Consensus Building for Ethernet Metadata Services

David Ofelt (Juniper Networks/HPE), Kent Lusted (Synopsys),
Adee Ran (Cisco), David Law (HPE),
Eugene Opsasnick (Broadcom), Gary Nicholl (Cisco),
John D'Ambrosia (Futurewei, US Subsidiary of Huawei),
Kapil Shrikhande (Marvell), Mark Nowell (Cisco)

2025-07-22 802.3 CFI Consensus Meeting

Contributors

- Adam Healey (Broadcom)

Supporters

- Shawn Nicholl (AMD)
- Arthur Marris (Cadence)
- Xiang He (Huawei)
- David Estes (Spirent)
- Yan Zhuang (Huawei)
- Jeffery Maki (Juniper Networks/HPE)
- Mabud Choudhury (Lightera)
- Adee Ran (Cisco)
- David Law (HPE)
- Eugene Opsanick (Broadcom)
- Gary Nicholl (Cisco)
- John D'Ambrosia (Futurewei, US Subsidiary of Huawei)
- Kent Lusted (Synopsys)
- Kapil Shrikhande (Marvell)
- Mark Nowell (Cisco)
- David Ofelt (Juniper Networks/HPE)
- David Malicoat (Malicoat Networking Solutions)
- James Weaver (Arista Networks)
- Weiqiang Chen (China Mobile)
- Jieyu Li (China Mobile)
- Eric Maniloff (Ciena)
- Nathan Tracy (TE)

Supporters (2)

- Stephan Kehrer (Belden)
- Tom Huber (Nokia)
- Howard Heck (TE Connectivity)
- Li Xu (Huawei)
- Marco Mascitto (Nokia)
- Shimon Muller (Enfabrica)
- John Calvin (KeySight Technologies)
- Ed Nakamoto (Spirent)
- Mike Dudek (Marvell)
- Marcel Kiebling (Beckhoff Automation)
- Rick Rabinovich (Keysight Technologies)
- Yuki Murakmai (1Finity)
- Scott Sommers (Molex)
- Andy Moorwood (Keysight Technologies)
- Toshiaki Sakai (Socionext)
- Sam Kocsis (Amphenol)
- Jose Castro (Panduit)
- Ray Nering (Cisco)
- Michael He (Terahop)

Goals

- To measure the interest towards a CFI to address:
  - Ethernet Metadata Services
- We do **not** need to:
  - Fully explore the problem
  - Debate strengths and weaknesses of solutions
  - Choose a solution
  - Create a PAR or 5 Criteria
  - Create a standard
- Anyone in the room may vote or speak
- RESPECT … give it, get it

Agenda

- Background
- AI/ML Market
- AI/ML Systems
- Motivating Example
- What can IEEE 802.3 do?
- Proposal

What is Metadata?

- Metadata is data about data!

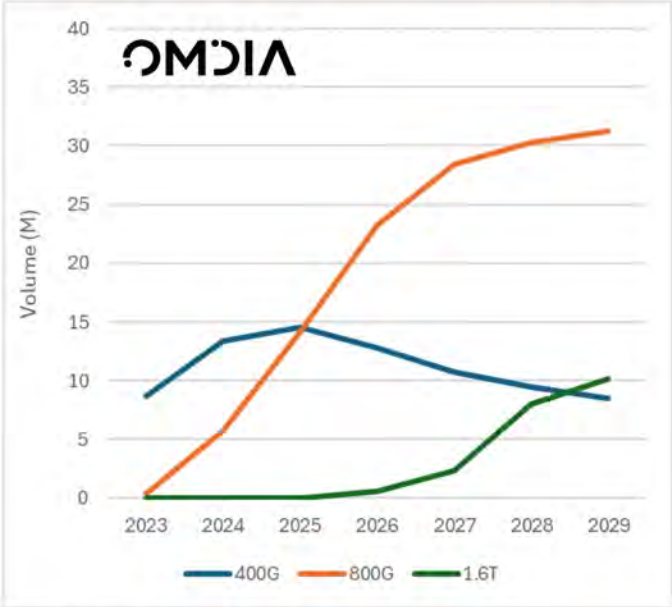
- Example: EXIF information attached to digital photos
  - Not the picture itself
  - But information about where the picture was taken, exposure, etc

