Physical Layer Aggregation update

Howard Frazier
Shimon Muller
Drew Perkins
November, 2006
Dallas, TX

Outline

- Application needs
- Resources
- 802.3ad Link Aggregation
- 802.3ah PME Aggregation
- Aggregation at the Physical Layer (APL)
- Configuration examples
- Application with various PHYs
- Summary

Application needs

Backplane

- ✓ Scalable speed
- ✓ Resiliency
- ✓ Many ports in small space
- ✓ Trade-off cost of signaling rate vs. number of channels

Copper (data center)

- ✓ Scalable speed
- ✓ Resiliency
- ✓ Relatively inexpensive medium
- ✓ Cable characteristics limit signaling rate

Short Haul Fiber

- ✓ Scalable speed
- ✓ Resiliency
- ✓ Relatively inexpensive medium
- ✓ Cable characteristics limit signaling rate

Long Haul Fiber

- ✓ Scalable speed
- ✓ Resiliency
- ✓ Relatively expensive medium
- ✓ Economics limits signaling rate

Resources

Backplane

- 4 x 2.5 G common today
- Variable lane widths common
  - XAUI, IB, PCI-e
- 10GBASE-KR nearing completion
- Signaling rates > 10 G will be challenging

Copper (data center)

- 10GBASE-T, 10GBASE-CX4
- With time, cost and power consumption will decline
- Signaling rates > 10 G will be challenging

Short Haul Fiber

- 10GBASE-SR, 10GBASE-LX4, 10GBASE-LRM
- Multimode fiber at 850 & 1300 nm
- Signaling rates > 10 G will be challenging

Long Haul Fiber

- 10GBASE-LR, 10GBASE-ER, 10GBASE-LW, 10GBASE-EW
- Single mode fiber at 1300 & 1550 nm
- Signaling rates > 10 G will be expensive

Resources

Backplane

- 25 G serial may require multilevel signaling, higher cost channels, more signal processing
  - higher power, cost and complexity
- n x 25 G will not be economically feasible any time soon

Copper (data center)

- Quad (octal?) 10GBASE-T PHYs drive growth of multi-port switches
- Packaging in multiples of four fits well with aggregation in multiples of four

Short Haul Fiber

- Parallel ribbon fiber
- VCSEL arrays demonstrated
- Multi-port (quad?) 10GBASE-LRM

Long Haul Fiber

- Photonic Integrated Circuits
  - 10 x 10 G demonstrated
- Declining cost of WDM

802.3ad Link Aggregation

- Specified in Clause 43
- LAG is performed above the MAC
- LAG assumes all links are:
  - full duplex
  - point to point
  - same data rate
- Provides graceful recovery from link failures
- Traffic is distributed packet by packet
- All packets associated with a given “conversation” are transmitted on the same link to prevent mis-ordering

802.3ad Link Aggregation

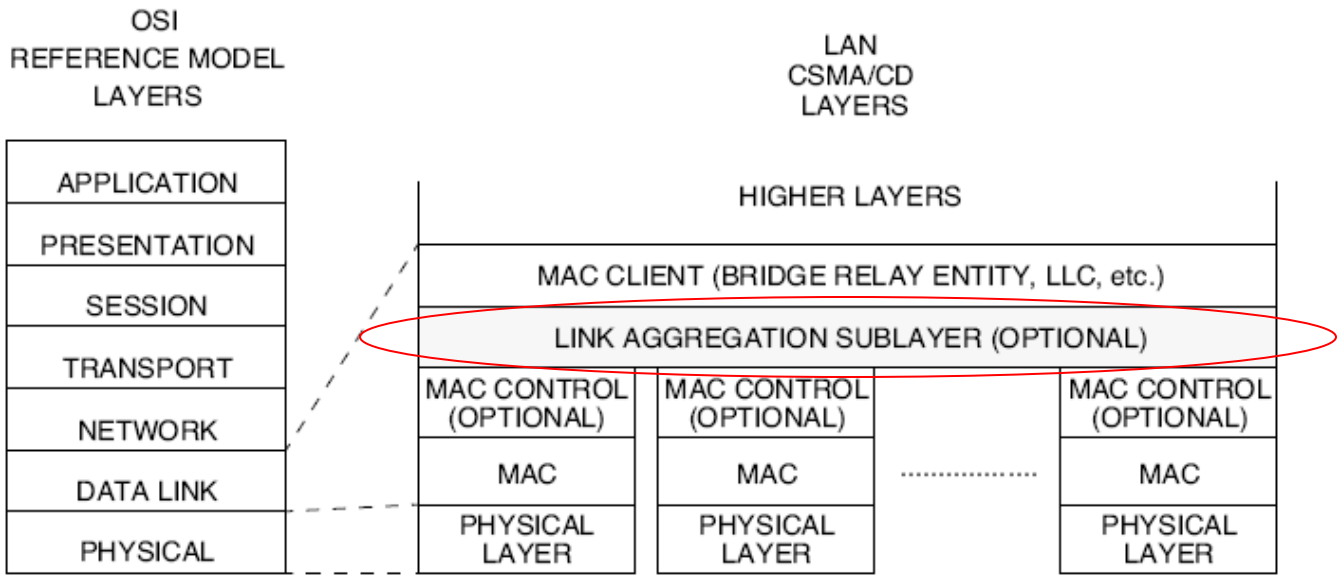
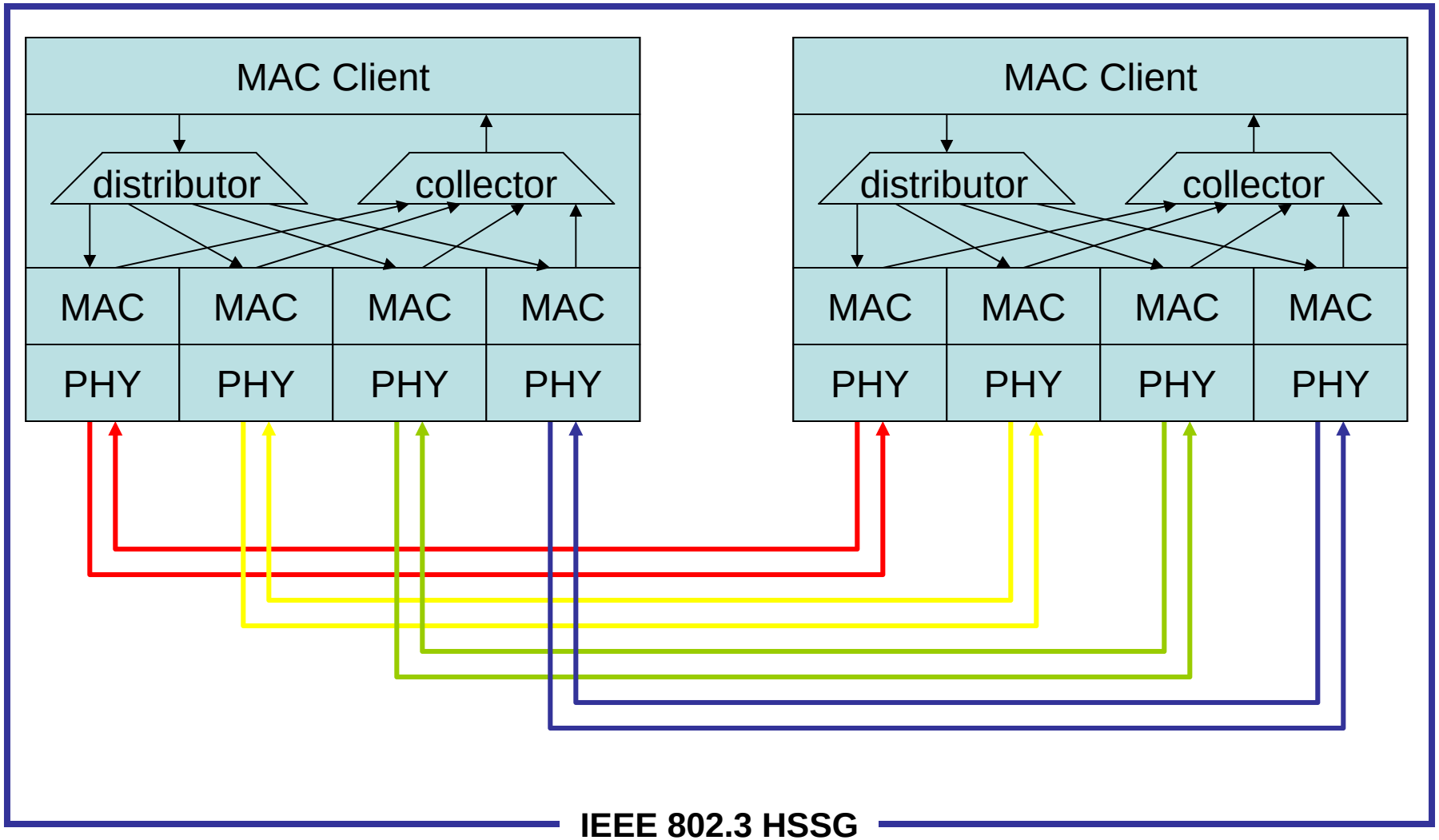


Figure 43–1—Architectural positioning of Link Aggregation sublayer

802.3ad Link Aggregation



802.3ad Link Aggregation

43.2.4 Frame Distributor

…

This standard does not mandate any particular distribution algorithm(s); however, any distribution algorithm shall ensure that, when frames are received by a Frame Collector as specified in 43.2.3, the algorithm shall not cause

a) Mis-ordering of frames that are part of any given conversation, or

b) Duplication of frames.

The above requirement to maintain frame ordering is met by ensuring that all frames that compose a given conversation are transmitted on a single link in the order that they are generated by the MAC Client; hence, this requirement does not involve the addition (or modification) of any information to the MAC frame, nor any buffering or processing on the part of the corresponding Frame Collector in order to re-order frames.

