Proposal – “802.3 Ethernet Interconnect for AI” Assessment

John D’Ambrosia
Futurewei, U.S. Subsidiary of Huawei

20 January 2025

IEEE 802.3 New Ethernet Applications Ad hoc

IEEE 802.3 Jan 2025 Interim, Phoenix, AZ, USA

Contributors

- Adam Healey, Broadcom
- Mark Nowell, Cisco
- Dave Ofelt, Juniper
- Kent Lusted, Synopsys
- David Law, HPE

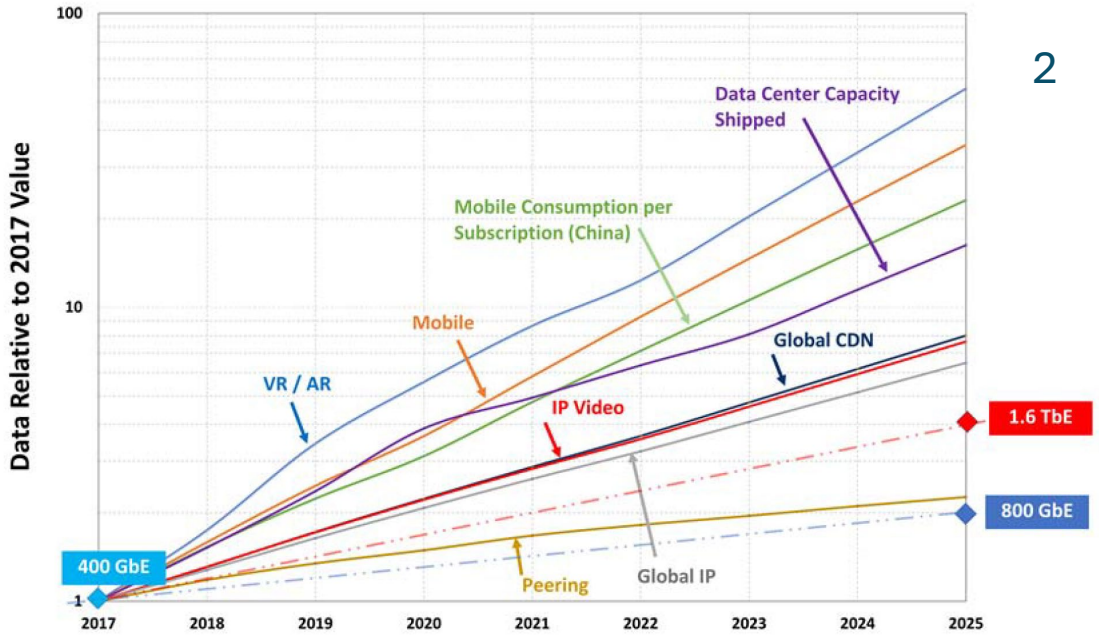
Introduction

- This presentation proposes the IEEE 802.3 NEA Ad hoc **initiate an assessment** of "802.3 Ethernet Interconnect for AI" with an emphasis on beyond 200 Gb/s signaling.
  - **This presentation is NOT a call-for-interest.**
- It has been my experience:
  - Bandwidth assessments has been utilized effectively to help efforts that targeted initiation of major new efforts, i.e. new speeds of Ethernet
    - BWA1[1] > 400 GbE Study Group
    - BWA2[2] > Beyond 400 GbE Study Group
  - These efforts took time to solicit input, gather data, analyze, and build consensus
  - These assessments were of great value to the subsequent efforts.
- This presentation is asking questions, and I will refrain from expressing any technical opinions.

1. IEEE 802.3 Industry Connections Ethernet Bandwidth Assessment Ad hoc, https://www.ieee802.org/3/ad_hoc/bwa/index.html
2. IEEE 802.3 Industry Connections NEA Ad Hoc Ethernet Bandwidth Assessment, Part II, https://www.ieee802.org/3/ad_hoc/bwa2/index.html

IEEE 802.3 2020 Ethernet Bandwidth Forecast, Part II [1]

- During the development of the bandwidth forecast for the report, there was discussion whether a future effort should target 800 GbE and 1.6 TbE.
- This forecast was subsequently used by the IEEE 802.3 Beyond 400 GbE Study Group to justify selection of 800 GbE and 1.6 TbE objectives.



2

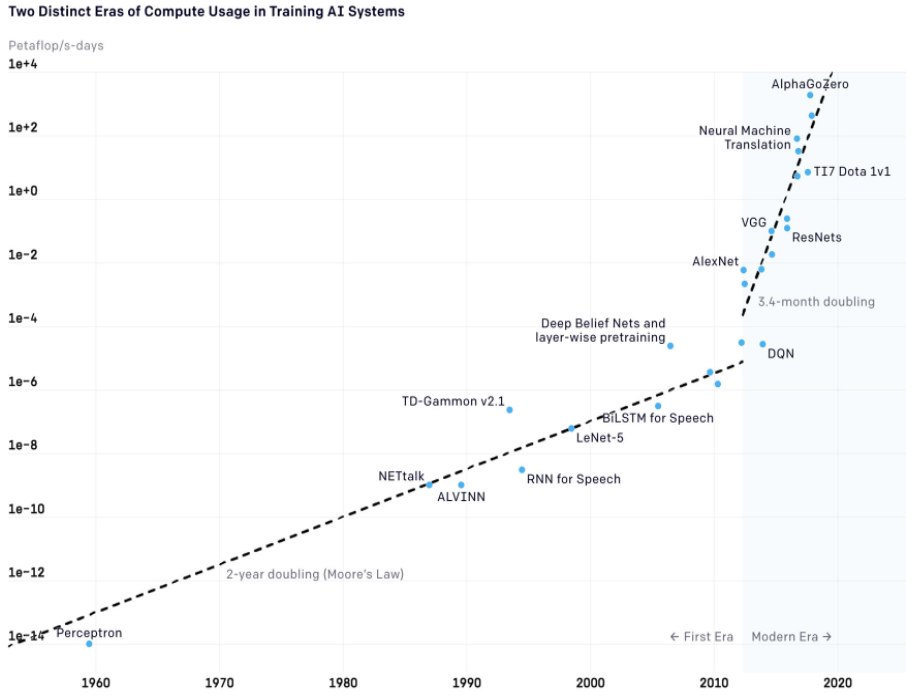
Figure 52—Bandwidth Curves (2017-2025)

1 - IEEE 802.3 Industry Connections NEA Ad Hoc Ethernet Bandwidth Assessment, Part II, https://www.ieee802.org/3/ad_hoc/bwa2/index.html
2 - IEEE 802.3 Industry Connections NEA Ad Hoc Ethernet Bandwidth Assessment Report, Part II, https://www.ieee802.org/3/ad_hoc/bwa2/BWA2_Report.pdf

Prior Influence of AI in IEEE 802.3

ARTIFICAL INTELLIGENCE & COMPUTE

- First Era (Before 2012)
  - Moore's Law – 2-year doubling
  - Uncommon to use GPUs for machine learning
- Modern Era (2012 and later)
  - 2012 – 2014: most results used 1-8 GPUs rated at 1-2 TFLOPS
  - 2014 – 2016: large-scale results used 10-100 GPUs rated at 5-10 TFLOPS
  - 2016 – 2017: greater algorithmic parallelism (huge batch sizes, architecture search, expert iteration), specialized hardware (TPUs), faster interconnects



