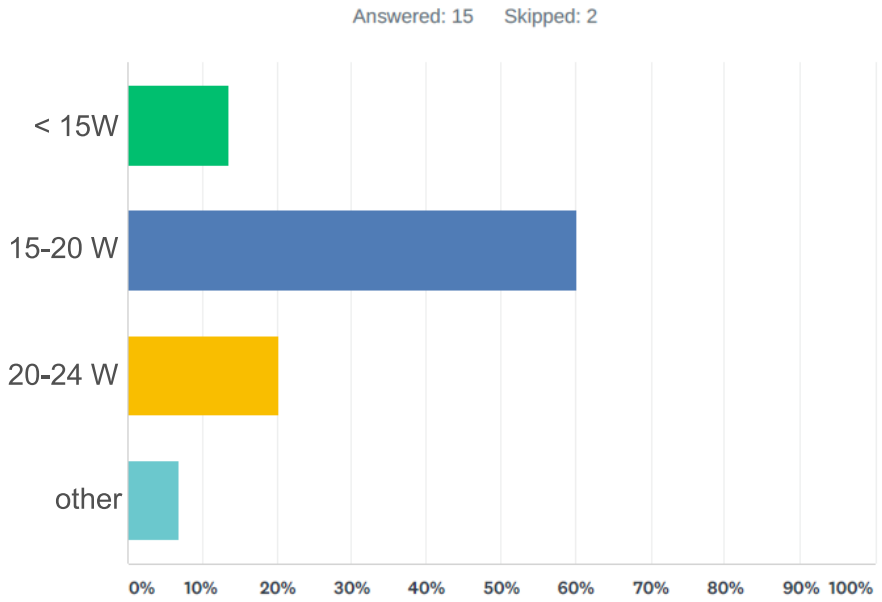
Is there demand for 800G?

- Yes!
- 2020 OIF Network Operator survey on *Beyond 400G**
- But: **power**…

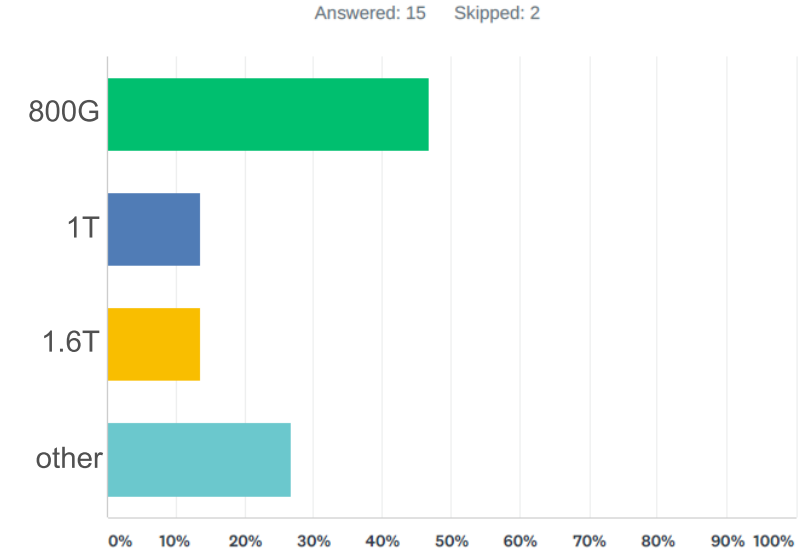
Q8 When are you expecting to need a standard coherent line side with a capacity larger than 400G/lambda?



Q11 What is the maximum electrical power per port you anticipate supporting?



Q9 What router port speed do you plan to deploy after 400G?



* OIF contribution oif2020.094

Microsoft Azure

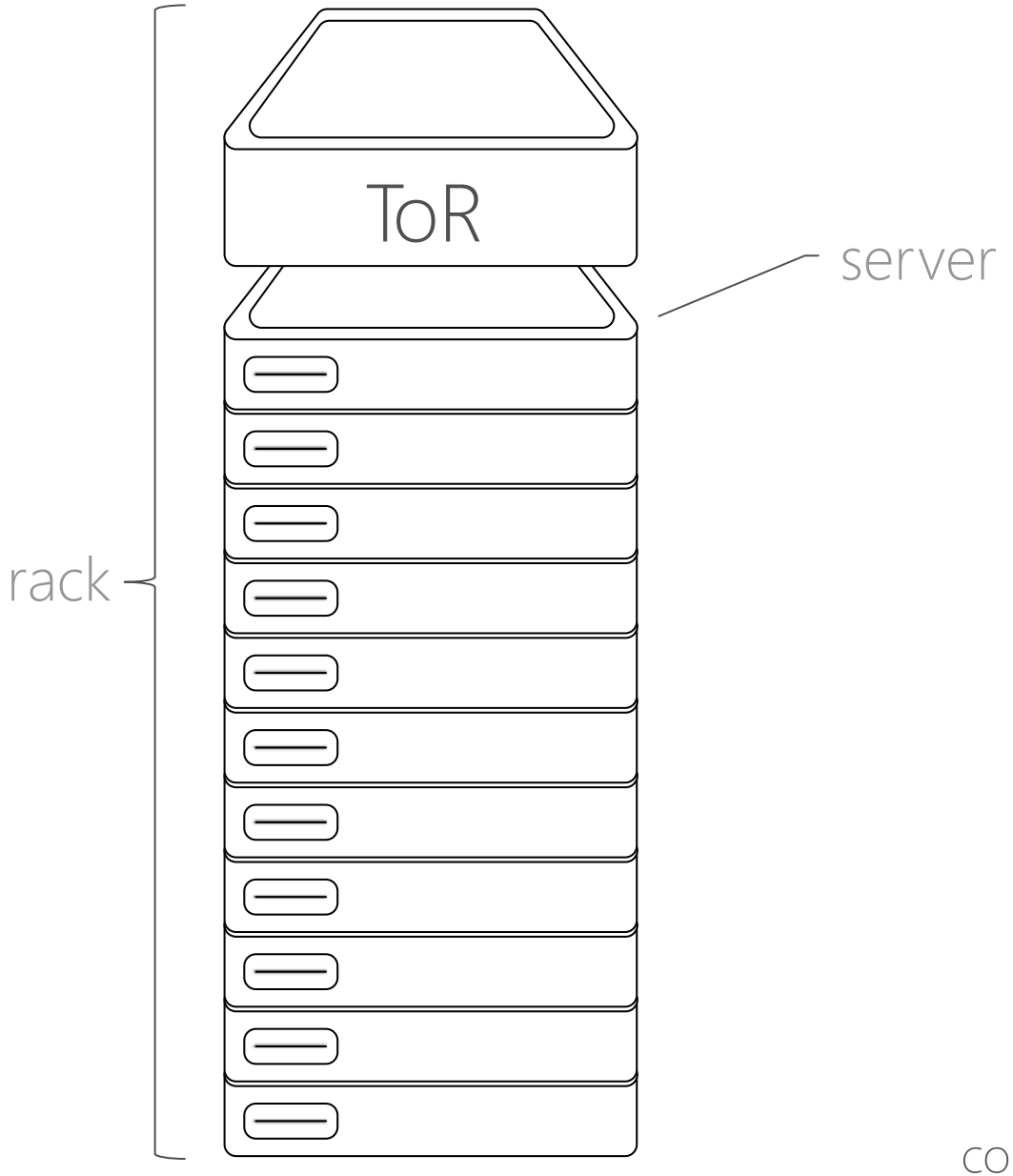
Retrospective: 100G to 400G

How do we build data center networks today?