…

802.3ad Link Aggregation

- Does not change packet format
  - No added headers or sequence numbers
  - Type/Length interpretation unchanged
- Does not require added buffers
  - No fragmentation or reassembly
- Does not re-order or mis-order packets
- Does not add significant latency
- Does not increase the bandwidth for a single conversation
- Achieves high utilization only when carrying multiple simultaneous conversations
- Is not transparent to some 802.1 sub-layers

802.3ad Link Aggregation

- Can link aggregation be “fixed”?
- Not so easy to accomplish
  - Inspect headers deep into packet

    or
  - Add sequence number to packet
    - Change the packet format, start a food fight with dot1

    and
  - Add LARGE buffers to receiver
  - Add LONG delay
- Still will not achieve linear scaling

802.3ad Link Aggregation

- Is a very good thing
  - It does what it was intended to do
  - It is relatively easy to implement and use
- Does not always provide a linear multiple of the data rate of a single link
  - N aggregated links usually do not provide N times the bandwidth
- Incurs a linear multiple of the cost of a single link
  - N aggregated links cost N times as much as a single link, because everything must be replicated
- Appears to the user as N individual links, which must be individually managed

802.3ah PME Aggregation

- Specified in Clause 61
- Supports multi-pair bonding for 10PASS-TS and 2BASE-TL
- Performed in the Physical Layer
- Aggregates up to 32 Physical Medium Entities (PMEs)
- Ensures low packet latency and preserves packet sequence
- Scalable and resilient to PME failure

802.3ah PME Aggregation

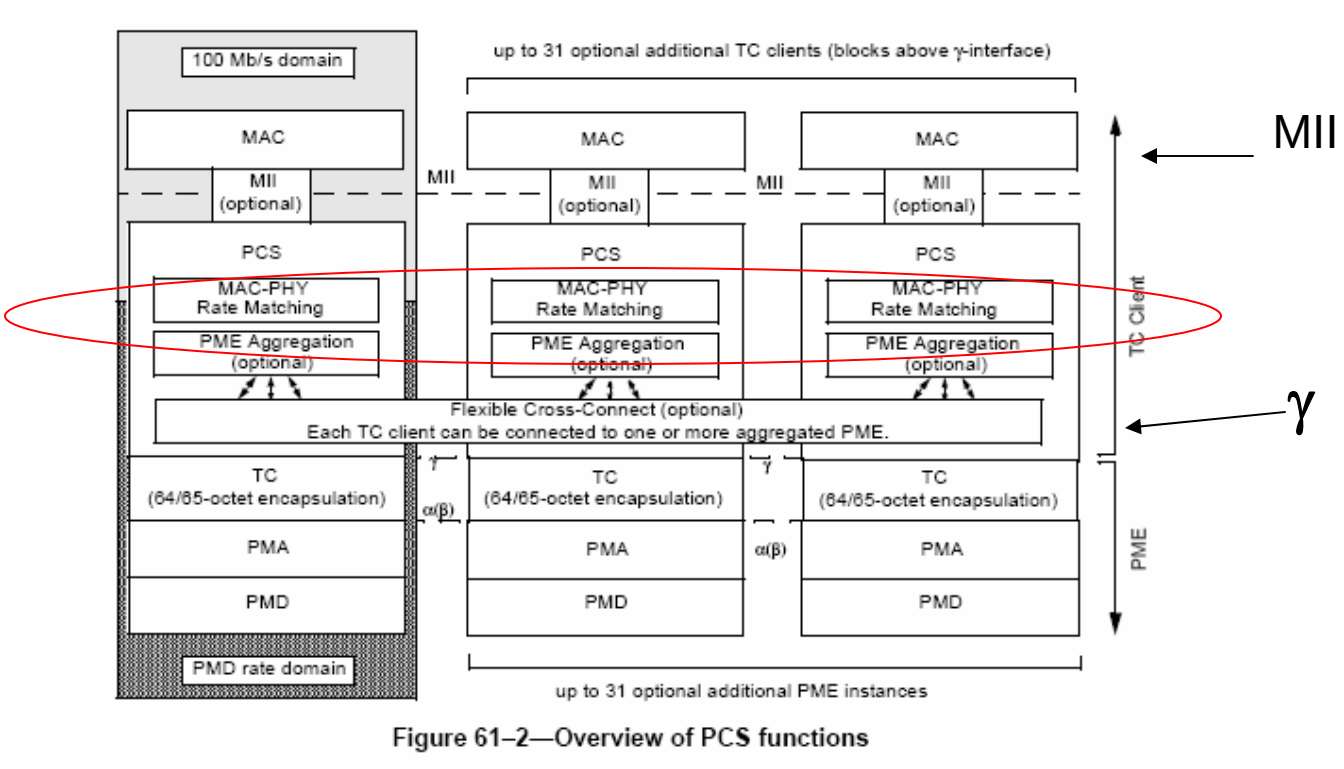
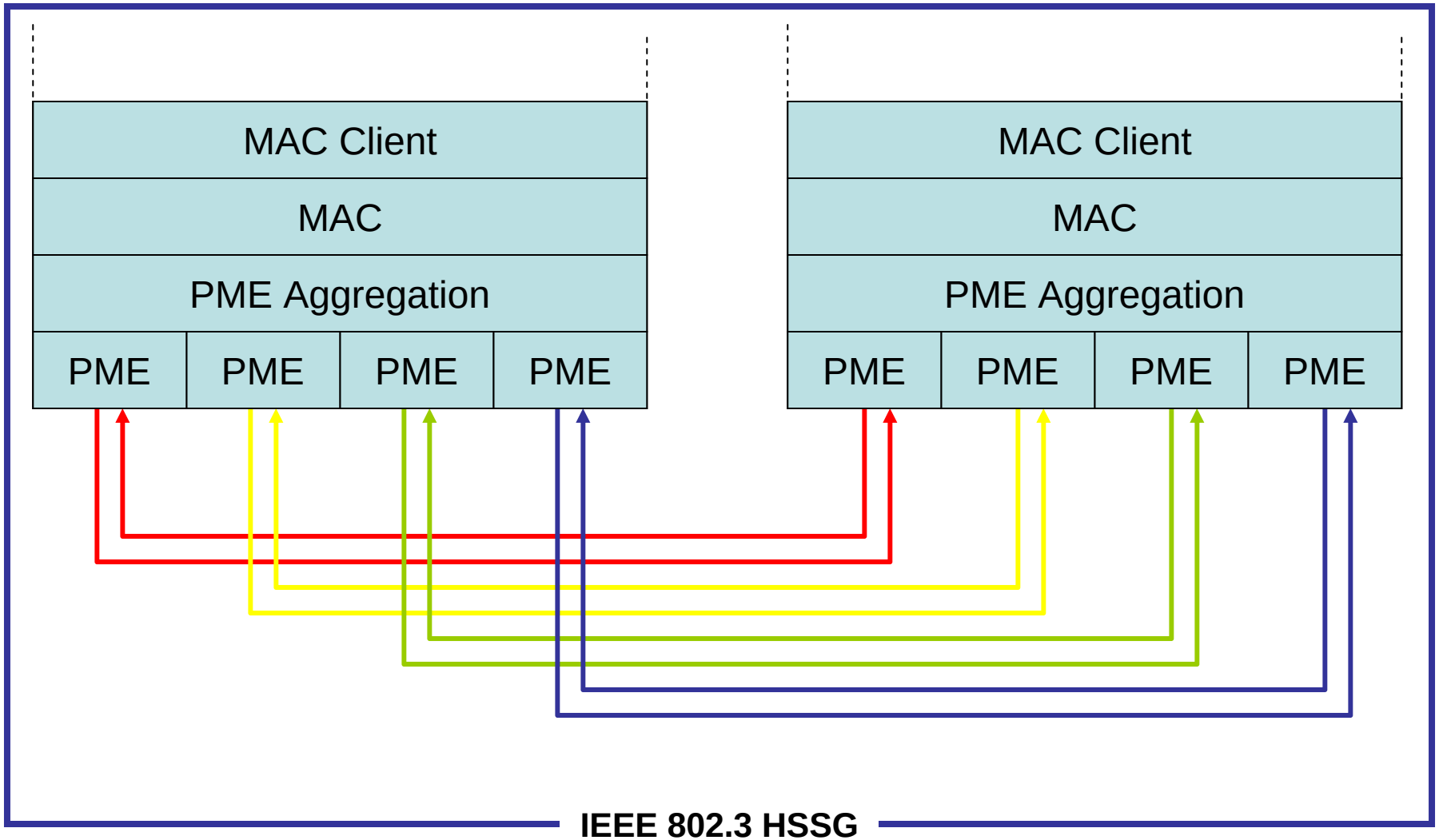


