Maintenance channel for 400GPL links

Adee Ran, Cisco

Introduction

- Ethernet links are complicated dynamic systems
- We have been relying on adaptive DSP receivers for several generations
- In 802.3dj we defined the ILT/RTS functions that enable receivers to control some transmitter parameters during “training”
  - And reduce the uncertainty in the link/path startup
- After the path is up, adaptation is performed only by the receiver
  - Transmitter parameters in all ISLs are fixed
- Tools for link monitoring and maintenance are limited
  - The only information exchanged about the link state is the “Degraded SER” bit (between PCSs)
  - There is no per-ISL information
  - Some observability may be available locally (e.g. through CMIS)

Opportunity

- What if 400GPL supports a low-rate in-band **maintenance channel** alongside the mission data?
- Possible uses:
  - Exchange of performance metrics between devices
  - Transmitter parameter change requests
  - Accurate delay measurement
  - Remote management
- Related PHY maintenance mechanisms exist in other specifications
  - 802.3 subclause 149.3.9: MultiGBASE-T1 PCS-level OAM (Operations, Administration, and Maintenance) and others
  - ITU-T PLOAM (Physical Layer Operations, Administration, and Maintenance) used in access networks
- This presentation provides an overview of how a maintenance channel can be implemented within the BASE-R architecture
  - An introduction to a possible future P802.3dv proposal

General idea: repurpose alignment markers

- Alignment markers (AMs) are inserted periodically by the BASE-R PCS. Their purpose is to enable PCS lane reordering and deskew.
- In 802.3dj, the SM-PMA and inner FEC sublayers also use the AMs to align the data stream to the FEC symbol boundary.
- After the initial alignment, the AMs are no longer needed, but are still inserted periodically by the PCS Tx and deleted by the PCS Rx.
  - Loss of AM lock indicates the link is down, but in practice this will be detected much earlier by uncorrected codewords.
- A link between two PMAs can use the AM bandwidth for an in-band maintenance channel (MC).

Example: MC between optical modules

(Can be mapped to Chip-to-module and other ISL types)

Module-to-module link startup

- Consider future modules using:
  - 400 Gb/s/lane PMDs
  - SM-PMAs similar to those defined in 802.3dj, possibly with 2:1 muxing ratio (“gearbox”)
- Module startup begins with ILT/RTS (Annex 178B) on both interfaces (electrical and optical).
- When training completes, the module’s PMA may perform AM lock (176.4.2.2) on each of its two interfaces.
  - In 802.3dj, this function is **required** for some PMAs, for example, 200GBASE-R 8:1 SM-PMA. It is not required for an n:n PMA, e.g. “retimer module” with 200G on both electrical and optical interfaces, but may still be implemented (e.g., to enable RS-FEC monitoring).
  - Let’s assume the module is “AM aware” in both directions.

Module-to-module link startup

- Completion of AM lock in the transmit direction (from the AUI) is communicated to the PMD by the PMA:IS_SIGNAL.request service interface primitive.
  - Similarly, AM lock in the receive direction (from the PMD) is communicated to the AUI by PMA:IS_SIGNAL.indication.
- Inside the PMA:
  - AM lock in the **transmit** direction (from the AUI) is a prerequisite to passing data from the host to the link partner
  - AM lock in the **receive** direction (from the PMD) is a prerequisite to passing data from the link partner to the host
  - If the peer (module) is sending AMs, it must have achieved AM lock on its AUI interface.
- Once AM lock is achieved in both interfaces, the module PMA is in data mode. From this point and on, AMs are not required for alignment or lane identification anymore.
  - As long as the Ethernet link is up, **the AM locations and content will not change**.

AM payloads (example from 800G PCS)

- There are 15 bytes per AM. The initial 3 bytes CM0-CM2 are common, and their bit-inverted versions CM3-CM5 are also common. The remaining 9 bytes are unique per lane.
- In the 800GBASE-R PCS, “flow 1” (additional 16 lanes) has the same content in the common bytes CM0-CM5 and different content in the remaining 9 bytes.
- 400G and 1.6T have a similar structure – 6 fixed bytes and 9 unique bytes per AM.
- The unique bytes are used only for identifying the received PCSL numbers (for reordering – a one-time operation).