DC architecture

Rack

- 10s of servers/rack
- server to ToR via < 2m DAC cables @ 50G

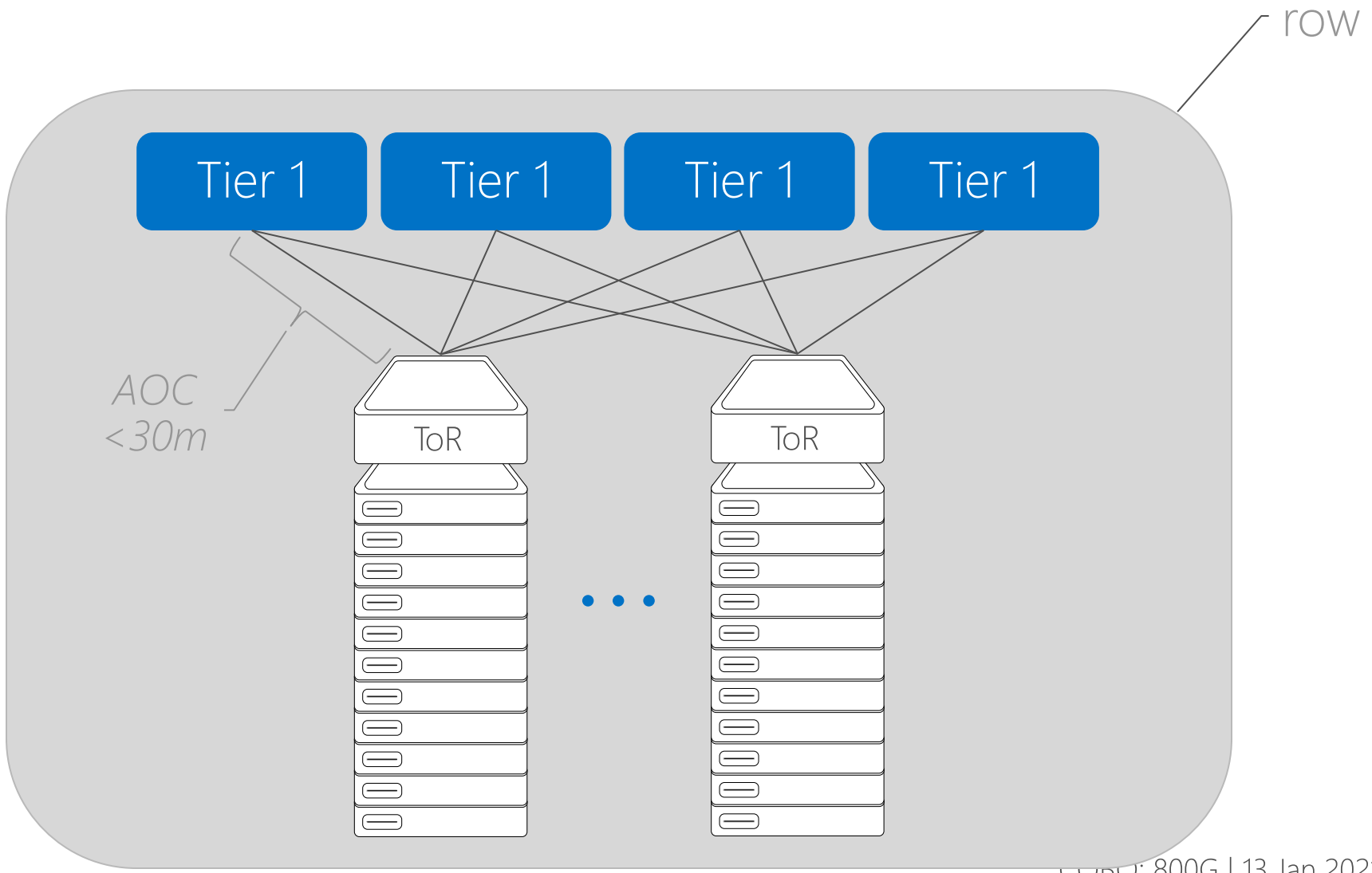


Microsoft Azure

DC architecture

Row

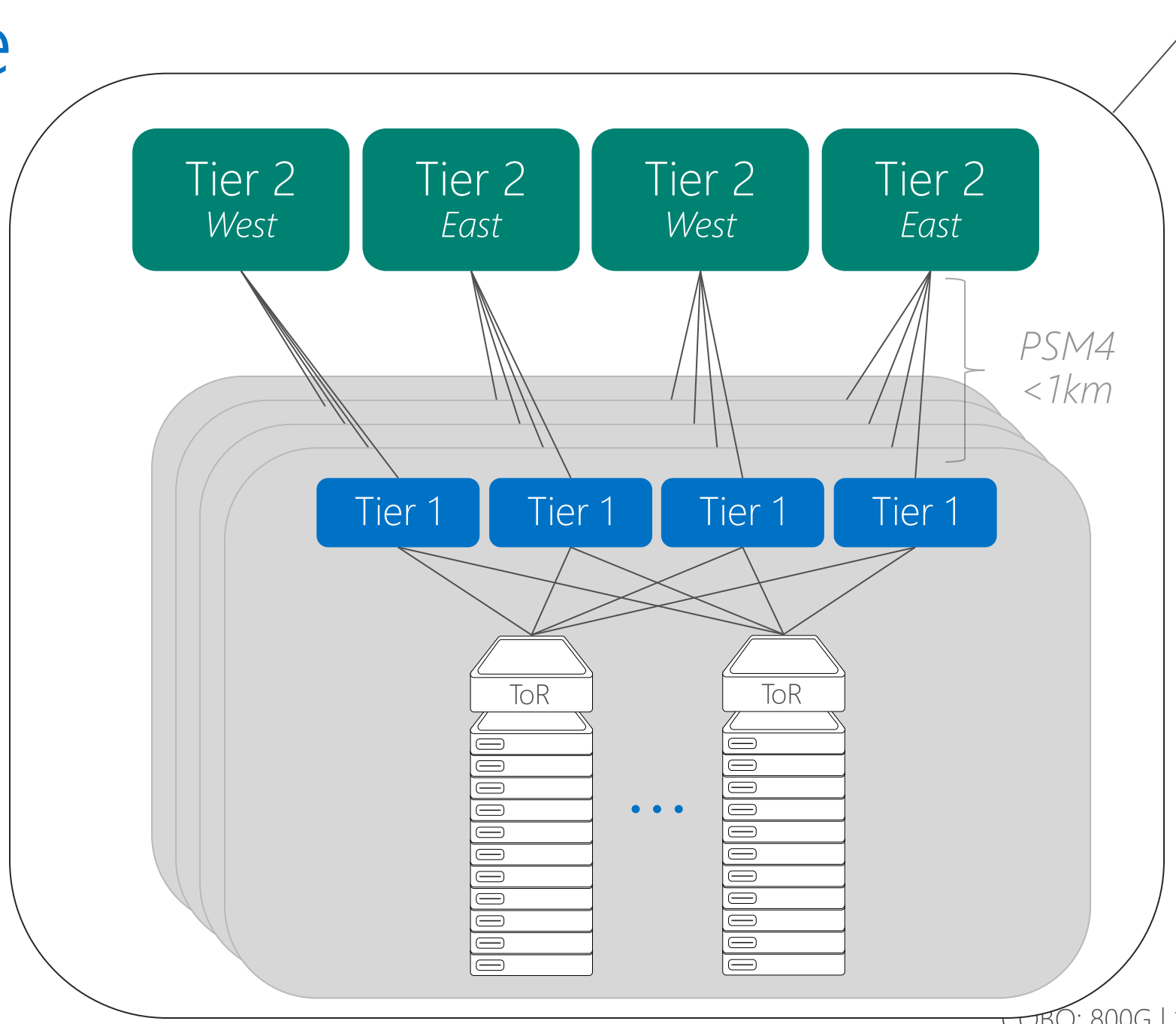
- 10s of racks / row
- ToR to Tier 1 clos fabric via < 30m AOC @ 100G
- 100G AOC power = 2.0-2.5 W



