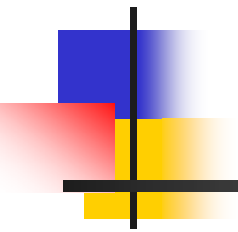


IEEE 802.3 architecture and 40/100GbE

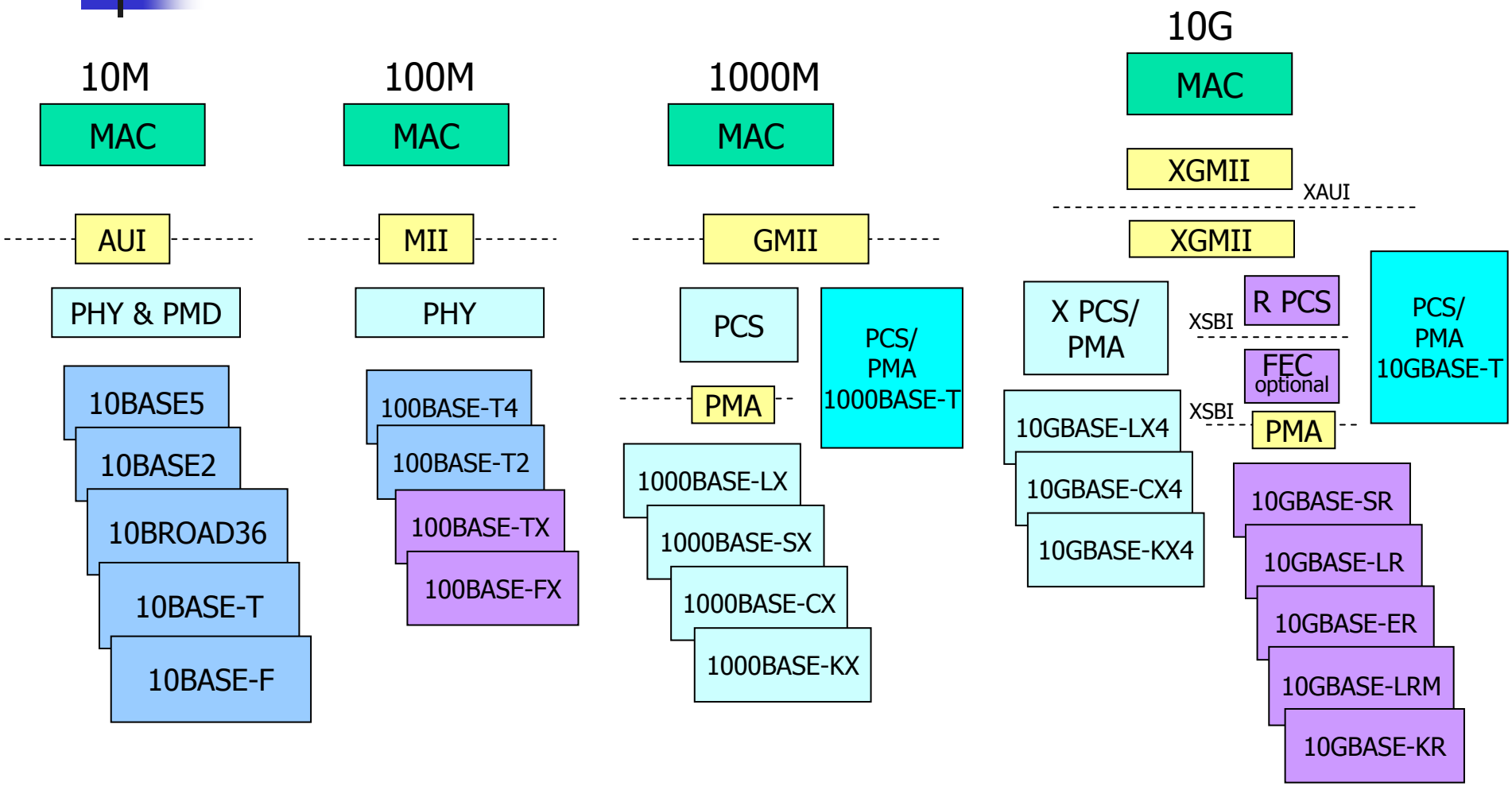
Ilango Ganga, Intel
John D'Ambrosia, Force10 Networks

Nov 13, 2007

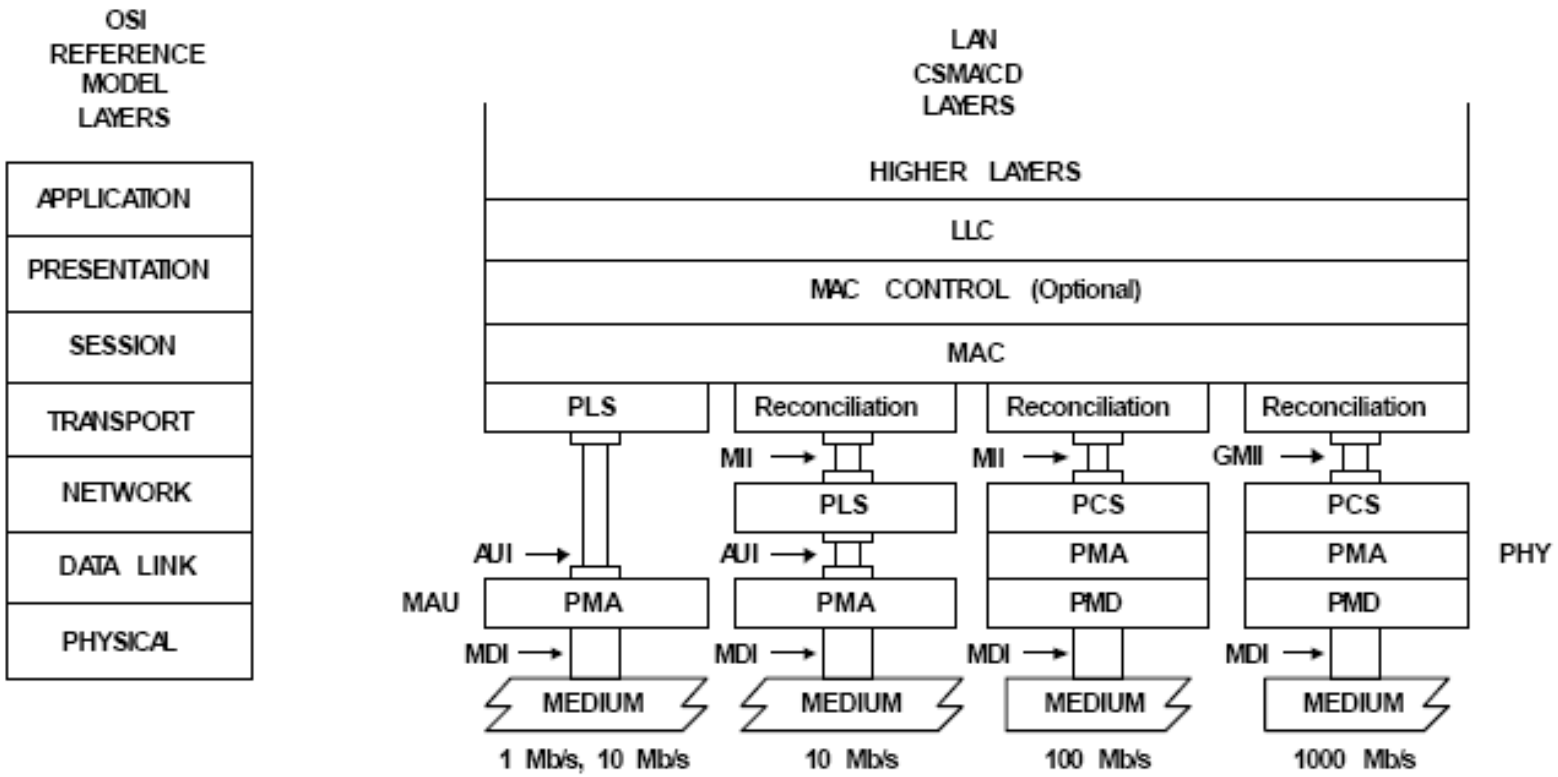
Agenda

- Ethernet architecture evolution
- 10/100/1000M architecture
- 10G architecture
- Management
- Inter-sublayer interfaces
- Backplane architecture
- Architecting 40G and 100G
- Interface considerations
- Decisions required

Ethernet architecture evolution



10/100/1000M architecture



AUI = Attachment Unit Interface
MDI = Medium Dependent Interface
MII = Media Independent Interface
GMII = Gigabit Media Independent Interface
MAU = Medium Attachment Unit

PLS = Physical Layer Signaling
PCS = Physical Coding Sublayer
PMA = Physical Medium Attachment
PHY = Physical Layer Device
PMD = Physical Medium Dependent

10/100/1000M interfaces

- No standardized instantiation of PMD service interface
- GMII is standardized instantiation of PCS interface
- TBI is standardized instantiation of PMA interface
- Industry specified instantiation of PCS interface
  - 10/100M examples: SMII, RMII
  - 1000M examples: RGMII, SGMII

