Link Layer Retransmission (LLR) and Its Impact on Next-Generation 400 Gb/s/lane Standardization

IEEE 802.3 400Gb/s/lane Signaling Study Group

Runlong Hu(China Mobile), Haojie Wang(China Mobile), Weiqiang Cheng(China Mobile)

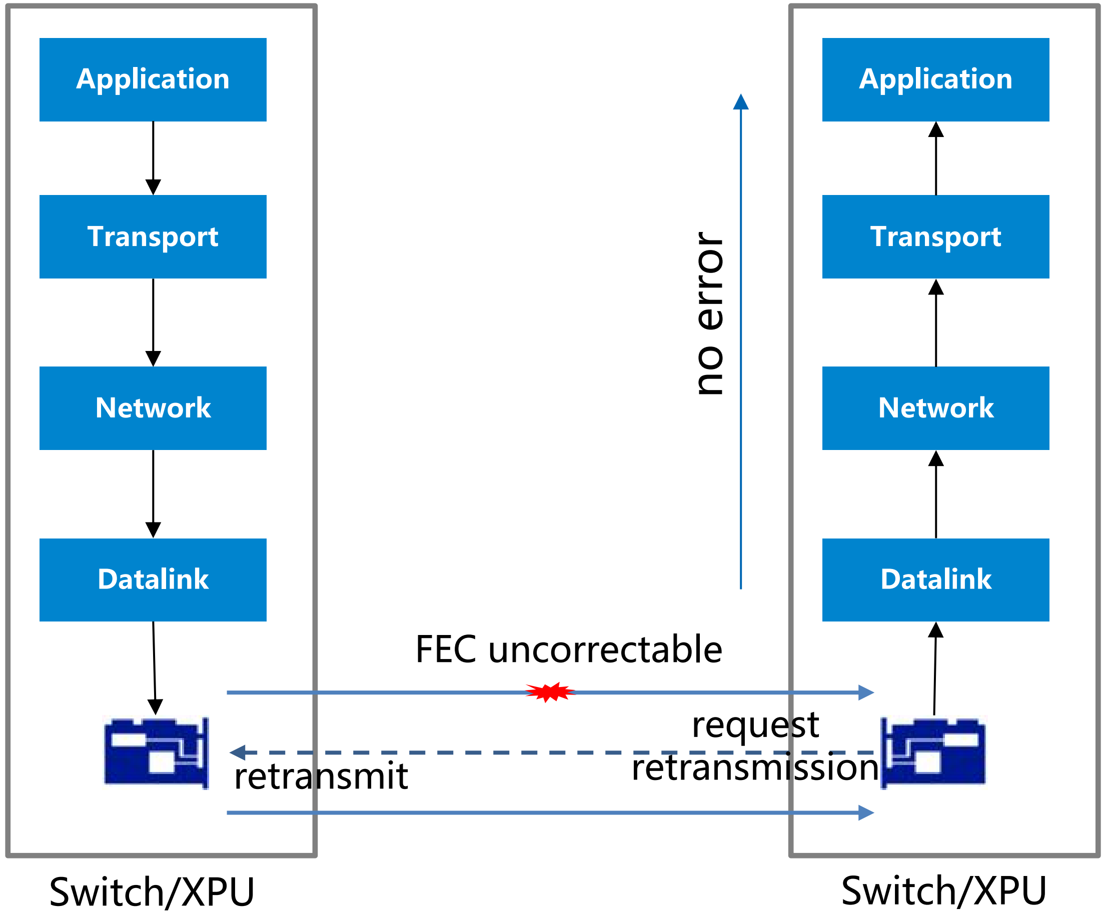
July 30, 2026

Motivation: 400 Gb/s/lane Is Pushing the Limit

- Nyquist frequency approaches ~100 GHz for PAM4; channel loss, reflection, and crosstalk become severe.
- Closing the link with FEC alone demands tighter PHY margins and/or stronger FEC.
- AIDC networks need both high reliability and controlled latency.

Question: Is there a complementary tool beyond heavier FEC?

