A 400GbE Architectural Option

IEEE P802.3bs 400 Gb/s Ethernet Task Force

July 2014 San Diego

Pete Anslow - Ciena
John D'Ambrosia – Dell
Mark Gustlin – Xilinx
Adam Healey – Avago
David Law – HP
Gary Nicholl - Cisco
Dave Ofelt – Juniper
Steve Trowbridge - ALU

Contributors

Special thanks for all of the work done by Hugh Barrass on these architectural concepts

Supporters

Rick Rabinovich - ALU

Agenda

- Proposed 400G architecture
- FEC options
- OTN reference point thoughts

What Needs to be Supported in the Architecture?

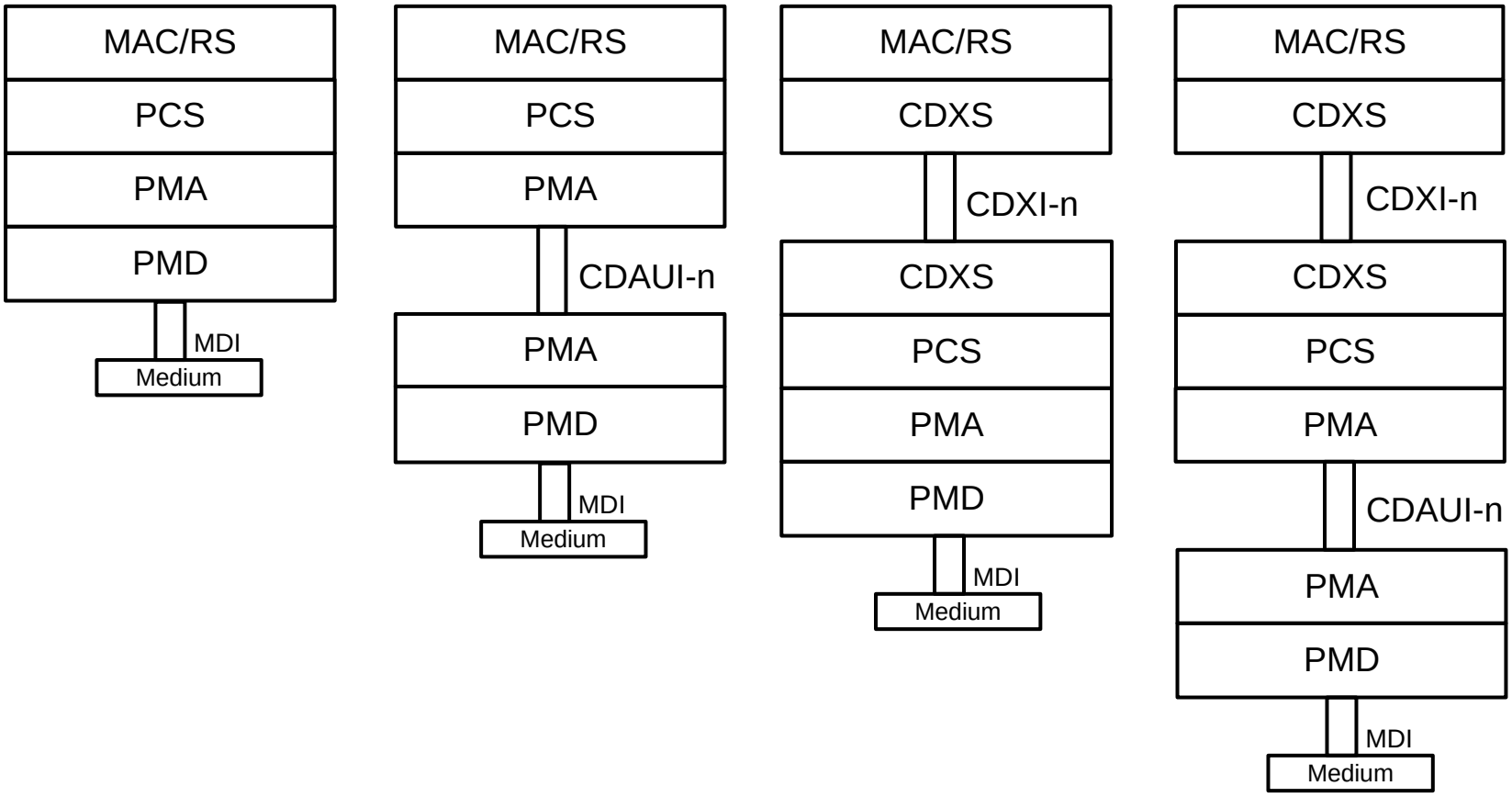
- The coding needs of the electrical interface may vary independently from the PMD interface
- The requirements for each interface can be different, both the FEC, modulation and number of lanes can change over time for each interface
- We need a single high level architecture which can support the evolving requirements of the interfaces over time
  - This does not mean it is requires a complicated implementation