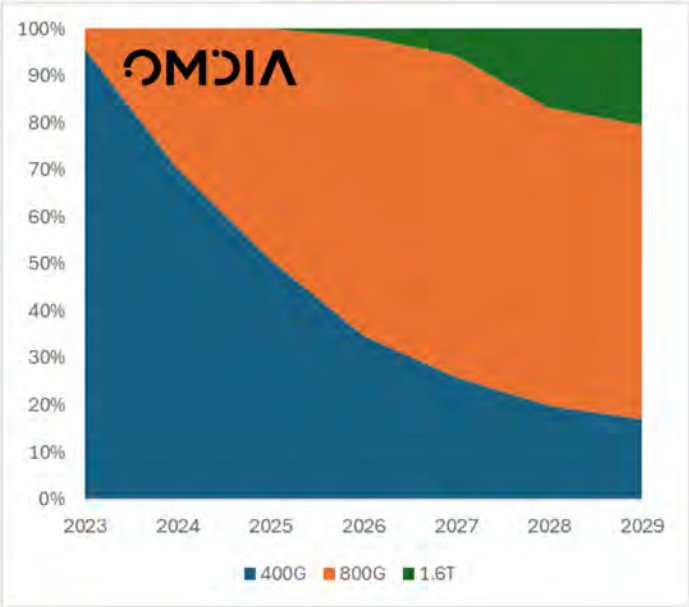
- For Ethernet-
  - Can be information associated with a packet
  - Can be information associated with the Ethernet link

- For Ethernet, new applications such as AI/ML networking can see utility in defining and using some Ethernet Metadata Service