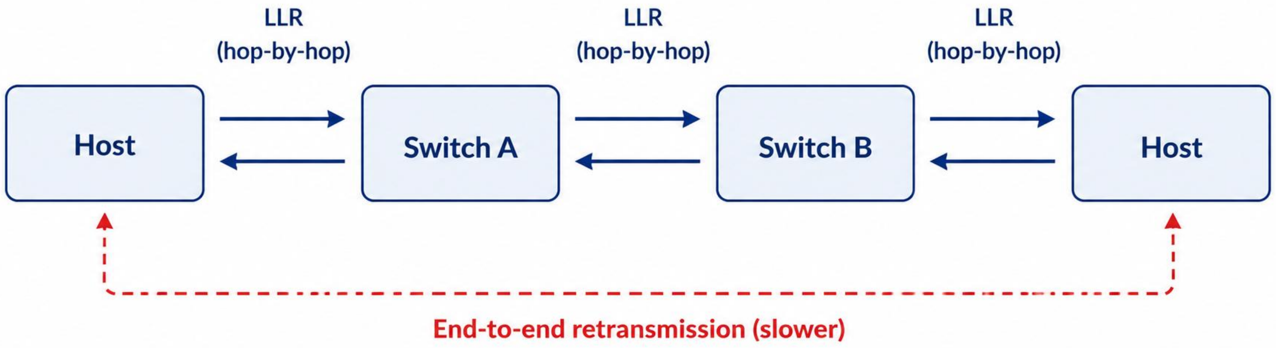
Introduction to LLR



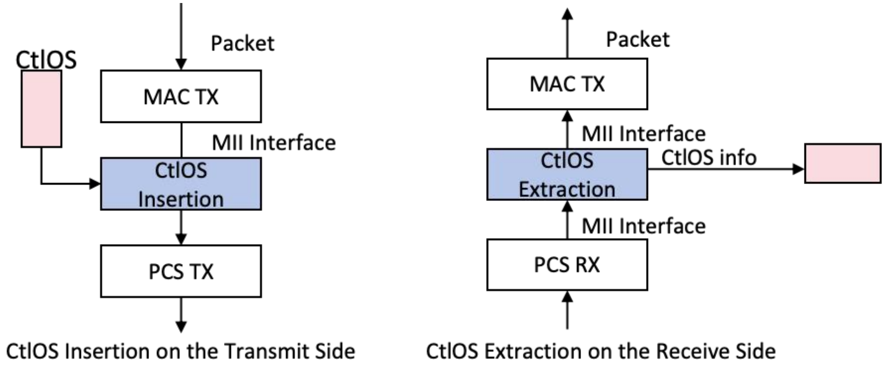
- Link Layer Retransmission: a hop-by-hop, data-link-layer recovery mechanism.
- Transmitter buffers frames until acknowledged; corrupted frames are retransmitted.
- Scope is between two adjacent link-layer entities; it does not span an end-to-end network path with multiple switching hops.

Credit: Ultra Ethernet Consortium (UEC) https://ultraethernet.org/ultra-ethernet-consortium-uec-launches-specification-1-0-transforming-ethernet-for-ai-and-hpc-at-scale/

LLR Is Latency-Efficient



LLR vs. End-to-End Retransmission



Transmit Side

										CtlOS
	8B_0	8B_1	8B_2	8B_3	8B_4	8B_5	8B_6	8B_7		
CtlOS Insertion	8B_0	8B_1	8B_2	8B_3	8B_4	8B_5	8B_6	CtlOS		
Next Beat Output	8B_7	8B_0	8B_1	8B_2	8B_3	8B_4	8B_5	8B_6		
Encoder	66b_0	66b_1	66b_2	66b_3	66b_4	66b_5	66b_6	66b_7		

Receive Side

										CtlOS
	66b_0	66b_1	66b_2	66b_3	66b_4	66b_5	66b_6	66b_7		
Decoder	8B_0	8B_1	8B_2	8B_3	8B_4	8B_5	8B_6	CtlOS		
CtlOS Extraction	8B_0	8B_1	8B_2	8B_3	8B_4	8B_5	8B_6			CtlOS