DC architecture

Datacenter

- “lots” of rows / DC
- Tier 1 to Tier 2 connected < 1km with parallel fiber via PSM4
- massively parallel Tier 2
- 100G PSM4 / CWDM4 power = 3.0 – 3.5 W

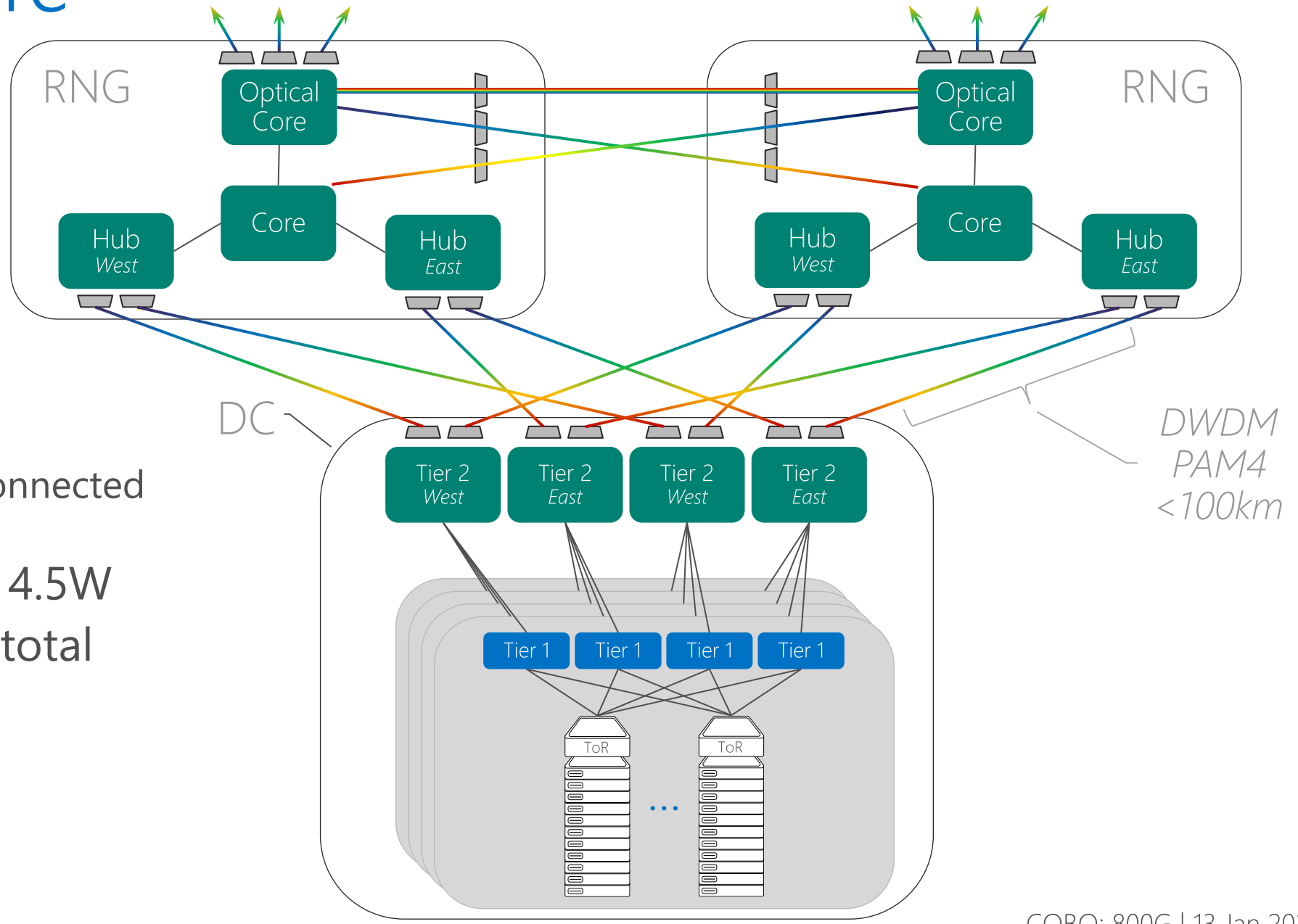


Microsoft Azure

DC architecture

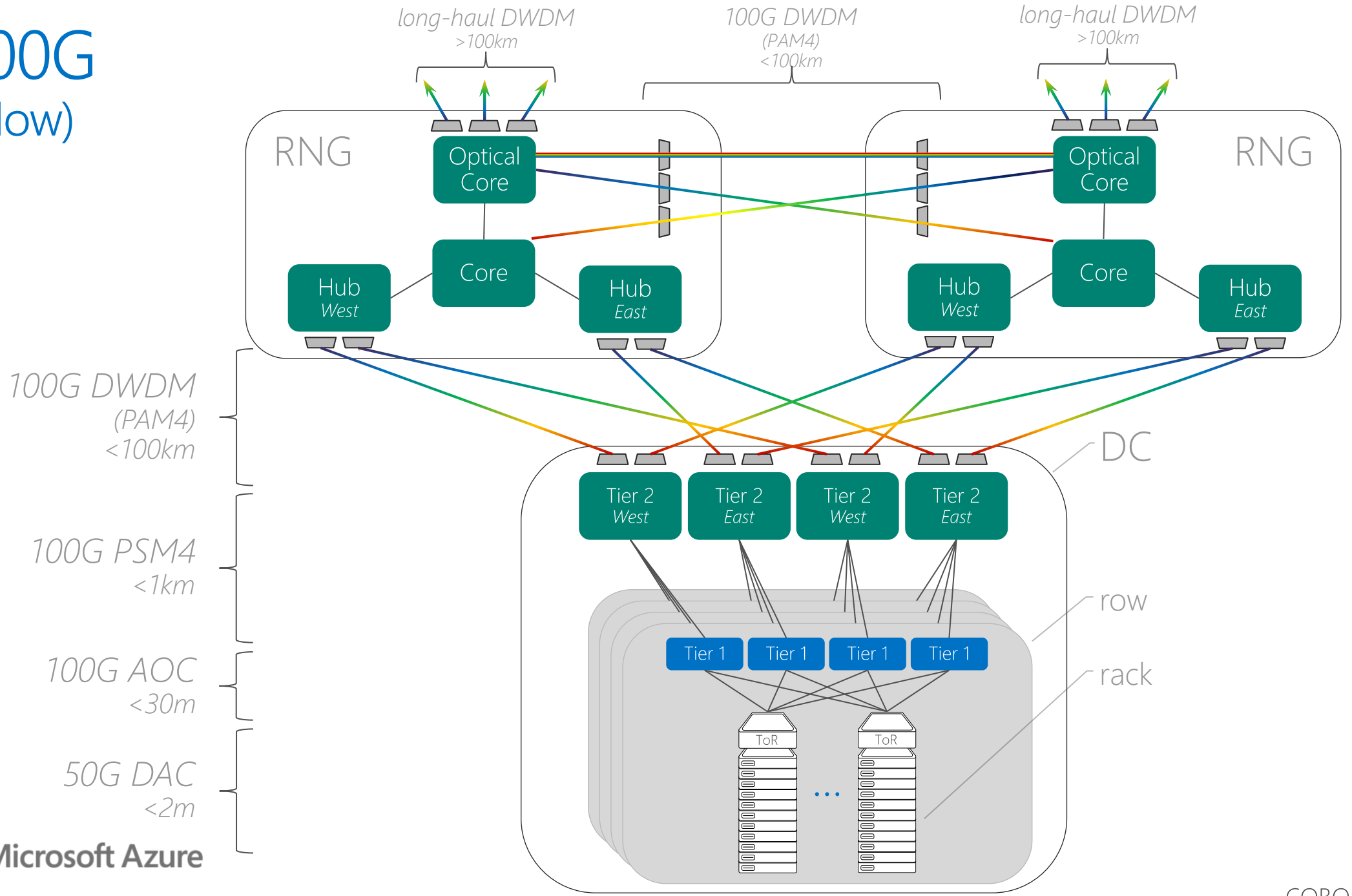
