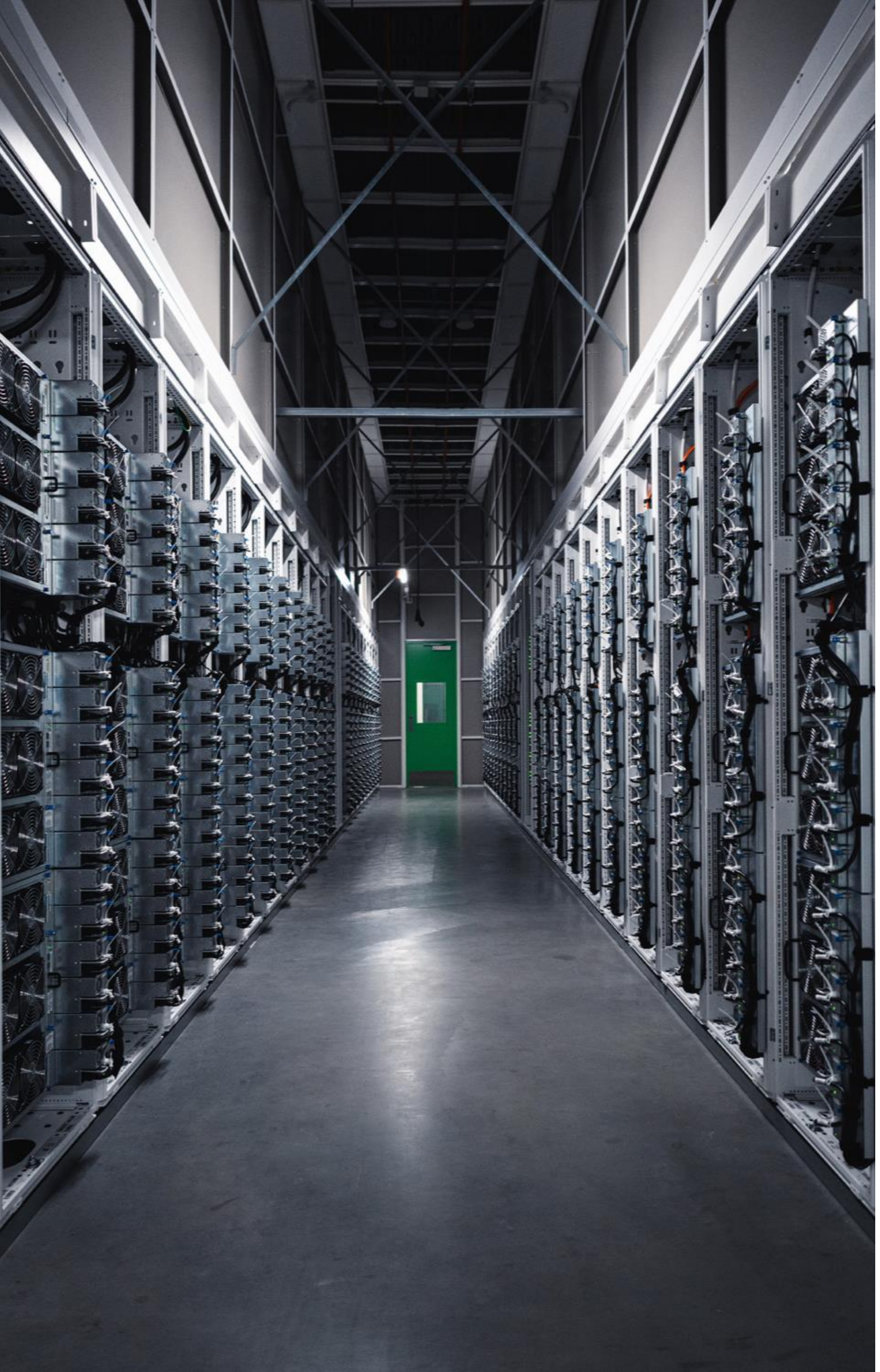
What’s beyond 400G ?