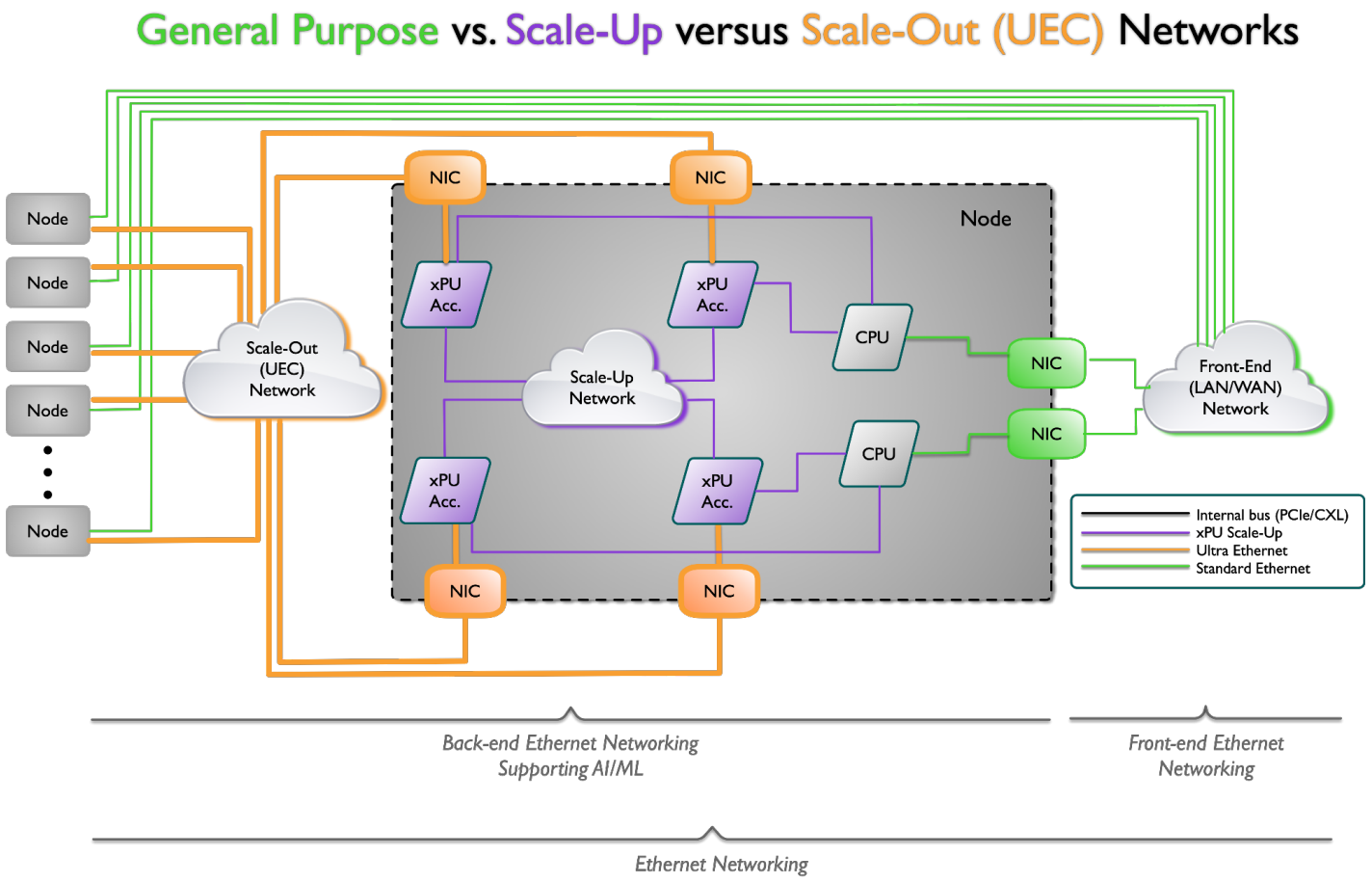
Source – OpenAI blog post 'AI and Compute' addendum 'Compute used in older headline results' posted 7th November 2019 by Girish Sastry, Jack Clark, Greg Brockman and Ilya Sutskever <https://openai.com/blog/ai-and-compute/>.

29 Oct 2020 IEEE 802.3 NEA Ad hoc - Beyond 400 Gb/s Ethernet CFI Consensus Presentation Page 14

- AI was cited as a factor influencing bandwidth growth in the "Beyond 400GbE" CFI

Source: https://www.ieee802.org/3/ad_hoc/ngrates/public/calls/20_1029/CFI_Beyond400GbE_Rev7_201029.pdf

What are AI Networks?



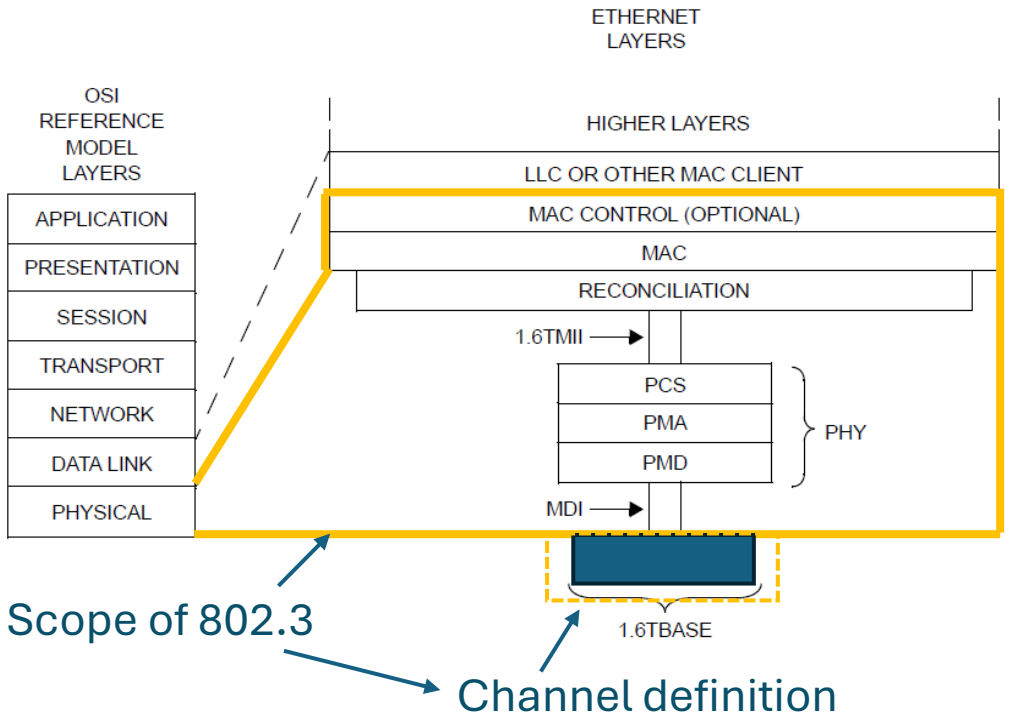
Source: Ultra Ethernet Consortium

- The author is aware that there are different representations of different implementations of AI Networks.
- The key takeaway is there are three types of networks for AI:
  - Front-end / traditional Ethernet
  - Back-end networks
    - Scale-up
    - Scale-out

Recent Relevant Liaisons to IEEE 802.3 WG

- Nov 2024 – UA Link
  (https://www.ieee802.org/3/minutes/nov24/incoming/UALink%20liaison%20to%20IEEE802.3%2010-2024%20v4.pdf)
  - "…an industry body focusing on exploring and specifying optimizations for networks based on the 802.3 physical layer and supporting Artificial Intelligence (AI) and Machine Learning (ML) **scale-up networks** and workloads."
- July 2024 – Ethernet Alliance TEF
  (https://www.ieee802.org/3/minutes/jul24/incoming/Liaison_EA_to_802d3_240718_final.pdf)
  - "…there are calls to begin exploration of 400Gb/s per lane electrical and optical signaling to support the AI networks of the future…."
- Sept 2023 – Ultra Ethernet Consortium
  (https://www.ieee802.org/3/minutes/sep23/incoming/UEC%20liaison%20to%20802d3%20-%20Signed_Redacted.pdf)
  - "We wish to inform the IEEE 802.3 Working Group of the formation of the Ultra Ethernet Consortium (UEC), an industry body focusing on exploring and specifying optimizations for Ethernet-based networks supporting Artificial Intelligence (AI) and Machine Learning (ML) workloads."
  - Note –white paper / URL referenced in liaison changed –
    - https://ultraethernet.org/wp-content/uploads/sites/20/2023/10/23.07.12-UEC-1.0-Overview-FINAL-WITH-LOGO.pdf
  - Focus for D1.0 – **Scale-out Networks**