Region

- "lots" of DCs / region possible
- DC-RNG and RNG-RNG connected ≤ 100km via DWDM PAM4
  - Some campus builds connected via bulk fiber (< 2km)
- 100G PAM4 power = 4.5W
- single percentage of total server BW in DCI



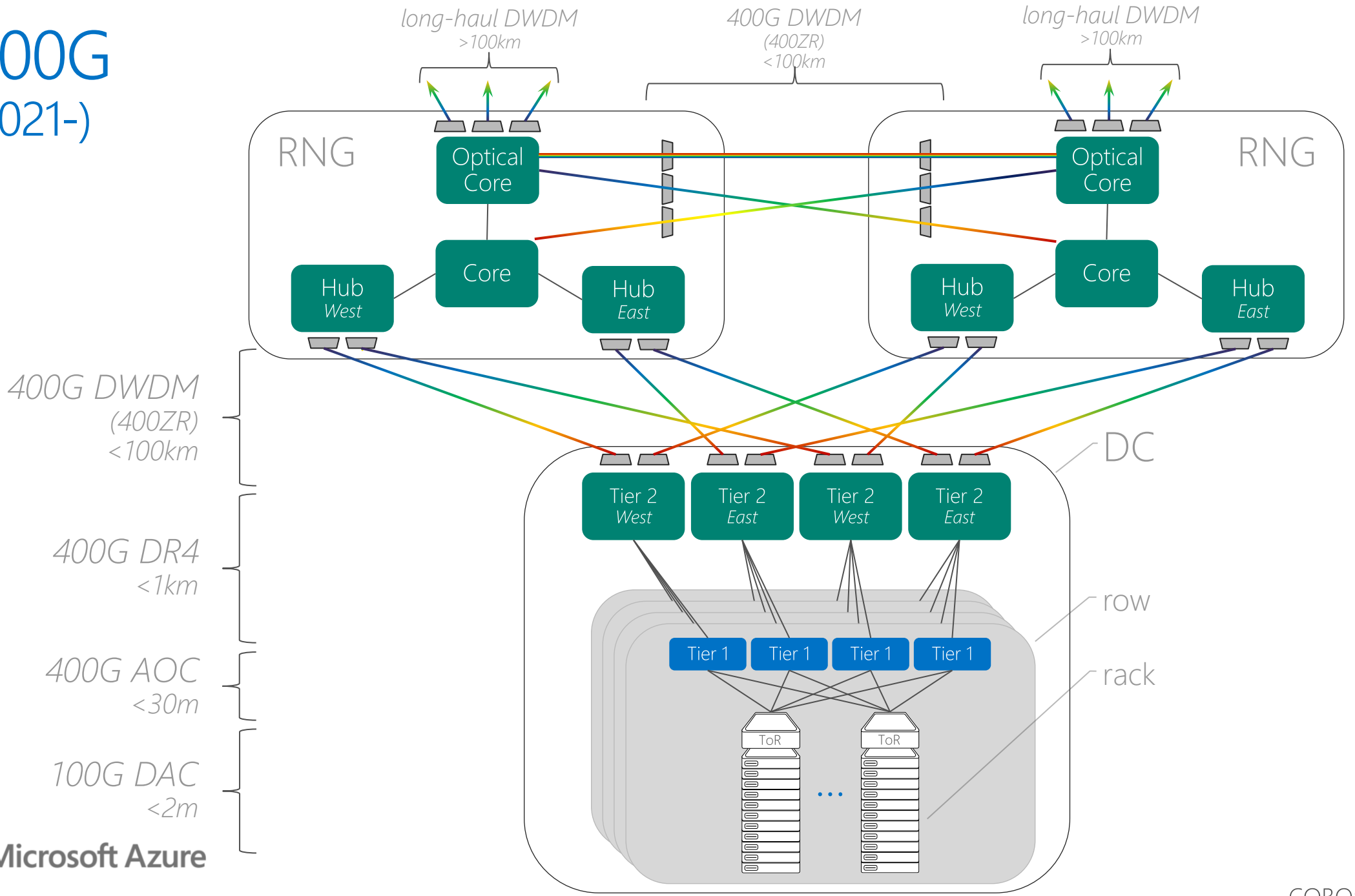
Microsoft Azure

100G
(Now)