10GbE architecture

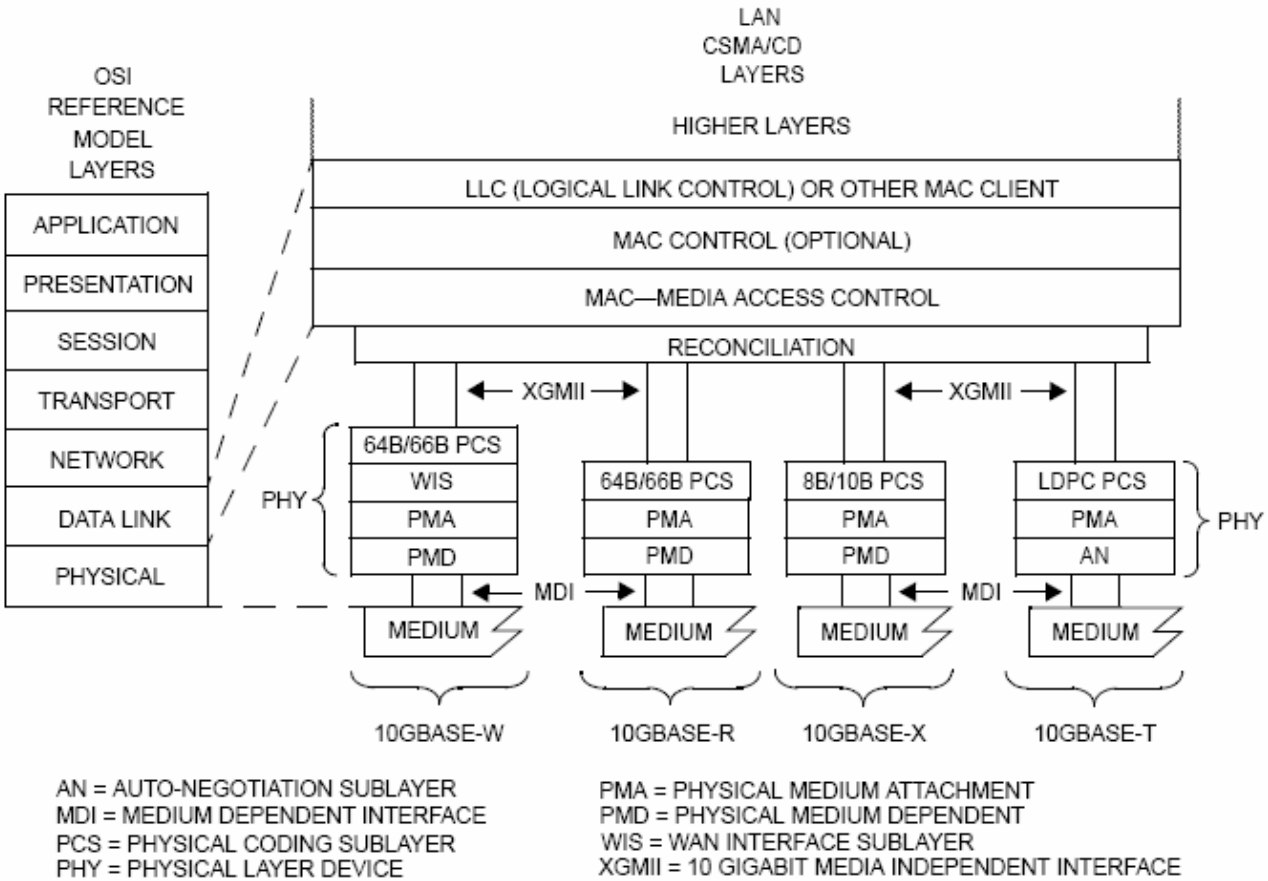


Figure 44–1—Architectural positioning of 10 Gigabit Ethernet

10G blocks overview (1)

- MAC
  - Data Encapsulation, Ethernet framing, Addressing
  - Error detection (e.g. CRC)
  - Speed independent
- RS (Reconciliation sublayer)
  - The RS converts the MAC serial data stream to the parallel data paths of XGMII interface and provides word alignment of the beginning frame, while maintaining total MAC transmit IPG.
- XGXS (XGMII Extender sublayer) and XAUI
  - The purpose of the XGMII Extender is to extend the operational distance of the XGMII and to reduce the number of interface signals.
  - XGXS converts bytes on an XGMII lane into a self clocked, serial, 8B/10B encoded data stream.
  - Each of the four XGMII lanes is transmitted across one of the four XAUI lanes
- 10GBASE-X PCS (Physical Coding sublayer)
  - Encodes 32bit data & 4 bit control of XGMII to 10bit code groups for communication with PMA (8b/10b encoding)
  - Synchronization, deskew and decoding of code groups boundaries
  - Management interface to control PCS & reporting of link status

10G blocks overview (2)

- 10GBASE-R PCS (Physical Coding sublayer)
  - Encodes 32bit data & 4 bit control of XGMII to 66bit code groups for communication with PMA (64b/66b encoding)
  - Transfers 66-bit encoded data through an optional 16-bit interface called XSBI (10G sixteen bit interface)
  - Management interface to control PCS & reporting of link status
- PMA (Physical Medium Attachment)
  - Serialization /de-serialization of bits from PCS to the underlying serial PMD
  - Provides loopback function at the PMA sublayer
  - Can be used to connect to different PMDs (like 10GBASE-KR, 10GBASE-SR, 10GBASE-LR etc.,)
- PMD (Physical Medium Dependent)
  - Transmission/reception of bit streams to/from the underlying medium
  - Examples: 10GBASE-SR fiber PMD to transmit/receive over 850nm fiber, 10GBASE-LR fiber PMD to communicate over 1310nm fiber
  - Provides signal detect and fault function to detect fault conditions

