

# Future DC Network Considerations

Brad Booth, Microsoft

IEEE Beyond 400G Study Group

May 24, 2021


# Supporters

- Andy Bechtolsheim, Arista
- Ali Ghiasi, Ghiasi Quantum
- Jeff Maki, Juniper Networks
- Cedric Lam, Google
- Mark Filer, Microsoft

# Ethernet in the Data Center

---

- Customer technology adoption
  - Readiness of technology
  - Intercept with customer requirements
  - Technology ingestion rate
  - Co-dependencies
- Standards development
  - Technical feasibility
  - Economics feasibility
  - Interoperable ecosystem
  - Timeliness

A 3D maze with a yellow arrow pointing towards the exit. The maze is constructed from dark grey and yellow blocks. The arrow is yellow and points towards the right side of the maze. The maze is set against a dark background.

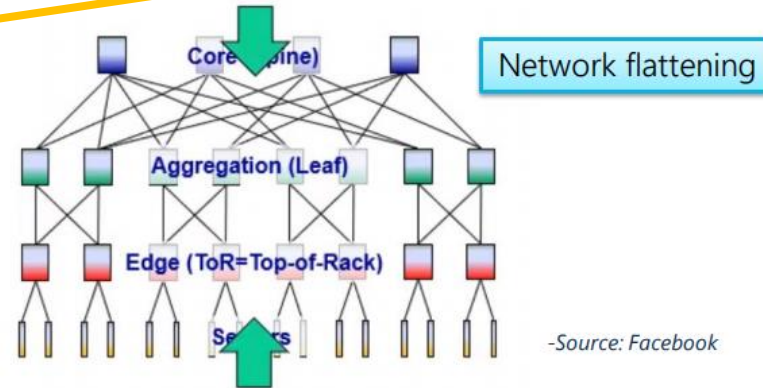
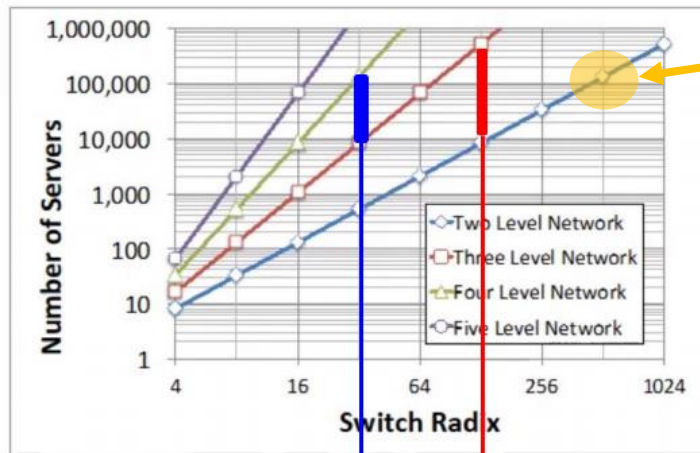
A Difficult Maze

# Maze Examples

- Microsoft 400G data center deployment is gated by 400ZR
  - Currently at 100G, “skipping” 200G generation
  - Looking at 800G modules (2x400G), but may skip 800G MAC
- Facebook topology drives a “radix” deployment rate
  - Switch silicon SERDES enables incremental bumps
- Google more aggressive on speed adoption
  - Advocate for Ethernet Technology Consortium 800G Ethernet specification
  - Preparing to consume 800G and investigating 1.6T (with 200G electrical)\*
- AWS at 400G... next step??
- Others...

# Topology Examples

## MSFT: Investigating Radix 512 networks



- Fewer tiers = decreased latency, lower power
- Volume of servers vs. power grid

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

Optical interconnects

Other CSPs use lower radix

A red pushpin is pinned to a calendar page. The calendar shows dates 15, 22, 23, and 24. The pushpin is positioned over the date 15.