Microsoft Azure

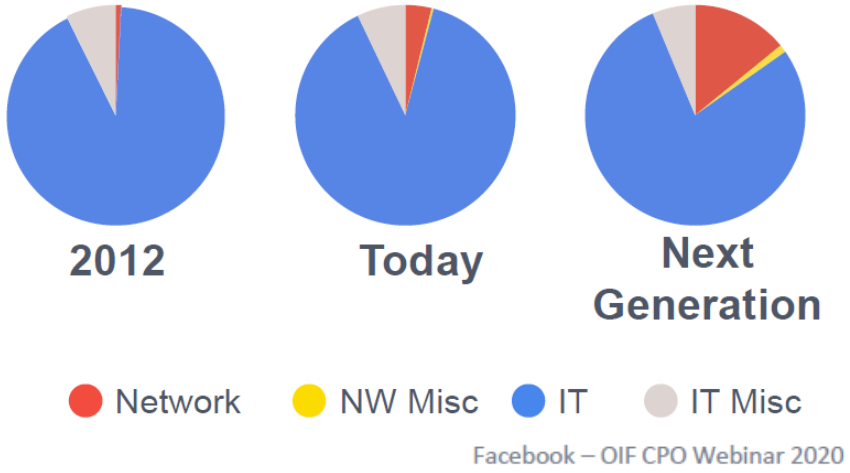
400G
(2021-)



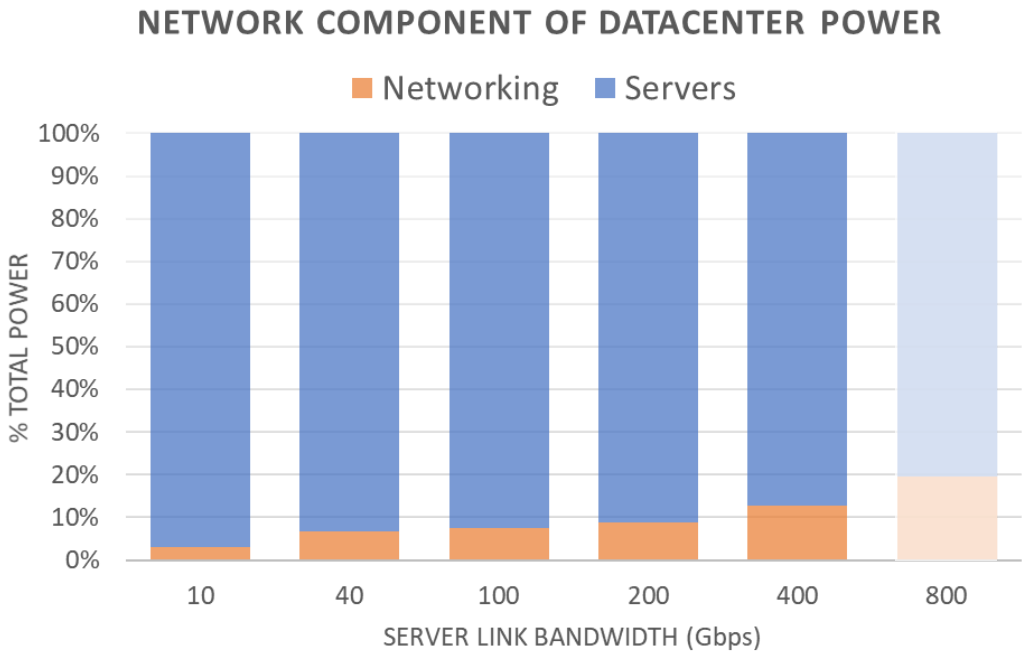
Microsoft Azure

Elephant in the room… Power

- Equipment power consumption at 400G is already problematic!
- Switches projected @ 3x power of 100G
- Optics projected @ 3-4x power of 100G
- Challenges power envelopes of facilities
- Uses power that could be generating revenue (lost server capacity)
- Costs $$$ and not green
- Trajectory makes transition to >400G appear all but impossible