Table 172–2—800GBASE-R flow 0 alignment marker encodings

PCS lane number	Encoding[a]														
	CM0	CM1	CM2	UP0	CM3	CM4	CM5	UP1	UM0	UM1	UM2	UP2	UM3	UM4	UM5
0	0x9A	0x4A	0x26	0xB6	0x65	0xB5	0xD9	0xD9	0xFE	0x71	0xF3	0x26	0x01	0x8E	0x0C
1				0x04				0x67	0xA5	0xDE	0x7E	0x98	0x5A	0x21	0x81
2				0x46				0xFE	0xC1	0xF3	0x56	0x01	0x3E	0x0C	0xA9
3				0x5A				0x84	0x79	0x80	0xD0	0x7B	0x86	0x7F	0x2F
4				0xE1				0x19	0xD5	0x51	0xF2	0xE6	0x2A	0xAE	0x0D
5				0xF2				0x4E	0xED	0x4F	0xD1	0xB1	0x12	0xB0	0x2E
6				0x3D				0xEE	0xBD	0x9C	0xA1	0x11	0x42	0x63	0x5E
7				0x22				0x32	0x29	0x76	0x5B	0xCD	0xD6	0x89	0xA4
8				0x60				0x9F	0x1E	0x73	0x75	0x60	0xE1	0x8C	0x8A
9				0x6B				0xA2	0x8E	0xC4	0x3C	0x5D	0x71	0x3B	0xC3
10				0xFA				0x04	0x6A	0xEB	0xD8	0xFB	0x95	0x14	0x27
11				0x6C				0x71	0xDD	0x66	0x38	0x8E	0x22	0x99	0xC7
12				0x18				0x5B	0x5D	0xF6	0x95	0xA4	0xA2	0x09	0x6A
13				0x14				0xCC	0xCE	0x97	0xC3	0x33	0x31	0x68	0x3C
14				0xD0				0xB1	0x35	0xFB	0xA6	0x4E	0xCA	0x04	0x59
15				0xB4				0x56	0x59	0xBA	0x79	0xA9	0xA6	0x45	0x86

[a] Each octet is transmitted LSB to MSB.

available bytes in the AM block

Maintenance channel using AMs

- After the PMA achieves AM lock in both directions, the unique AM content can be repurposed for a maintenance channel (MC):
  - In the transmit direction: overwrite with “MC data”.
  - In the receive direction: extract the “MC data” and restore the original content.
  - The common bytes should be preserved to keep the AM lock functionality.
- Considerations:
  - AMs are sent by the PCS at the same time on all PCSLs (“AM group”).
  - The AMs are distributed between physical lanes; the distribution is fixed. Each physical lane sees a part of the AM group.
  - MC data is not FEC protected and needs to be “randomized”.
  - The original payloads of the received AMs can be stored during the initial AM lock, and restored in the output to the PMA client (to ensure compatibility with existing PMAs and PCSs).

Maintenance channel using AMs

- The number of AMs per physical lane varies as shown in the table.

PHY type	400GBASE-R		800GBASE-R		1.6TBASE-R	
Number of AMs	16		32		16	
Time between AM groups [1]	~105 µs		~105 µs		~105 µs	
SerDes rate [2]	400G	200G	400G	200G	400G	200G
# of PMA lanes	1	2	2	4	4	8
AM group size per PMA lane	16	8	16	8	4	2
Message bytes per PMA lane	144	72	144	72	36	18