Figure 61–2—Overview of PCS functions

802.3ah PME Aggregation



802.3ah PME Aggregation

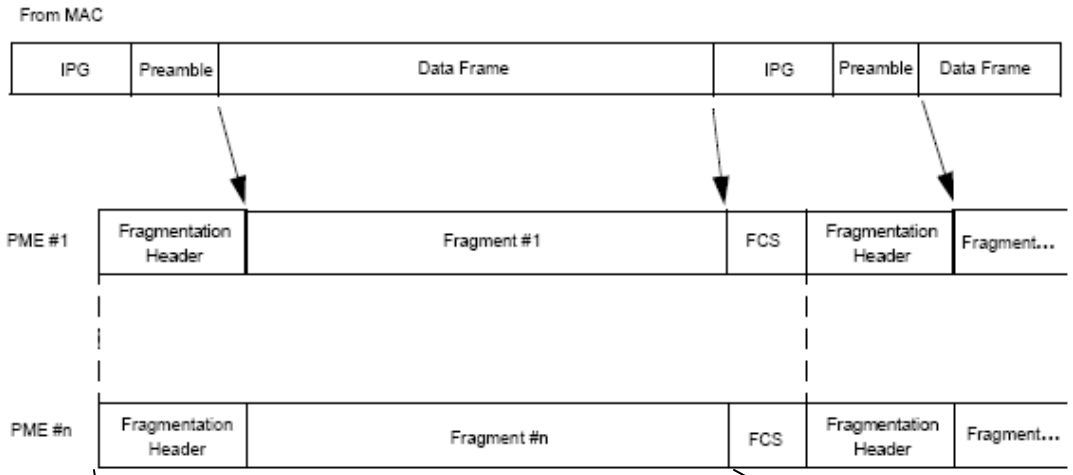


Figure 61–9—Data frame fragmentation

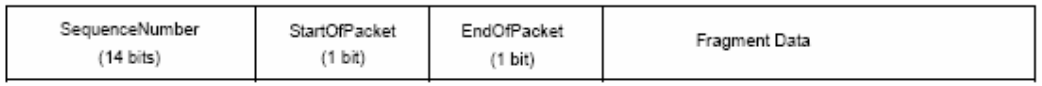


Figure 61–10—Fragment format

802.3ah PME Aggregation

- Fragments can vary in size from 64 to 512 octets
- 16 or 32 bit Fragment Check Sequence (FCS) provided by the underlying TC sublayer
- Data rate can vary by 4:1 between fastest and slowest PME in an aggregation
- Differential delay restricted to 15,000 bits
- The combination of fragment size, differential delay, and speed ratio limits the size of the reassembly buffer to 2^{14} bits (2 kB) per PME

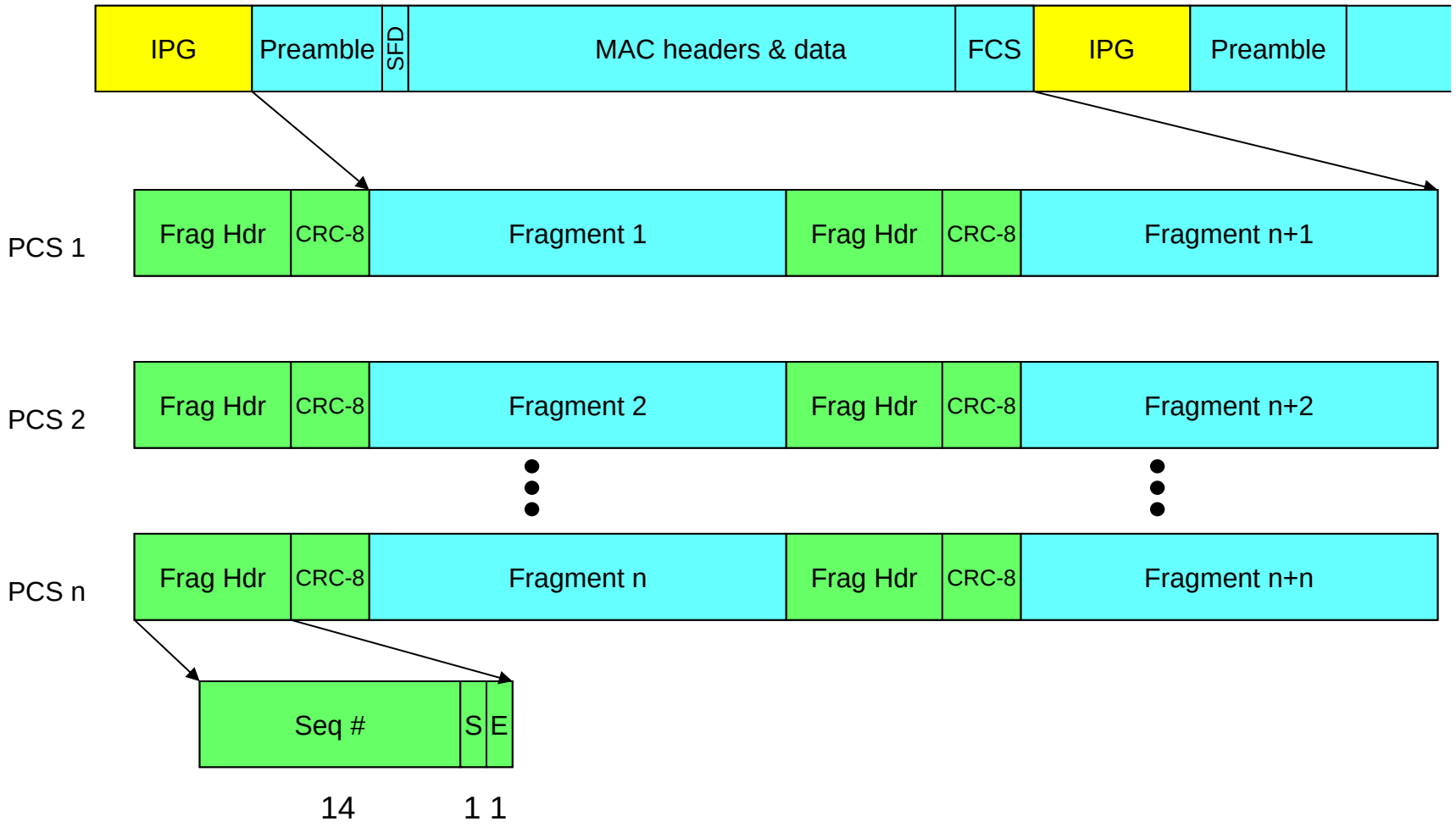
802.3ah PME Aggregation

- Fragmentation and reassembly supports much higher link utilization compared to 802.3ad
- Traffic in a single conversation is not limited to traversing a single link
- Provides a nearly linear increase in bandwidth for a (less than) linear increase in cost

Aggregation at the Physical Layer (APL)

- The PME aggregation concept can be used with existing 10GBASE PHYs
- Will need a fragment CRC (or equivalent) to protect against fragment loss or corruption
  - must protect fragmentation header at the very least
- Fragmentation header can be changed if desired
- Preamble should be propagated, not stripped
  - various standards-based and proprietary uses
- maxFragmentSize, minFragmentSize, maxDifferentialDelay and maxSpeedRatio can be changed as needed to balance efficiency vs. buffer size and latency

APL fragment format



Fragment format considerations

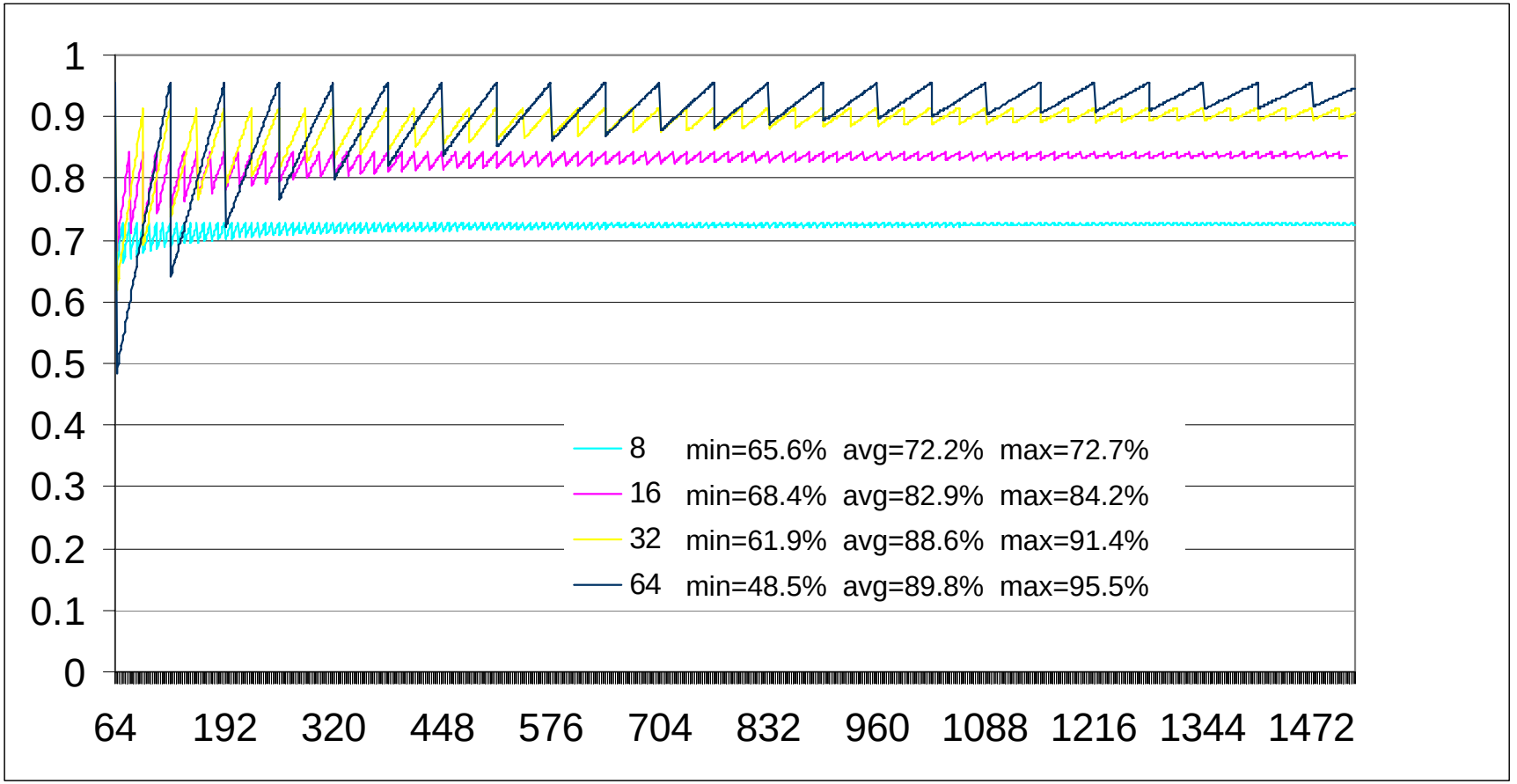
- Shorter fragment header + longer fragment length = lower overhead
- Longer fragment sequence # = greater skew tolerance
  - Estimate of required skew tolerance = 10 us