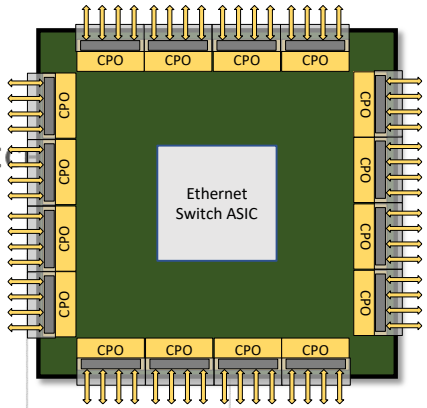
Facebook – OIF CPO Webinar 2020

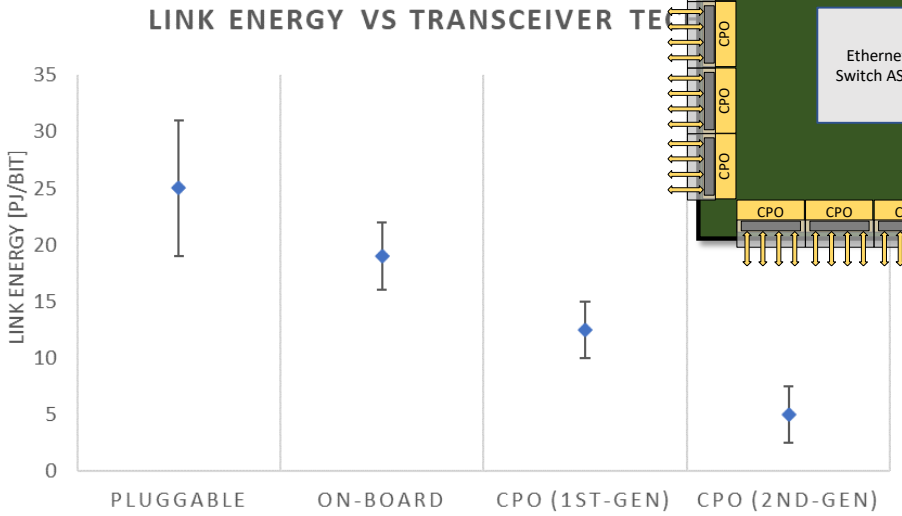


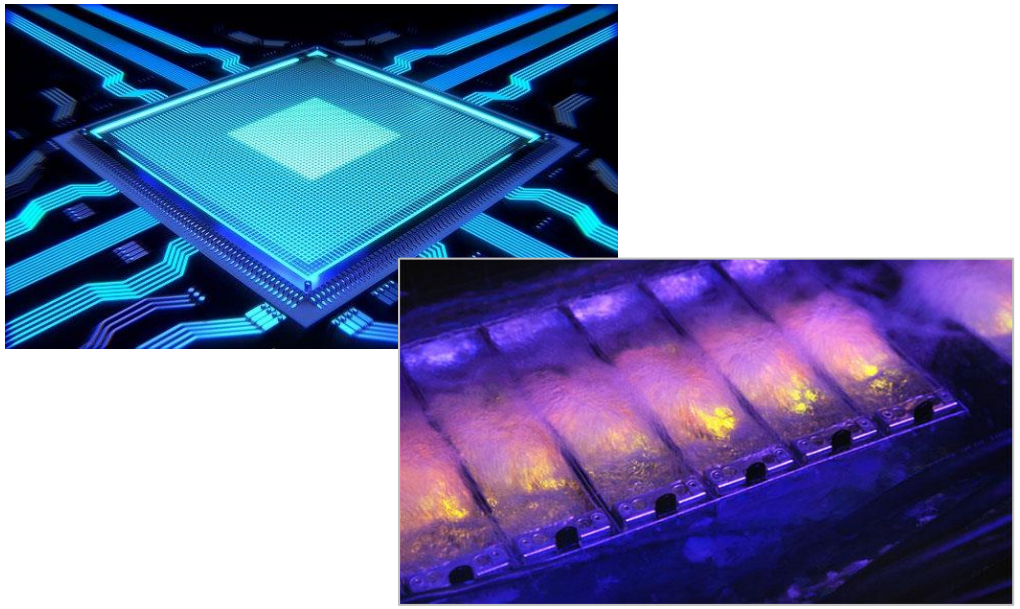
Microsoft Azure

Possible Solutions

- Photonics
  - Co-Packaged Optics (CPO)
  - Novel optical approaches
- Network architecture + HW changes
  - Collapsed tiers with multi-homed NICs (fanning out horizontally)
  - Simplified forwarding requirements → cooler ASICs
  - Additional integration, e.g. encryption on switch ASIC
  - Liquid cooling

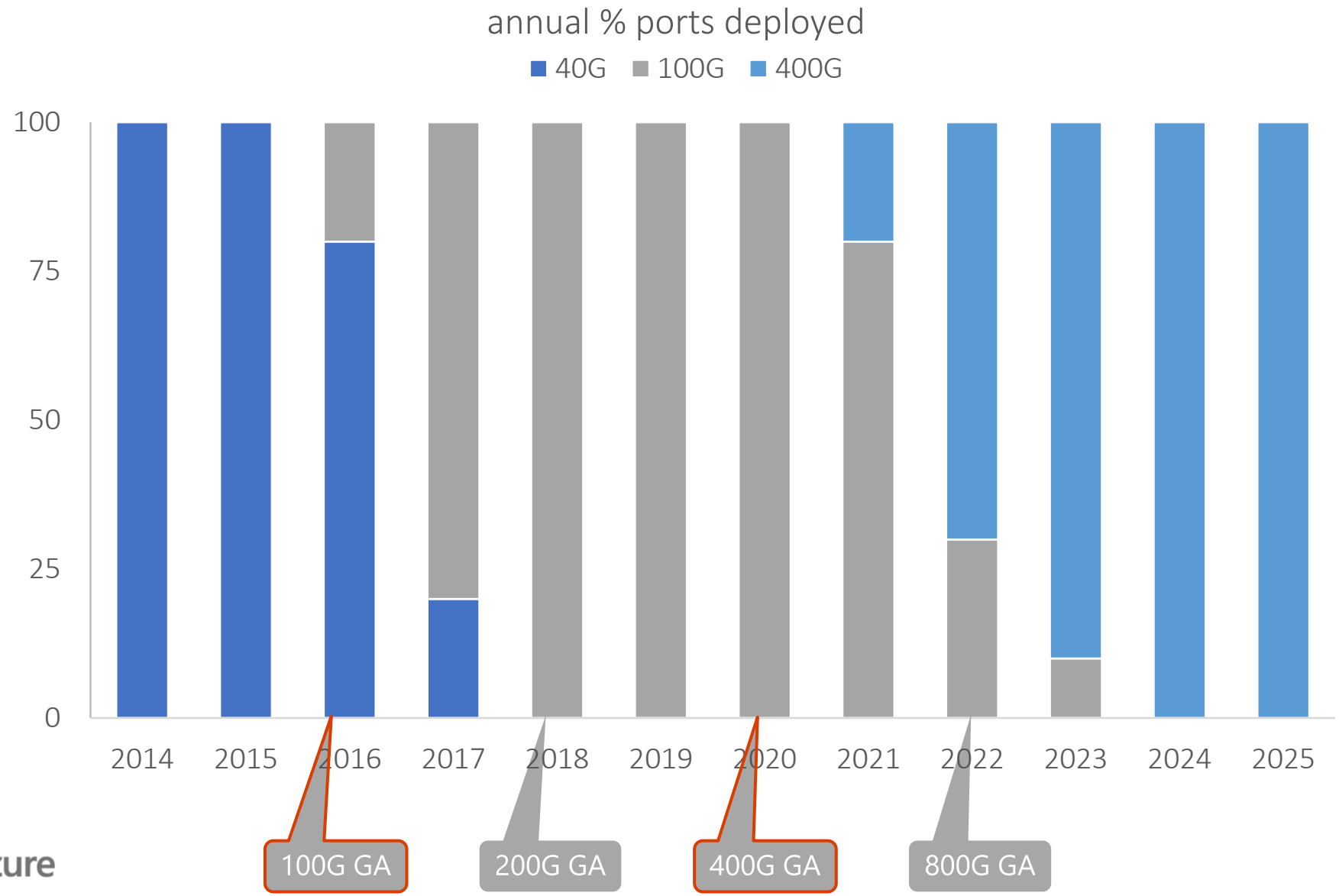






Takeaway: we can't just keep scaling link bandwidths… "next gen" systems will require all of the above