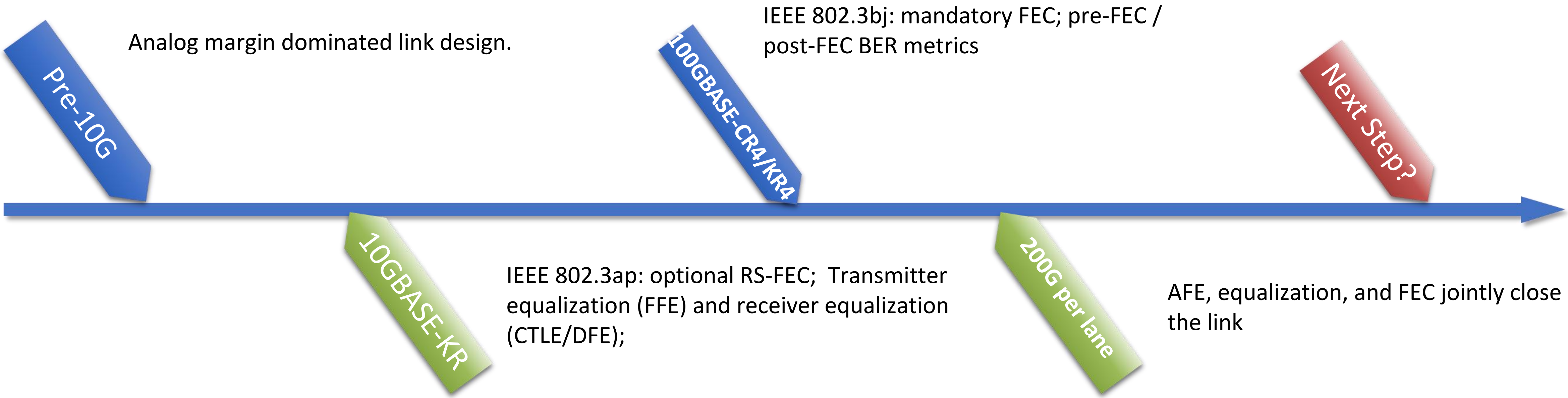
CtlOS preempting data stream

Credit: IEEE 802.3-li_ems_01a_251022
https://www.ieee802.org/3/EMS/public/25_1022/li_ems_01a_251022.pdf

The retransmission penalty can be small enough for latency-sensitive AI traffic

- Short-reach links have small propagation delay
  - One hop without crossing the full network path that has multiple switching stages.
  - Hide packet loss from upper layers, avoiding time-consuming protocols such as TCP timeouts and RDMA retries.
  - The round-trip time is dominated by retimer/DSP processing, and the cable flight time can be ignored.
- Control Ordered Sets (CtlOS) can carry LLR control messages (ACK/NACK/INIT) with low overhead in the PCS lane stream.
  - LLR control ordered sets can preempt in-flight data without queuing behind long Ethernet frames.

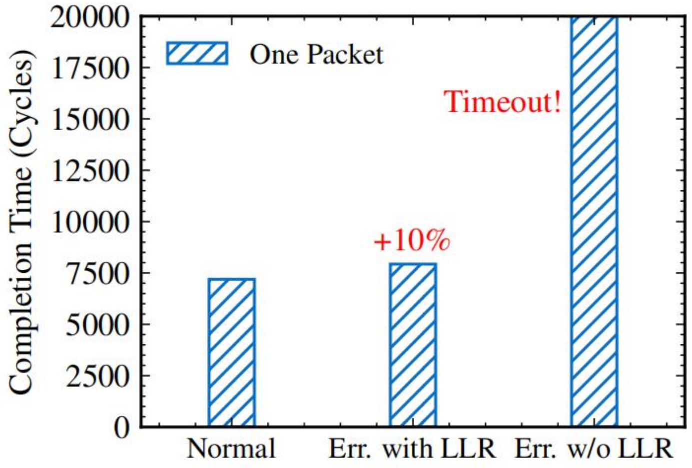
Historical Analogy

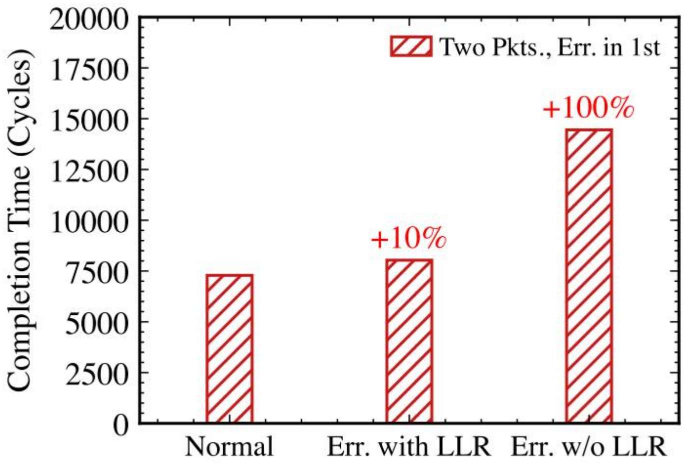


- Lower-rate links: Analog design alone left enough margin.
- Higher-rate links: Channel impairments including insertion loss, ISI, and crosstalk consumed the analog margin; digital recovery measures — equalization, coding, and FEC — were added to keep the link closed.
- 400G per lane: As Nyquist frequency rises, channel impairments worsen and raw link quality degrades. To close the link, heavier FEC is needed, but that brings higher latency and power. LLR offers an alternative way to recover residual errors without pushing FEC further.