10GBASE-X interfaces

XGMII Interface

Optional XGXS with XAUI

PMA interface abstract

PMD interface abstract

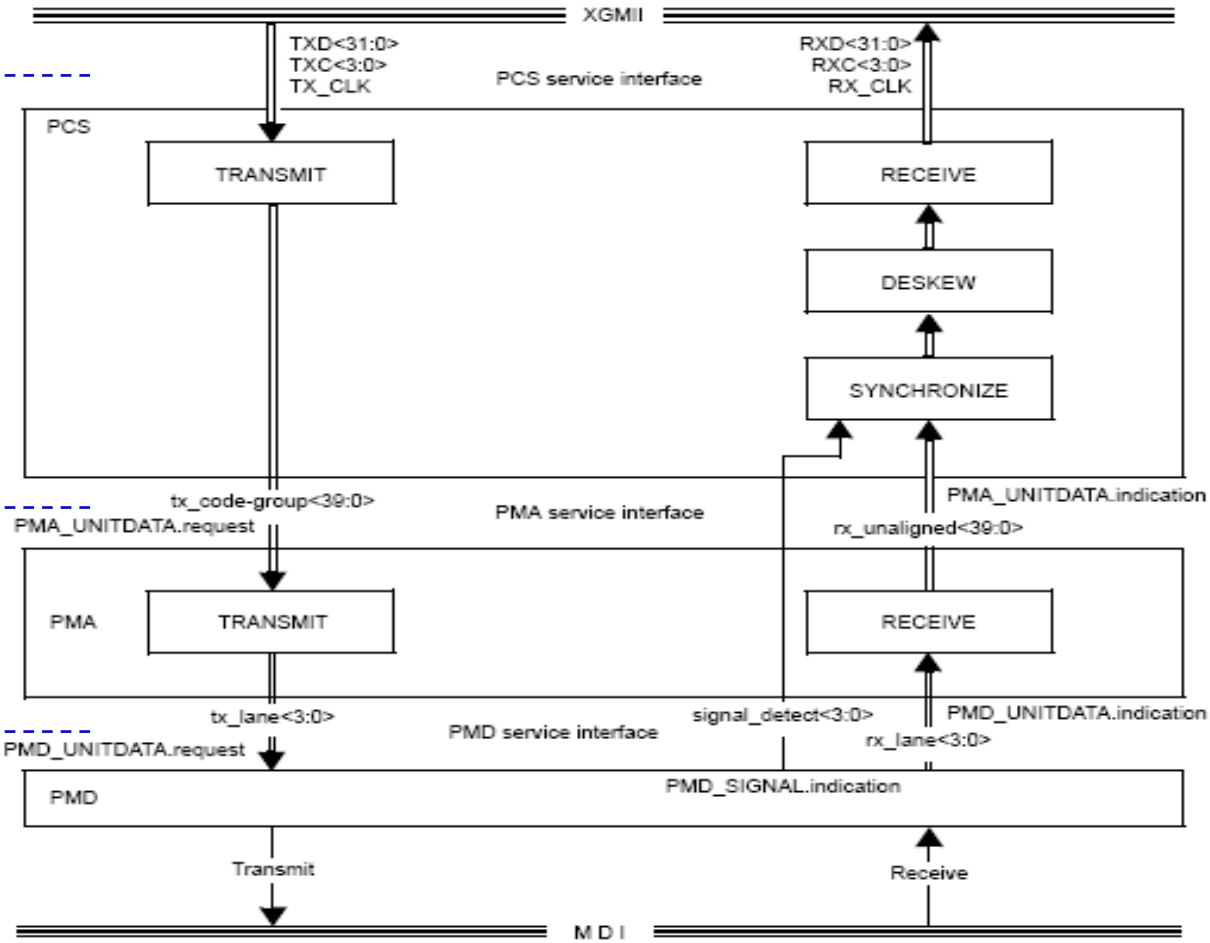


Figure 48–2—Functional block diagram

PCS: 8b/10b encoding

4 parallel lanes

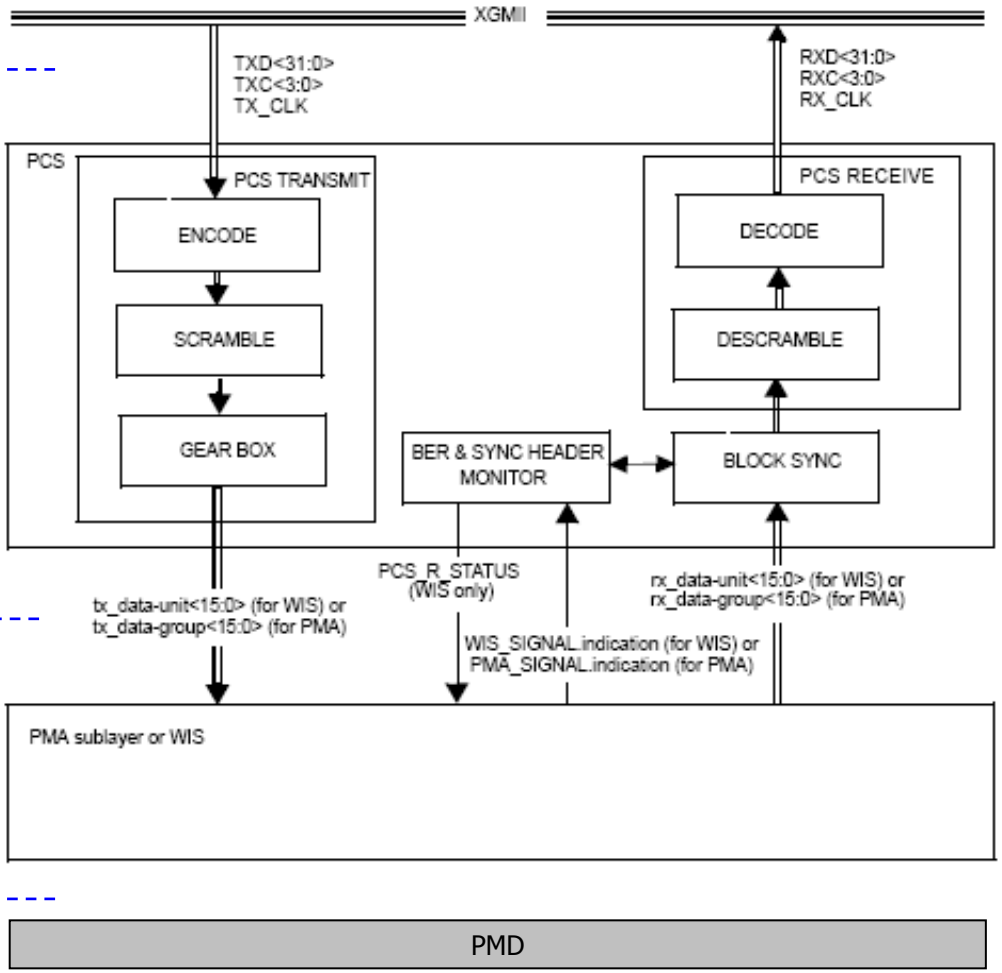
10GBASE-R interfaces

XGMII Interface

Optional XGXS with XAUI

Optional XSBI Interface

Industry Standard Interfaces e.g: XFI, SFI

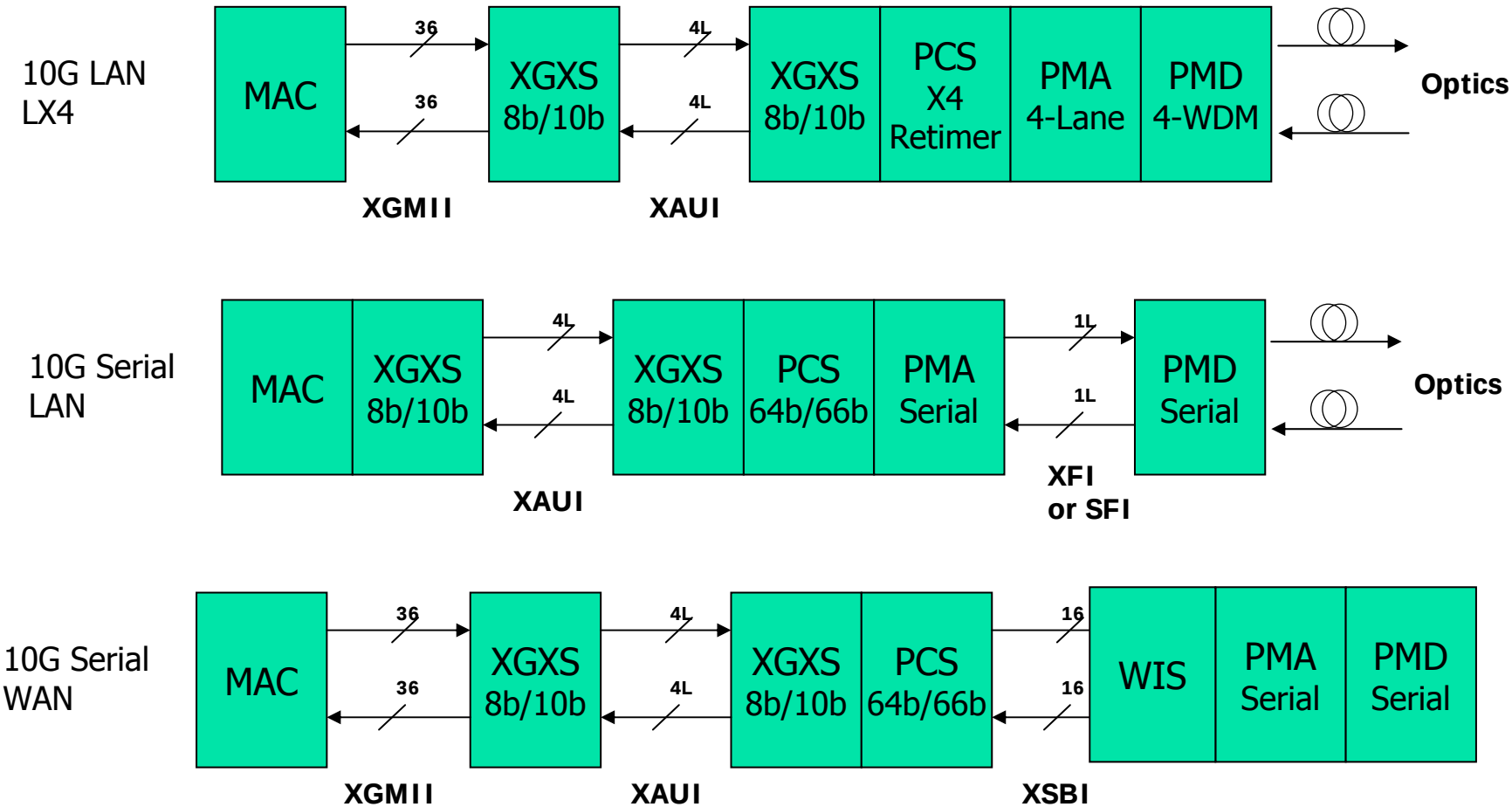


PCS: 64b/66b encoding

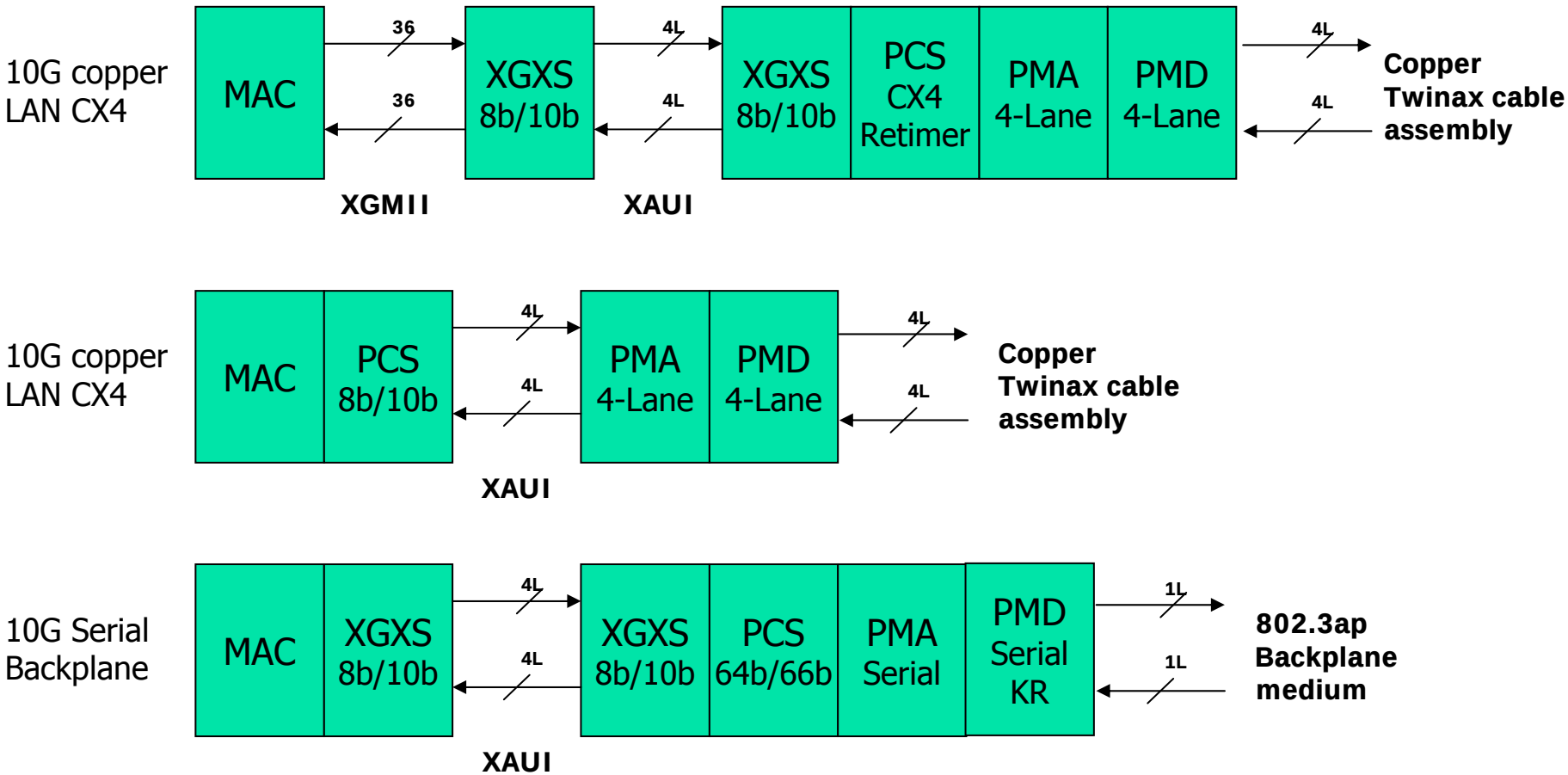
10G Service interfaces

- XGMII is standardized instantiation of PCS interface (Clause 46)
- XAUI is standardized instantiation of XGMII Extender (Clause 47)
  - In practical implementations physical interface between MAC and PHY
- XSBI is an optional physical instantiation of PMA service interface for 10GBASE-R and 10GBASE-W (Clause 51)
  - 16-bit bidirectional interface with source synchronous clock
  - LVDS electricals, 400mV differential output
  - 644.53125 MHz for 10GBASE-R and 622.08 MHz for 10GBASE-W
- PMD service interface is defined as an abstraction without implying physical implementations
- Industry specified physical instantiation of PMD service interface for 10GBASE-R and 10GBASE-W
  - Examples: XFI, SFI