AI/ML Market

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



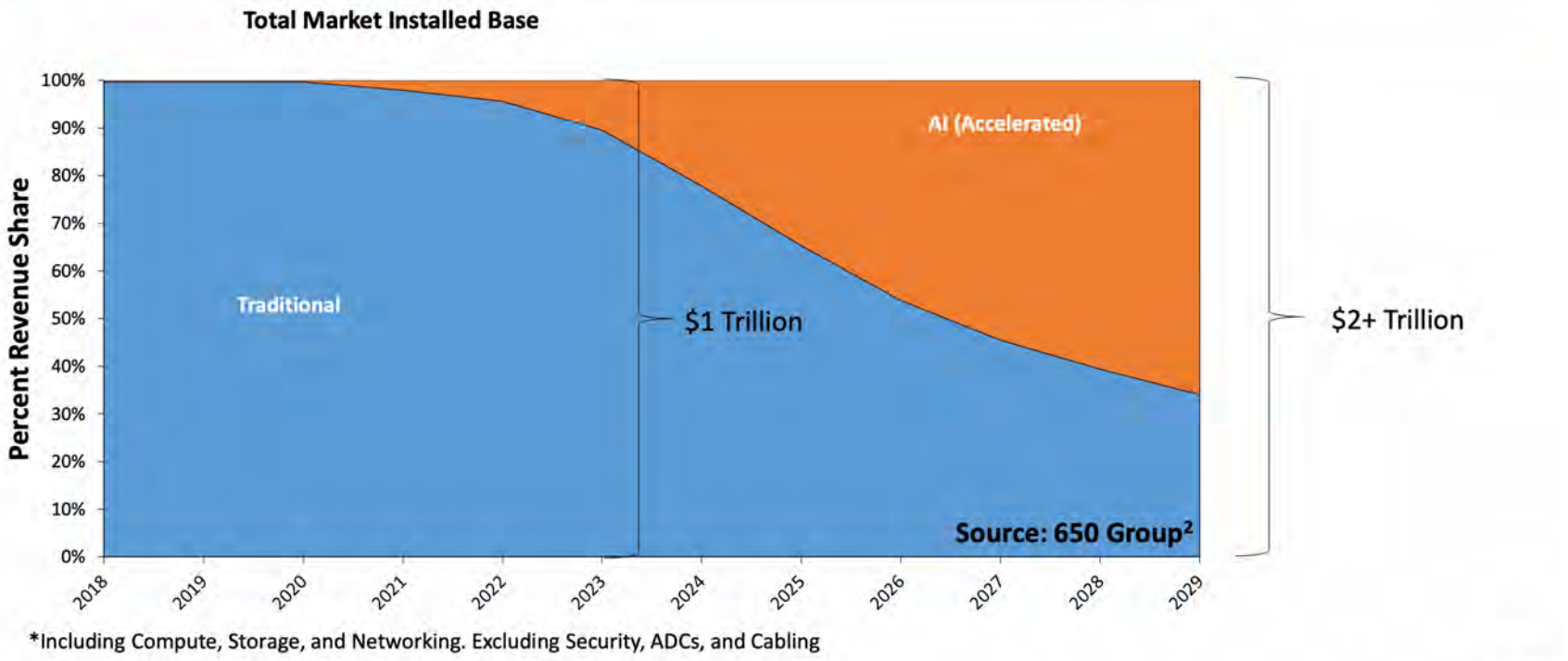


Data Source: Provided by and used with permission by Omdia.
Note – Update by Omdia currently underway and 800G and 1.6T modules expected to increase in 2025

https://www.ieee802.org/3/ad_hoc/ngrates/public/25_01/dambrosia_nea_01a_2501.pdf

Total Data Center (IT) Equipment: Total Market Installed Base

650 GROUP
MARKET INTELLIGENCE RESEARCH



3/21/2025

Page 5

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

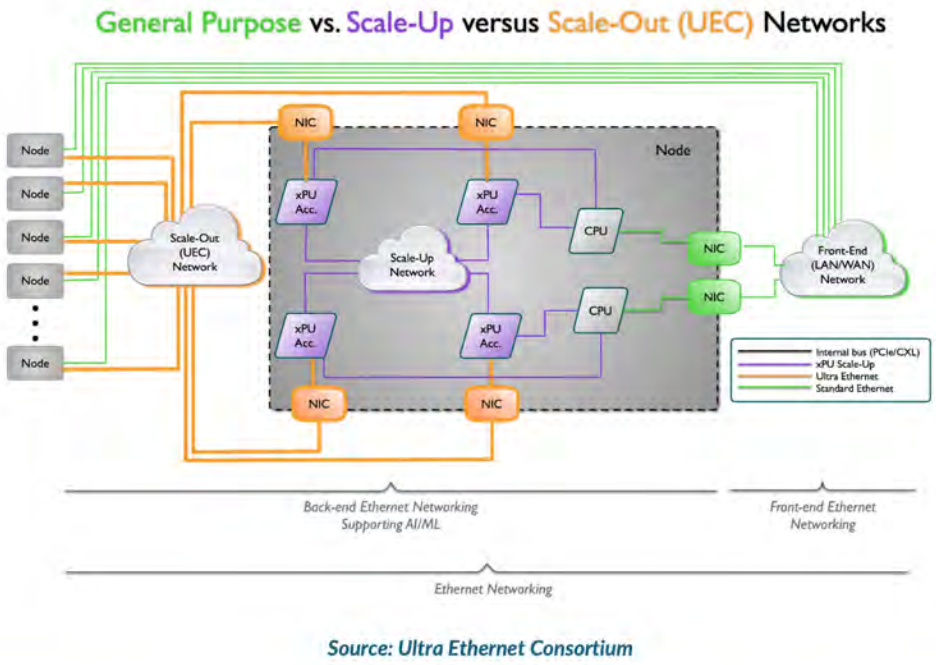
From: https://www.ieee802.org/3/ad_hoc/E4AI/public/25_0327/weckel_e4ai_01_250327.pdf

Recent Industry Activity

- Ethernet Alliance TEF
  - https://ethernetalliance.org/tef-2024-ethernet-in-the-age-of-ai-presentations-form/
- OIF Workshop
  - https://www.oiforum.com/meetings-events/448gbps-signaling-for-ai-workshop-apr-2025/

AI/ML Systems

What are AI Networks?