Important –
Scope of 802.3



Industry Events

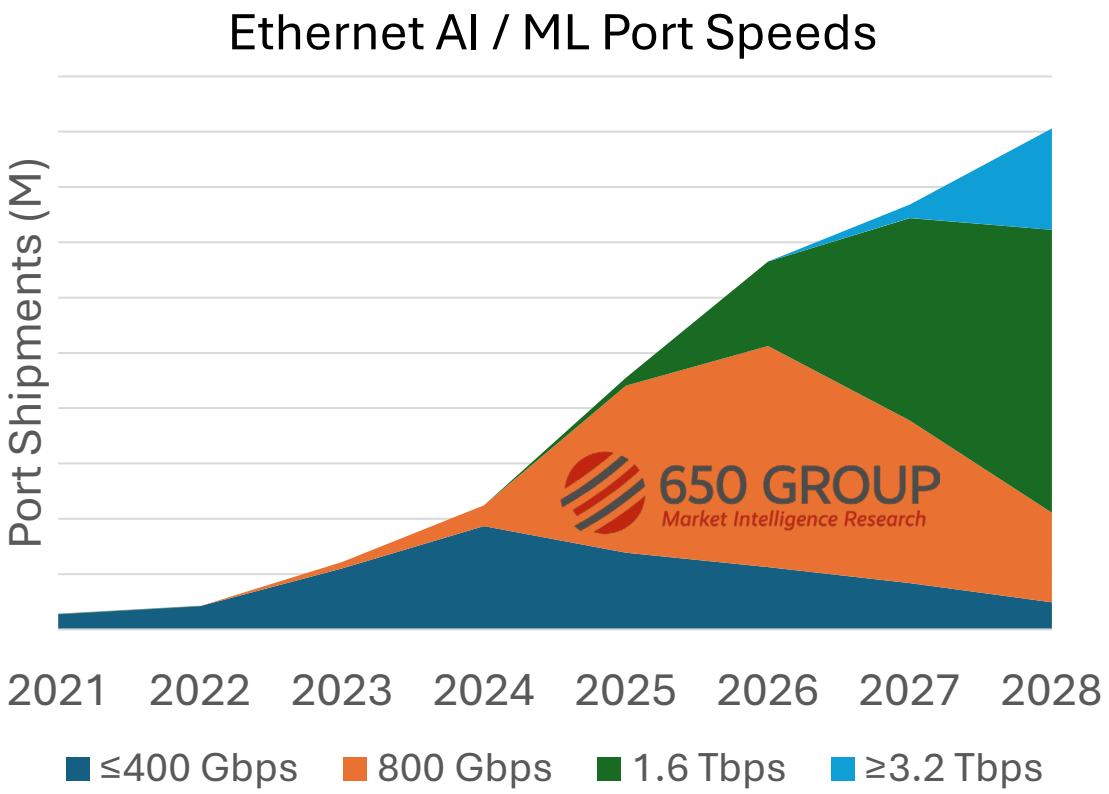
- Ethernet Alliance – “Ethernet in the Age of AI” TEF
  - Reference - https://ethernetalliance.org/tef-2024-ethernet-in-the-age-of-ai/
  - Event Proceedings – https://ethernetalliance.org/tef-2024-ethernet-in-the-age-of-ai-presentations-form/
  - Keynotes
    - Ram Huggahalli (Microsoft) - “AI-Centric Datacenters and their Diverse Network Requirements”
    - Moray McLaren (Google) - “The Future of Networking for AI in Hyperscale data centers”
    - Nicolaas Viljoen (Meta) - “Ethernet-The foundation of AI @ Meta”
  - Presentations focused on 400 Gb/s electrical / optical signaling
    - Halil Cirit (Meta) – “System Overview and Exploring Alternative FEC Techniques for 400G Performance Enhancement”
- SFF / SNIA – “400G AI Workshop”
  - Focus on 400G copper channel requirements and designs for AI applications
  - Reference - https://www.eventcreate.com/e/aiworkshopbysniasff

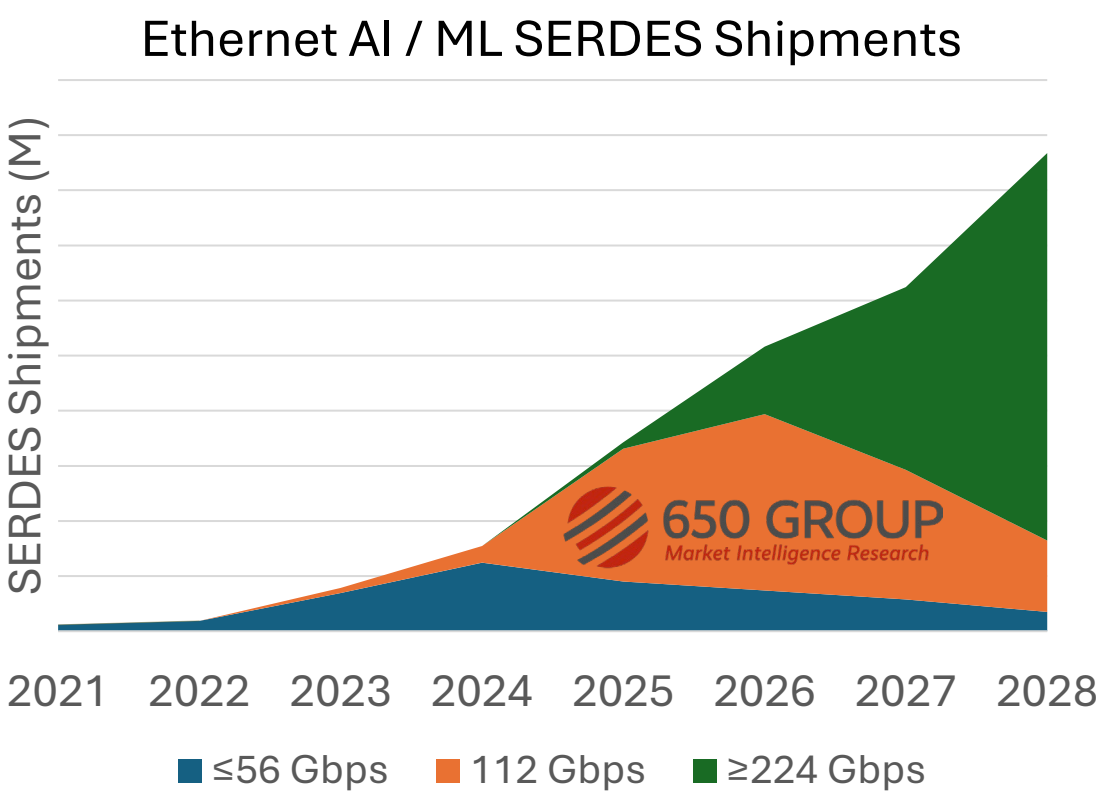
AI Market Potential

A Smattering of AI in the 2024 News

1. 18 Dec 2024 – Dell'Oro Press Release – "Hyperscale Capex Surges 82 Percent in 3Q 2024, Fueled by AI Infrastructure Spending, According to Dell'Oro Group"“
   - https://www.delloro.com/news/hyperscale-capex-surges-82-percent-in-3q-2024-fueled-by-ai-infrastructure-spending/
2. 12 Dec 2024 – Omdia: "Semiconductor market posts strong Q3, set for significant growth in 2024"
   - https://omdia.tech.informa.com/pr/2024/dec/omdia-semiconductor-market-posts-strong-q3-set-for-significant-growth-in-2024
3. 04 August 2024 – Lightwave – "Ethernet optical transceivers for AI Growth to double in 2024"
   - https://www.lightwaveonline.com/home/article/55130661/ethernet-optical-transceivers-for-ai-to-double-in-2024
4. 04 June 2024 – 650 Group Press Release – "Data Center AI Networking to Surge to Nearly $20B in 2025, According to 650 Group"
   - https://650group.com/press-releases/data-center-ai-networking-to-surge-to-nearly-20b-in-2025-according-to-650-group/