The pattern: When analog margin alone is no longer enough, the industry adds a digital recovery mechanism.

Experimental Validation - Completion Time Under Packet Errors





* Test Environment: A single-switch topology with one network hop. GPU-A <-> SW <-> GPU-B emulated by FPGAs. Perform memory-semantics write operations, and measure RTT and goodput at the sender application.

One Error Packet

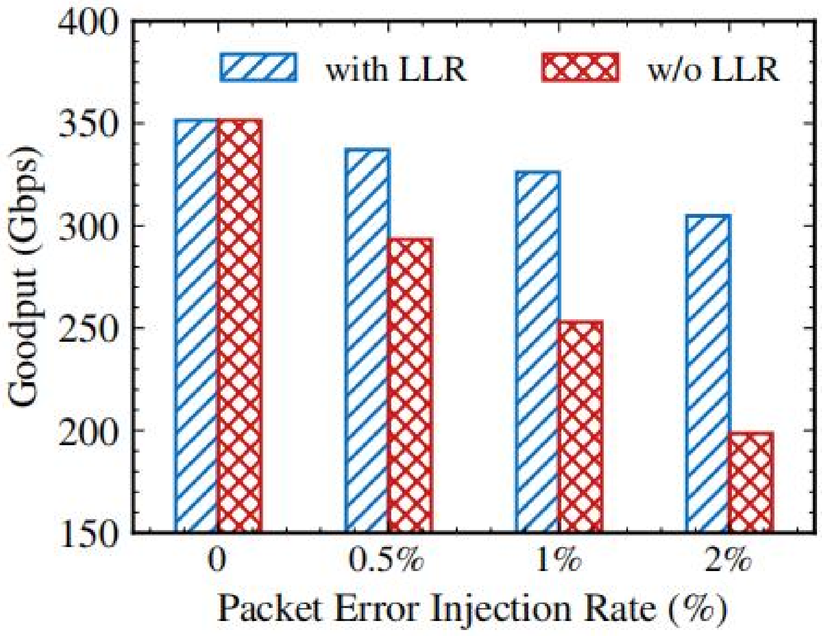
- Without LLR: Packet loss causes a timeout at the sender GPU, leading to a large increase in transmission completion time.
- With LLR: Link-layer retransmission adds only a ~10% increase in completion time.

Two Packets, Error Injected into the First

- Without LLR: The second packet arrives out of order at the receiver GPU, triggering a NAK and a sender-side retransmission. This doubles the completion time.
- With LLR: The local retransmission greatly reduces packet retransmission delay by ~90%.

Credit: Kefei Liu, Ruixue Wang, Tianrun Jiang, Runlong Hu, et. al. Evaluating Link-level Lossless Mechanisms in AI Networks. In SIGCOMM Workshop 2026

Experimental Validation - Goodput Under Packet Corruption



* The injected packet errors model the residual errors that remain after physical-layer equalization and FEC.

- With LLR, ACK packets consume only ~1% bandwidth, so baseline goodput is essentially unchanged.
- At 2% packet error injection, goodput with LLR stays near 305 Gbps, versus ~198 Gbps without LLR.
- LLR reduces goodput degradation from ~43% to ~13% at 2% error rate.

- PER can be translated into BER:
  - For independent bit errors, packet error rate = 1 - (1 - BER)^L ≈ L × BER, where L is the packet length.
  - For a 1500-byte packet (12,000 bits), 2% PER maps to about 1.7×10^{-6} BER.
  - The x-axis can therefore be read as a proxy for the post-FEC bit error rate seen by the link layer.

- Goodput captures the effective application throughput after retransmission penalties. Without LLR, a dropped packet triggers upper-layer timeout/retransmission; with LLR, recovery is local and fast.

Credit: Kefei Liu, Ruixue Wang, Tianrun Jiang, Runlong Hu, et. al. Evaluating Link-level Lossless Mechanisms in AI Networks. In SIGCOMM Workshop 2026

Impact on 400G/lane Standardization

1. Cross-layer coupling

- LLR is a data-link feature, but it changes the error-rate target the PHY must meet.
- PHY standardization therefore needs to leave room for LLR-aware operation.

2. Operating modes

- Define LLR-capable operation as an option, not a requirement.
- Distinguish “LLR-capable” and “non-LLR” modes.