10GbE interface implementation examples (1)



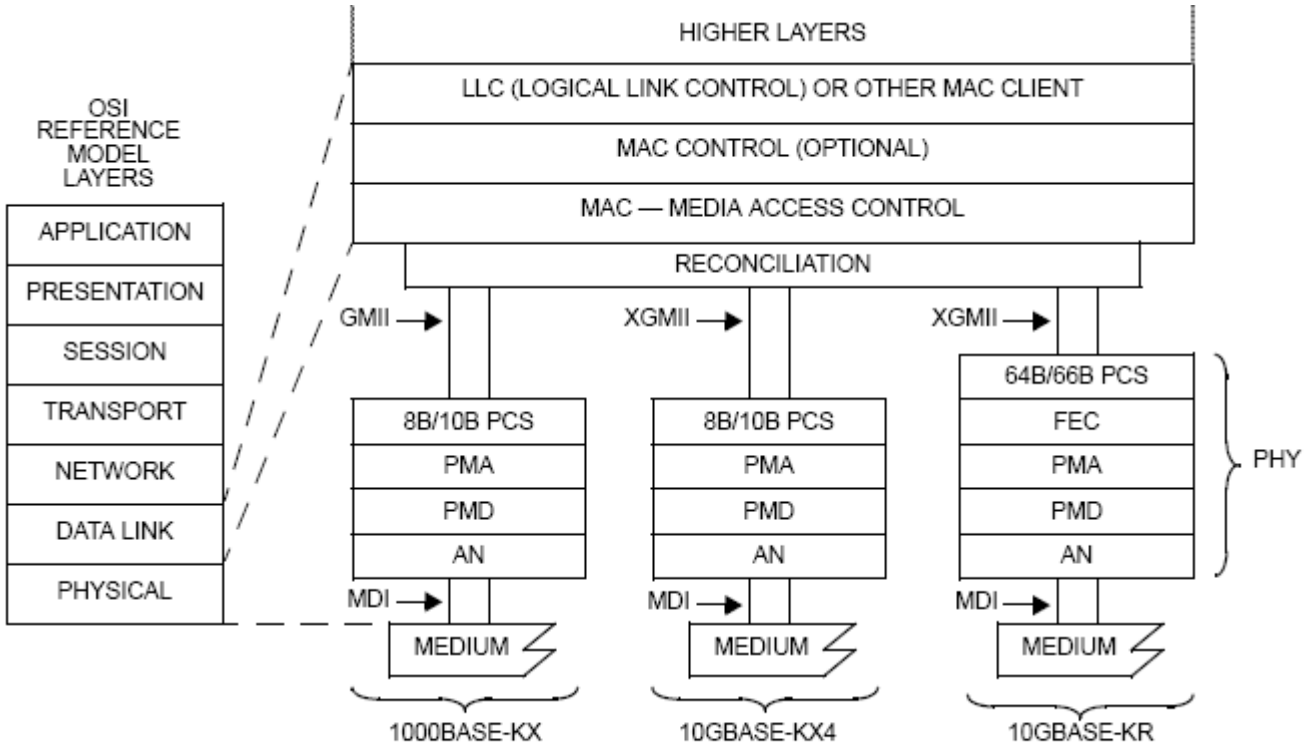
10GbE interface implementation examples (2)



Backplane Ethernet overview

- IEEE Std 802.3ap-2007 Backplane Ethernet defines 3 PHY types
  - 1000BASE-KX : 1-lane 1 Gb/s PHY (Clause 70)
  - 10GBASE-KX4: 4-lane 10Gb/s PHY (Clause 71)
  - 10GBASE-KR : 1-lane 10Gb/s PHY (Clause 72)
- Forward Error Correction (FEC) for 10GBASE-R (Clause 74) – optional
  - Optional FEC to increase link budget and BER performance
- Auto-negotiation (Clause 73)
  - Auto-Neg between 3 PHY types (AN is mandatory to implement)
  - Parallel detection for legacy PHY support
    - Automatic speed detection of legacy 1G/10G backplane SERDES devices
  - Negotiate FEC capability
- Clause 45 MDIO interface for management
- Channel
  - Controlled impedance (100 Ohm) traces on a PCB with 2 connectors and total length up to at least 1m.
  - Channel model is informative as defined in Annex 69B
- Support a BER of 10^{-12} or better

Backplane Ethernet architecture



AN = AUTO-NEGOTIATION
GMII = GIGABIT MEDIA INDEPENDENT INTERFACE
MDI = MEDIUM DEPENDENT INTERFACE
PCS = PHYSICAL CODING SUBLAYER
FEC = FORWARD ERROR CORRECTION

PHY = PHYSICAL LAYER DEVICE
PMA = PHYSICAL MEDIUM ATTACHMENT
PMD = PHYSICAL MEDIUM DEPENDENT
XGMII = 10 GIGABIT MEDIA INDEPENDENT INTERFACE

Figure 69–1—Architectural positioning of Backplane Ethernet

Management interface (1)

- MDIO logical and electrical interface defined in Clause 45
- 10GbE devices and new generation PHYs use Clause 45 MDIO
  - EFM and Backplane devices also use Clause 45 MDIO
- Backward compatibility to access Clause 22 registers
  - 10/100/1000M devices use Clause 22 MII management interface/registers
- 40G and 100G could use Clause 45 MDIO logical and electrical interface

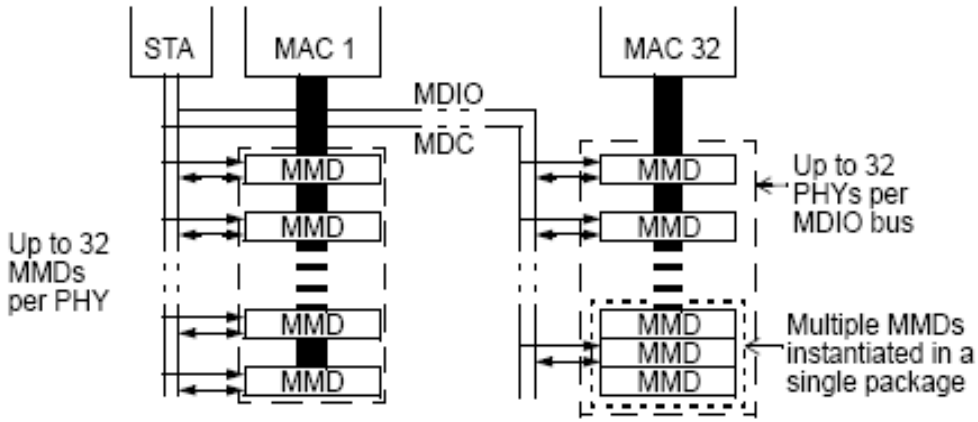


Figure 45–1—DTE and MMD devices

Management interface (2)

- Management Interface between Station Management (STA) and the PHY
  - A manageable PHY is called MDIO manageable device (MMD)
- One STA can manage up to 32 MMDs per MDIO bus
  - 32 register space in Clause 22 based MMD
  - 64K register space in Clause 45 based MMD
    - Ample expansion capability to add 40 and 100GbE management
- MDIO is a 2 wire single ended interface
  - MDC – Clock (unidirectional)
  - MDIO – Data (bidirectional)
  - Max clock rate = 2.5Mhz (400ns)
- Management interface electrical specifications
  - MII Management interface (Clause 22 based)
    - TTL compatible IO, 5V driver or 3.3V drivers that can tolerate 5V inputs
  - MDIO interface (Clause 45 based)
    - 1.2V IO compatible with CMOS devices that use 1.2V supply voltage
    - Clause 22 devices cannot be connected directly to Clause 45 devices

MDIO manageable device addresses

- Up to 32 MMDs per PHY
  - MMD1 used for PMA/PMD
  - MMD3 used for PCS
  - MMD7 used for Auto-Neg
  - Registers m.5, m.6 to indicate which MMDs are present in package
- For example 10GBASE-T PHY uses 3 MMDs
  - MMD1, MMD3, MMD7

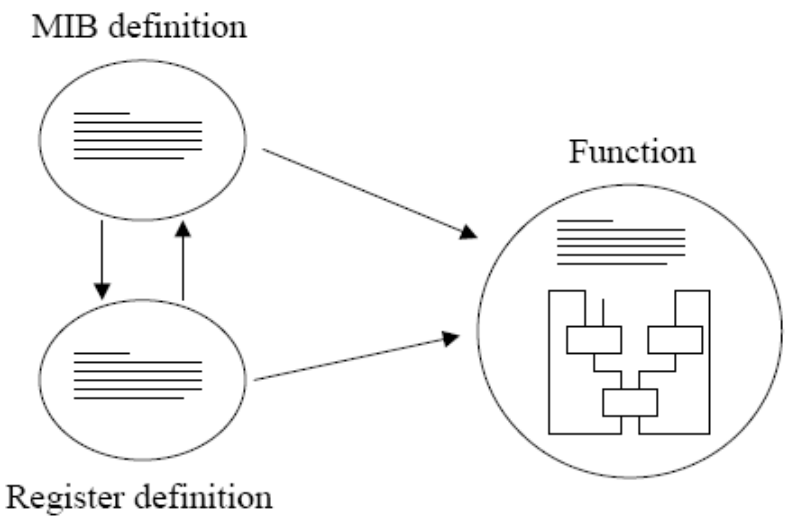
Table 45–1—MDIO Manageable Device addresses

Device address	MMD name
0	Reserved
1	PMA/PMD
2	WIS
3	PCS
4	PHY XS
5	DTE XS
6	TC
7	Auto-Negotiation
8 through 28	Reserved
29	Clause 22 extension
30	Vendor specific 1
31	Vendor specific 2

Managed objects

- Clause 30 defines layer management framework and managed objects
- Identifies which objects are applicable to which 802.3 components

MIB, Registers and Function



Reference:
http://www.ieee802.org/3/ae/public/sep00/law_2_0900.pdf

Managed objects - example

- oMAU
  - MAU managed object class contains attributes, actions and notifications of MAU sublayers as a group including PMA and PMD.
  - Example: Attribute aFECUncorrectableBlocks in MAU object class is given below

30.5.1.1.16 aFECUncorrectableBlocks

ATTRIBUTE

APPROPRIATE SYNTAX:

Generalized nonresetable counter. This counter has a maximum increment rate of ~~1 600 000~~ 10 000 counts per second for 10 Mb/s implementations, ~~500 000~~ 1 200 000 counts per second for 1000 Mb/s implementations, and 5 000 000 counts per second for 10 Gb/s implementations.