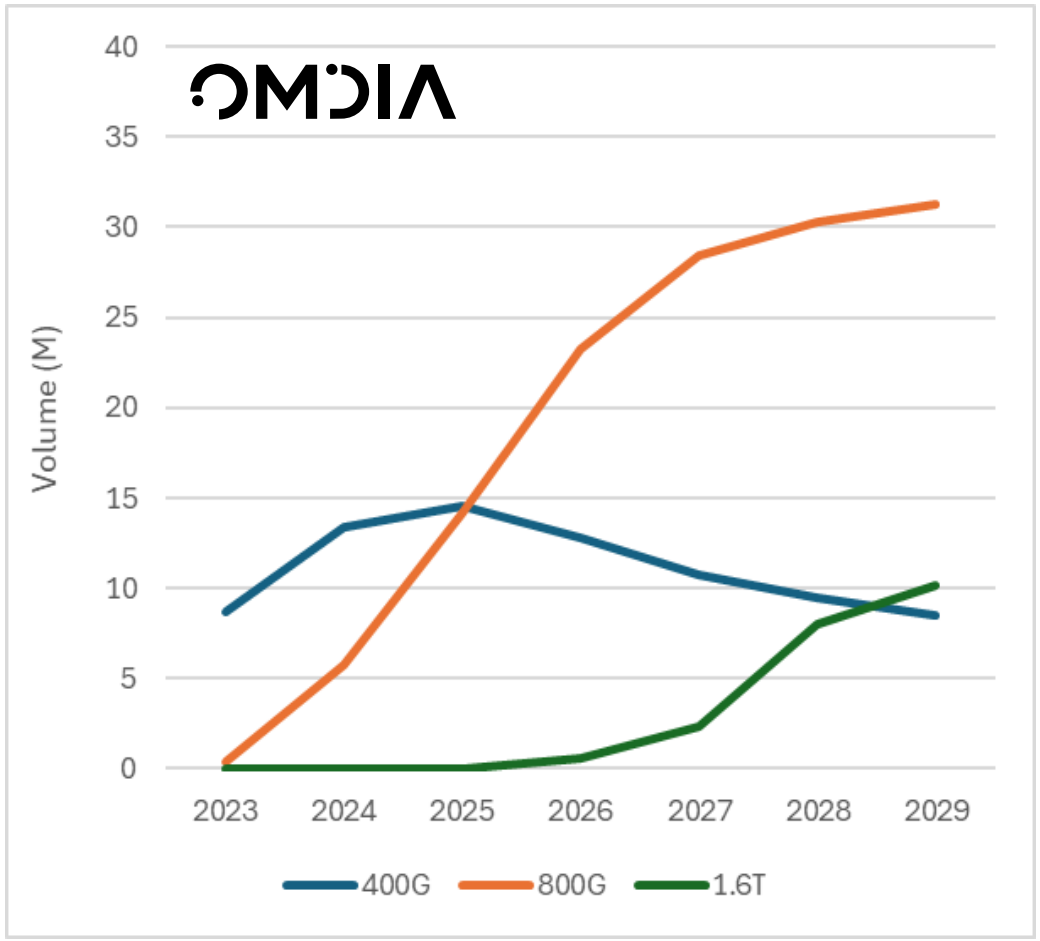
Ethernet Switch: Data Centers

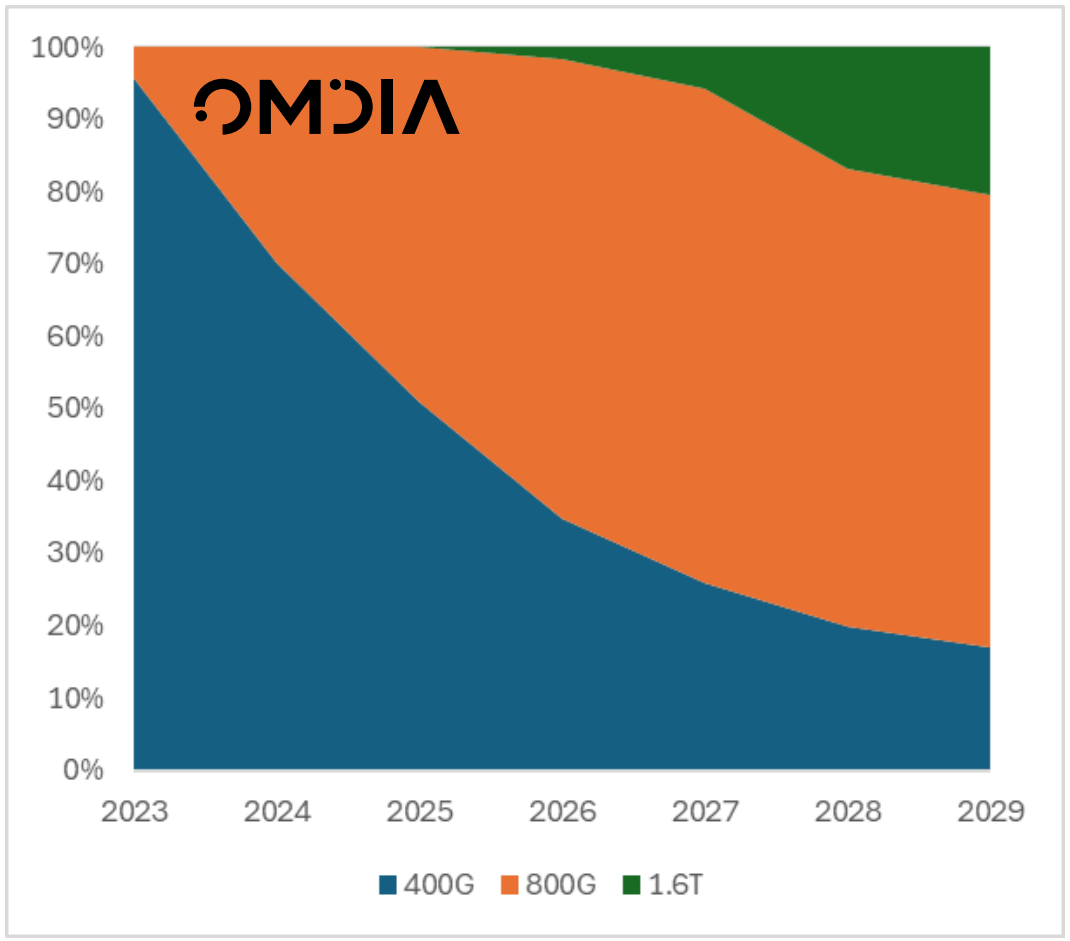




Data Source: Provided by and used with permission by Alan Weckel, 650 Group.