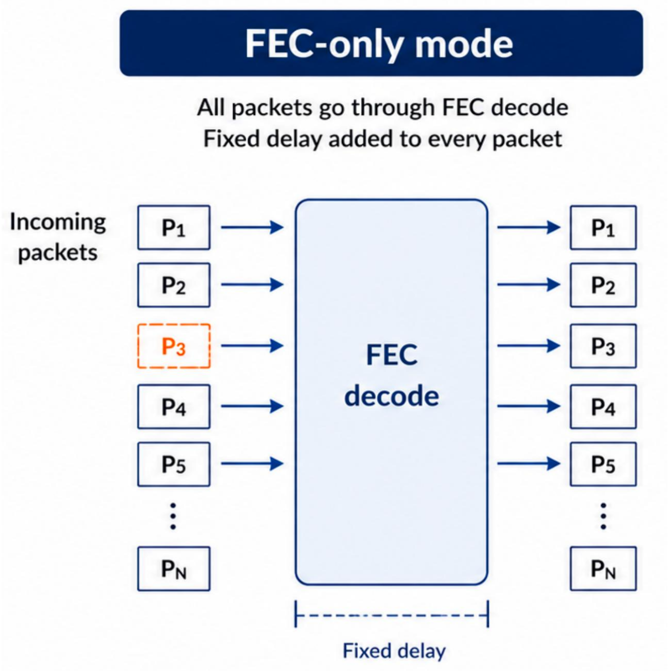
3. PHY/FEC budget impact

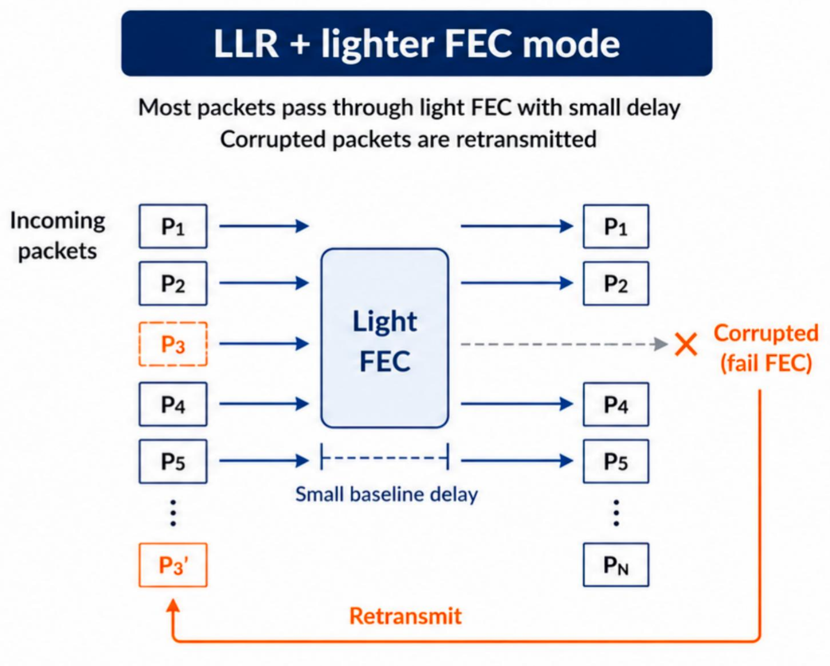
- With LLR, the end-to-end FLR/BER target can be met without pushing the raw PHY budget as hard.
- LLR can enable a lighter FEC while maintaining the same reliability target.
- Allow separate PHY/FEC budgets for LLR and non-LLR modes.

4. Interface & test requirements

- Specify interoperable test points for both modes.
- Provide a low-latency control-message carriage, e.g., via CtlOS. (accomplished by IEEE 802.3dt)

Impact on 400G/lane Standardization





5. Trade-off between LLR and FEC

- Heavier FEC adds deterministic latency and complexity to every frame; LLR can relax FEC, reducing that baseline latency and power.
- In normal operation, LLR avoids the latency cost of stronger FEC. But when retransmission is triggered, LLR still consumes time, even though its overhead is low.
- Unlike FEC — where pre-FEC/post-FEC BER gives a unified, continuous quality metric — LLR latency depends on how often retransmission occurs. Integrating LLR is therefore not as straightforward as simply replacing FEC with a single BER threshold.
- We need to weigh the latency saved by a lighter FEC against the latency cost of occasional retransmissions.

Summary

- 400G/lane repeats a historical pattern: when analog margin runs out, add a digital recovery mechanism.
- LLR is an optional feature that can relax PHY/FEC requirements for short-reach links.
- PHY standardization should consider LLR’s cross-layer impact on error budgets, latency, and FEC choices.