BEHAVIOUR DEFINED AS:

For 1000BASE-PX PHYs or 10GBASE-R PHYs, a count of uncorrectable FEC blocks. This counter will not increment for other PHY types.

Increment the counter by one for each FEC block that is determined to be uncorrectable by the FEC function in the PHY.

If a Clause 45 MDIO Interface to the PCS is present, then this attribute will map to the FEC uncorrectable blocks counter (see 45.2.7.6 and 45.2.1.87).;

40GbE objectives

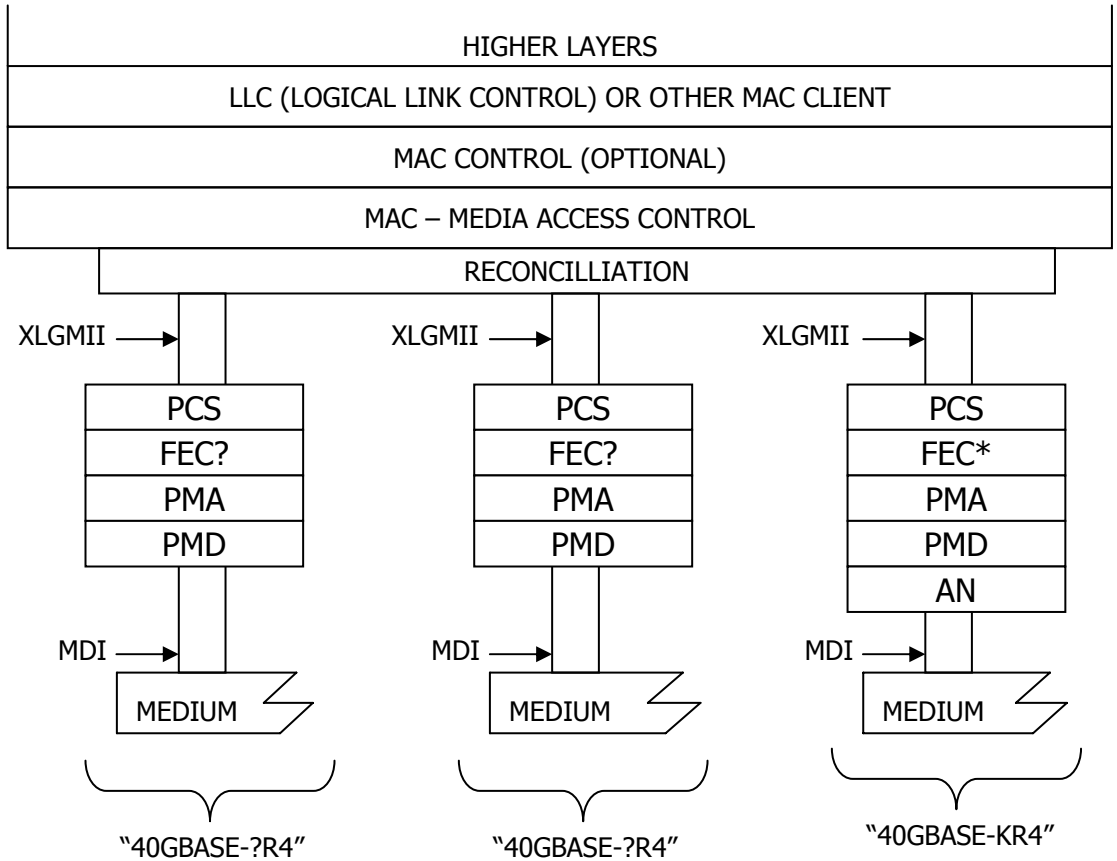
HSSG Objectives

- Support full-duplex operation only
- Preserve the 802.3 / Ethernet frame format utilizing the 802.3 MAC
- Preserve minimum and maximum FrameSize of current 802.3 standard
- Support a BER better than or equal to 10^{-12} at the MAC/PLS service interface
- Provide appropriate support for OTN
- Support a MAC data rate of 40 Gb/s
- Provide Physical Layer specifications which support 40 Gb/s operation over:
  - at least 100m on OM3 MMF
  - at least 10m over a copper cable assembly
  - at least 1m over a backplane
- Support a MAC data rate of 100 Gb/s
- Provide Physical Layer specifications which support 100 Gb/s operation over:
  - at least 40km on SMF
  - at least 10km on SMF
  - at least 100m on OM3 MMF
  - at least 10m over a copper cable assembly

Adopted by HSSG and approved by 802.3 at July 2007 Plenary

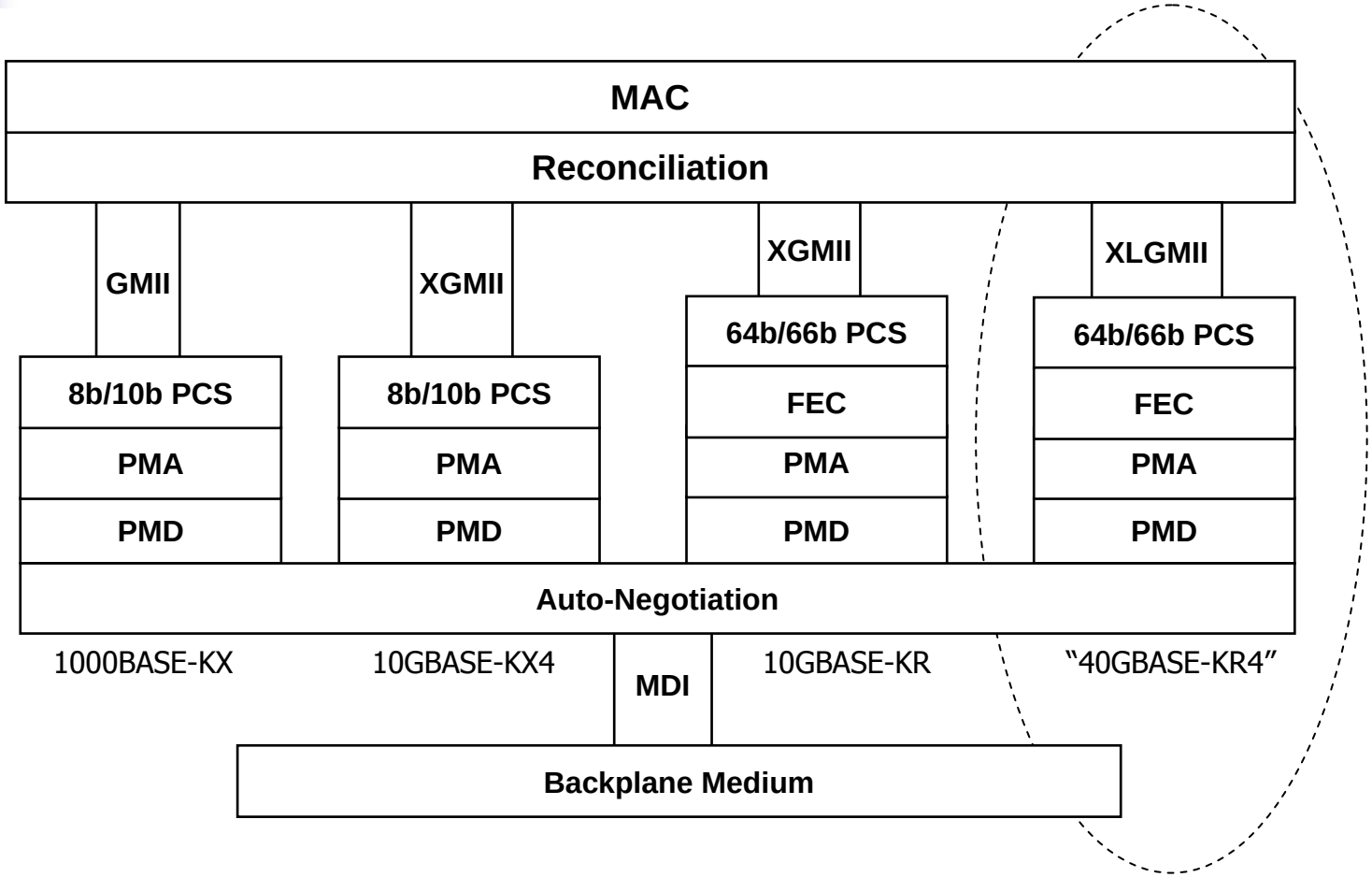
Possible 40GbE architecture

- Desire to leverage 10GBASE-R technology expressed in HSSG
- Assumptions
  - 4 lanes of 10G
  - R PCS (64b/66b coding)