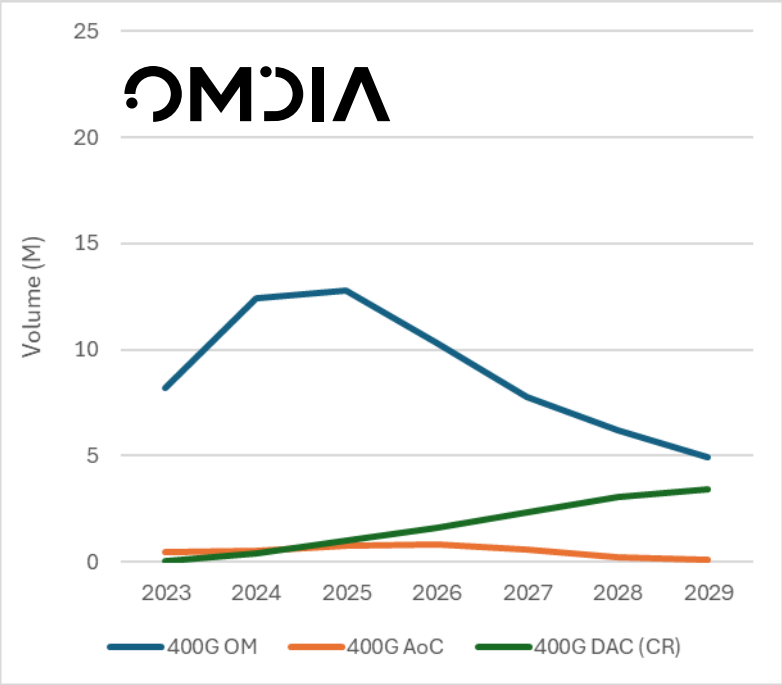
Total Volume Shipments (Scale-up / out Networks) For Considered Rates

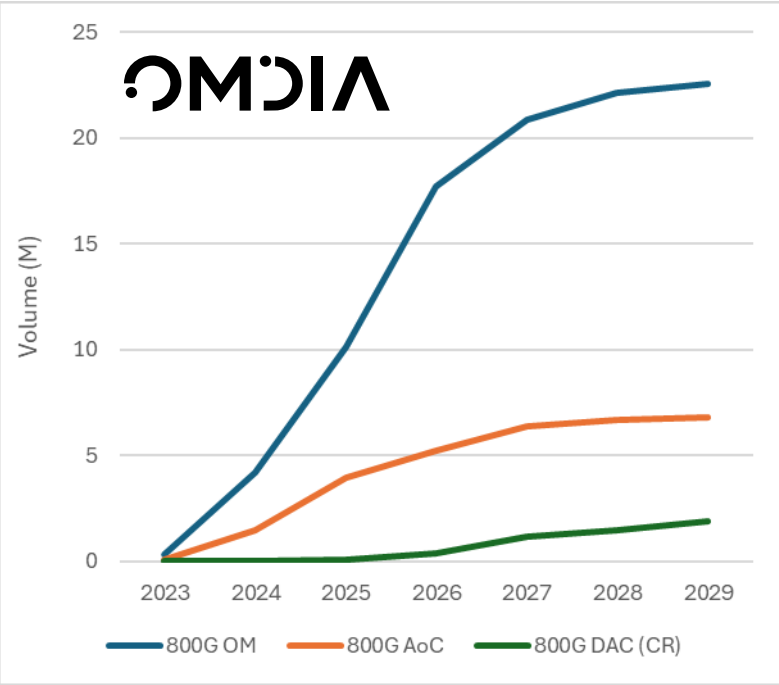


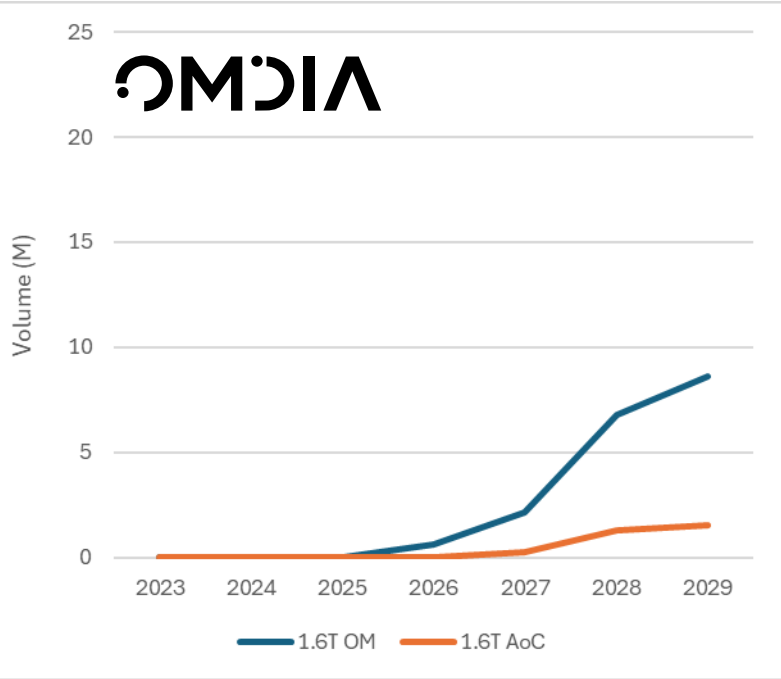


Data Source: Provided by and used with permission by Omdia.
Note – Update by Omdia currently underway and 800G and 1.6T modules expected to increase in 2025

Volume Shipments (Scale-up / out Networks)





