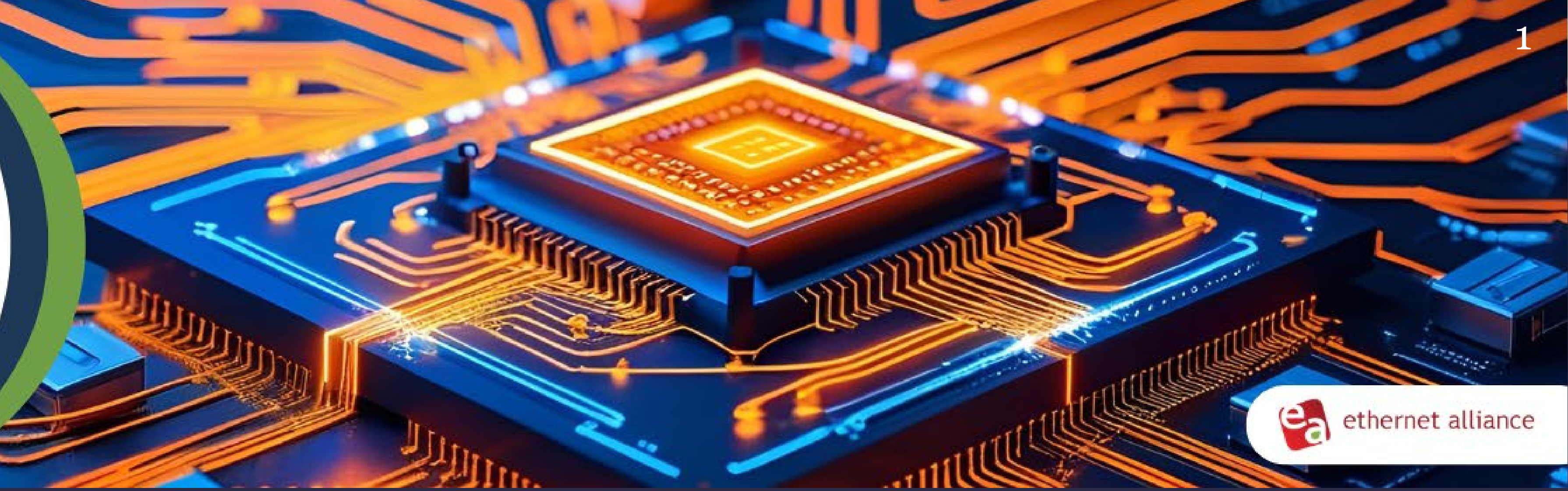
TEF 2025
Ethernet for AI

December 2-3, 2025
Hyatt Centric Mountain View, CA, USA



Ultra Ethernet: AI Evolution & Insights for 400G

December 2-3, 2025

This presentation is prepared for the Ethernet Alliance TEF 2025, and is intended to educate and promote the exchange of information.
Opinions expressed during this presentation are the views of the presenter, and should not be considered the views or positions of the Ethernet Alliance or the Ultra Ethernet Consortium.

Ultra Ethernet: AI Evolution & Insights for 400G

Adee Ran
Principal Hardware Engineer, Cisco
Co-chair, Ultra Ethernet Consortium Physical Layer Working Group

TEF 2025
Ethernet for AI

December 2, 2025

Outline

- Ultra Ethernet Consortium Physical Layer Working Group introduction
- Activities toward next generation signaling
- Summary of findings
- Focus area: minimum latency effect on Ultra Ethernet applications

UEC Physical Layer WG introduction

- The PHY WG is focused on enabling reliable operation of the physical layer for Ultra Ethernet applications
  - These applications require essentially lossless and error-free communication
  - In massive parallel processing, packet loss is intolerable; tails of statistical distributions dominate performance; “Tail Latency”
- Main development areas:
  - Providing guidance for estimation of **mean time between PHY errors (MTBPE)** in large networks
  - Defining PHY-level mechanisms to support **link-layer retry (LLR)** and **credit-based flow control (CBFC)**, enabling 100% reliability in UE links

IEEE 802.3 recently started the Ethernet Metadata Services Study Group to provide extensions to support UE PHY mechanisms in the Ethernet architecture

Activities toward next generation signaling

- The UEC membership is expected to be among the early adopters of next-generation Ethernet
- We want to help drive the PHY technology
  - But realized quickly that many SDOs working in parallel is inefficient
- The PHY WG conducted surveys within its participants to find the important areas to focus on
  - The intent was to influence objectives for future IEEE projects
  - Results were less conclusive than we hoped

Summary of findings

- Ultra Ethernet members are interested in 400G per lane for
  - Both scale-up and scale-out
  - Both single-lane (high radix) and multi-lane (high bandwidth) ports
  - Both electrical and optical media
  - Multiple network topologies
  - With and without retimers in the path

  … bottom line, a wide range of applications!
- One topic in agreement – **minimum latency is important**

Minimum latency?