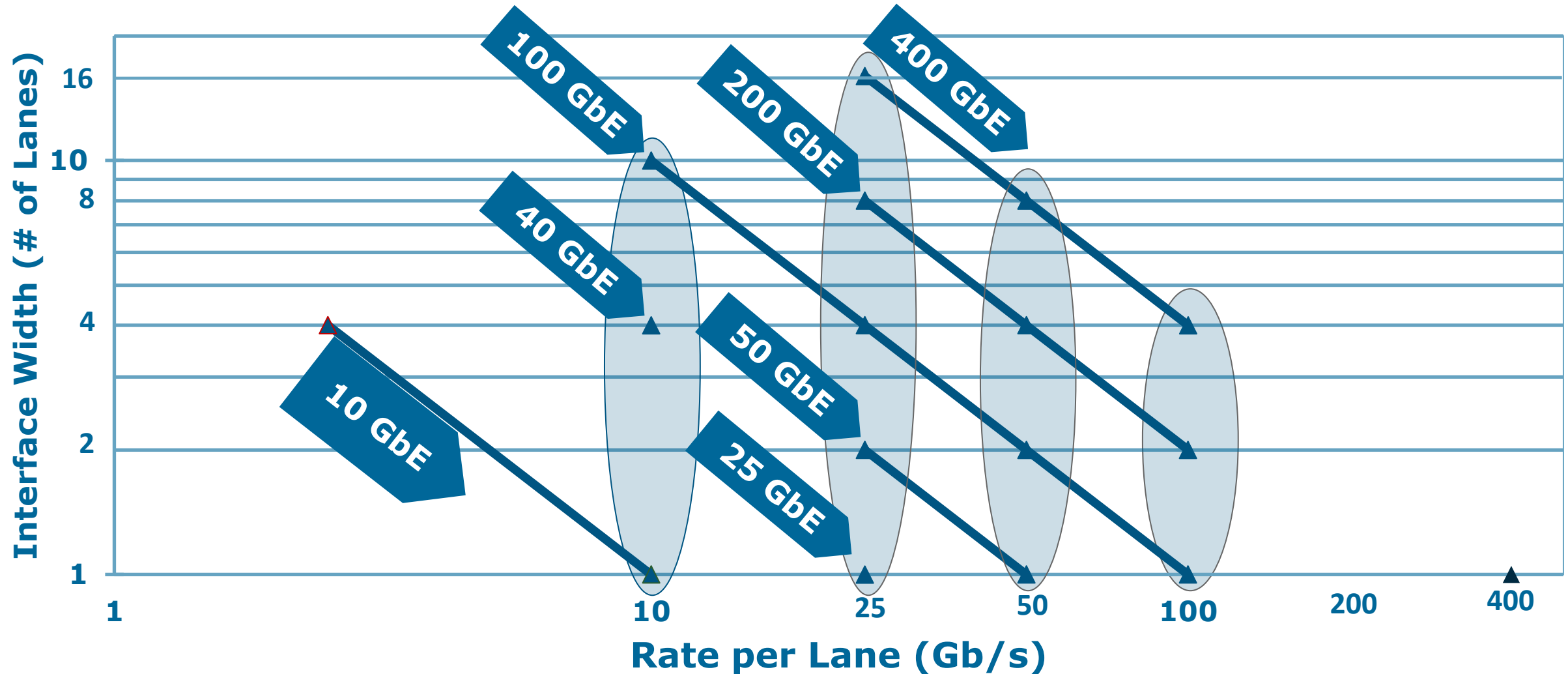
# Standards Timeliness

---

- 802.3 can be fast with standards development
  - Focused project
  - Building off existing or deployed technology
- Otherwise...
  - Study Group: typically, between 8-12 months
  - Getting to Working Group ballot: 16-24 months
  - WG ballot phase: 8-12 months
  - LMSC ballot phase: 8-12 months
  - Best case scenario: 40 months (just over 3 years)
  - Typical: 4-5 years