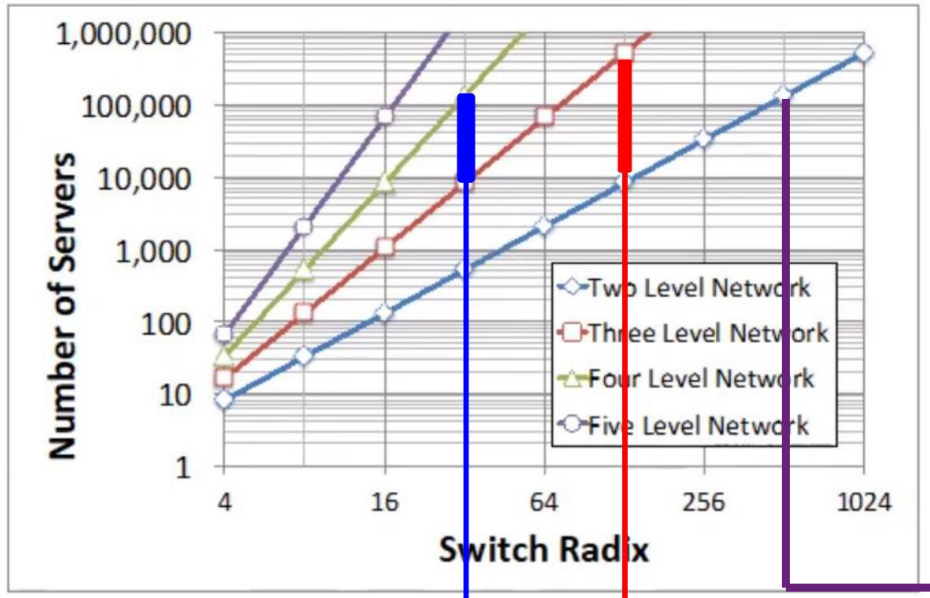
Microsoft DC ecosystem technology life cycles

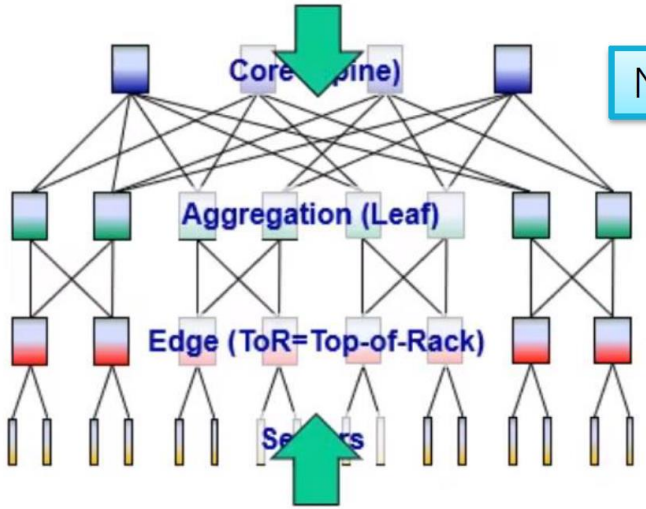


Microsoft Azure

Radix argument revisited

Why do we care about the datacenter radix?