- not “minimize latency”
- “minimum latency” is the inherent delay created by the specification and the physics

Latency effect on Ultra Ethernet applications

- Minimum latency has a quantifiable effect in UE, due to LLR
- A port that supports LLR (a MAC client layer) must store packets locally for possible retransmit, until their reception is acknowledged by the link partner
- LLR is intended to be used in short to mid-range links – with a reach goal of ~150 m (not a formal requirement)

Latency effect on Ultra Ethernet applications

- LLR requires buffering on each port
- Math is simple:

$$Port\ buffer\ size \geq\ (Roundtrip\ delay) \times (Effective\ port\ rate)$$

- Alternately, with fixed buffer size, there is a maximum supportable roundtrip delay:

$$Roundtrip\ delay \leq\ (Buffer\ size)/(Effective\ port\ rate)$$

Latency effect on Ultra Ethernet applications

- The speed of light defines a minimum for roundtrip delay for a given length of media
  - The speed of light in optical fiber is ~5 ns/m ➔ roundtrip delay in the medium is ~10 ns/m

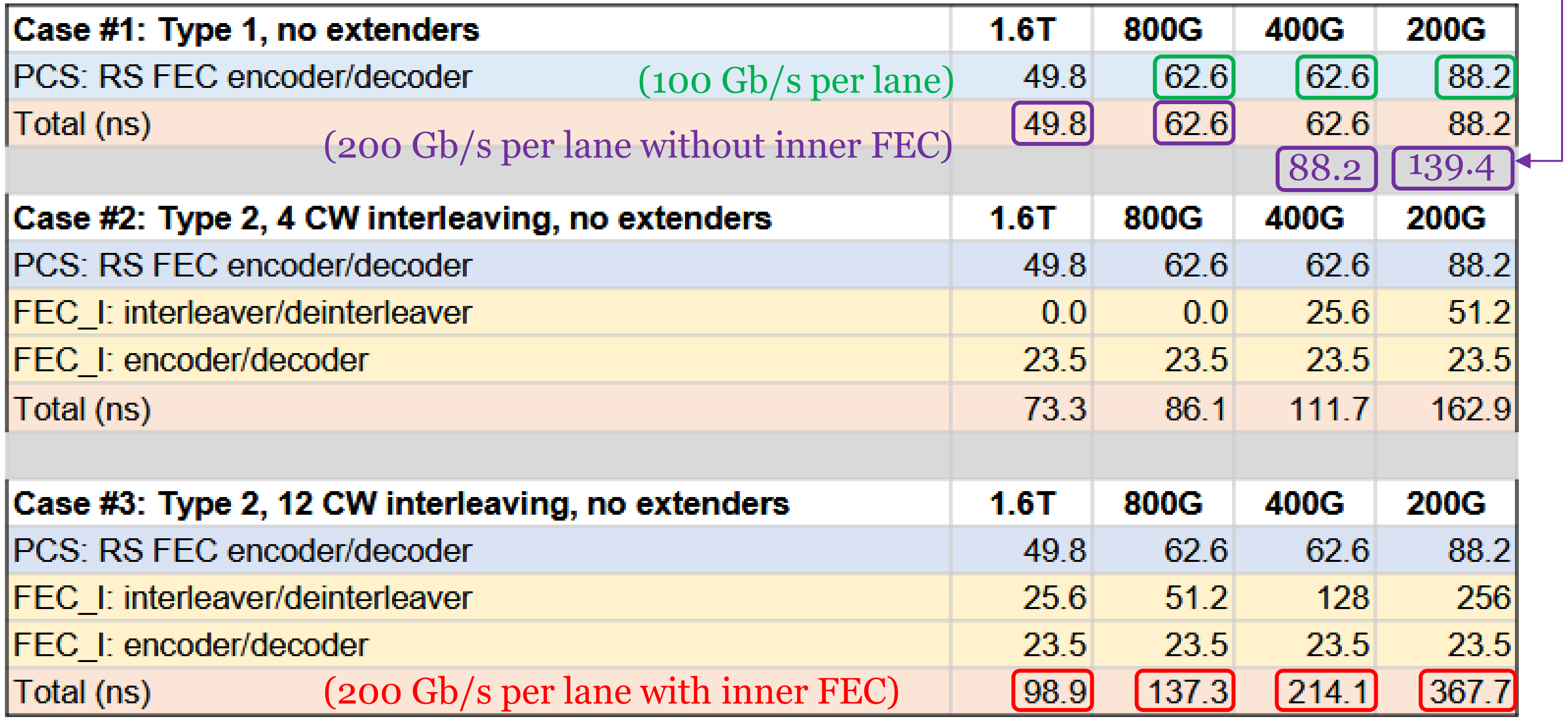
Not negotiable!