Frag size	Eff (%)*
8	73
16	84
32	91
64	95.5

* assuming 14 bit sequence #

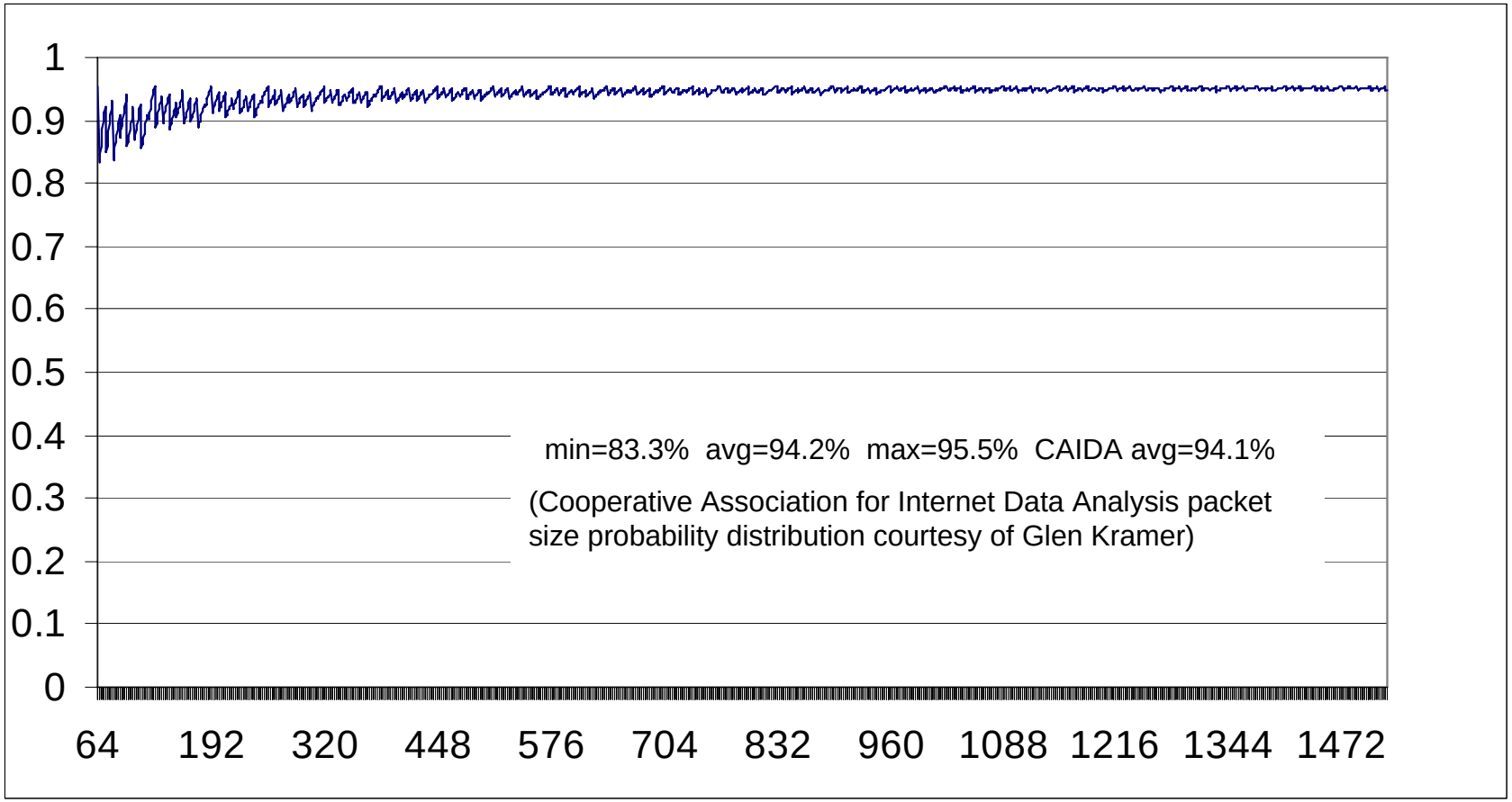
- 16 or 32 byte fragment size seems reasonable
- 14 bit sequence number seems reasonable
- Fragments are delimited just like packets on any given interface

Fixed vs. variable fragment size



Efficiency as a function of packet length for fixed size fragments

Fixed vs. variable fragment size

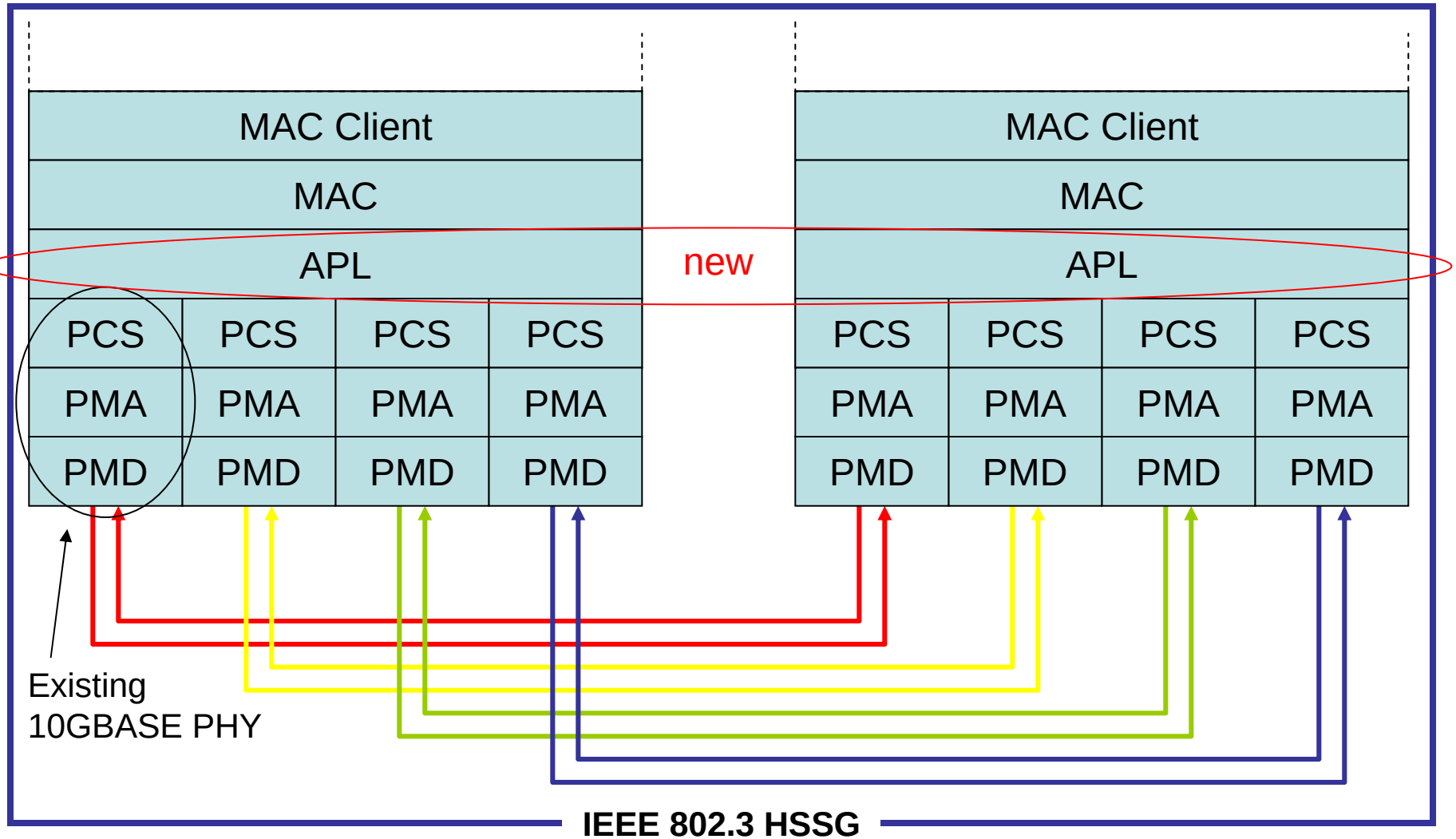


Efficiency as a function of packet length for variable size (64, 32, 16, 8) fragments

Fragment size summary

- Decimation into variable size fragments yields the highest efficiency for all packet sizes
- The bulk of a packet is always decimated into the largest size fragments
- The tail of a packet is decimated into the optimal combination of fragments
- Will need to indicate fragment size in the fragment header

APL location



APL summary

- Packet fragmentation, distribution, collection and reassembly similar to 802.3ah PME aggregation
- Accommodates reasonable differential delay
- Assumes equal speed links
- Assumes point to point, full duplex links
- Resilient and scalable

APL summary

- Ensures ordered delivery
- Detects lost or corrupted fragments
- Minimal added latency
- Fits well with multi-port (quad/octal) PHYs
- Line code independent thus providing compatibility with all existing PHYs
- Uses existing compatibility interfaces (e.g. XGMII, XAUI)

