Addressing PTP timestamp accuracy for 800GBASE-ER1 with an 800GMII Extender

Peter Sinn - Alphawave

Gary Nicholl, Mike Sluyski - Cisco

Ulf Parkholm - Ericsson

Xiang He - Huawei

Dave Ofelt - Juniper Networks

Andras Dekoos - Microchip

Tom Huber - Nokia

IEEE 802.3 May 2024 Interim, Annapolis, MD, USA

Supporters

- Jerry Pepper – Keysight
- Matt Brown – Alphawave
- Eugene Opasnick - Broadcom
- Steve Gorshe - Microchip
- Atul Srivastava - NTTD America
- Markus Weber - Cisco
- Jeffery Maki - Juniper Networks
- Sean Nicholl – AMD
- Ted Sprague – Infinera
- Antonio Tartaglia - Ericsson

References

Recommendation ITU-T G.8273/Y.1368 - Framework of phase and time clocks (06/2023). https://www.itu.int/rec/T-REC-G.8273/en

IEEE 1588-2019 – Standard for a Precision Clock Synchronization Protocol for Network Measurement and Control Systems. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9120376

IEEE 802.3cx; Improving PTP Timestamping Accuracy on Ethernet Interfaces: https://www.ieee802.org/3/ad_hoc/ngrates/public/calls/19_0702/tse_nea_01_190702.pdf

IEEE P802.3df: Incoming: MOPA: Time synchronization error in PTP networks. https://www.ieee802.org/3/minutes/nov23/incoming/MOPA_to_IEEE_802p3_231102_Redacted.pdf

Updated logic baseline for an 800GbE coherent PHY based on oFEC/C-band. https://grouper.ieee.org/groups/802/3/dj/public/23_07/nicholl_3dj_02a_2307.pdf

Consideration of timestamp accuracy with MII-extender in coherent 800GBASE-ER1

https://www.ieee802.org/3/dj/public/23_11/parkholm_3dj_01_2311.pdf

Introduction

- PTP accuracy is becoming more important in a broader set of networks
  - Reference liaison letter from MOPA: https://www.ieee802.org/3/minutes/nov23/incoming/MOPA_to_IEEE_802p3_231102_Redacted.pdf
- Ethernet PHYs, including those being developed within 802.3dj, can address timestamping accuracy by following Clause 90. The exception to this is the currently adopted 800GBASE-ER1 baseline with an 800GMII Extender (at either end).
- This contribution proposes an update to the current 800GBASE-ER1 baseline to address this limitation.