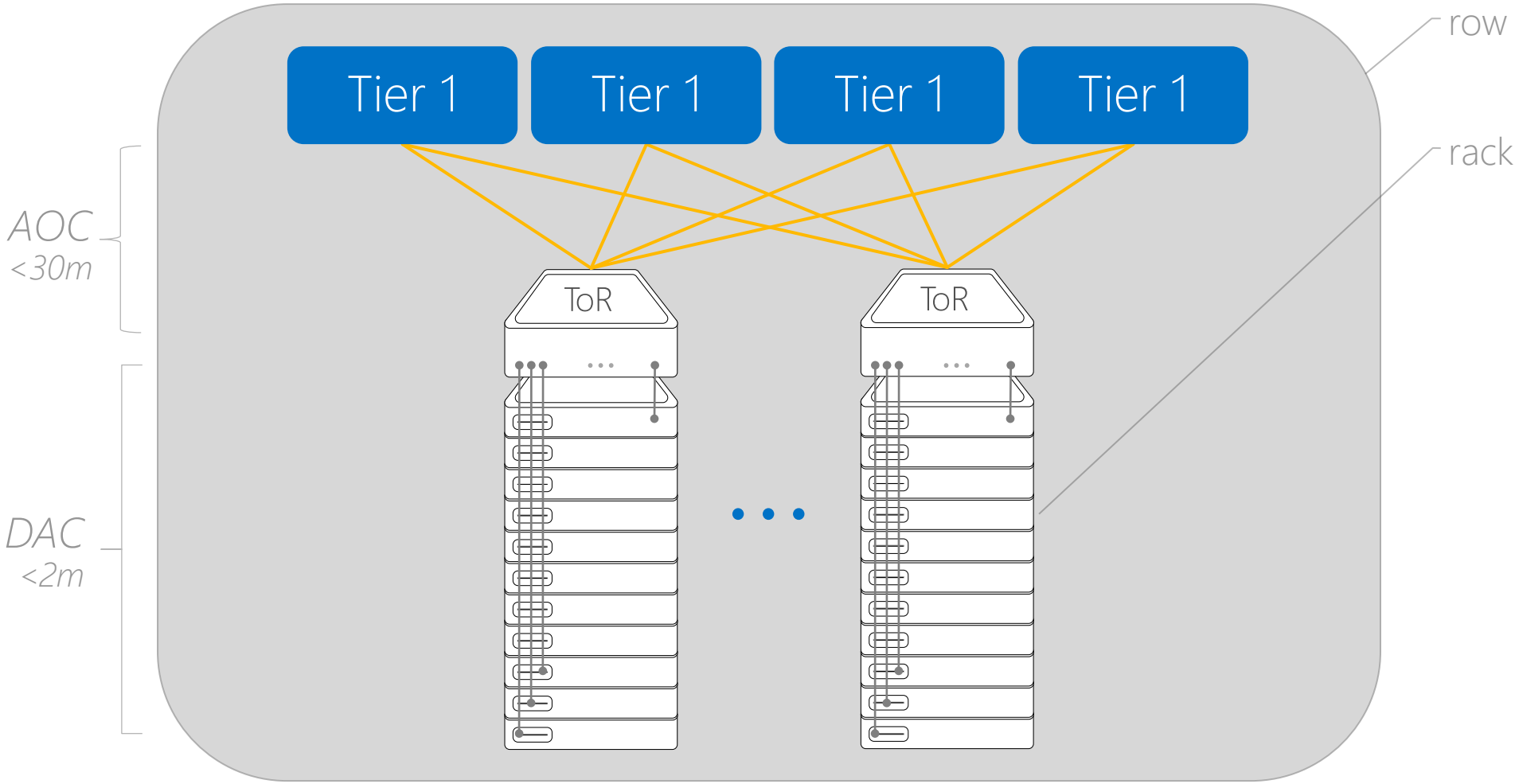
Network flattening

100G is native electrical lane speed
1 MAC = 1 elec lane = 1 optical lane

-Source: Facebook

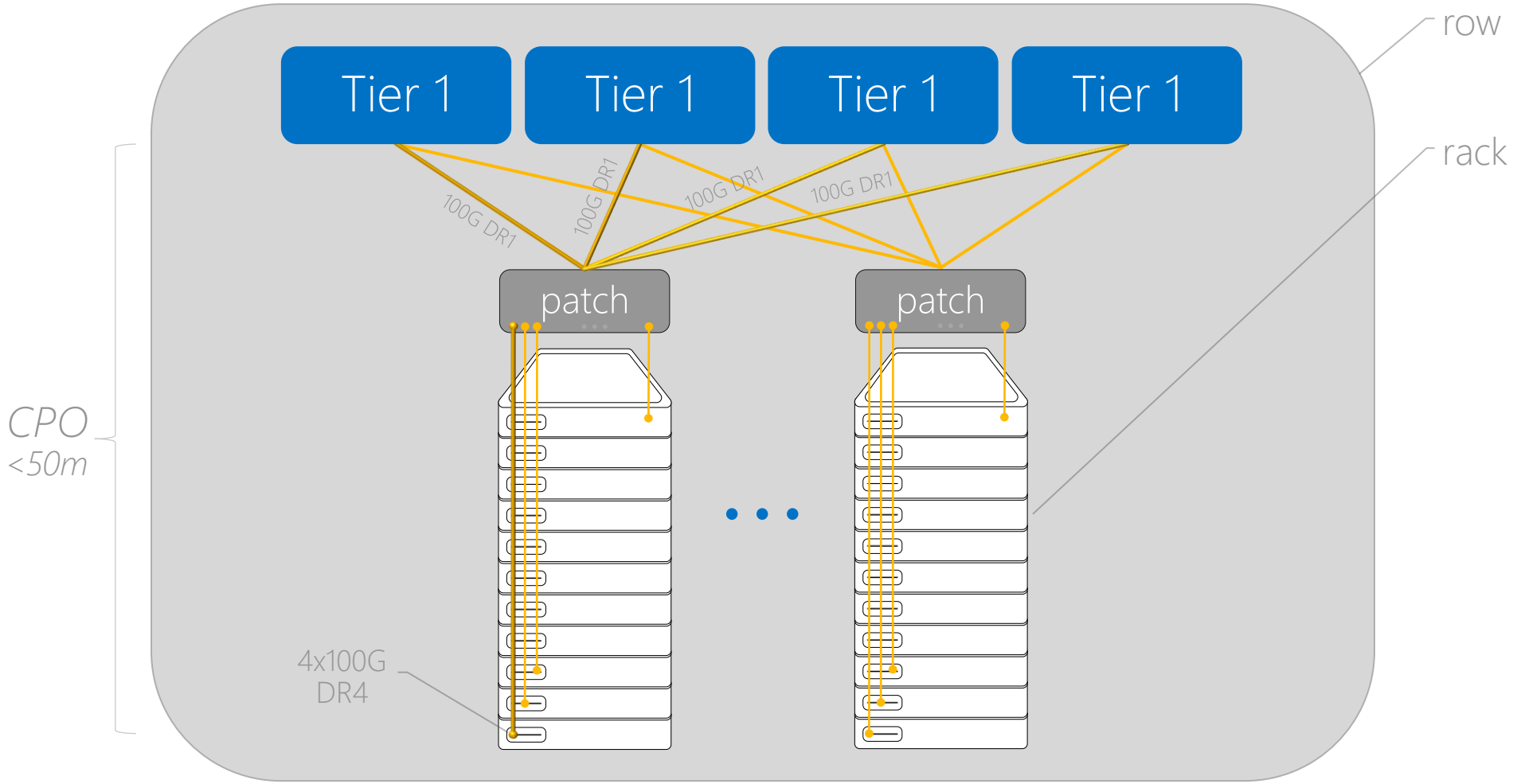
Switch Generation	Radix = 32	Radix = 64	Radix = 128	Radix = 512
12.8T	400G	200G	100G	25G
25.6T	800G	400G	200G	50G
51.2T	1.6T	800G	400G	100G

Server-ToR-Tier1 topology



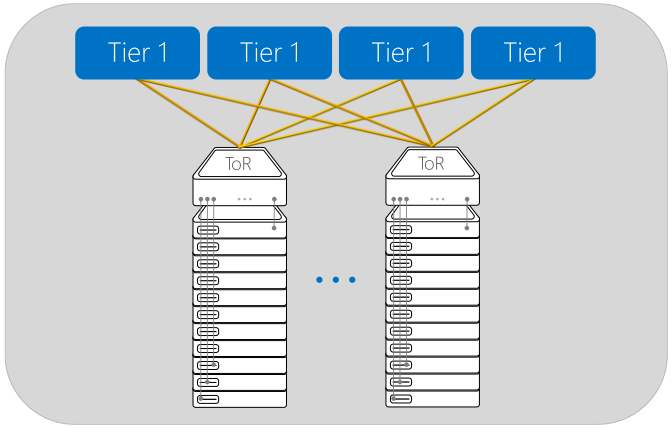
Microsoft Azure

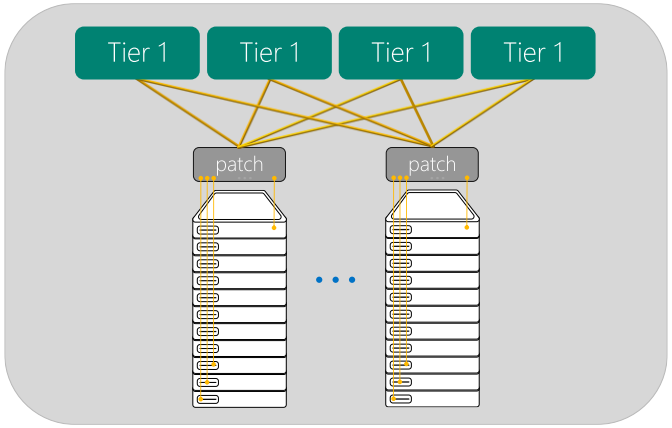
ToR bypass – multi-homed NIC



Microsoft Azure

ToR bypass efficiencies (100G lane speeds)





	Tier1-ToR-server	**ToR-bypass**
failure domain	ToR is SPOF for rack	no SPOF – multi-homed NIC
switch ASIC count	4X-8X	1X
switch space + power	baseline	reduced space and ~**1/3 power**
switch radix	can't leverage higher radix chips (stranded capacity at ToR)	can leverage full switch radix (multi-chip T1 box)
oversubscription	3:1 typical	fully non-blocking in row
reach limits	DAC < 3m; AOC < 30m	1m-2km+

Microsoft Azure

Summary

- Power is the main limiter for “beyond 400G” data centers
- We can’t continue to simply scale link bandwidths while building networks exactly as we do today
- Historical ecosystem life cycles would indicate we won’t be ready for “800G” when the industry is (32x100G CPO will suit our needs better)
- 100G electrical lanes will be a foundational building block for power-efficient data center designs for the foreseeable future
- Future data center networks will require a combination of photonic innovation (e.g., CPO), optimized network architectures, and advanced hardware implementations

Thank you.

mark.filer@microsoft.com