- Minimum PHY latency (Tx+Rx) is an additional term
  - In recent generations of Ethernet, the minimum PHY latency is strongly affected by RS-FEC interleaving
  - As of IEEE 802.3dj (D2.3):
    - PCS 4-way RS-FEC interleaving in 800GBASE-R and 1.6TBASE-R (inherent in the PCS)
    - 2-way RS-FEC interleaving in 200GBASE-R and 400GBASE-R, increased to 4-way by the PMA (see he_3dj_02a_2307)
    - Increased to 12-way RS-FEC interleaving with inner FEC (for reach >500 m)
  - **UE currently limits its scope to PHYs without inner FEC**

Minimum PHY latency analysis

- The numbers marked in purple /red are the estimated minimum latencies for 802.3dj PHYs (200 Gb/s per lane) in ns.
- The calculations assumed core clock frequency of 1 GHz.
- Implementations can have additional delays due to buffers, MAC processing, etc. – these are independent of FEC choice and media length.

Corrected for 4-way interleaving by the PMA



Case #1: Type 1, no extenders	1.6T	800G	400G	200G
PCS: RS FEC encoder/decoder (100 Gb/s per lane)	49.8	62.6	62.6	88.2
Total (ns)	49.8	62.6	62.6	88.2
(200 Gb/s per lane without inner FEC)			88.2	139.4
Case #2: Type 2, 4 CW interleaving, no extenders	**1.6T**	**800G**	**400G**	**200G**
PCS: RS FEC encoder/decoder	49.8	62.6	62.6	88.2
FEC_I: interleaver/deinterleaver	0.0	0.0	25.6	51.2
FEC_I: encoder/decoder	23.5	23.5	23.5	23.5
Total (ns)	73.3	86.1	111.7	162.9
Case #3: Type 2, 12 CW interleaving, no extenders	**1.6T**	**800G**	**400G**	**200G**
PCS: RS FEC encoder/decoder	49.8	62.6	62.6	88.2
FEC_I: interleaver/deinterleaver	25.6	51.2	128	256
FEC_I: encoder/decoder	23.5	23.5	23.5	23.5
Total (ns) (200 Gb/s per lane with inner FEC)	98.9	137.3	214.1	367.7