Mark Filer

Principal Engineer, Azure Hardware Architecture (AHA)

Microsoft Azure

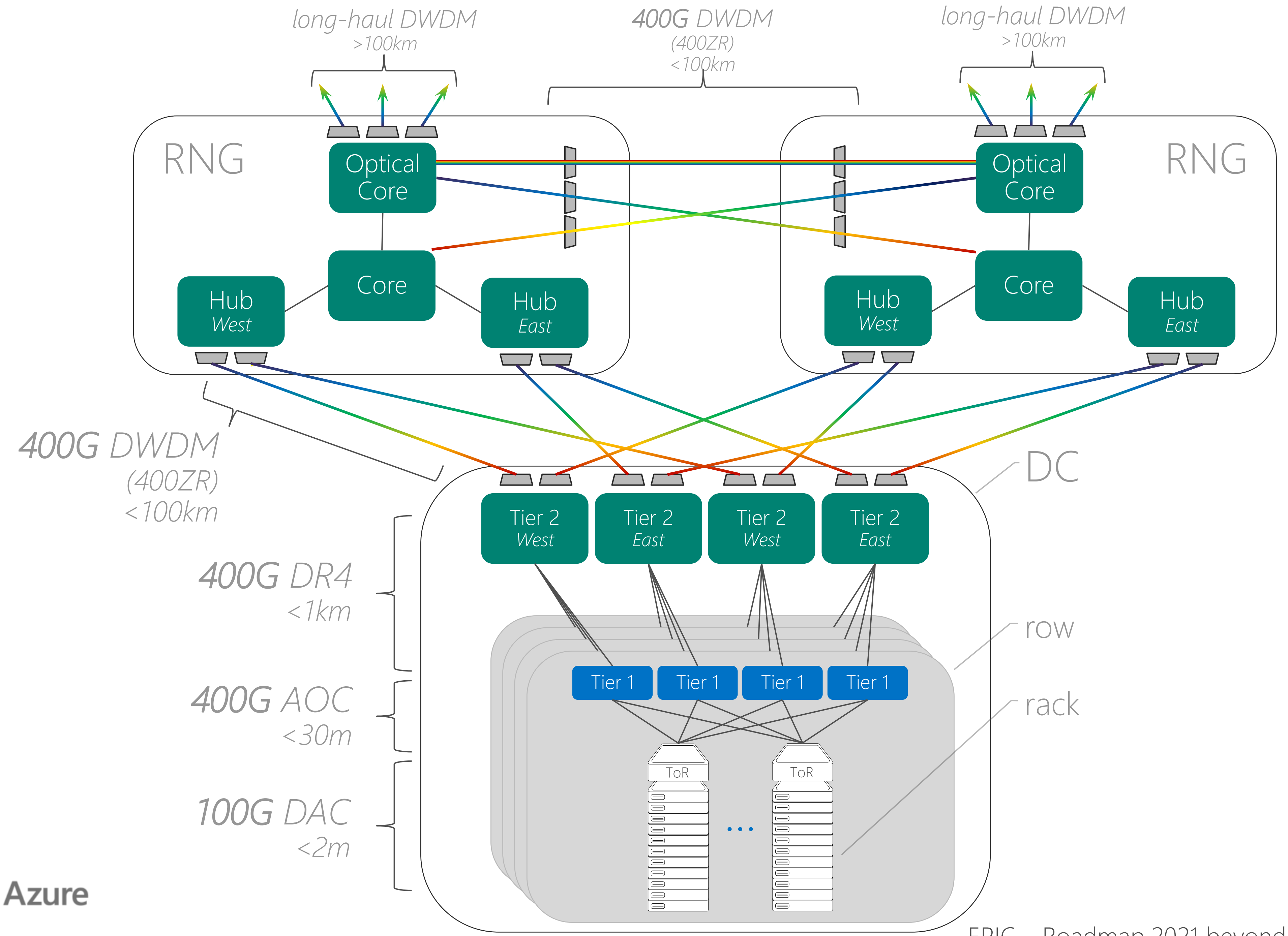


Microsoft Azure

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

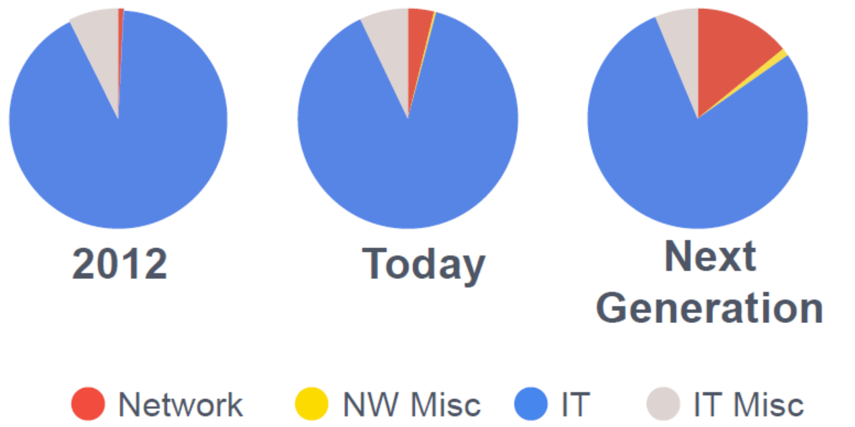
400G
(2021-)