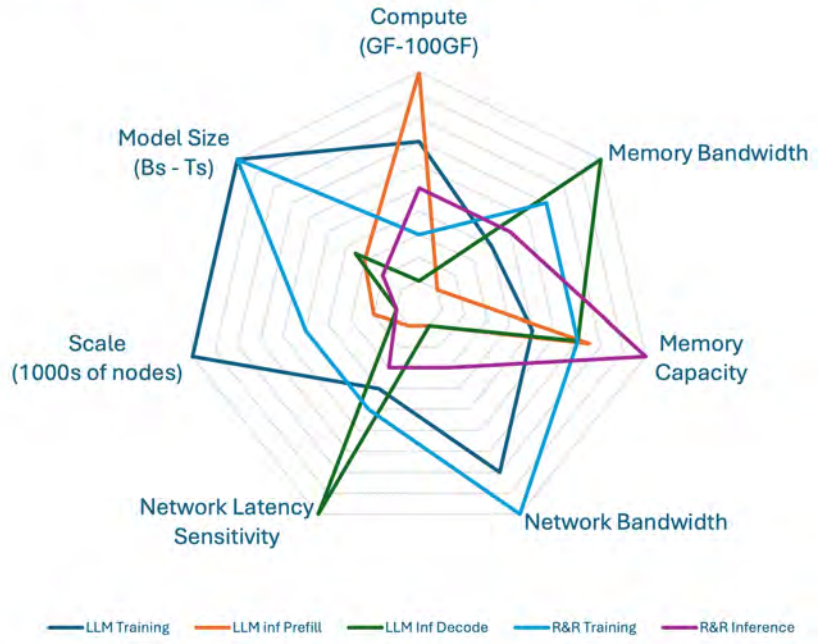


Source: Ultra Ethernet Consortium

- 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

https://www.ieee802.org/3/ad_hoc/ngrates/public/25_01/dambrosia_nea_01a_2501.pdf

AI Application Requirements



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

Ethernet Priorities (2 – 5 years)

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

Latency can be traded off for resilience, reach and power

https://www.ieee802.org/3/ad_hoc/ngrates/public/25_01/dambrosia_nea_01a_2501.pdf

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)*

https://www.ieee802.org/3/ad_hoc/ngrates/public/25_01/dambrosia_nea_01a_2501.pdf

Why are we talking about Metadata?

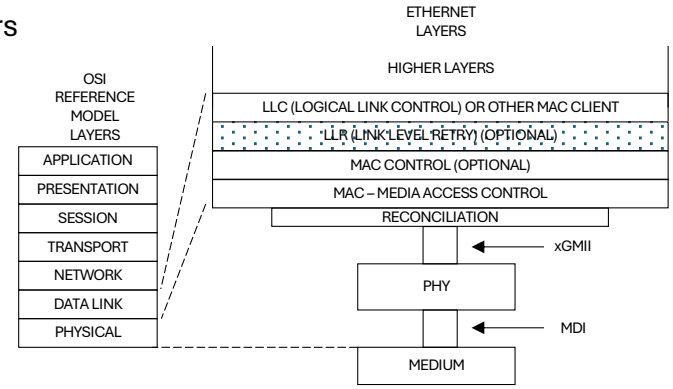
- AI/ML is driving innovations across the industry
  - Some of those are additions and modifications to Ethernet
  - Some of these are valuable and should be encouraged
- Two examples from the Ultra Ethernet Consortium[1]
  - LLR – Link Level Retry – improves link reliability
  - CBFC – Credit-Based Flow Control – improved flow control feature set
- These improve Ethernet's ability to target AI/ML/HPC applications
- Both live above the MAC but rely on PHY-level metadata features

[1] https://ultraethernet.org/uec-1-0-spec

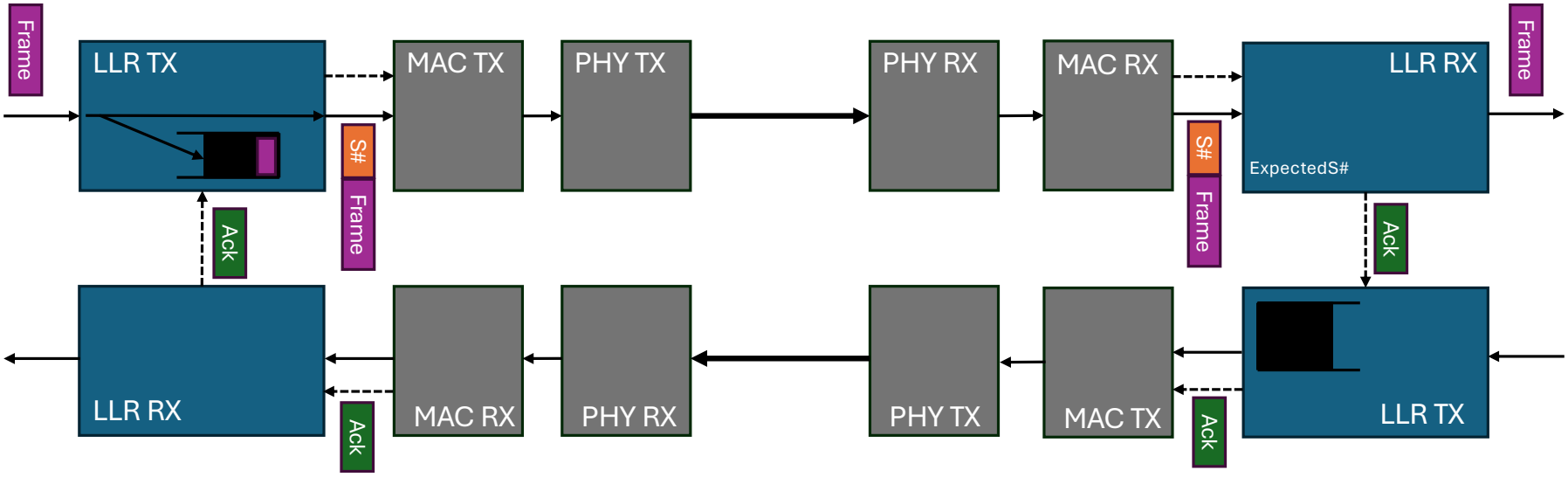
Motivating Example - Link Level Retry (LLR)