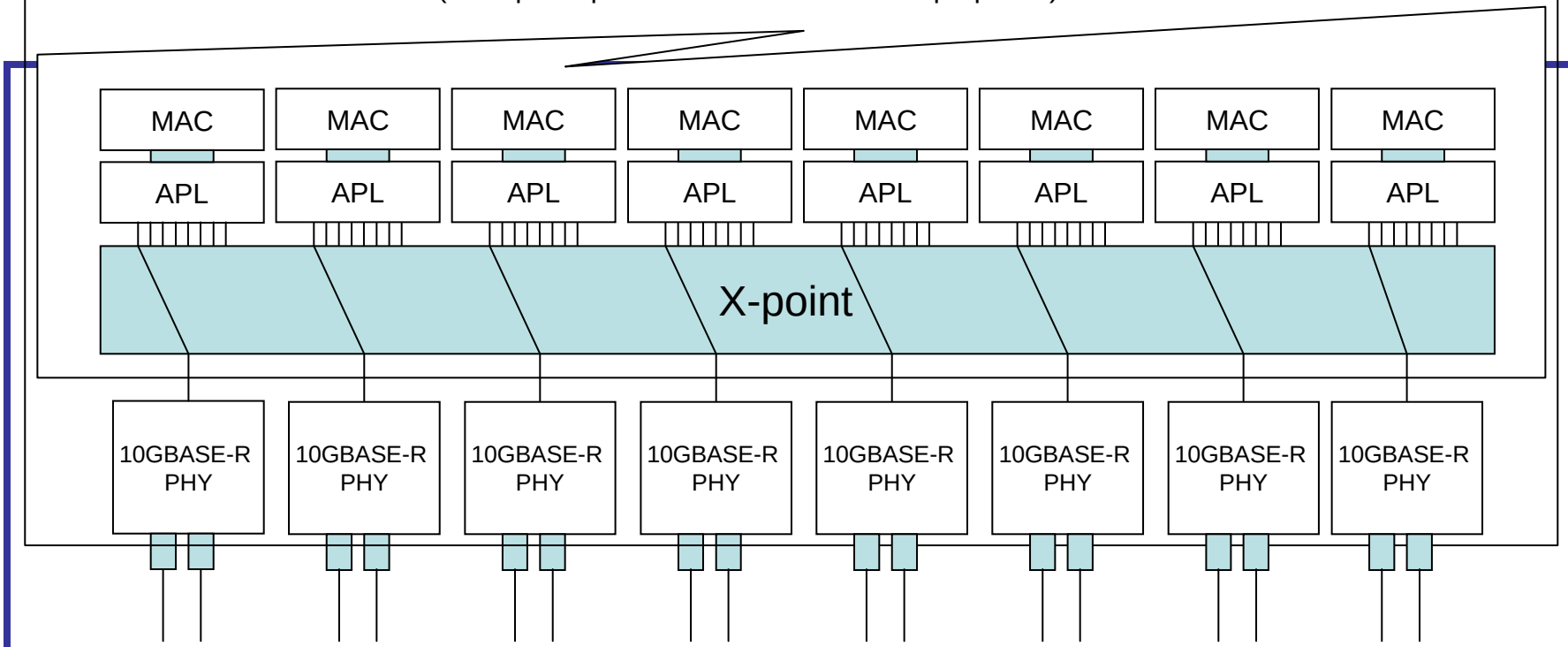
APL control protocol

- Will need a protocol to:
  - detect aggregation link partners
  - configure aggregations
  - detect failures
  - reconfigure
- Link Aggregation Control Protocol (LACP) uses Slow Protocol frames for these functions
  - Existence proof, see 43.4
- Clause 61 PME Aggregation Discovery relies on remote PHY register access
  - Probably do not want to assume this

Configuration Examples

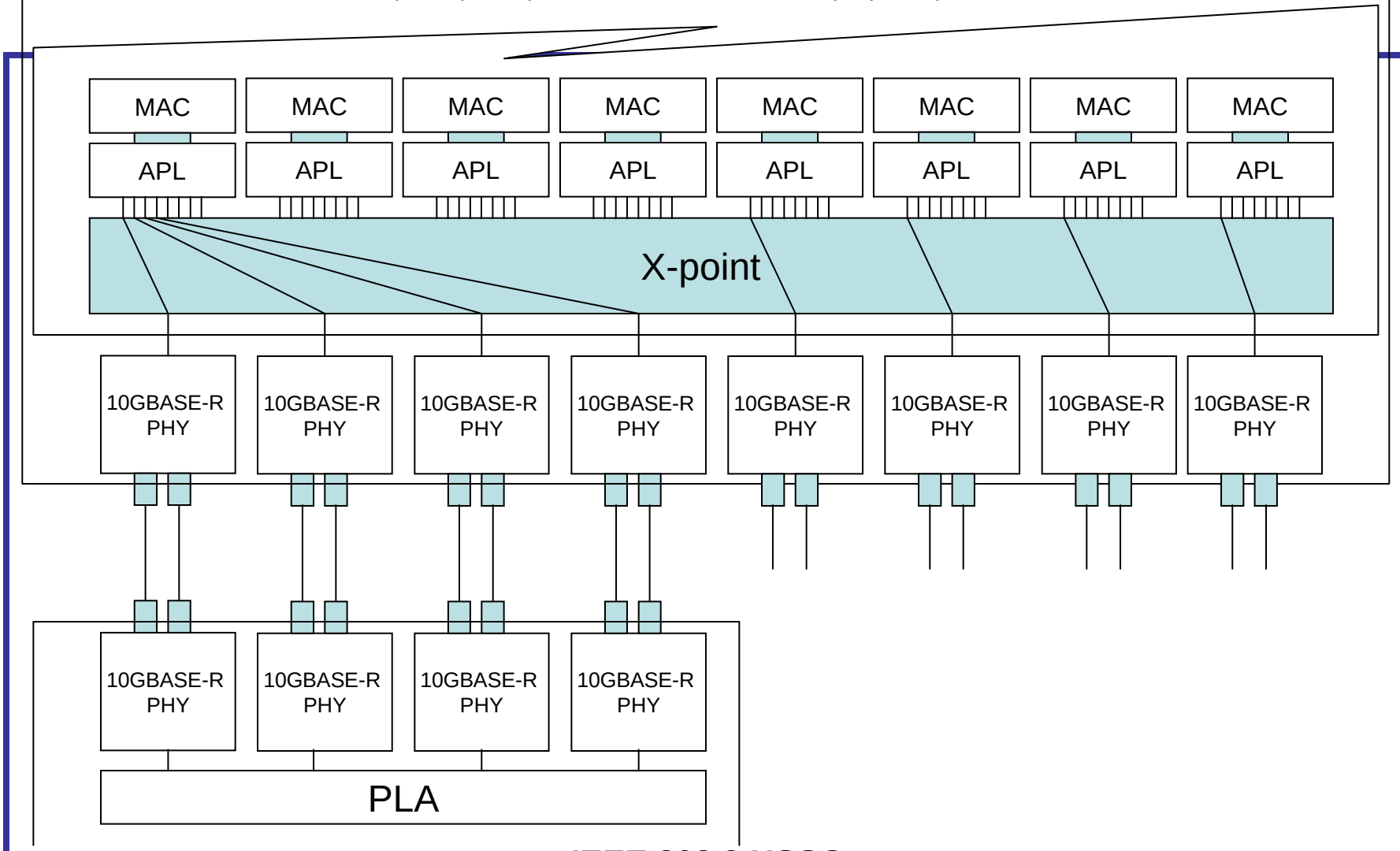
Default configuration

(example implementation for illustrative purposes)



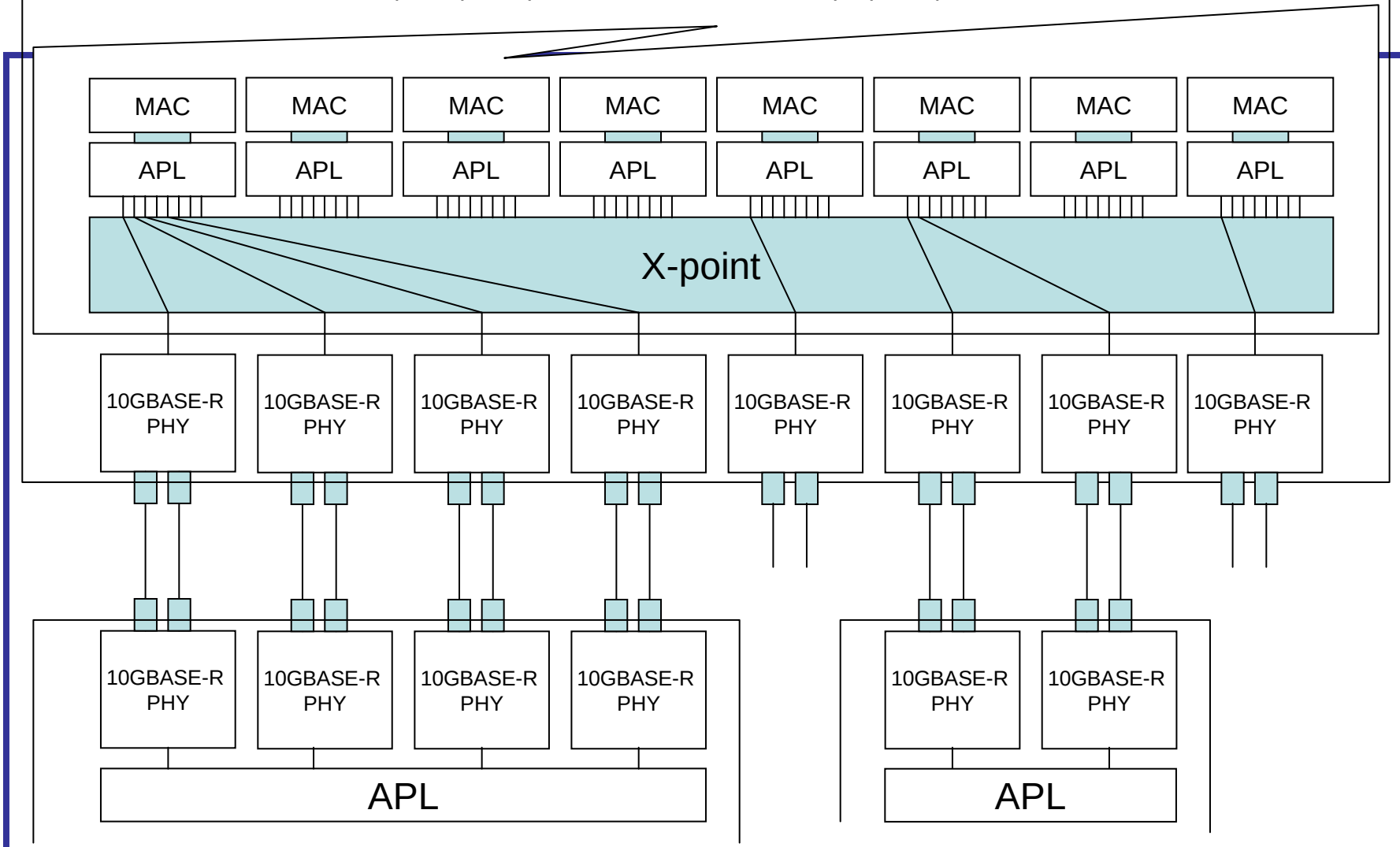
40Gb/s Agg link connected

(example implementation for illustrative purposes)