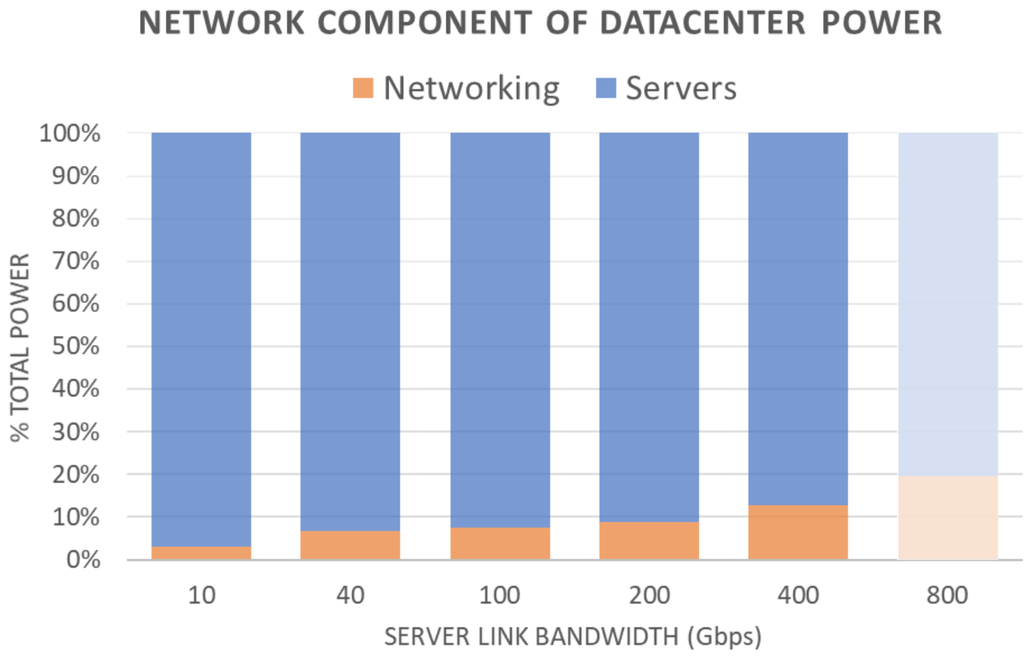
Microsoft Azure

Power limits future DC scaling

- 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

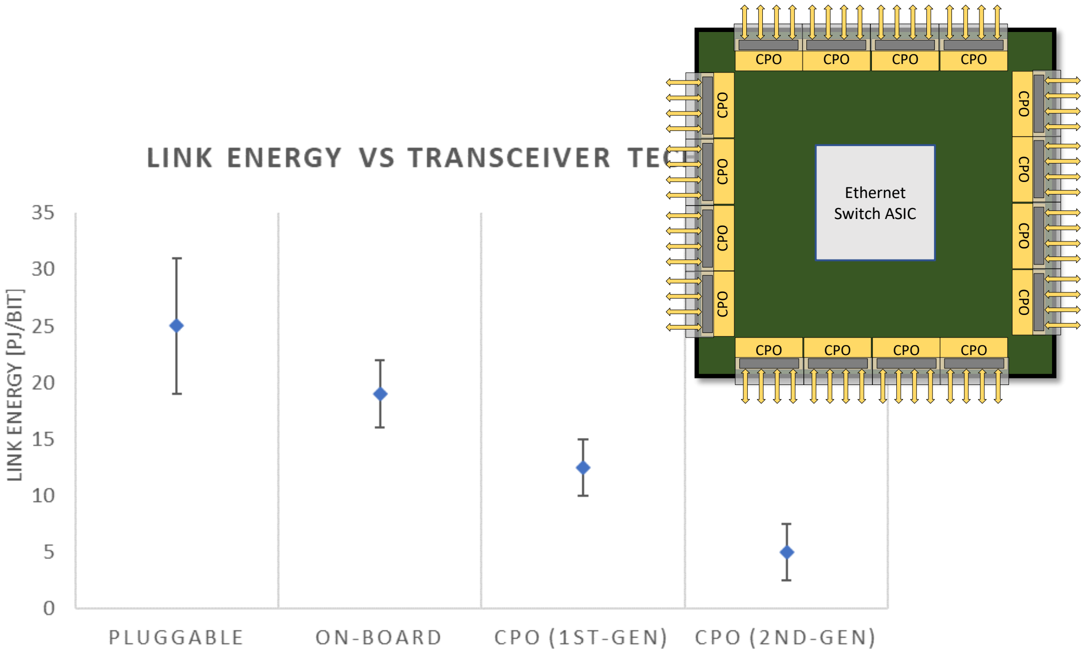


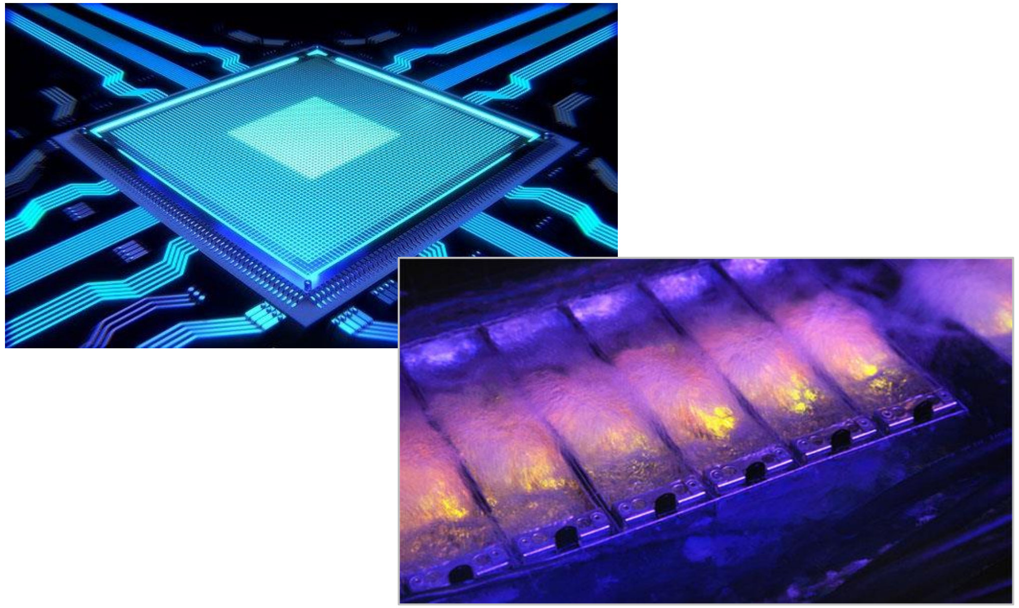