Link Level Retry (LLR)

- LLR compensates for frame loss by retransmitting lost frames
  - Example uses a simple go-back-N retry approach
  - Frame retransmission is transparent to the upper layers
  - Increases link resilience (e.g., reduced frame loss) presented to upper layers
- At the transmitter, each frame is:
  - Marked with a sequence number by the transmitter
  - Saved in a retransmit buffer
  - Per-frame metadata used to carry sequence number between endpoints
- When a frame successfully arrives at the receiver
  - An ACK is sent if the current frame is the next expected sequence number
  - A NACK is sent if it isn't
  - The ACK/NACK is sent using a frame-independent metadata channel
- When the original transmitter gets an:
  - ACK – it removes the ACKed frame from the retransmit buffer
  - NACK – it starts retransmitting frames from the NACKed one onward



Example of successful Frame transmission



Example Metadata Implementation

- Per-frame Metadata
  - Can use some of the preamble bits to encode the sequence number
- Frame-independent Metadata
  - Can use ordered sets to send the ACK/NACKs

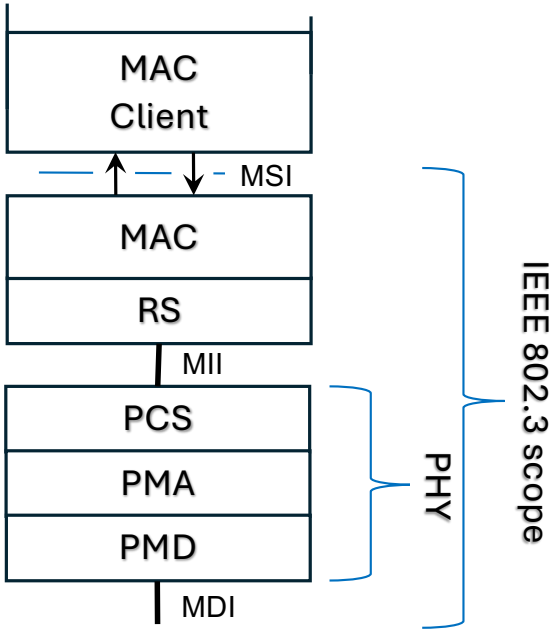
How to add features to IEEE 802.3 Ethernet?

Two pieces to a new feature

- Need a way of getting information in and out of the Ethernet stack
- Need a way of transmitting information between endpoints (changes to the PHY)