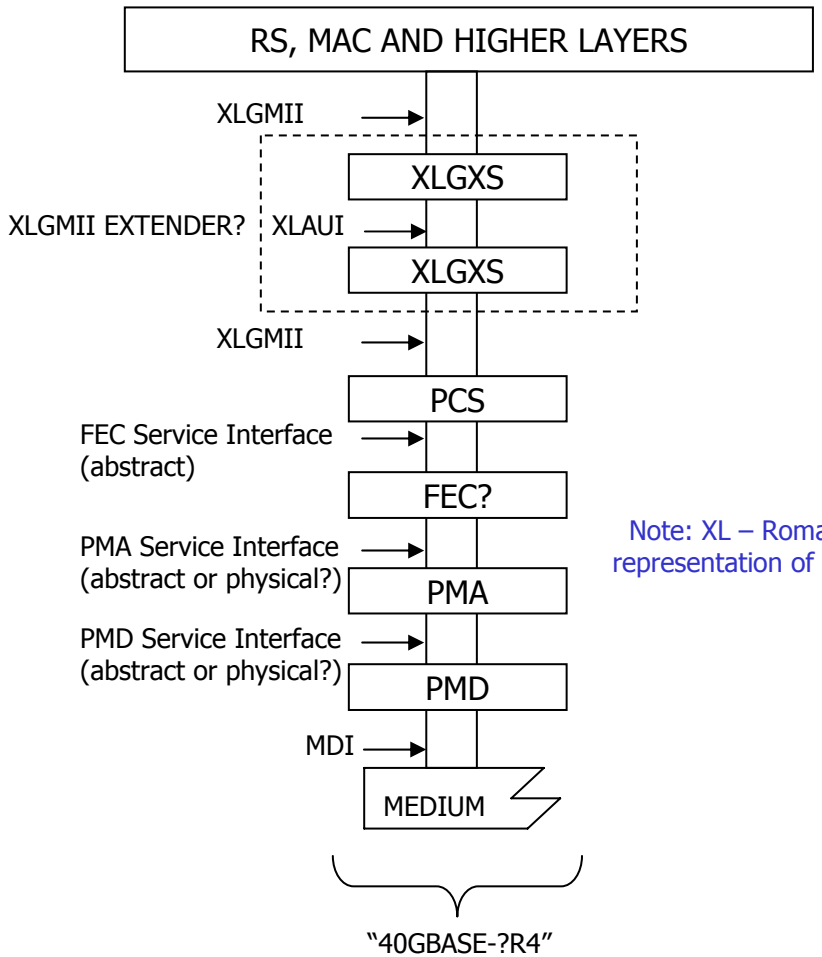
Note: * optional

Possible 40G backplane layer stack



40GbE interfaces

- All interfaces need to be discussed in terms of
  - Type of definition
    - Physical?
    - Abstract?
  - If physical
    - Optional?
    - Signaling?
    - Channel?
  - Common for both 40G and 100G?
  - Number of lanes
    - 4 x 10G solution assumed based on discussions in HSSG
- Necessary for a technically complete document



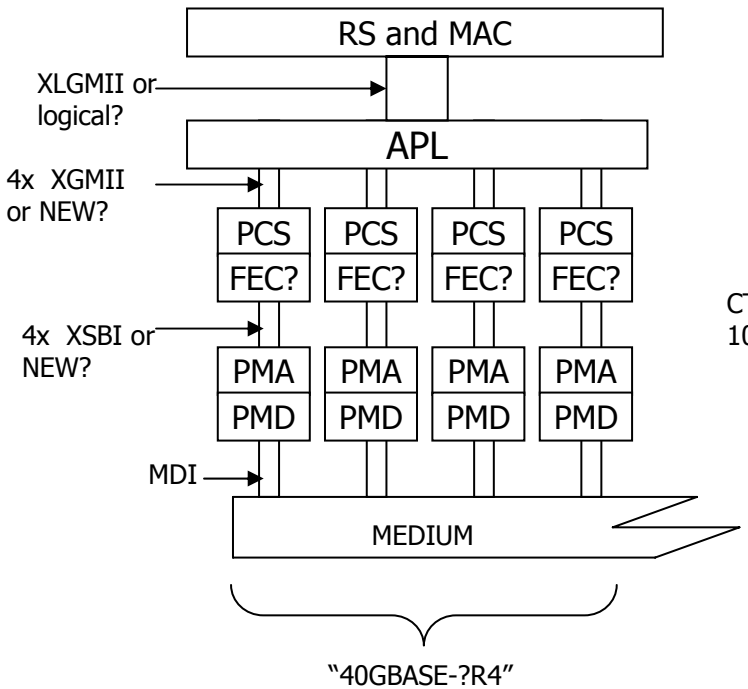
Note: XL – Roman numeral representation of 40

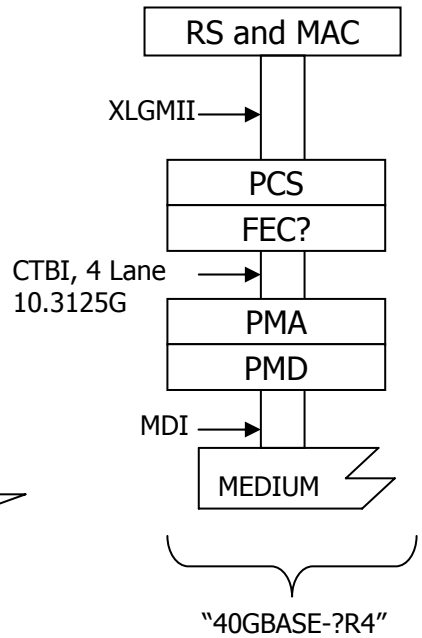
Interface solutions required

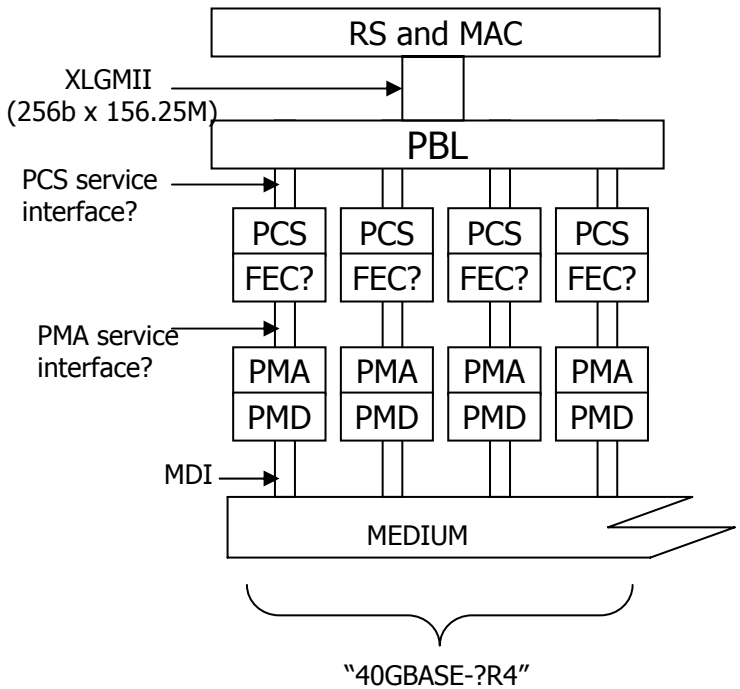
- XLGMII (Forty Gigabit MII) – PCS interface
  - Definition, data width, clock frequency, electricals
  - Logical abstraction or physical interface?
  - Should be same as 100 Gigabit MII (CGMII) or different?
- XLGMII extender sublayer required?
  - We cannot assume sublayers will always be implemented by co-locating with MAC
  - We have to plan for physical instantiation of an interface similar to XAUI for 10GbE
- PMA Service interface
  - Logical abstraction or physical instantiation ?
  - 4 - lane serial interface assumed based on discussions in HSSG
  - Electrical definition, channel characteristics, same board or with connector(s)
  - Common for all 40GbE PHY types
- FEC Service interface
  - Can be a logical abstraction interface
- PMD Service interface
  - Physical instantiation or abstraction?
  - Historically this interface has been standardized outside of IEEE 802.3

Architectural options discussed in HSSG

- Proposals for 4-lane solution for 40GbE (4 lane x 10.3125G)
  - by leveraging existing 10GbE architecture/technology where possible
- Options discussed in HSSG (listed in alphabetical order)
  - APL (Aggregation of physical layers)
  - CTBI (100G ten bit interface)
  - PBL (Physical bundling layer)







40GbE summary

- MAC, mostly speed independent, update MAC parameters table
- Desire to reuse 10GBASE-R technology expressed
  - 4-lane solution for all PHY types?
  - PCS solution to be selected
    - Block coding to be selected (64b/66b coding?)
    - Lane striping solution to be selected
- Decision on physical instantiation of inter-sublayer interfaces
  - PCS service interface to be defined (physical or abstraction?)
  - XLGMII extender?
  - Physical instantiation of PMA service interface?
  - Physical instantiation of PMD service interface?
    - Industry standard or IEEE standard?
- Could leverage 10GBASE-KR architecture including FEC, Auto-Neg and PMA/PMD
- Short reach cable solution can leverage backplane architecture
- Solutions to comprehend MTTFPA
- Impact of EEE objective related to 10GBASE-KR