Facebook – OIF CPO Webinar 2020



Possible Solutions

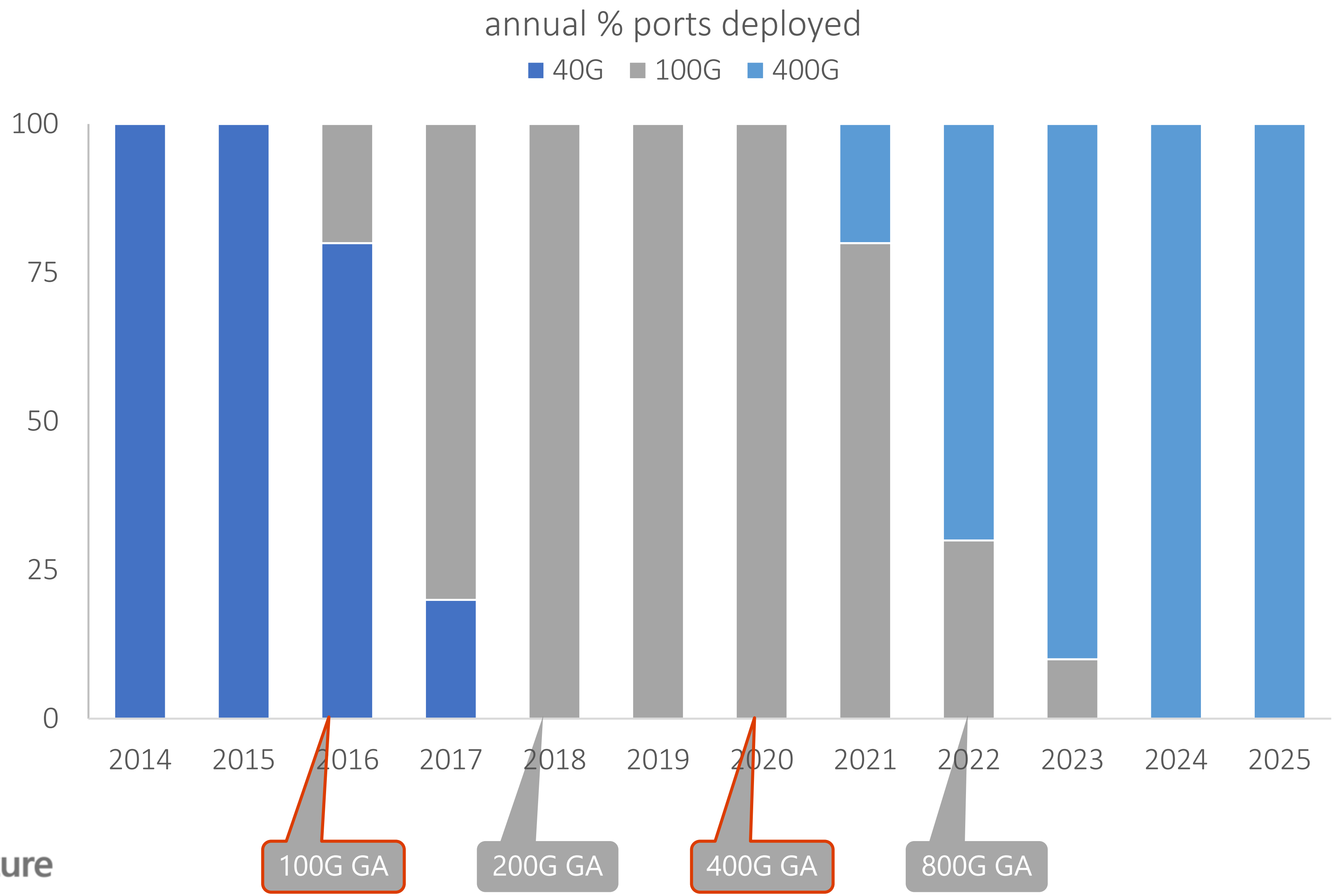
- Photonics
  - Co-Packaged Optics (CPO)
  - Novel optical approaches
- Network architecture + HW changes
  - Collapsed tiers with multi-homed NICs (fanning out horizontally)
  - Simplified forwarding requirements → cooler ASICs
  - Additional integration, e.g. encryption on switch ASIC
  - Liquid cooling