MAC Service Interface

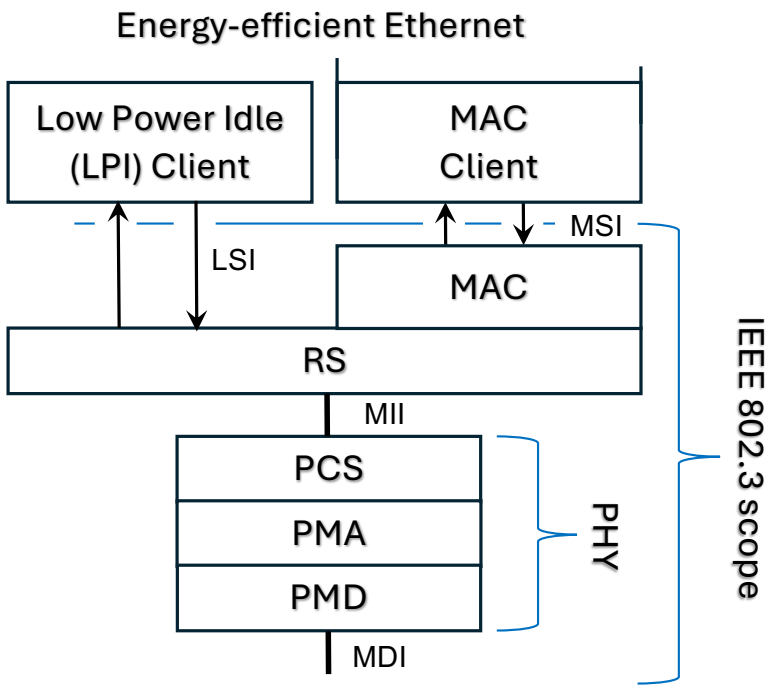
- MAC service interface specification
  - MAC Client can only supply:
    - destination address
    - source address
    - mac service data unit (the rest of the frame)
    - (optional) frame check sequence
  - Contract with the rest of the world
- To support the extensions:
  - Need to provide new ‘services’ to upper layer
- If there is a desire to support these features within the IEEE 802.3 Ethernet Standard, they would need to fit into the IEEE 802.3 Ethernet architecture



MSI: MAC abstract service interface
MII: Media Independent Interfaces
MDI: Medium Dependent Interfaces

Existing approaches

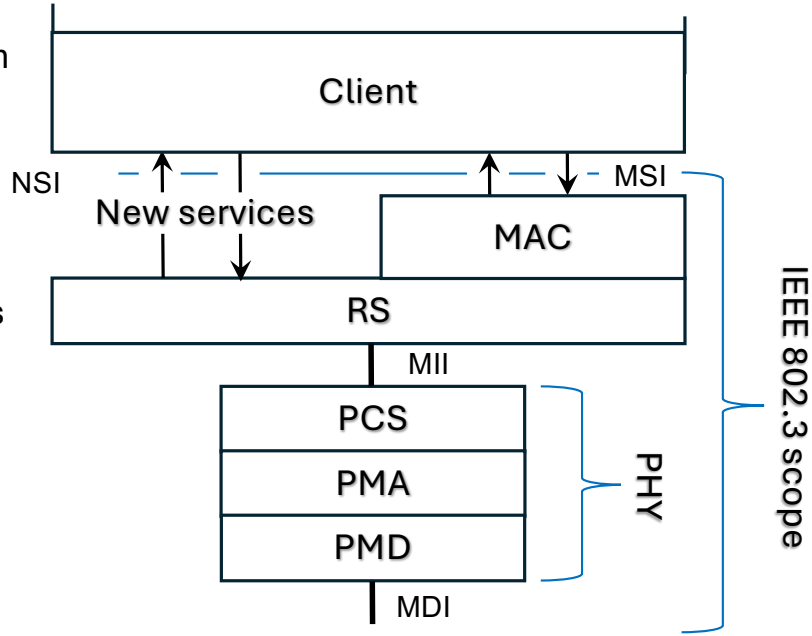
- IEEE 802.3 Ethernet Working Group approach
  - Avoid changing the IEEE 802.3 Ethernet MAC
    - MAC is the 'core' of IEEE 802.3 Ethernet standard
  - Don't change the MAC service interface
    - Contract with the rest of the world (e.g. IEEE 802.1)
- But new 'services' have been added!
  - Support for time synchronisation protocols
  - EPON Multipoint MAC Control
  - Energy-efficient Ethernet
  - Packet pre-emption
- So, how does this work without changing the MAC or MAC Service Interface?
  - Adding new services in parallel to MAC service interface
- Some things would require changes to the MAC
  - The recent cut-through switching discussion
  - The current set of ideas being proposed aren't a problem



MSI: MAC abstract service interface
MII: Media Independent Interfaces
MDI: Medium Dependent Interfaces
LSI: LPI abstract service interface