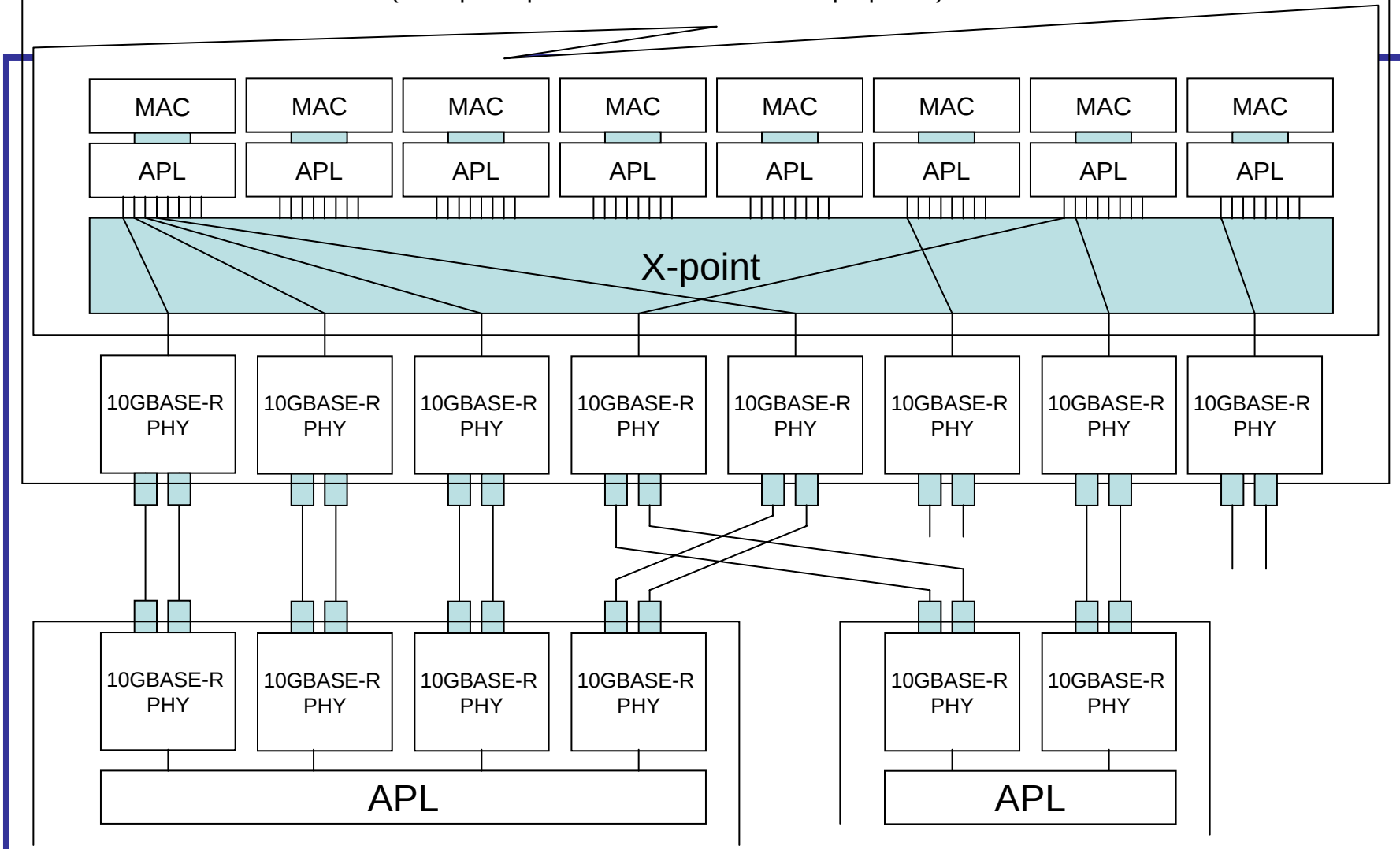
20Gb/s Agg link added

(example implementation for illustrative purposes)



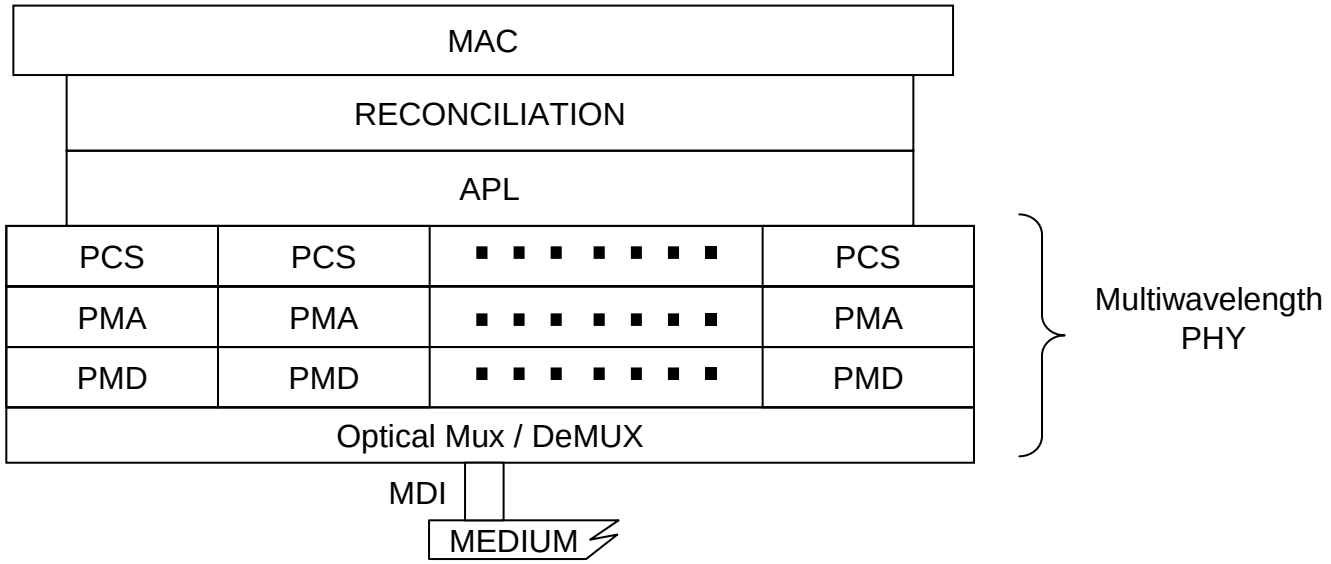
Port reconfigured

(example implementation for illustrative purposes)