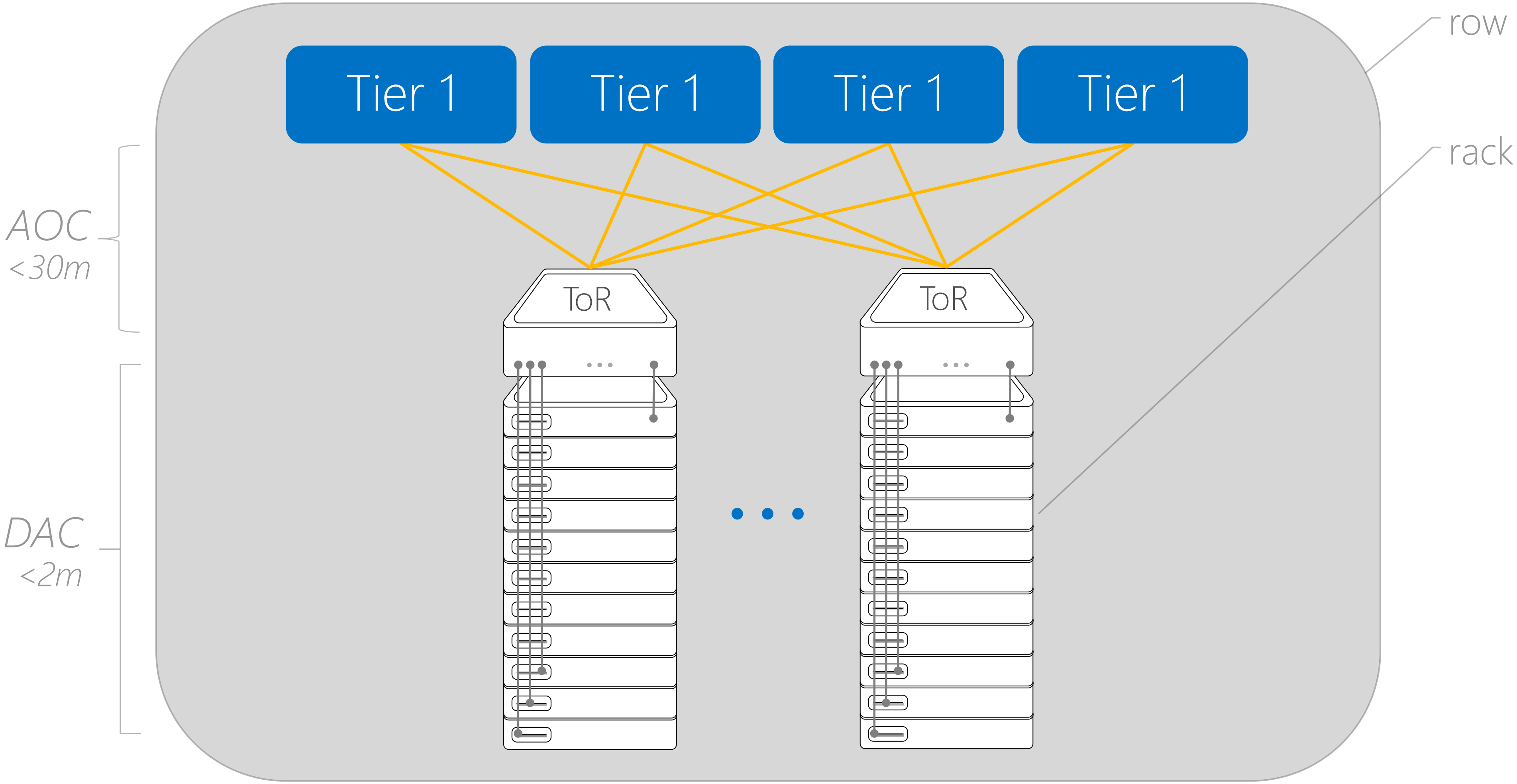
Takeaway: we can't just keep scaling link bandwidths… "next gen" systems will require all of the above