[1] See gustlin_3dj_01b_230206, page 10

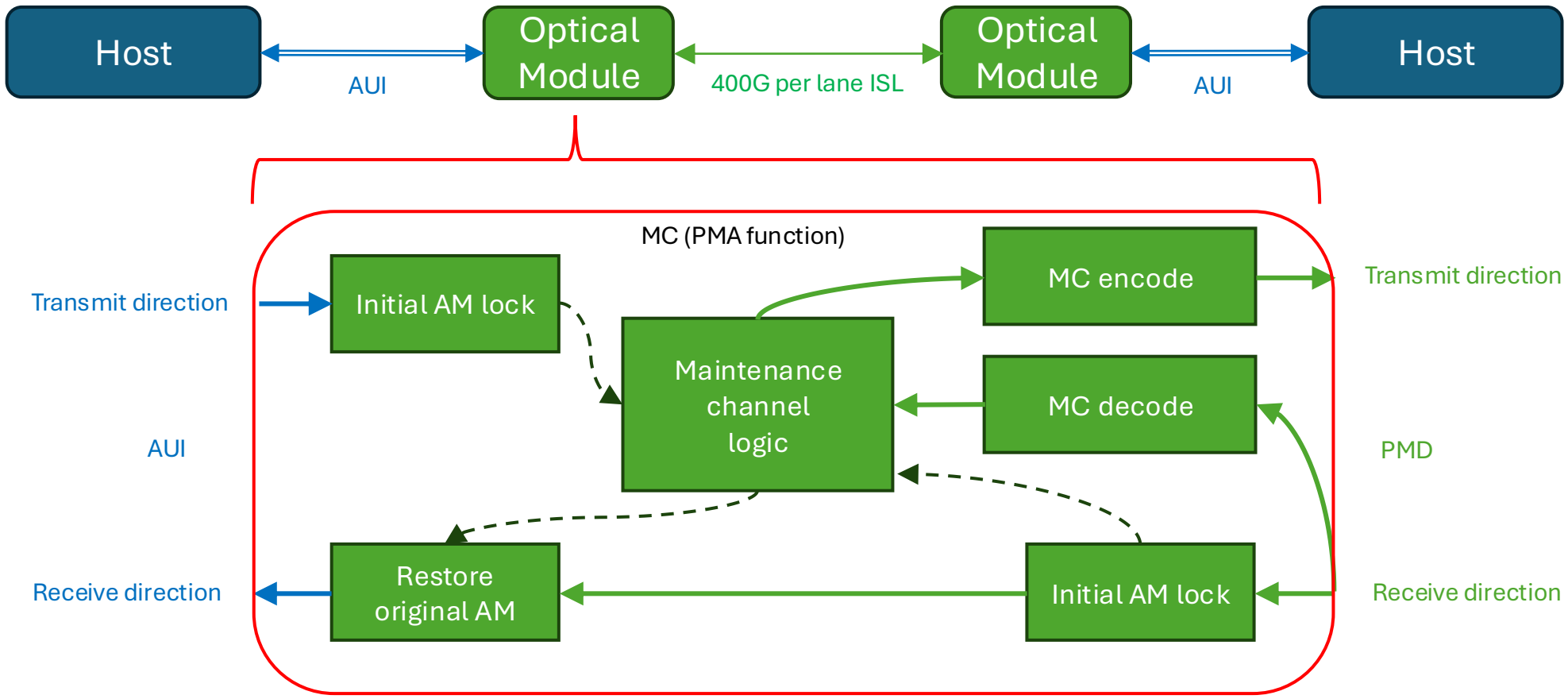
[2] 200G/lane PMAs are likely out of scope for 802.3dv

The minimum message size is 36 bytes per lane for 400GPL PMAs. Depending on PHY type, longer messages can be used.

Additional considerations

- PMA-to-PMA communication is preferable – it enables module-level functionality and PCS-independent operation
  - May be extended to cross PMA boundary, e.g., PCS to PCS
- Per-lane communication is preferable – it simplifies decoding logic
  - Some messages are inherently lane-specific
  - Per-lane messages may be aggregated to form per-link (multi-lane) messages
- Firmware-friendly protocol is preferable
  - 36 bytes every 105 µs creates a raw bandwidth of ~2.7 Mb/s
  - The receiver should be able to throttle the message rate
- MC function should enable bidirectional communication
  - Usage may be asymmetric

Block diagram



Future work

- Details of message format, encoding and decoding processes
- Handshake and acknowledgement mechanism
- State diagrams, flowcharts, etc.
- Optional or mandatory?
- Establishment procedure
- Message types for actual use cases

Summary

- What was presented:
  - A possible way to create a maintenance channel (MC) for 400GPL by repurposing alignment markers (AMs)
  - Recommended as a PMA function for per-ISL, per-lane communication
  - AM functionality outside of the MC is presereved
- Usage models include
  - Exchange of performance metrics between devices
  - Transmitter parameter change requests
  - Accurate delay measurement
  - Remote management
- A future detailed proposal may be provided in P802.3dv

That’s all!

Questions?