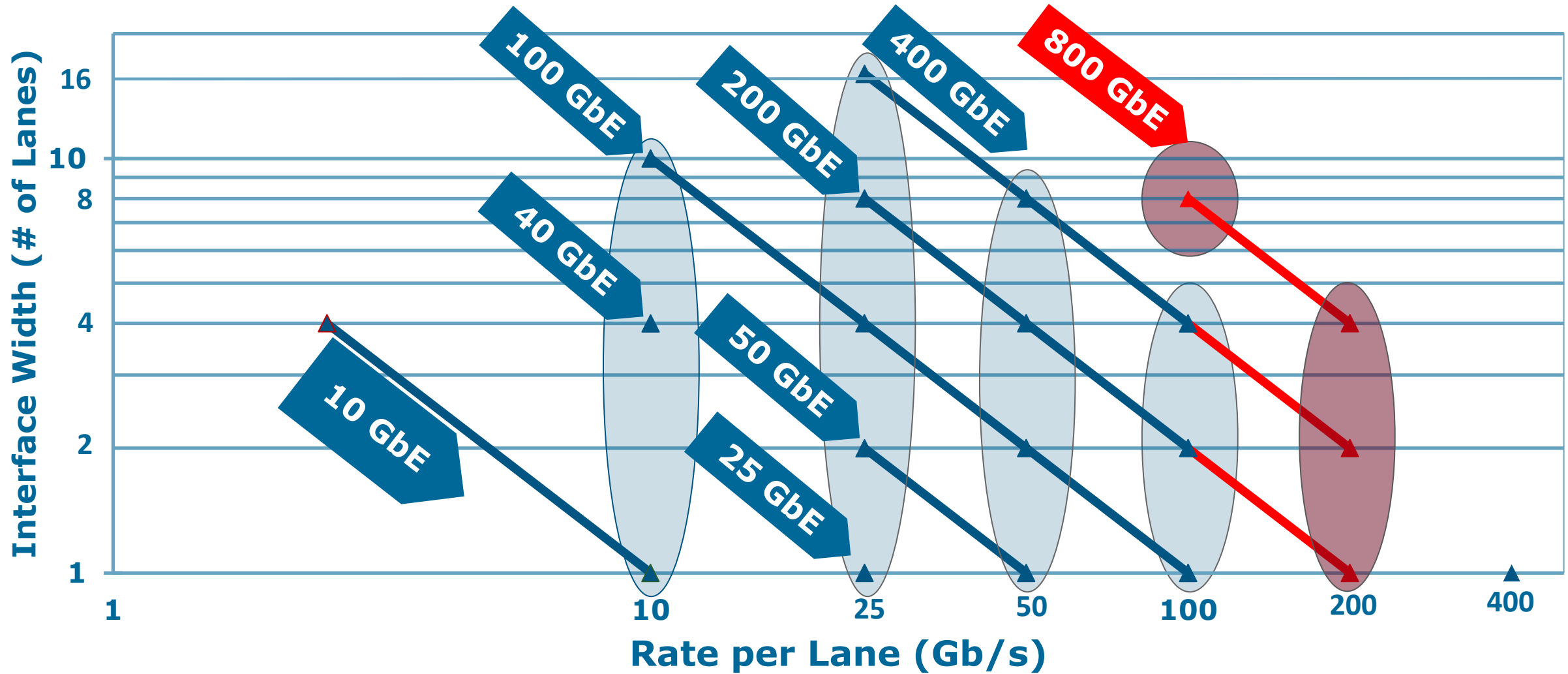
# Current Ethernet MAC Rates vs Signaling Rates

Slide courtesy of John D'Ambrosia, Futurewei a US Subsidiary of Huawei



# B400G SG Adopted Ethernet MAC & Signaling Rates

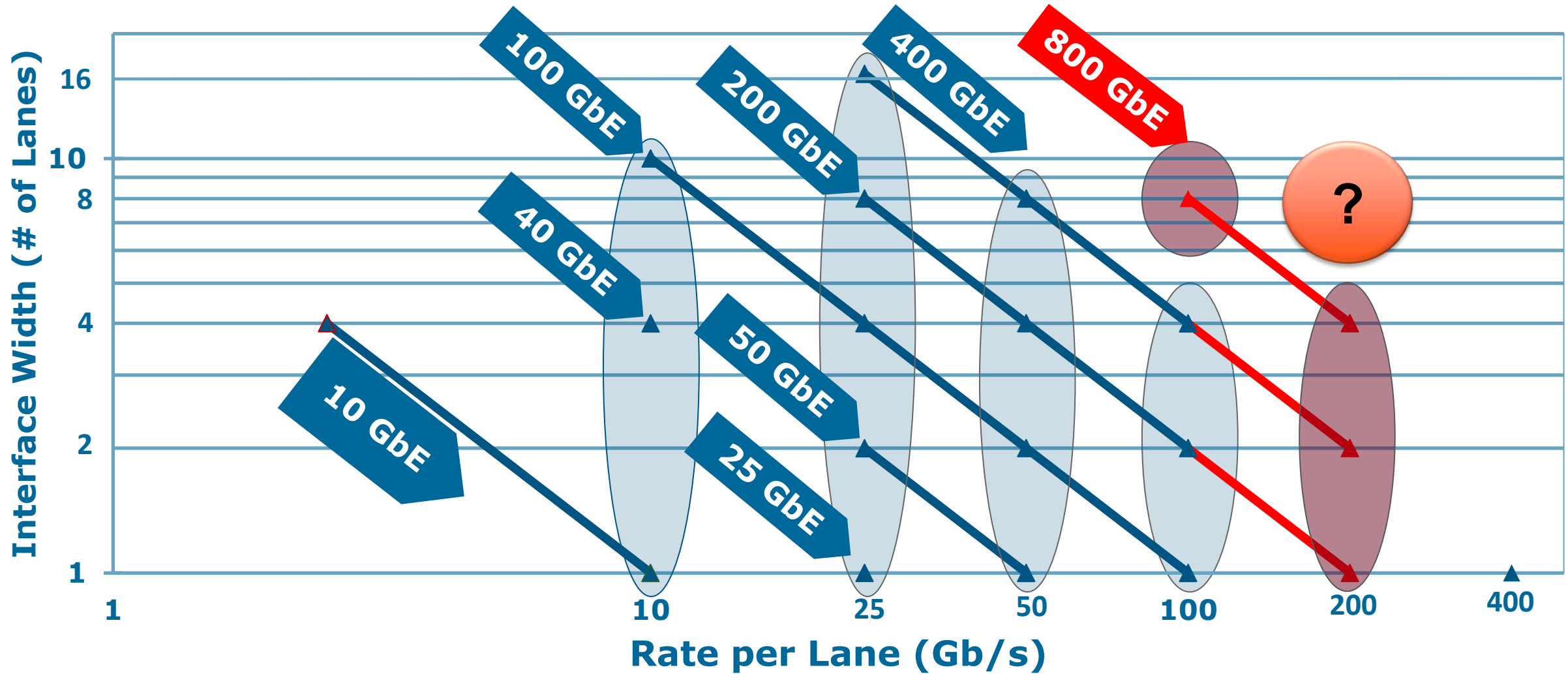
Original slide courtesy of John D'Ambrosia, Futurewei a US Subsidiary of Huawei