Microsoft DC ecosystem technology life cycles



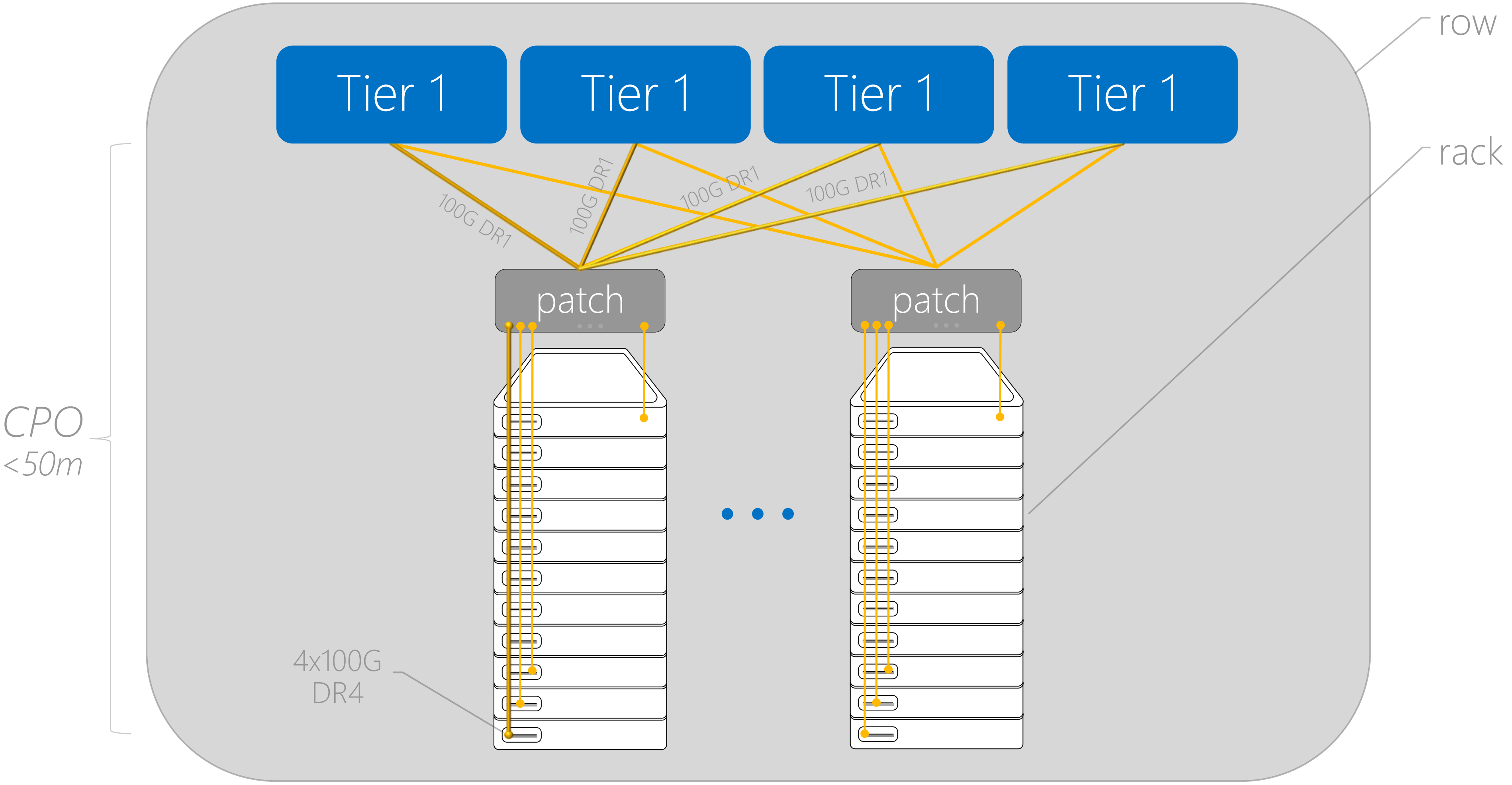
Microsoft Azure

Server-ToR-Tier1 topology



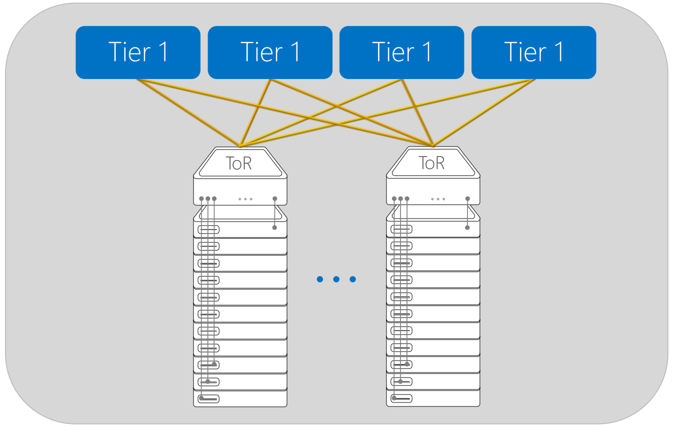
Microsoft Azure

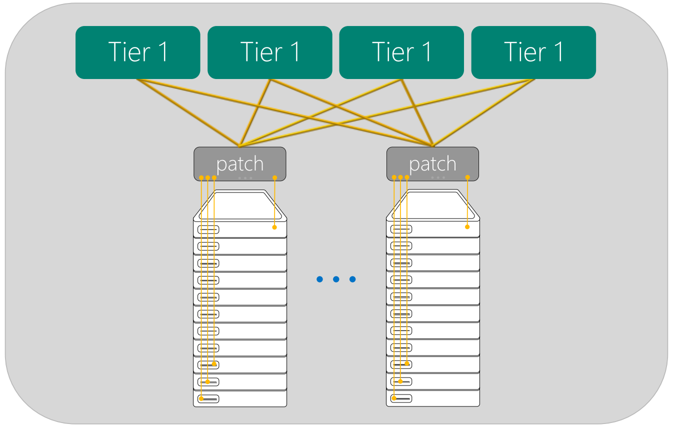
ToR bypass – multi-homed NIC



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)	leverages full switch radix
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