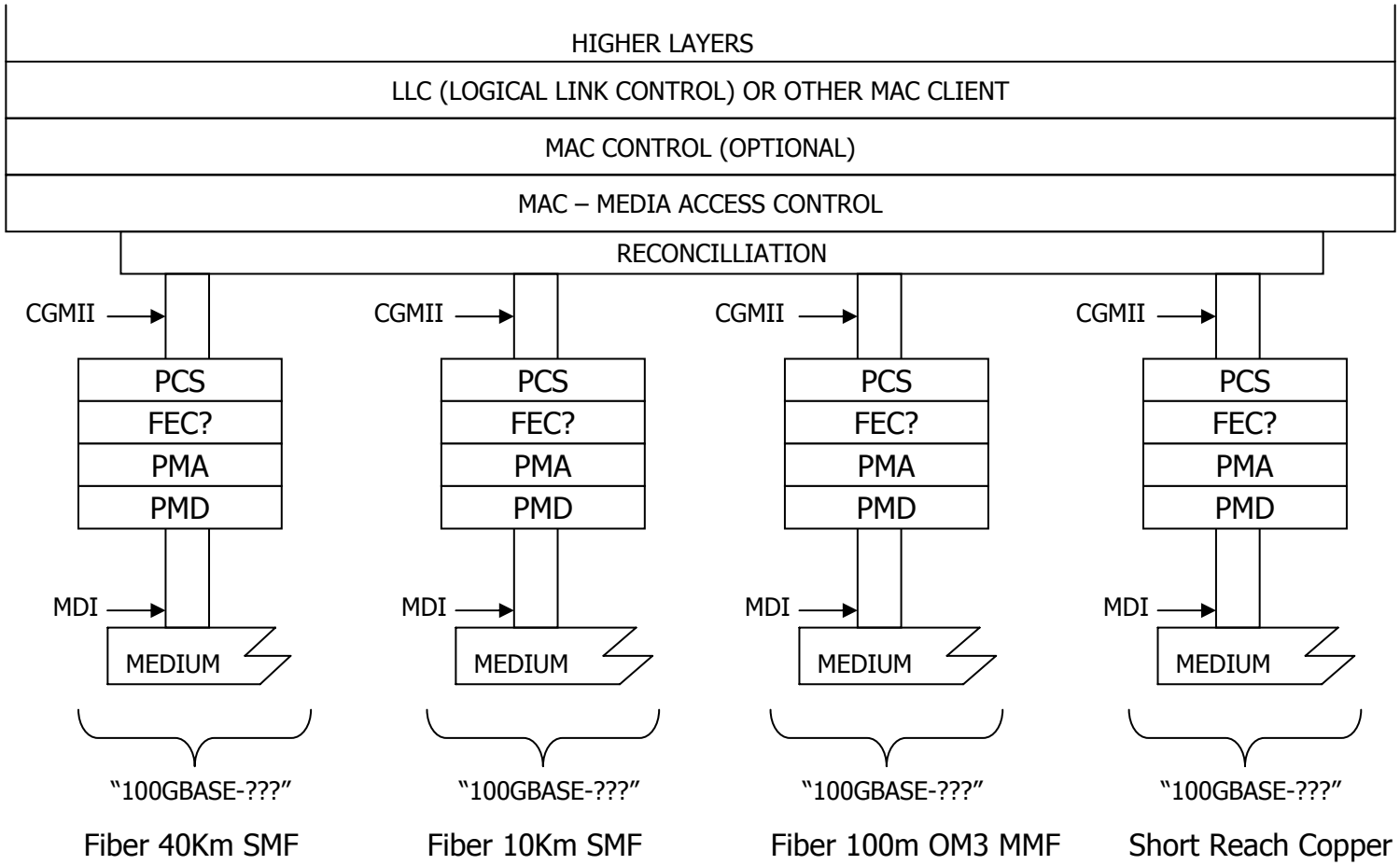
100GbE objectives

HSSG Objectives

- Support full-duplex operation only
- Preserve the 802.3 / Ethernet frame format utilizing the 802.3 MAC
- Preserve minimum and maximum FrameSize of current 802.3 standard
- Support a BER better than or equal to 10^{-12} at the MAC/PLS service interface
- Provide appropriate support for OTN
- Support a MAC data rate of 40 Gb/s
- Provide Physical Layer specifications which support 40 Gb/s operation over:
  - at least 100m on OM3 MMF
  - at least 10m over a copper cable assembly
  - at least 1m over a backplane
- Support a MAC data rate of 100 Gb/s
- Provide Physical Layer specifications which support 100 Gb/s operation over:
  - at least 40km on SMF
  - at least 10km on SMF
  - at least 100m on OM3 MMF
  - at least 10m over a copper cable assembly

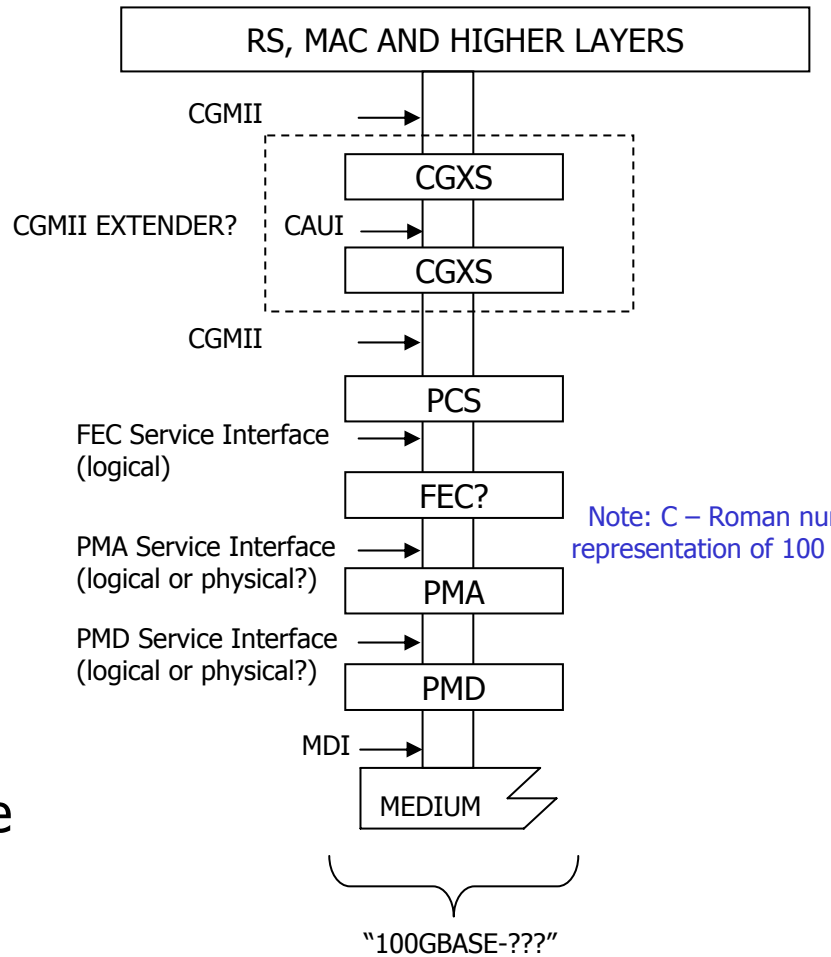
Adopted by HSSG and approved by 802.3 at July 2007 Plenary

Possible 100GbE architecture



100GbE interfaces

- All interfaces need to be discussed in terms of
  - Type of definition
    - Physical?
    - Abstract?
  - If physical,
    - Optional?
    - Signaling?
    - Channel?
  - Common for both 40G and 100G?
  - Number of lanes?
  - Rate per lane?
- Necessary for a technically complete document

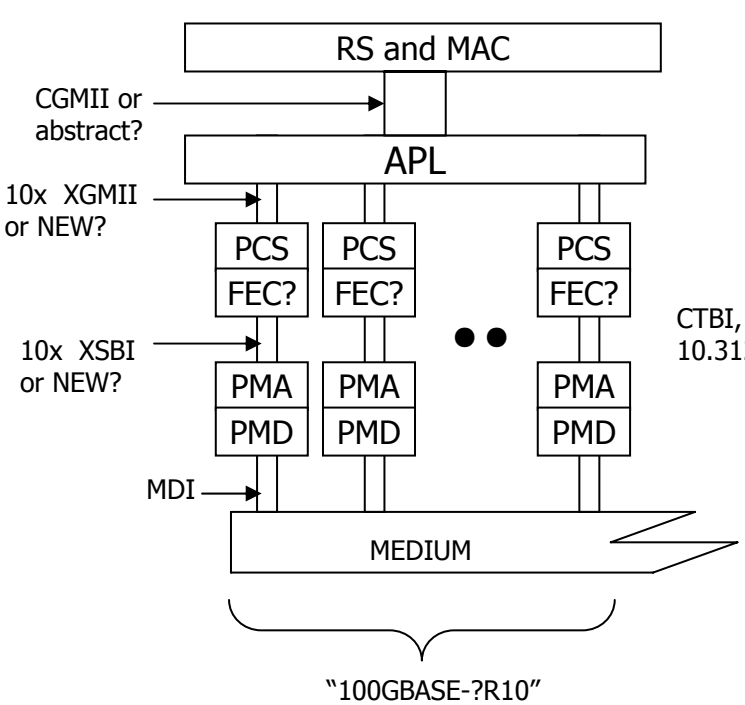


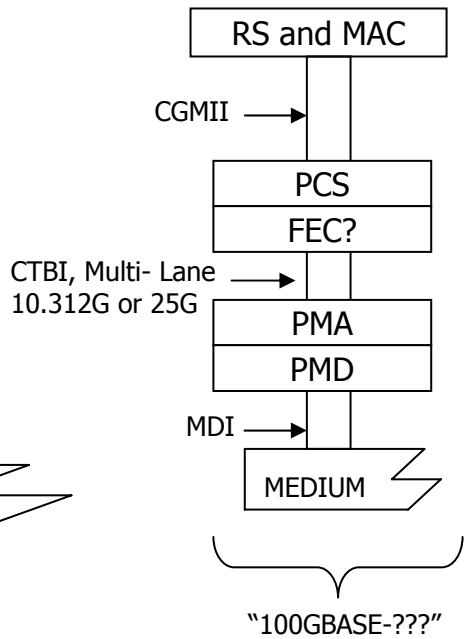
Interface solutions required

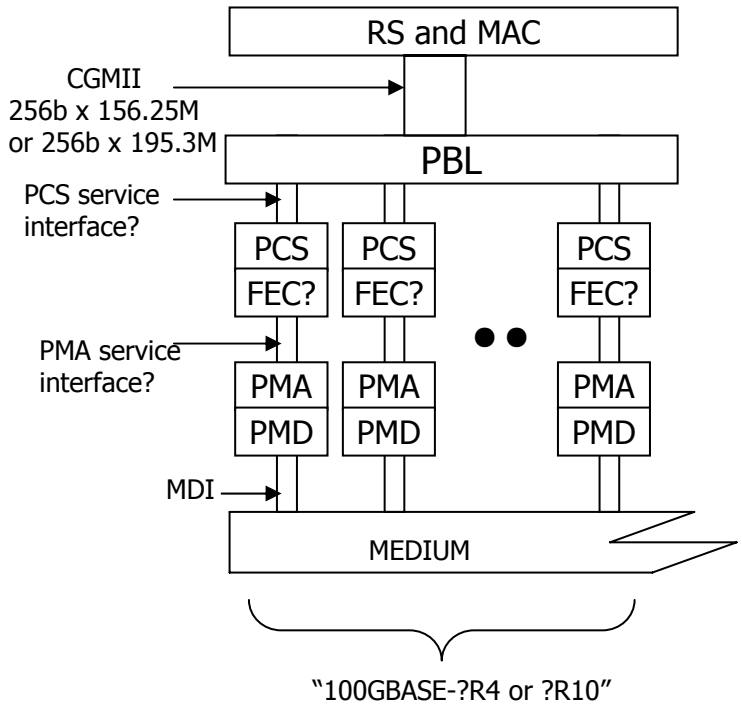
- CGMII (100 Gigabit MII) – PCS interface
  - Should CGMII be logical abstraction or physical interface
  - Definition, data width, electricals, clock frequency
  - Should CGMII support both 100Gb/s and 40Gb/s operation?
- CGMII Extender sublayer required?
  - We cannot assume PCS will always be implemented by co-locating with MAC
  - Multilane serial interface (10 lane x 10G or 4 lane x 25G or N lane x ?G)
- PMA Service interface
  - Physical instantiation required?
  - Multilane serial interface (10 lane x 10G or 4 lane x 25G or N lane x ?G)
  - Electrical definition, channel characteristics, same board or with connector(s)
  - Can this be common across all 100GbE PHY types (and 40GbE PHY types?)
- FEC Service interface
  - Is FEC layer required?
  - Can be similar to PMA Service interface
- PMD Service interface
  - Physical instantiation or abstraction?
  - Historically this interface has been standardized outside of IEEE 802.3