# Glaring Absence

Original slide courtesy of John D'Ambrosia, Futurewei a US Subsidiary of Huawei



# Why is 1.6T Missing?

- Speed per lane
  - 200G already incorporated as part of the objectives
    - over 4 pairs of SMF with lengths up to at least 500 m \*
    - over 4 pairs of SMF with lengths up to at least 2 km \*
    - over 4 wavelengths over a single SMF in each direction with lengths up to at least 2 km \*
- Number of lanes
  - Eight lanes... been there, doing that
    - over 8 pairs of MMF with lengths up to at least 50 m \*
    - over 8 pairs of MMF with lengths up to at least 100 m \*
    - over 8 pairs of SMF with lengths up to at least 500 m \*
- MAC data rate
  - Just one extra entry in Table 4-2

# Table 4-2

- Pile-on column
  - The core of full duplex Ethernet
- 40G+ use the same NOTE
  - Unlikely to change going forward
  - Note could easily be simplified (and there's precedence)

NOTE 7—For 40 Gb/s, 100 Gb/s, 200 Gb/s, and 400 Gb/s operation, the received interpacket gap (the spacing between two packets, from the last bit of the FCS field of the first packet to the first bit of the Preamble of the second packet) can have a minimum value of 8 BT (bit times), as measured at the XLGMII, CGMII, 200GMII, or 400GMII receive signals at the DTE due to clock tolerance and lane alignment requirements.

Table 4-2—MAC parameters

Parameters	MAC data rate			
	Up to and including 100 Mb/s	1 Gb/s	2.5 Gb/s, 5 Gb/s, 25 Gb/s, 40 Gb/s, 100 Gb/s, 200 Gb/s, and 400 Gb/s	10 Gb/s
slotTime	512 bit times	4096 bit times	not applicable	not applicable
interPacketGap <sup>a</sup>	96 bits	96 bits	96 bits	96 bits
attemptLimit	16	16	not applicable	not applicable
backoffLimit	10	10	not applicable	not applicable
jamSize	32 bits	32 bits	not applicable	not applicable
maxBasicFrameSize	1518 octets	1518 octets	1518 octets	1518 octets
maxEnvelopeFrameSize	2000 octets	2000 octets	2000 octets	2000 octets
minFrameSize	512 bits (64 octets)	512 bits (64 octets)	512 bits (64 octets)	512 bits (64 octets)
burstLimit	not applicable	65 536 bits	not applicable	not applicable
ipgStretchRatio	not applicable	not applicable	not applicable	104 bits



# Shifting Focus

---

- Study Group development of 5 Critters (Criteria)
  - Very PHY centric
- A task force focuses primarily on PHY
  - MAC is simple table entry
- Even if PHY technology exists, next Ethernet speed gated by process
  - Ethernet Technology Consortium took advantage of this for 800G
  - Why would 802.3 want to leave that door open?
- Mark Nowell, “ignore the MAC”
- Get the MAC out of the way of progress... use process for PHYs

# Thoughts (Re-cap)

- PHY technology is driving higher speed projects
  - Beyond 10G, MAC exceeded PHY → started “fill-in” approach
  - Electronic & photonic mismatch → technologies progress at their own pace
  - New optics form factors support mismatch, data rate growth
- Time to modify 802.3’s approach to growth
  - Many other standards bodies have already shifted their approach
  - PHY per lane the building block (n, where n= 100G, 200G, etc.)
  - Permit MAC to scale seamlessly ( $n \cdot 2^m$ , where  $m=\{0, 1, 2, 3\}$ )
- Enables a smoother growth path w/ fewer “fill-in” projects
  - Avoids delays in “next speed” due to 802.3 process



# Recommendation

---

- Adopt an objective to simplify future Ethernet MAC rates
- Something like: “Preserve the Ethernet MAC Parameters for data rates  $\geq 800\text{G}$ ”



Thank you.