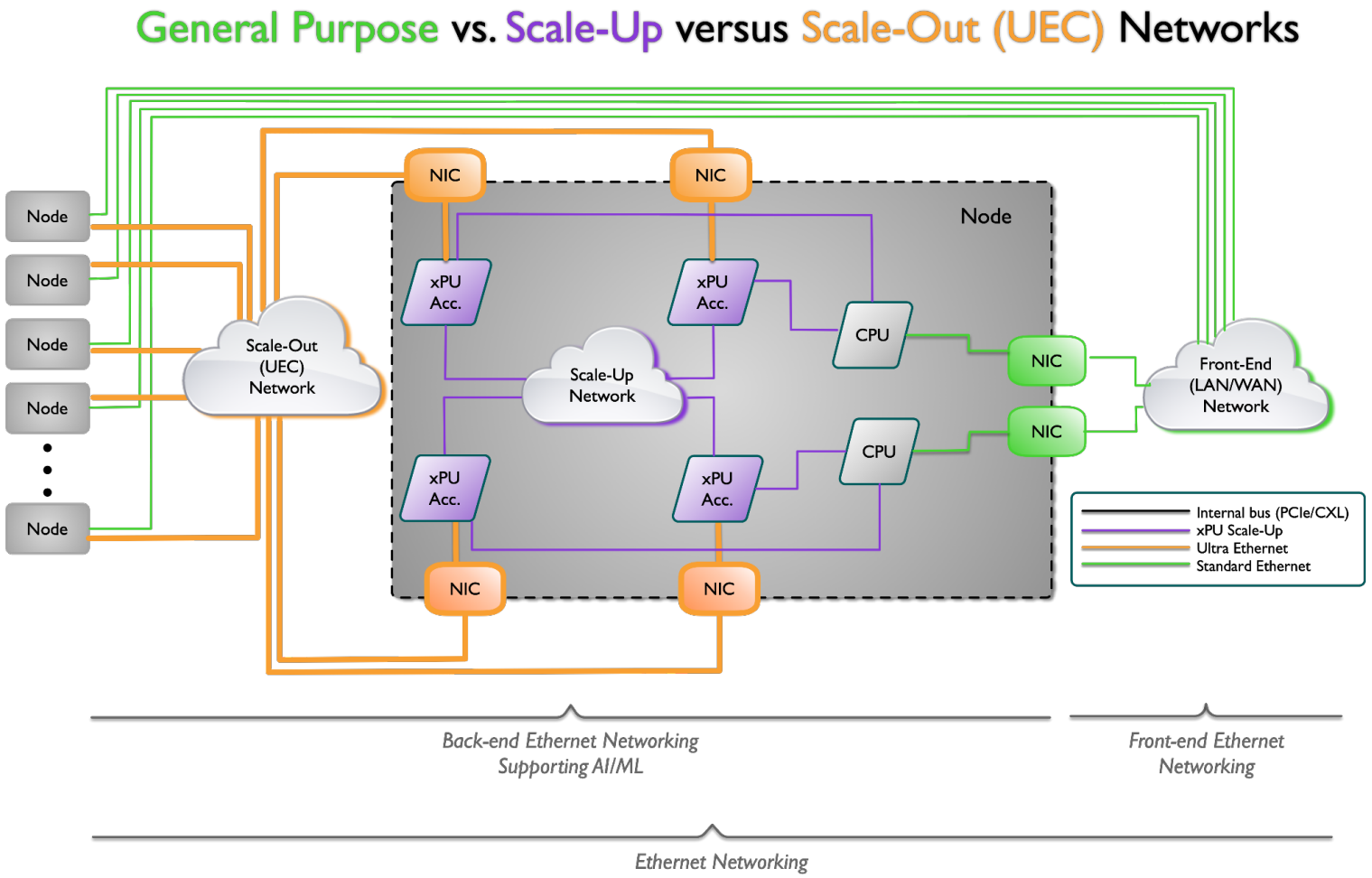


OM: Optical Modules: 400G – QSFP-DD/OSFP, 800G QSFP-DD/OSFP, 1.6T (not defined)
AoC: Active Optical Cables - 400G – QSFP-DD/OSFP, 800G QSFP-DD/OSFP, 1.6T QSFP-DD / OSFP
DAC (CR only, **does not include internal copper cables**): 400G: QSFP-DD, 800G: QSFP-DD/OSFP

Data Source: Provided by and used with permission by Omdia

Overview of AI Networks

AI Networks



Source: Ultra Ethernet Consortium

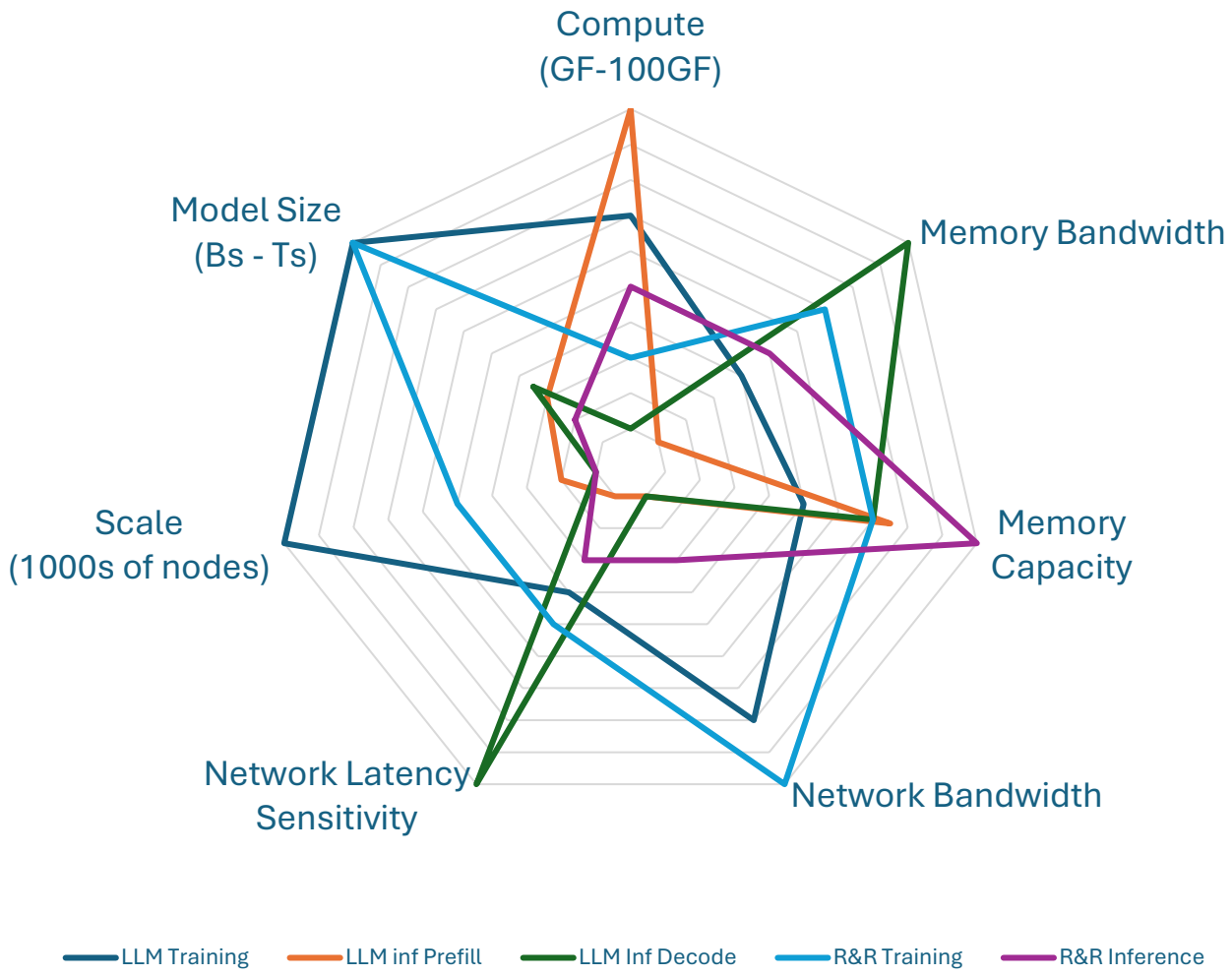
Common Messages Observed from EA TEF Keynotes

- Standards-based approaches desirable / “welcomed”
- Resilience important
- Power Critical
- Latency sensitivity depends on the application
- Connectivity innovation is key for future scale-up / scale-out
- Bandwidth density important (even more so than front-end)

Source: Ethernet Alliance TEF 2024 Keynotes –

1. *Ethernet – The foundation of AI @ Meta, Nic Viljoen (Meta)*
2. *AI Datacenters and their Diverse Network Requirements, Ram Huggahalli (Microsoft)*
3. *Future networking for AI in Hyperscale data centers, Moray McLaren (Google)*

AI Application Requirements



Ethernet Priorities (2 – 5 years)

- Resilience
- Reach
- Beachfront (I/O BW Limitation)
- Power

Latency can be traded off for resilience, reach and power

Source: Ethernet Alliance TEF 2024 Keynote – Ethernet – The foundation of AI @ Meta, Nic Viljoen

Questions to ask

- What are the interconnect requirements for the different AI networks?
- What are the performance requirements of these interconnects?
- What are the priorities for the development of these interconnects?
- What tradeoffs can be made between latency and resilience / reach / power?

- This is a dynamic scenario where all of the answers appear to be interdependent

Starting with IEEE P802.3dj

Initial Thoughts

IEEE P802.3dj represents a reasonable starting point for the future

In the opinion of the author (Chair of IEEE P802.3dj)

- IEEE P802.3dj is one of the largest projects in the past 25 years in IEEE 802.3
- Now is a moment to reflect on the decisions made to date and consider potential insight for future efforts.