Multi-Lane options discussed in HSSG

- Proposals for multi-lane solution for 100GbE
  - 10 lane x 10G
  - 4 lane x 25G
- Options discussed in HSSG (listed in alphabetical order)
  - APL (Aggregation of physical layers)
  - CTBI (100G ten bit interface)
  - PBL (Physical bundling layer)







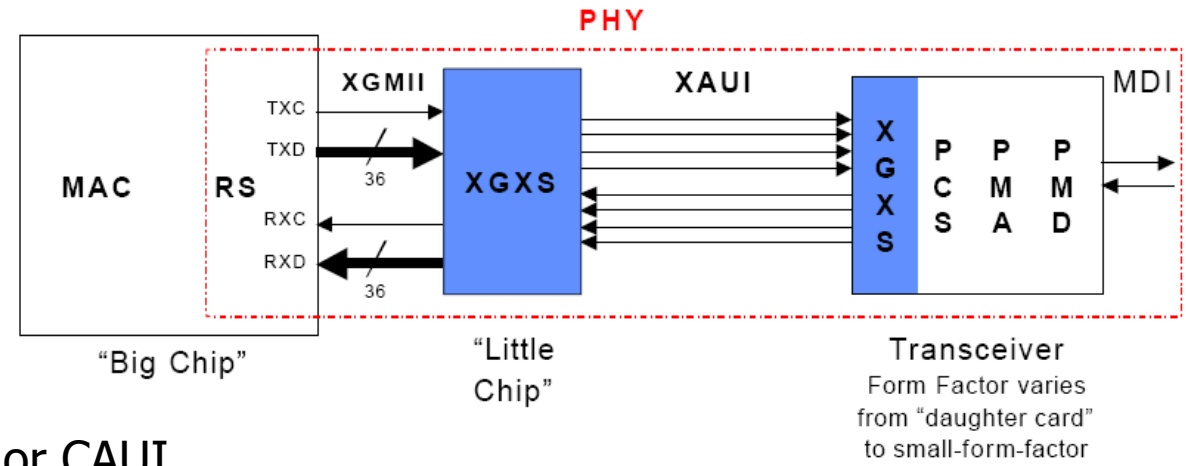
100GbE summary

- MAC, mostly speed independent, update MAC parameters table
- 1 to N lane solution for PHY types being discussed
- PCS solution to be selected
  - Block coding selection
  - Lane striping solution to be selected (APL, CTBI, PBL)?
  - Should the solution be common to 40GbE
    - 10GbE has 3 PCS solutions optimized for respective applications
- Decision on physical instantiation of inter-sublayer interfaces
  - PCS service interface to be defined (physical or abstraction?)
  - Physical instantiation of PMA service interface and electricals
  - Physical instantiation of PMD service interface and electricals
    - Industry standard or IEEE standard?
- Short reach cable architecture for 100GbE
- Solutions to comprehend MTTFPA

XLGMII/CGMII considerations

- Interface between MAC and PHY layers
- PCS Service interface
- Define Logical abstraction or physical interface?
- Need to decide if this interface be common to both 40G and 100G
- If defined physically -
  - How wide?
  - How fast?
- Example - XGMII
  - Defined interface - 74 signals
    - 32 bits data in each direction
    - 4 control signals in each direction
    - 1 clock in each direction
  - Optional
  - Optionally extended with XAUI
    - 4 pair differential lanes in each direction instead of 32-bit bus
    - Self clocked, solves skew issues with wide bus

XLAUI/CAUI considerations (1)



Reference:
XAUI / XGXS proposal, taborek_2_0700

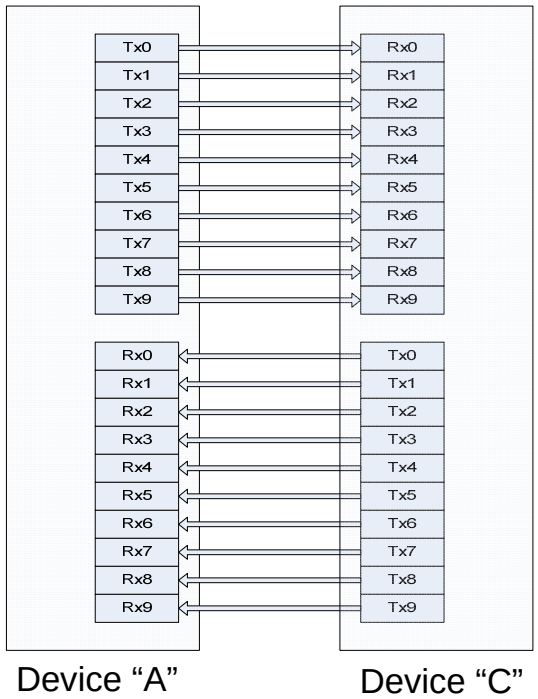
- XLAUI or CAUI
  - Possible interface:
    - To extend XLGMII or CGMII
    - Reduce width (routing issues)
  - If defined, need to decide to support 40G, 100G, or both
- Example - XAUI (optional)
  - Reduced 74 pin XGMII to 16 pins
  - Popular, but being overtaken by industry standard serial 10G interfaces
    - Less complex modules desired
    - Serial 10G deployment
- Possible examples
  - 4 lane x 10G
  - 10 lane x10G
  - 4 lane x 25G
- Possible evolution path
  - Current capabilities
  - Future proofing

XLAUI/CAUI considerations (2)

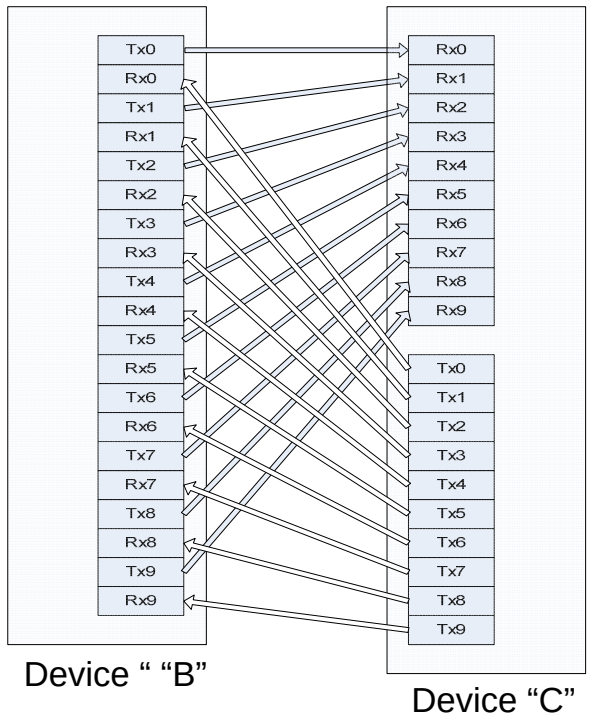
- Applications
  - Chip to chip
  - Chip to module?
- Input to module
  - Narrow interface is good
  - Can impact design complexity of module
- Channel definition - what type of channel?
  - Short reach MAC PHY interface – 8 to 12 inches (chip to chip interconnect)
  - Long reach – 30 to 40 inches (for backplane and module applications)
  - XAUI – 20 inches
- Routing concerns
  - 10 lane x10G
    - 20 differential pair
    - Signal integrity
  - 4 lane x 25G
    - 8 differential pair
    - Signal integrity

Routing considerations

Implementation “A”



Implementation “B”



- Classic Problems
  - Different implementations possible, different pinouts, routing
  - Wider interface: Cost and routing complexity grows with the width of interface
  - Narrow interface: Going faster has its own issues, e.g. signal integrity

Copper PMDs

- Objectives to support ≥ 10m over Cu Cable @ 40G and 100G
- # of Lanes?
- Rate per lane -
  - 40G: 4 x 10 assumed
  - 100G: Two options being discussed
    - 10 x 10G
    - 4 x 25G
      - This will be challenging
      - May require alternate modulation schemes
      - May be different solution than trying to do over a short reach channel
- Point of discussion – Can or should short reach and long reach channels be serviced by the same solution
      - Complexity / Power
      - Timing
      - Optimum solutions may be desired for short reach MAC to PHY interface (chip to chip interconnect)

Decisions required

- Interfaces not defined physically must be defined abstractly
  - Number of lanes and rate per lane for each interface
- XLGMII/CGMII extender
- PCS and lane striping
- Define specification for today's capabilities?
  - Future proofing can make a solution flexible but complex
  - Do we maximize flexibility at the expense of cost?
  - Or optimize for cost at the expense of flexibility?
- Common solutions for 40 and 100G, or solutions optimized for each rate