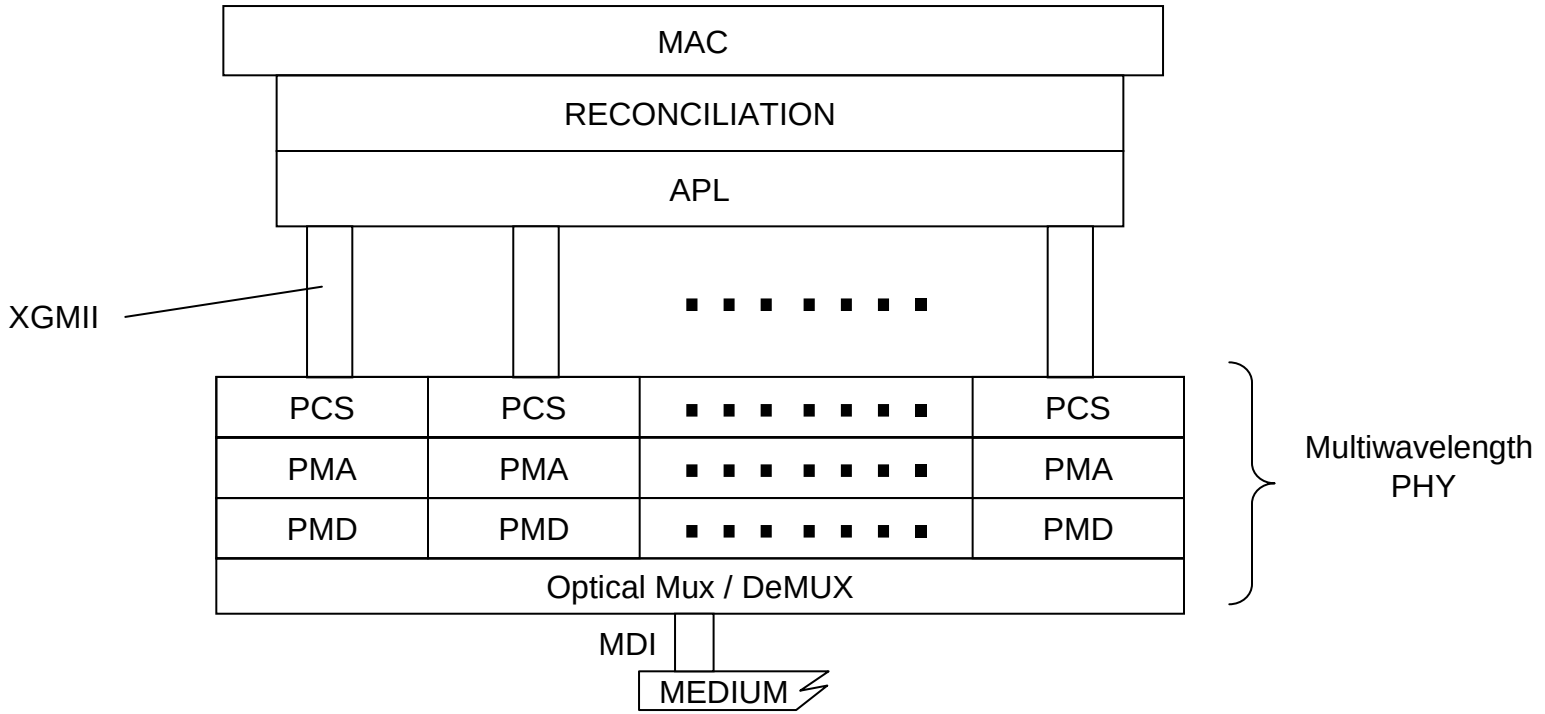


Application with various PHYs

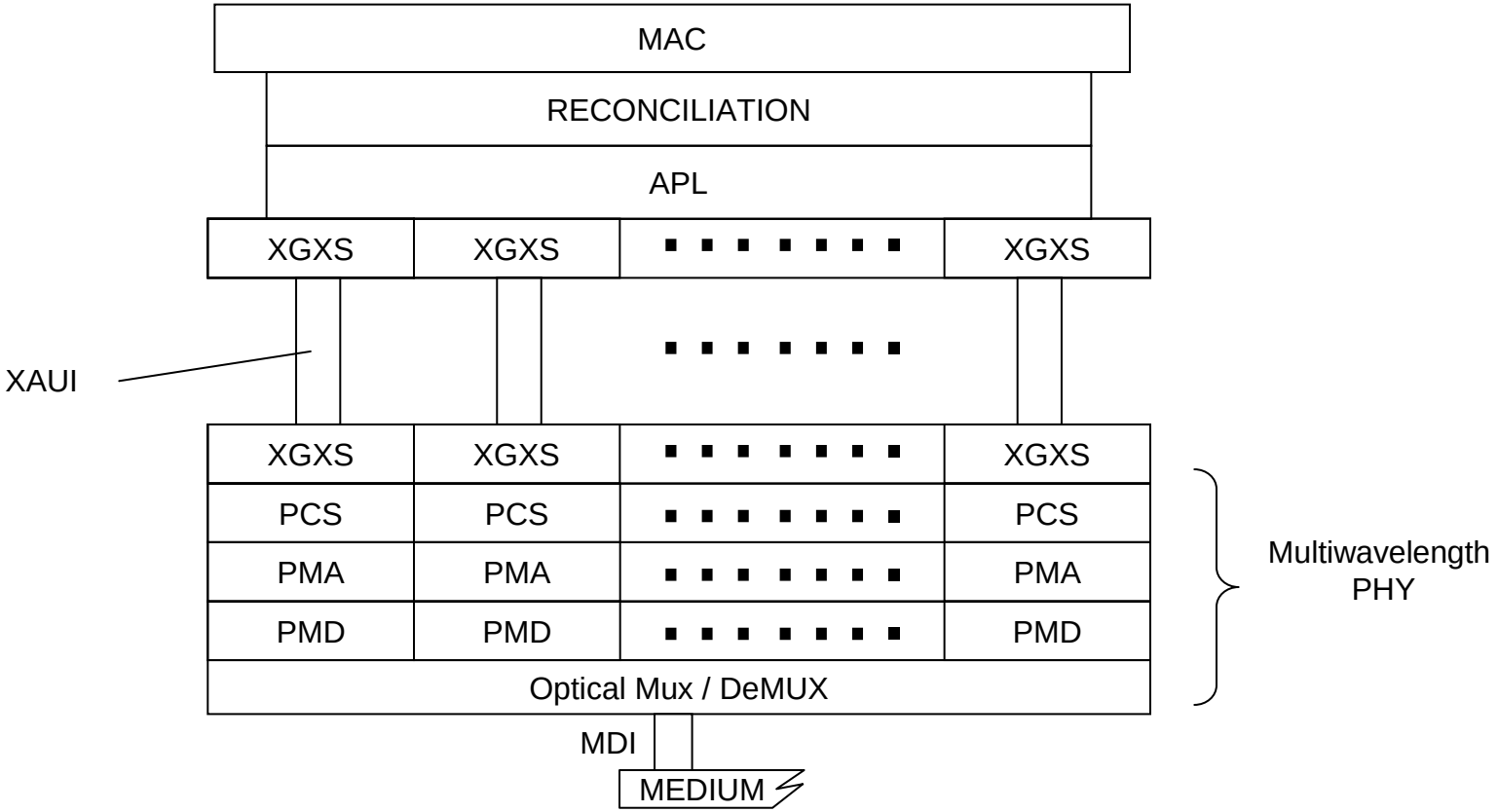
Multi-wavelength PHY



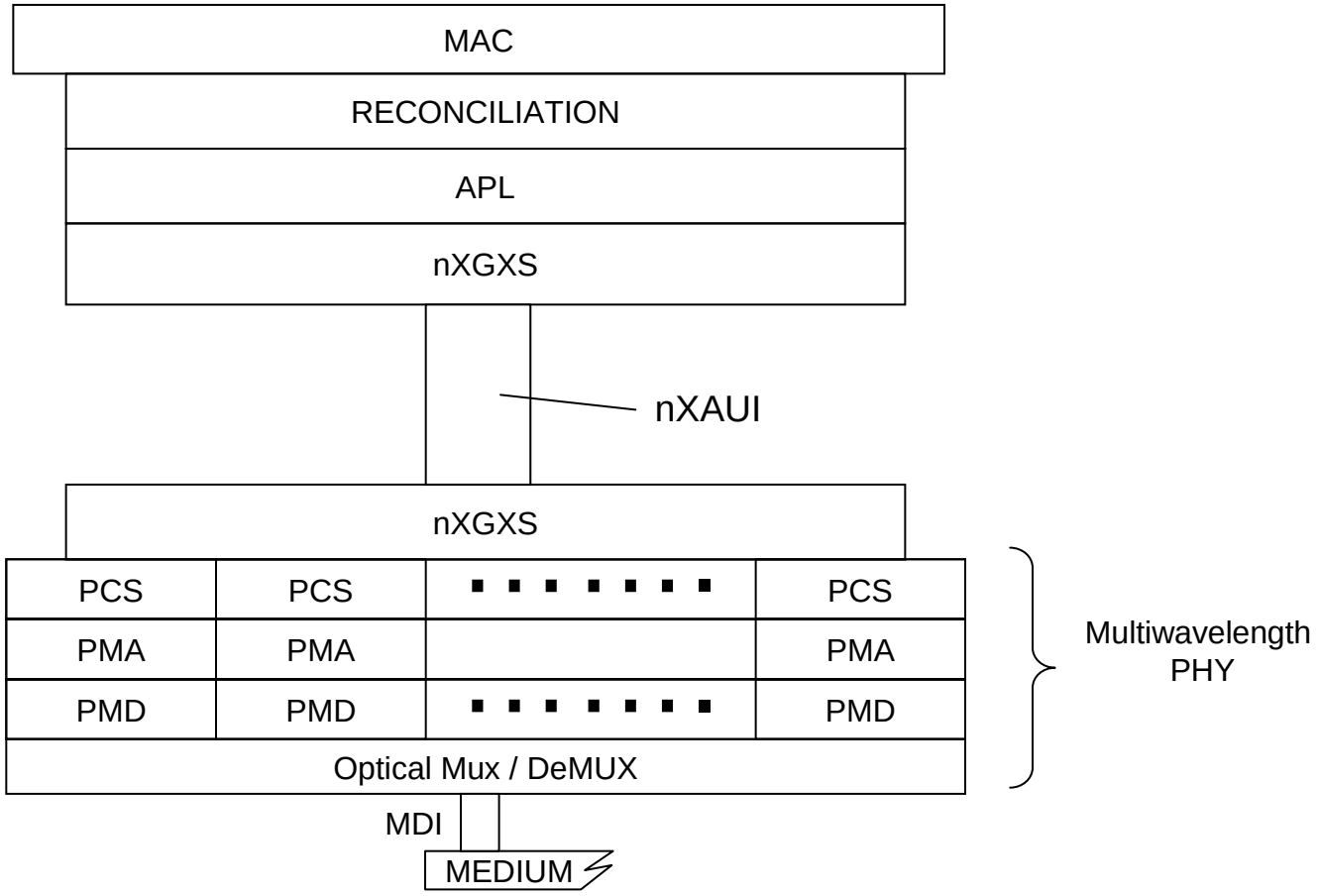
Multi-wavelength PHY, XGMII



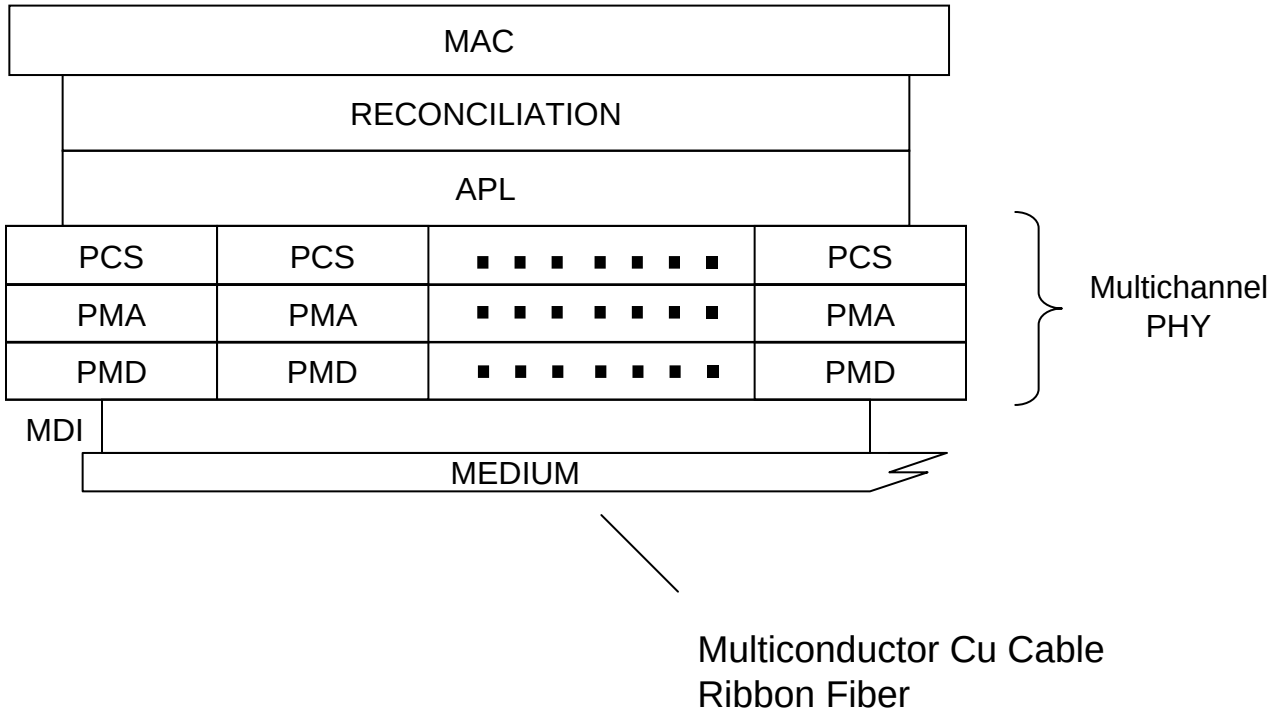
Multi-wavelength PHY, XAUI



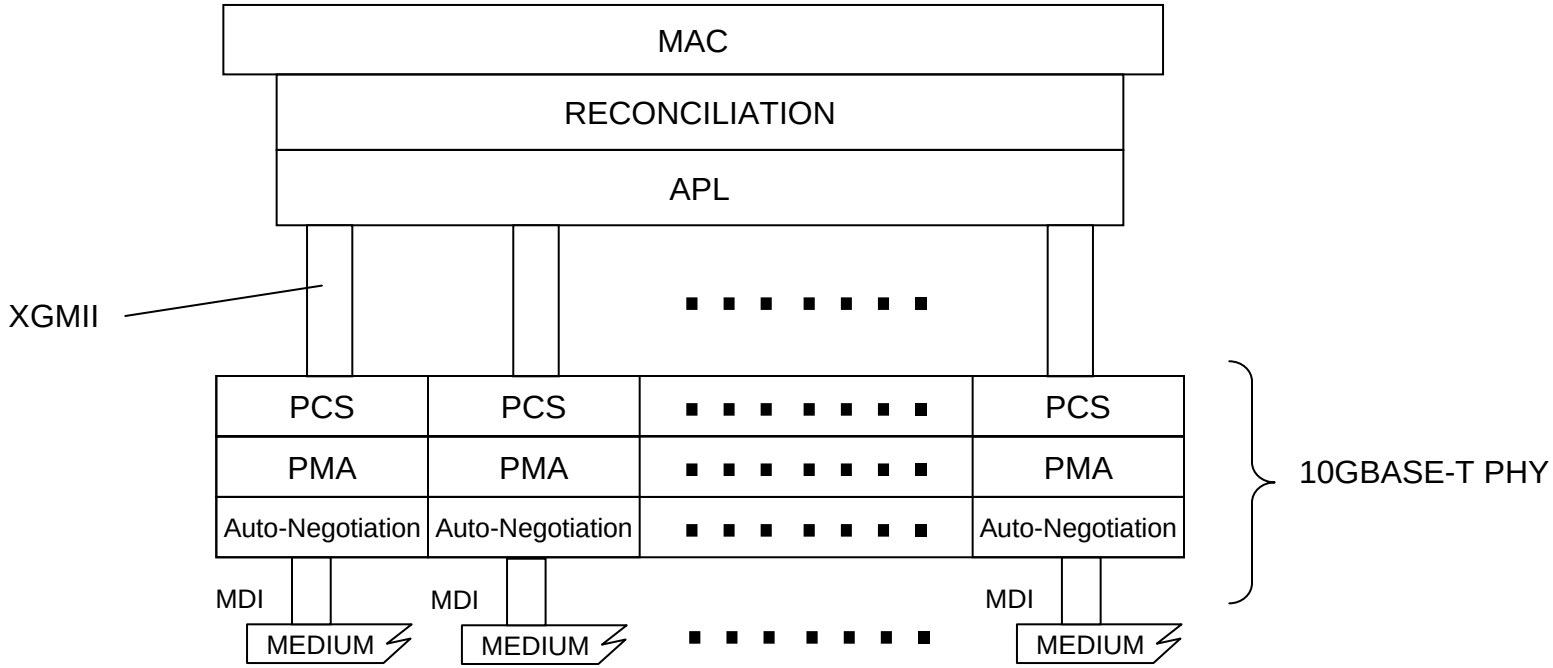
Multi-wavelength PHY, nXAUI



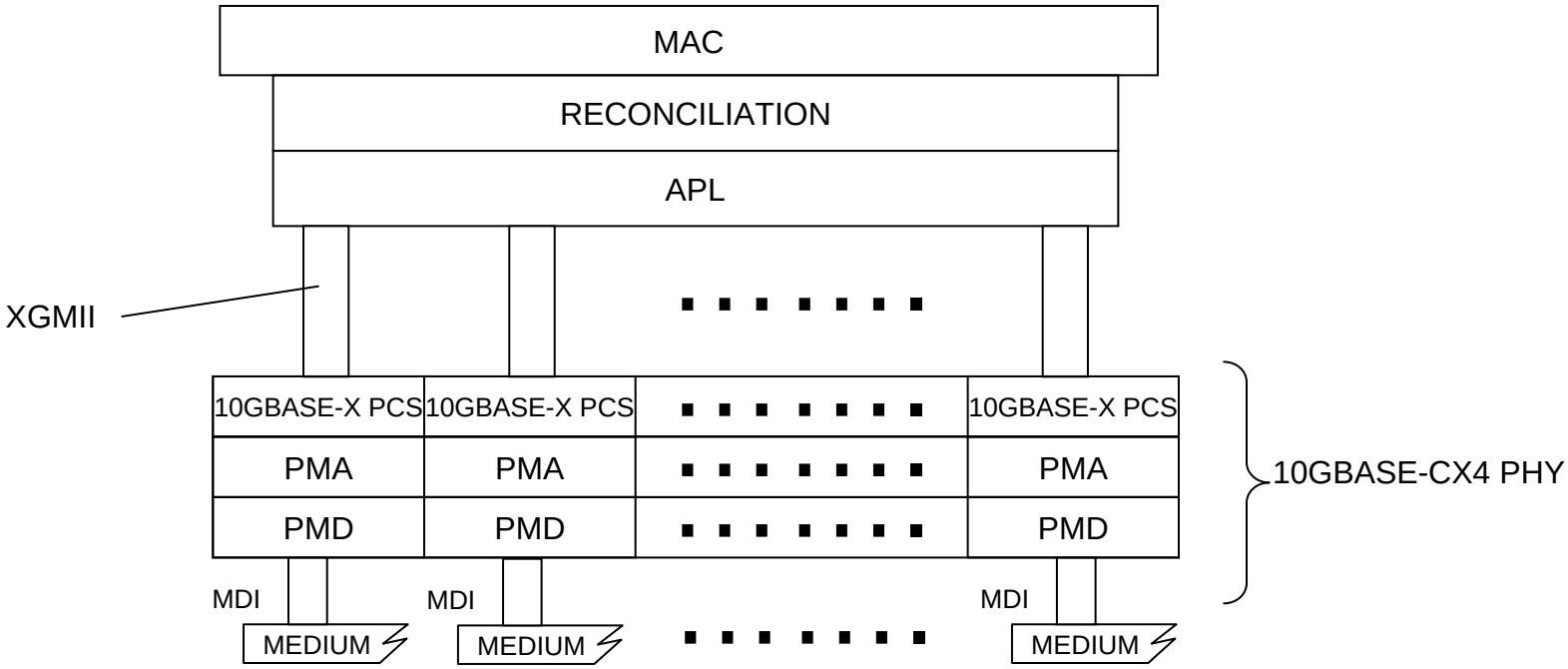
Multi-channel PHY



Multiple 10GBASE-T PHYs



Multiple 10GBASE-CX4 PHYs



Summary

- There is a broad set of applications for aggregated links
  - Backplanes, data centers, short and long haul optical
- Existing 10GBASE Physical layers can be aggregated
- 802.3ad LAG does not fully address the need
- 802.3ah PME aggregation is optimized for DSL-based links

Summary

- Concepts from 802.3ad LAG and 802.3ah PME aggregation can be re-used to define Aggregation at the Physical Layer (APL)
- APL can provide a scalable, resilient interface using existing 10G PHYs
- APL can also be used with newly defined PHYs that provide higher serial rates