IEEE P802.3dj Objectives

Ethernet Rate	Signaling Rate	AUI	Backplane	Cu Cable	SMF 500m	SMF 2km	SMF 10km	SMF 20km	SMF 40km
200 Gb/s	200 Gb/s	200GAUI-1 C2C C2M	200GBASE-KR1	200GBASE-CR1	200GBASE-DR1	200GBASE-DR1-2			
400 Gb/s	200 Gb/s	400GAUI-2 C2C C2M	400GBASE-KR2	400GBASE-CR2	400GBASE-DR2	400GBASE-DR2-2			
800 Gb/s	200 Gb/s	800GAUI-4 C2C C2M	800GBASE-KR4	800GBASE-CR4	1.800GBASE-DR4 2.800GBASE-FR4-500	1. 800GBASE-DR4-2 2. 800GBASE-FR4	800GBASE-LR4		
	800 Gb/s						800GBASE-LR1	800GBASE-ER1-20	800GBASE-ER1
1.6 Tb/s	100 Gb/s	1.6TAUI-16 C2C C2M							
	200 Gb/s	1.6TAUI-8 C2C C2M	1.6TBASE-KR8	1.6TBASE-CR8	1.6TBASE-DR8	1.6TBASE-DR8-2			

IEEE P802.3dj Modulation Approaches

Ethernet Rate	Signaling Rate	AUI	Backplane	Cu Cable	SMF 500m	SMF 2km	SMF 10km	SMF 20km	SMF 40km
200 Gb/s	200 Gb/s	200GAUI-1 C2C C2M	200GBASE-KR1	200GBASE-CR1	200GBASE-DR1	200GBASE-DR1-2			
400 Gb/s	200 Gb/s	400GAUI-2 C2C C2M	400GBASE-KR2	400GBASE-CR2	400GBASE-DR2	400GBASE-DR2-2			
800 Gb/s	200 Gb/s	800GAUI-4 C2C C2M	800GBASE-KR4	800GBASE-CR4	1. 800GBASE-DR4 2. 800GBASE-FR4-500	1. 800GBASE-DR4-2 2. 800GBASE-FR4	800GBASE-LR4		
	800 Gb/s						800GBASE-LR1	800GBASE-ER1-20	800GBASE-ER1
1.6 Tb/s	100 Gb/s	1.6TAUI-16 C2C C2M							
	200 Gb/s	1.6TAUI-8 C2C C2M	1.6TBASE-KR8	1.6TBASE-CR8	1.6TBASE-DR8	1.6TBASE-DR8-2			

PAM-4

DP-QAM16

IEEE P802.3dj FEC Approach Overview

Ethernet Rate	Signaling Rate	AUI	Backplane	Cu Cable	SMF 500m	SMF 2km	SMF 10km	SMF 20km	SMF 40km
200 Gb/s	200 Gb/s	200GAUI-1 C2C C2M	200GBASE-KR1	200GBASE-CR1	200GBASE-DR1	200GBASE-DR1-2			
400 Gb/s	200 Gb/s	400GAUI-2 C2C C2M	400GBASE-KR2	400GBASE-CR2	400GBASE-DR2	400GBASE-DR2-2			
800 Gb/s	200 Gb/s	800GAUI-4 C2C C2M	800GBASE-KR4	800GBASE-CR4	1. 800GBASE-DR4 2. 800GBASE-FR4-500	1. 800GBASE-DR4-2 2. 800GBASE-FR4	800GBASE-LR4		
	800 Gb/s						800GBASE-LR1	800GBASE-ER1-20	800GBASE-ER1
1.6 Tb/s	100 Gb/s	1.6TAUI-16 C2C C2M							
	200 Gb/s	1.6TAUI-8 C2C C2M	1.6TBASE-KR8	1.6TBASE-CR8	1.6TBASE-DR8	1.6TBASE-DR8-2			

End-to-End
RS (544,514)

Concatenated
BCH(128,120) or BCH (126,110)

Segmented
oFEC

Observations of IEEE P802.3dj

- Lots of objectives
- Successful in doubling signaling lane rate > electrical interfaces / electrical PHYs / optical IMDD PHYs
- Successful in doubling the new maximum Ethernet rate
- Modulation
  - All electrical interfaces / PHYs and optical IMDD PHYs > PAM4
  - Coherent optical PHYs > DP-QAM16
- Multiple FEC schemes & codes>
  - Common FEC scheme / code for all copper interfaces / PMDs
    - AUI / KRx / CRx > End-to-end > RS (544,514)
  - Multiple FEC schemes / codes for all IMDD optical PMDs
    - DR / FR4-500 > End-to-end > RS (544,514)
    - DRx-2 / FR4 / LR4 > Concatenated > RS (544,514) + BCH(126,110)
  - Multiple FEC schemes for coherent optical PMDs
    - LR1 > Concatenated > RS(544,514) + BCH(128,120)
    - ER1-20 / ER1 > Segmented > RS(544,514) / oFEC
- Interface widths
  - AUIs / Copper & Optical PHYs - x1, x2, x4, x8,
  - AUIs (Initial justification was for initial test equipment to support 1.6 TbE) – x16

Technical Questions to Ask Going Forward

The Basic Assumption…..

- The IEEE 802.3 Ethernet community can just turn the crank [again] and double the rate of the current Ethernet solutions [again]

- To quote Kent Lusted, IEEE P802.3dj Electrical Track Chair, regarding development of 200 Gb/s signaling for electrical PMD's -

"Optimization of PMD characteristics is key – every little bit matters - i.e. finding pennies in the sofa"