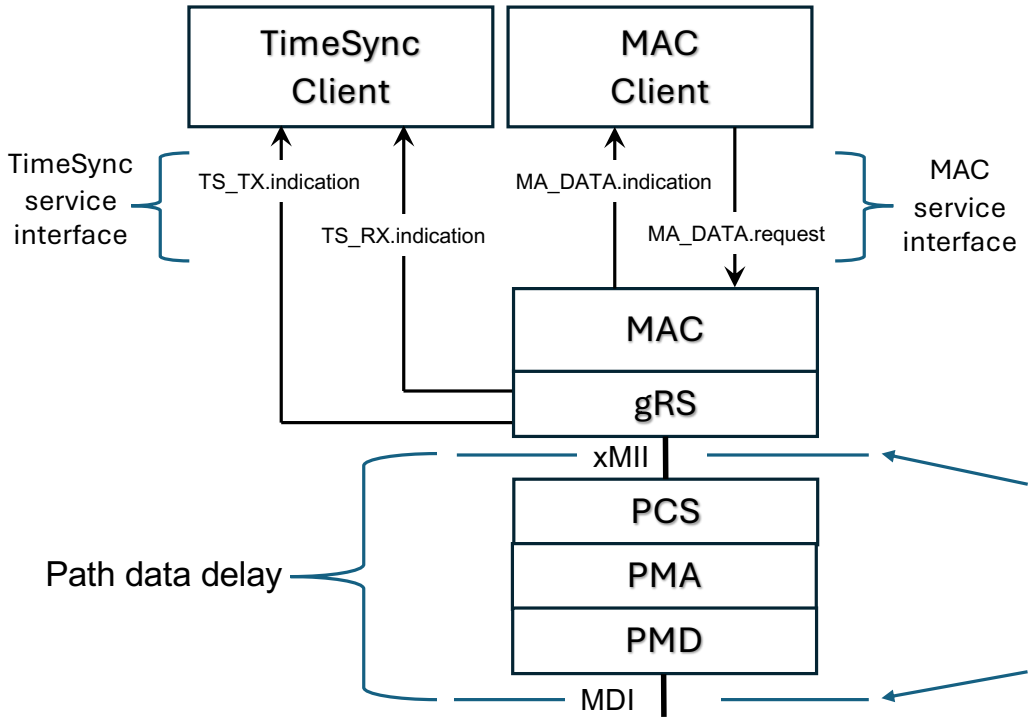
Existing approaches

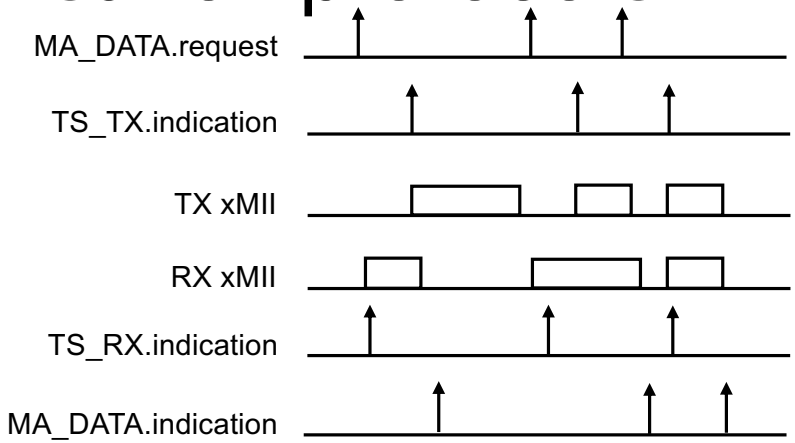
- Much of this discussion is about our standard's formalism rather than implementation
  - Additions need to fit into IEEE 802.3 architecture
- We can add extensions using additional client interfaces that live in parallel to the MAC service interface as before
- The PHY is within the IEEE 802.3 scope so adding features there is easy once we have an interface to the upper layers



MSI: MAC abstract service interface
MII: Media Independent Interfaces
MDI: Medium Dependent Interfaces
NSI: New Service Interface

Existing approach
Support for time synchronisation protocols





Ingress/egress time measured by TimeSync client at xMII

TimeSync calculates ingress/egress time at MDI by adjusting measured value by path data delay

Changes to the PHY

- The PHYs are owned by IEEE 802.3
  - Can make whatever changes that we think are useful
- Need to pick which PHYs are supported
- New features added to existing PHYs need to be optional

What can IEEE 802.3 do?