Names & definitions

- … the naming of things

Item	Name Used Temporarily	Function/definition
Extender sublayer	CDXS	Extends xMII (recovers raw 400G datastream) – used whenever a different coding or FEC is required further out in the PHY. Includes line code, FEC & timing required for extender interface.
Extender interface	CDXI-n	Interface between two CDXS, may be various widths
PMA interface	CDAUI-n	Physical instantiation of PMA service interface (similar to CAUI)

A Possible 400G Architecture



The PCS can be unique for each PMD!
Though we strive for commonalty where possible

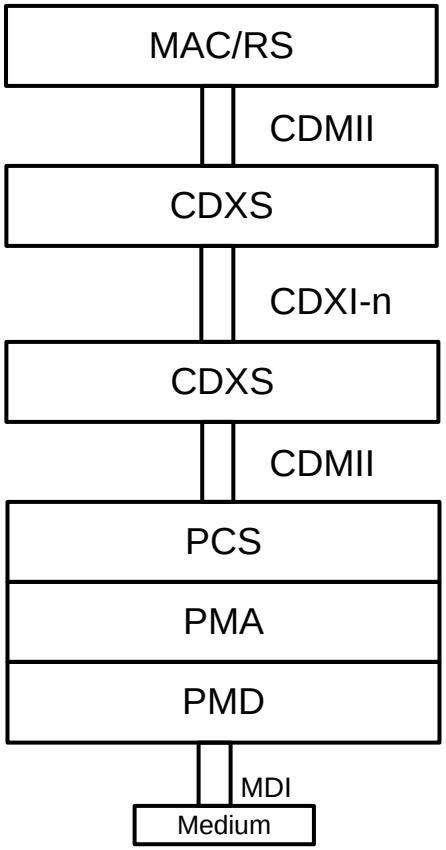
Sublayer Functions (at a high level)

Sublayer	10GbE	100GbE	400GbE (proposed)
MAC	Framing, addressing, error detection	Framing, addressing, error detection	Framing, addressing, error detection
Extender	PCS + PMA	N/A	PCS + PMA + FEC
PCS	Coding (8B/10B, 64B/66B), lane distribution, EEE	Coding (64B/66B), lane distribution, EEE	Coding, lane distribution, EEE, FEC
FEC	FEC, transcoding	FEC, transcoding, align and deskew	N/A?
PMA	Serialization, clock and data recovery	Muxing, clock and data recovery, HOM	Muxing, clock and data recovery, HOM??
PMD	Physical interface driver	Physical interface driver	Physical interface driver

Note that there are variations with a single speed, not all are captured in this table