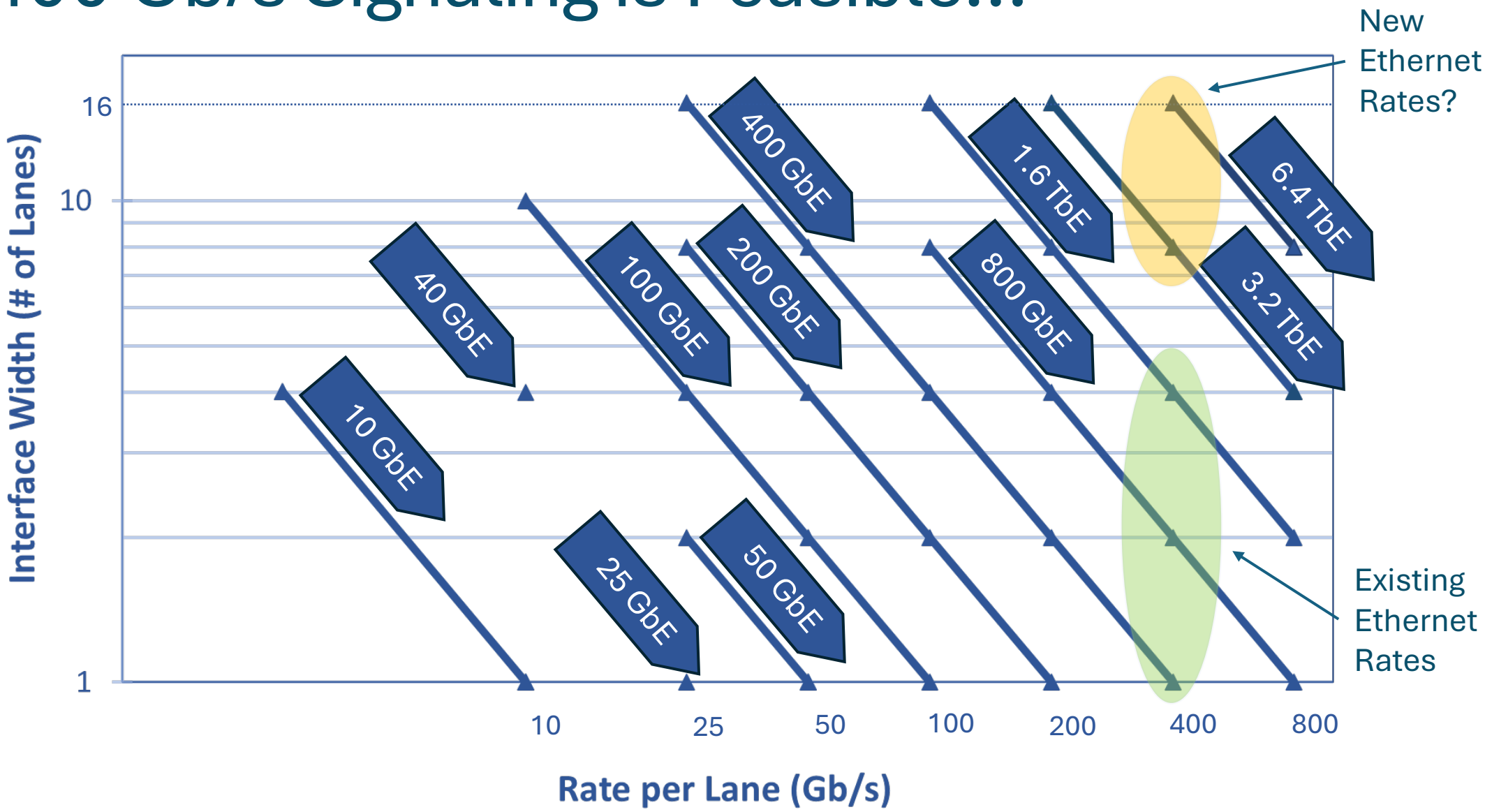
Challenging this Assumption

	Assumptions Based on P802.3dj	Questions to ask going forward
Target Objectives	IEEE 802.3 can handle projects with significant number of objectives	1. What are the objectives of each AI network? 2. Should IEEE 802.3 prioritize objectives based on nearest term needs?
Signaling Rate	For the past 10 years Ethernet has effectively doubled the signaling lane rate with each generation - 400 Gb/s is the next data rate on that path	Can doubling the effective data rate be done or should lane rates be optimized on a per lane capability basis?
Modulation Approach	1. Electrical interfaces / PHYs are based on the same modulation 2. Electrical interfaces / PHYs and optical IMDD PHYs based on the same modulation	One modulation approach for all target interconnects or choose the optimal modulation approach for each target objective?
FEC Scheme	FEC schemes / codes optimized for groupings of interfaces / PHYs	1. Different FEC schemes / codes for different groupings? 2. Different FEC schemes / codes for similar PHY objectives for different AI Networks?
Ethernet Interconnects	The same kinds of interconnects can be made to work, adjusted for rate of signaling. Examples - • Pluggable modules can still work • Passive copper cables are viable	Are new interconnect approaches necessary?
Interface / PHY widths	x1, x2, x4, x8, x16	If lane rate is not doubled – is it ok to consider different interface widths?

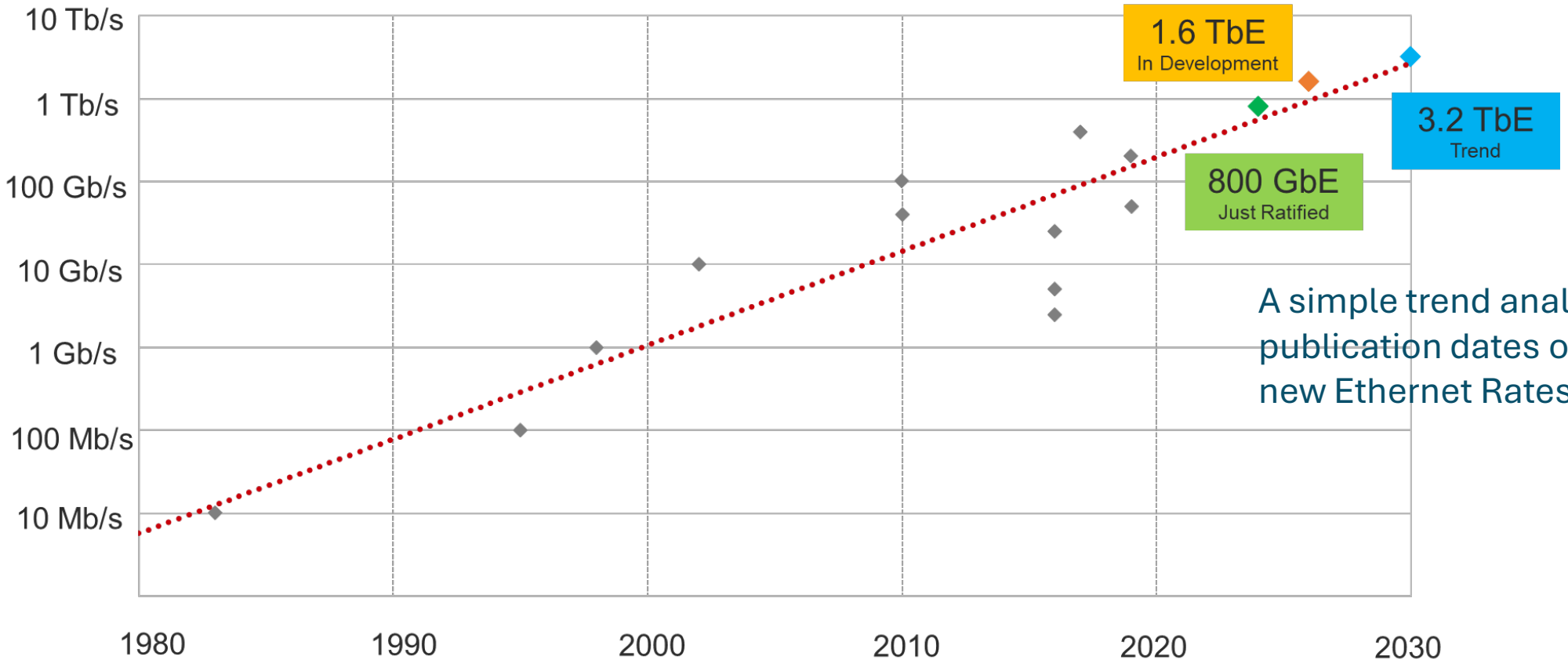
Exploring Technical Feasibility

- Any future effort in IEEE 802.3 will need to explore technical feasibility
- Addressing “Ethernet for AI” will require
  - Understanding application / performance requirements and potential trade-offs
  - Technical work
    - Signaling rate
    - Modulation
    - FEC
    - Channels
      - Will new approaches be necessary? Available?
  - Channel data for target interconnects
    - Sufficient frequency content to support analysis of effective data rate of 400 Gb/s, based on PAM4
- Exploration of technical feasibility may uncover items related to “broad market potential” or “economic feasibility” that need to be addressed

If 400 Gb/s Signaling is Feasible…



Just to be Provocative….



A simple trend analysis of publication dates of prior new Ethernet Rates

Year of publication for prior “New Ethernet Rates”

“Ethernet Interconnect for AI” Assessment

- Proposal: Assessment of “802.3 Ethernet Interconnect for AI” with an emphasis on >200 Gb/s signaling per lane
- Potential topics
  - AI Networks & Performance Requirements
  - Market Potential
  - What AI interconnects are target interconnects for Ethernet
  - Priorities
  - Exploring Technical Feasibility
  - Others…..
- Proposed Output
  - Records of meetings (Minutes & Presentations)
  - Consensus Presentation summarizing findings

Future Work

- Inform 802.3 WG on Thurs (23 Jan)
- Creation of Ad hoc page / reflector for “802.3 Ethernet Interconnect for AI” Assessment
- Focus on electronic meetings to progress effort
  - Teleconferences to be announced
- Reach out to organizations with potential input
  - ITU-T
  - OCP
  - OIF
  - SFF / SNIA
  - UALink
  - UEC
- General Call for information

THANKS