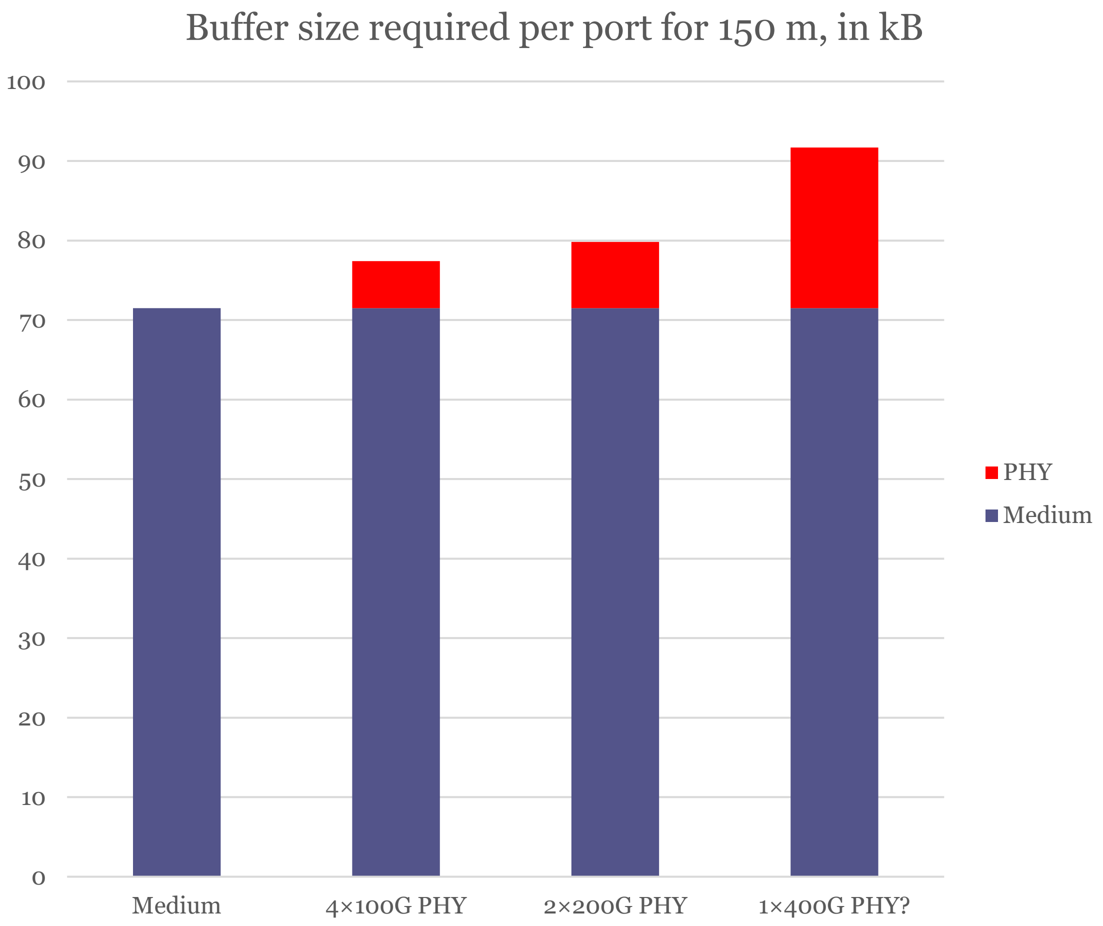
Source: Matt Brown, "MAC link latency considerations", IEEE 802.3dj (brown_3dj_optx_01c_230413, slide 6)

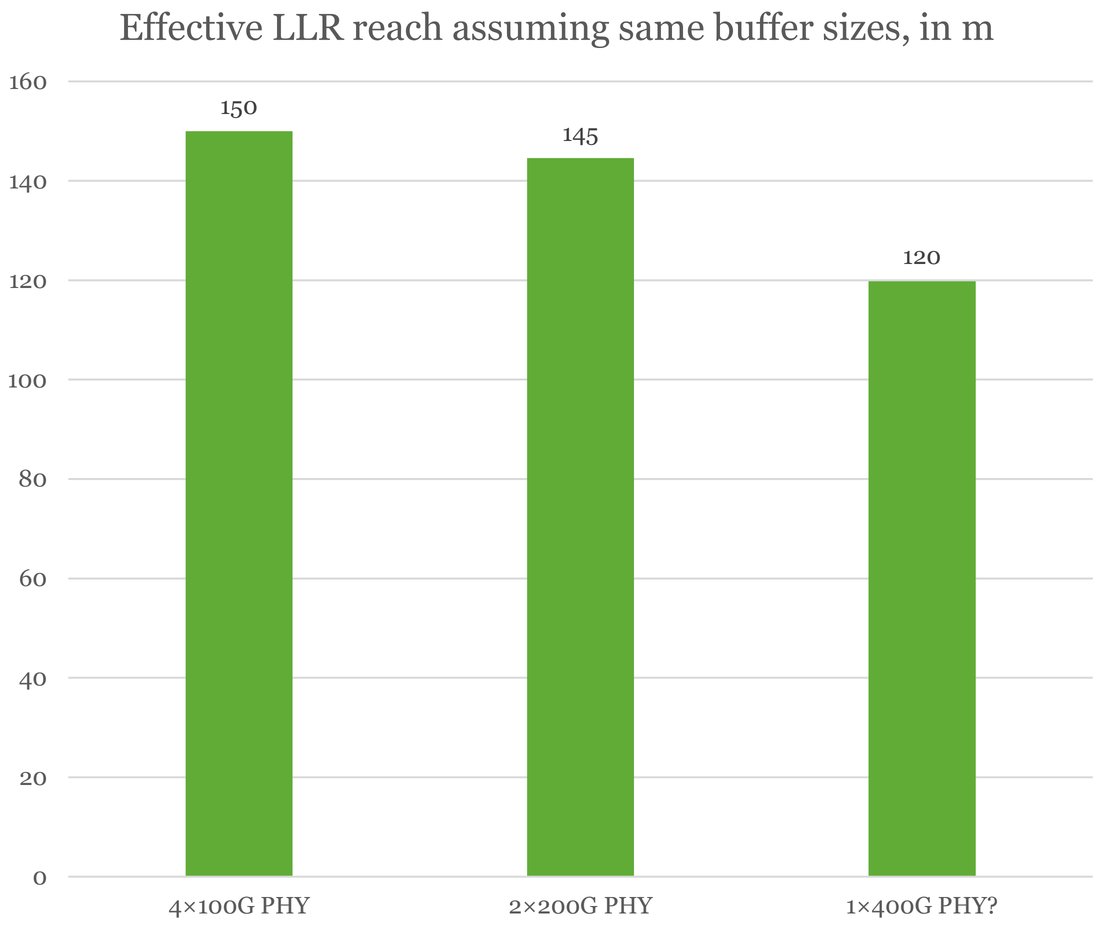
What does it mean?