Innovation outside of IEEE 802.3

- Extensions to Ethernet to better support AI/ML workloads are being proposed in groups outside the IEEE 802.3 Ethernet Working Group
- These extensions are built on:
  - A mechanism for per-packet metadata
  - A packet-independent metadata channel
- Supporting these features in IEEE 802.3 can
  - Help Ethernet address the current AI/ML market
  - Provide useful tools for other innovations using Ethernet
  - And do so in an extensible and multi-vendor interoperable way

Other uses of Metadata

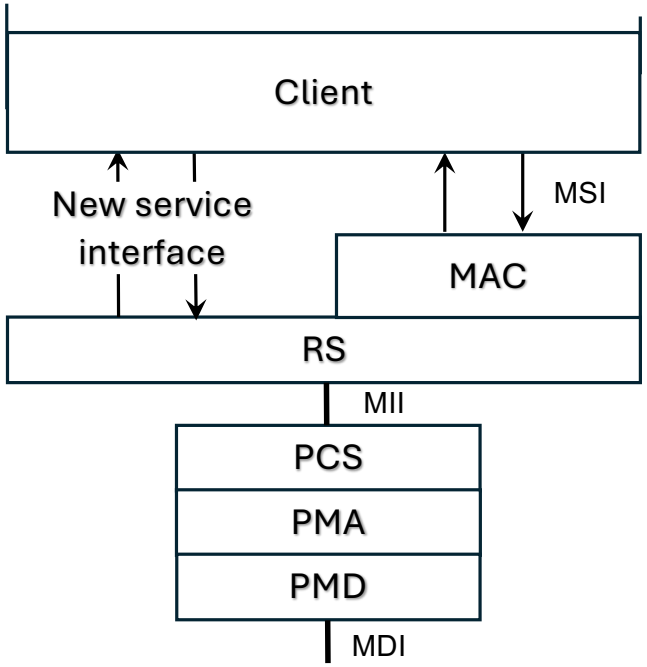
- Several examples of per-packet metadata in use
  - EPON (802.3)
  - Packet Preemption (802.3)
  - Proprietary channelization approaches
  - UEC Link Layer Retry (LLR)[1]
  - UEC Credit-Based Flow Control (CBFC)[1]
- Also examples of packet-independent metadata
  - Local/Remote FEC Degrade Signaling (802.3)
  - FlexE Message Channel
  - MAC Merge/Preemption (if you stare at it the right way) (802.3)

[1] As described in the *Ultra-Ethernet Consortium Ethernet Extensions* liaison letter: https://www.ieee802.org/3/minutes/mar25/incoming/2025.03.10%20UEC%20Ordered%20Sets%20Letter%20-%20signed.pdf

Proposal

Proposal

- Add support for Metadata Services with new service interfaces
  - IEEE 802.3 Ethernet MAC unchanged
  - MAC Service Interface unchanged
- Add support for Metadata Services to appropriate PHYs
  - Client can provide per-packet metadata
  - Client can provide packet-independent metadata channel
- PHY type (data rate and media) specific
  - Management entity queries the PHY type
    - Certain PHY types cannot support this service
  - Can be optional in all supported PHY types
  - Both ends need to support
  - Use can be negotiated with LLDP



MSI: MAC abstract service interface
MII: Media Independent Interfaces
MDI: Medium Dependent Interfaces

Summary

- The Ethernet market continues to grow with significant new opportunities coming from AI/ML
  - Ethernet needs to evolve to support these applications better than competing technologies
  - Providing the right building blocks will enable this support to be added
- This presentation introduced the concept of metadata services
  - These enable per-packet and packet-independent metadata
- Implementation of these ideas will involve:
  - Extensions to the RS to provide a new service interface
  - Extensions to the appropriate PHYs to support the new metadata services
- IEEE 802.3 Ethernet MAC and MAC Service Interface unchanged

Next Steps

- ~~CFI Consensus Building Meeting~~
  - ~~Teleconference the week before the July plenary in Madrid~~
- The actual CFI & Study Group Formation Vote
  - CFI during opening IEEE 802.3 meeting at July Plenary
  - Vote for Study Group formation at the closing IEEE 802.3 meeting

Straw Polls

- Should a study group be formed to develop a PAR, CSD responses, and objectives for “Ethernet Metadata Services”?
  - Yes: 49
  - No: 0
  - Abstain: 5

- If formed, will you participate in this Study Group?
  - Tally: 40

- Unique affiliations for those who indicated they’d participate (post processed)
  - Tally: 26

Thank You!

Questions?