A Possible 400G Architecture

- The interface between the CDXS and the MAC or PCS sublayer is always a CDMII



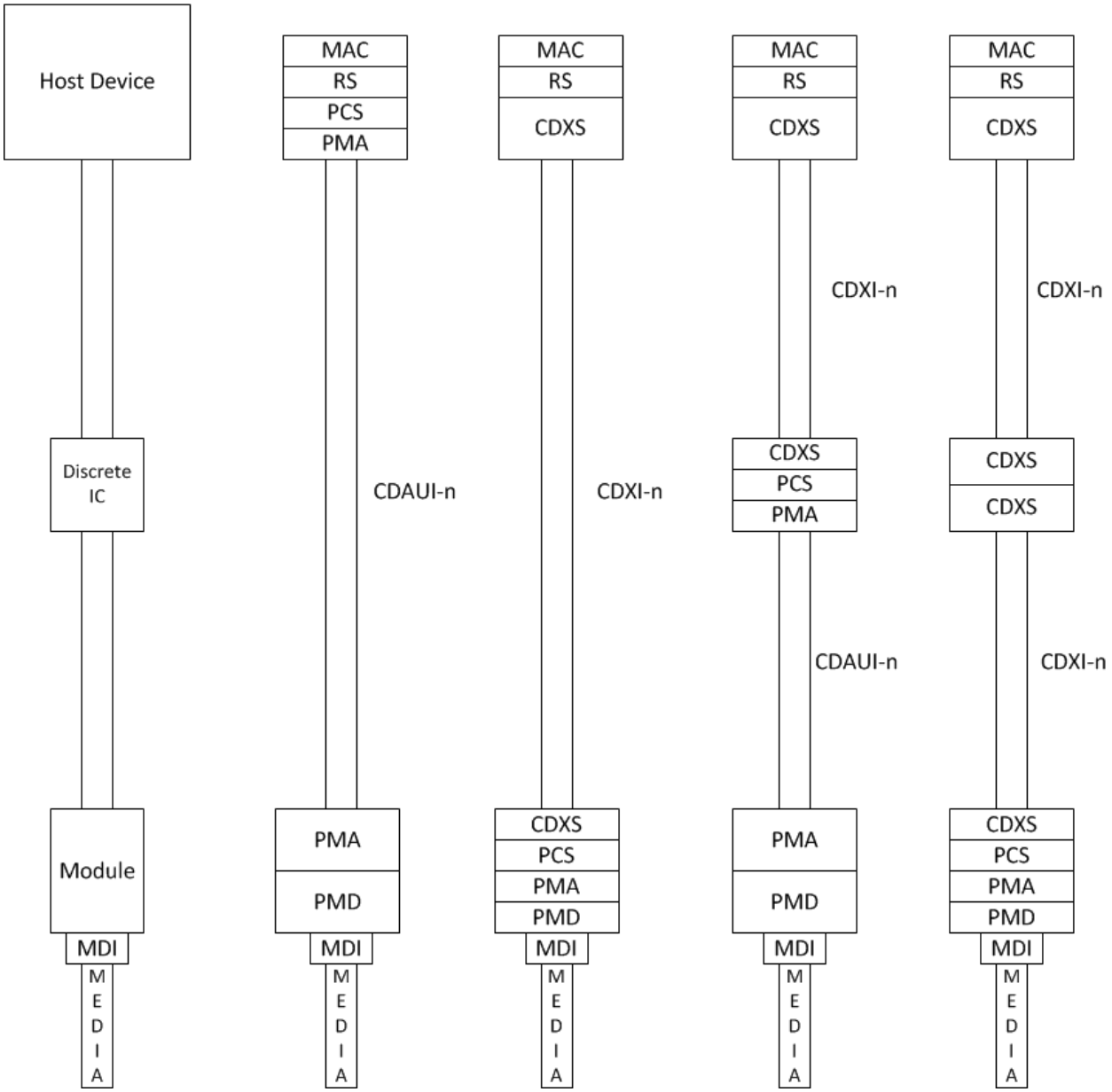
Line Encoding/FEC + PMA

Line Encoding/FEC + PMA

Line Encoding/FEC

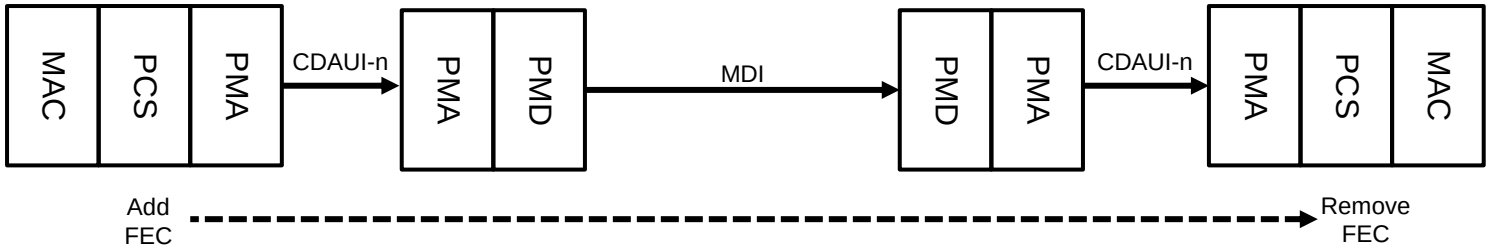
Muxing, CDR, HOM

400GbE Example Implementations



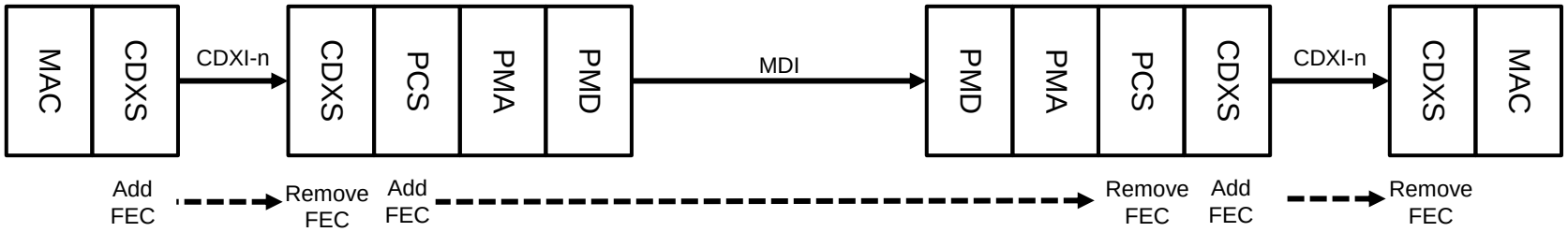
FEC Strategies: End to End

- End to End FEC pros and cons
  - \+ Simple, lowest overall complexity, latency and power
  - \- How to handle differentiation by application?
    - Short reach might require low latency, long reach can tolerate higher latency
  - \- How to handle the evolution of an electrical interface, legacy hosts etc.
    - Will mean in reality not having end to end FEC in some cases
  - \- How to allocate FEC error budget across multiple interfaces?
    - Works well if the BER contributed by the electrical interfaces is 0.1 x the BER from the PMD for instance



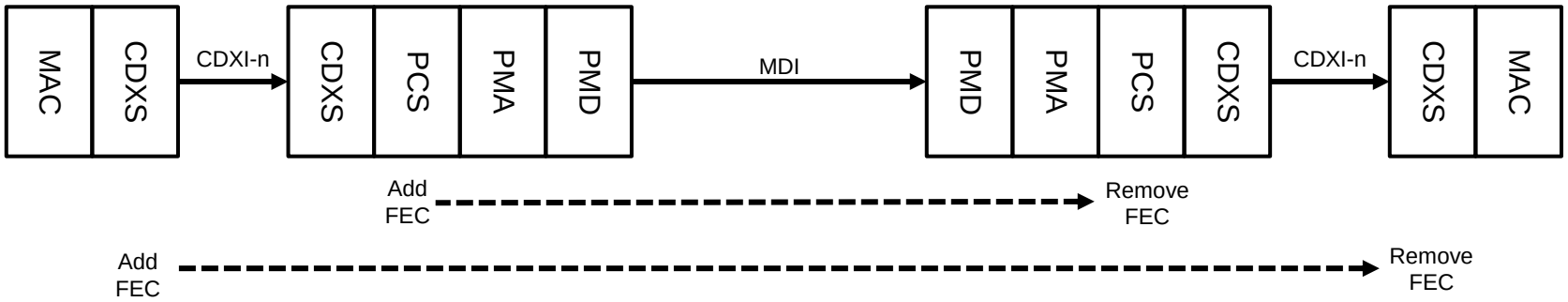
FEC Strategies: Segment by Segment

- Segment by Segment FEC pros and cons
  - \+ Most flexible, FEC is optimized for each application
  - \+ Easy to handle evolution of interfaces, legacy hosts etc.
  - \+ No issues with parsing BERs of multiple interfaces
  - \- Highest complexity, power, latency etc.



FEC Strategies: Encapsulated FECs

- Encapsulated FEC pros and cons
  - \+ Moderate complexity, latency and power
  - \+ Easier to handle evolution of interfaces, legacy hosts etc.
  - \- How to handle differentiation by application?
  - \- How to allocate FEC budget across multiple interfaces?
    - Up to 5 interfaces?
  - \- Bit rate might be higher than other options



Since the proposed architecture goes back to the MII between the CDXS and PCS layers, this FEC strategy is not possible!

OTN Reference Point

- Can the OTN reference point be the MII?
  - What if we want to carry end to end information (BIP or other stuff), how is that handled if the reference point is the MII?
- If the MII does not work, do we have to have the OTN reference point as a sublayer interface, or can we define an intra-sublayer interface for OTN?

Thanks!