- Consider a 400 Gb/s link with 90% utilization with a fiber length of 150 m
  - Roundtrip delay due to the medium is ~**1.5 μs**
  - Supporting LLR over this distance requires a buffer of **71.5 kB** on each side of the link
- Additional due to the PHY:
  - For 4x100G (2-way RS interleaving) the additional roundtrip delay due to the PHY is 2×62.6 ns ≅ **0.125 μs**
    - Equivalent to **12.5 m** of fiber
  - For 2x200G (4-way RS interleaving) it becomes 2×88.2 ns ≅ **0.18 μs**
    - Equivalent to **18 m** of fiber
  - **What about 1x400G?**
    - If it requires an inner FEC (as defined by 802.3dj, with 12-way RS interleaving), the PHY additional delay is 2×214.1 ns ≅ **0.43 μs**
    - Equivalent to **43 m** of fiber
- Additional buffers for the PHY on each side of the link:
  - +5.9 kB for 4x100G, total ~**77** kB (8% overhead)
  - +8.3 kB for 2x200G, total **80** kB (12% overhead)
  - +20.2 kB for 1x400G (with inner FEC), total ~**92** kB (28% overhead)
- **If the supported fiber reach is lower, the overhead is larger!**

The 802.3dj inner FEC with 12-way interleaving is used as an example. Other types of inner FEC with different delays due to interleaving or decoding may be considered.

Visually





TEF 2025 Ethernet for AI

Implication of limited LLR reach

- LLR is used within UE clusters with high internal connectivity to reduce the hop count
- Maximum LLR reach limits the physical size of the cluster and thus the number of nodes
- Example:
  - Assuming reduction of the reach from 150 m to 120 m – a factor of **0.8** (20% reduction)
  - If the cluster topology is mappable to a surface (2D): the number of nodes is reduced by a factor of 0.8^2=**0.64** (36% reduction)
  - If the cluster topology is mappable to a volume (3D): the number of nodes is reduced by a factor of 0.8^3=**0.512** (48.8% reduction)
  - The real effect is likely somewhere in between

What can we do?

- Increase buffer sizes?
  - In a radix-1024 switch with 200G per lane (4-way interleaving), LLR over 150 m requires 39 kB per port or **39 MB** in total
  - For a radix-1024 switch with 400G per lane and the same 4-way interleaving, the total is **80 MB**
  - If 400G per lane uses inner FEC, the total becomes **92 MB**
  - The relative impact is larger if we start from a lower supported reach
  - Acceptable?
- Decrease the inherent PHY latency?
  - If inner FEC is required, avoid additional interleaving (more likely, make it configurable)
  - Compromise the MTBPE – assuming LLR will compensate (but to what extent?)
  - Configuration/negotiation complexity
- Decrease the supportable reach?
  - Affects maximum cluster size
  - Acceptable?

Final notes

- In “The Economics of Latency” (ofelt_3dj_01_2305) it was suggested that “advanced knobs for experts” to reduce latency should be provided
  - This suggestion has not been implemented in 802.3dj
- With UE’s LLR we can navigate the tradeoff between minimum latency and FLR
  - Extensions to support the UE features are expected to be adopted into standard Ethernet
  - UE deployments are more “controlled” than typical front-end networks – advanced knobs can be more readily used
- Future Ethernet PHY specifications should provide such advanced knobs as standard features.

QUESTIONS?