Problem statement

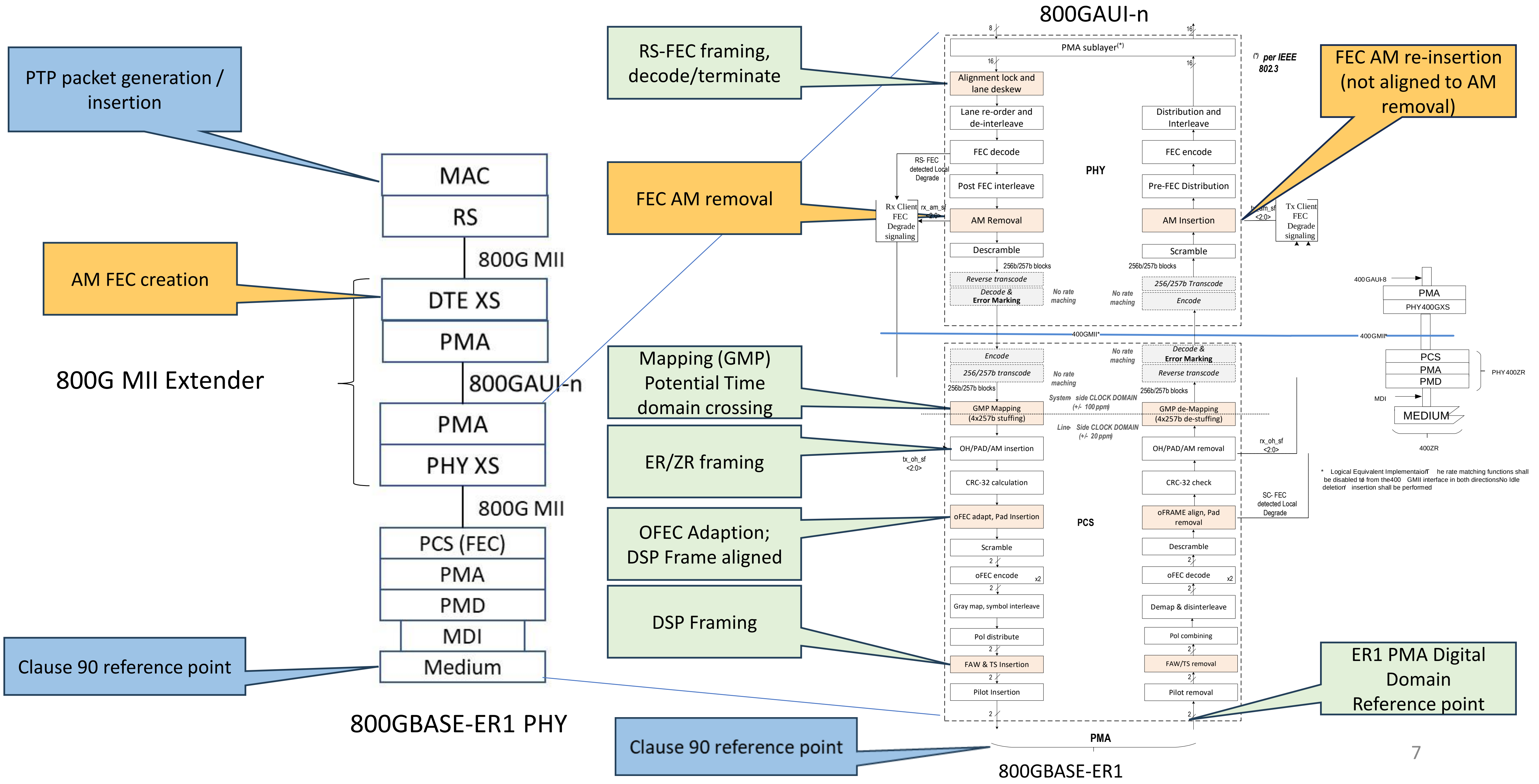
- In addition to the usual need for very careful design to achieve high PTP accuracy, 800GBASE-ER1 when used with an 800GMII Extender has an architectural constraint that limits the achievable accuracy (even after using Clause 90 compensation):
  - In the transmit path AMs are removed (by the PHY_XS) before mapping the Ethernet symbols into the GMP frame (in the 800GBASE-ER1 PCS).
  - In the receive path, after Ethernet symbols are de-mapped from the GMP frame (in the 800GBASE-ER1 PCS), AMs are then re-inserted at some random location (by the PHY_XS).
  - This changes the relationship between packets and introduces timestamping errors.
- Analysis has shown that the insertion/removal of the Alignment Mechanism (AM) fields can affect c|TE| in a measurable way (5.12 nS@ 800G).

$$1/(803125000000) \times 257 \times 16 = 0.00000000512$$

Solution overview

- A solution is proposed that ensures AMs are re-inserted in the same positions by the receive 800GMII Extender, relative to where they were removed by the transmit 800GMII Extender.
- This preserves the Host-to-Host timing relationship of the underlying PTP packet across the 800GBASE-ER1 link.
- Implementation of this function is considered optional and does not affect interoperability. It is assumed the 800G MII Extender choice is symmetrical across the 800GBASE-ER1 link.

800GBASE-ER1 Architecture with 800GMII Extender



PTP packet generation / insertion

AM FEC creation

800G MII Extender

Clause 90 reference point

800GBASE-ER1 PHY

RS-FEC framing, decode/terminate

FEC AM removal

Mapping (GMP) Potential Time domain crossing

ER/ZR framing

OFEC Adaption; DSP Frame aligned

DSP Framing

Clause 90 reference point

FEC AM re-insertion (not aligned to AM removal)

ER1 PMA Digital Domain Reference point

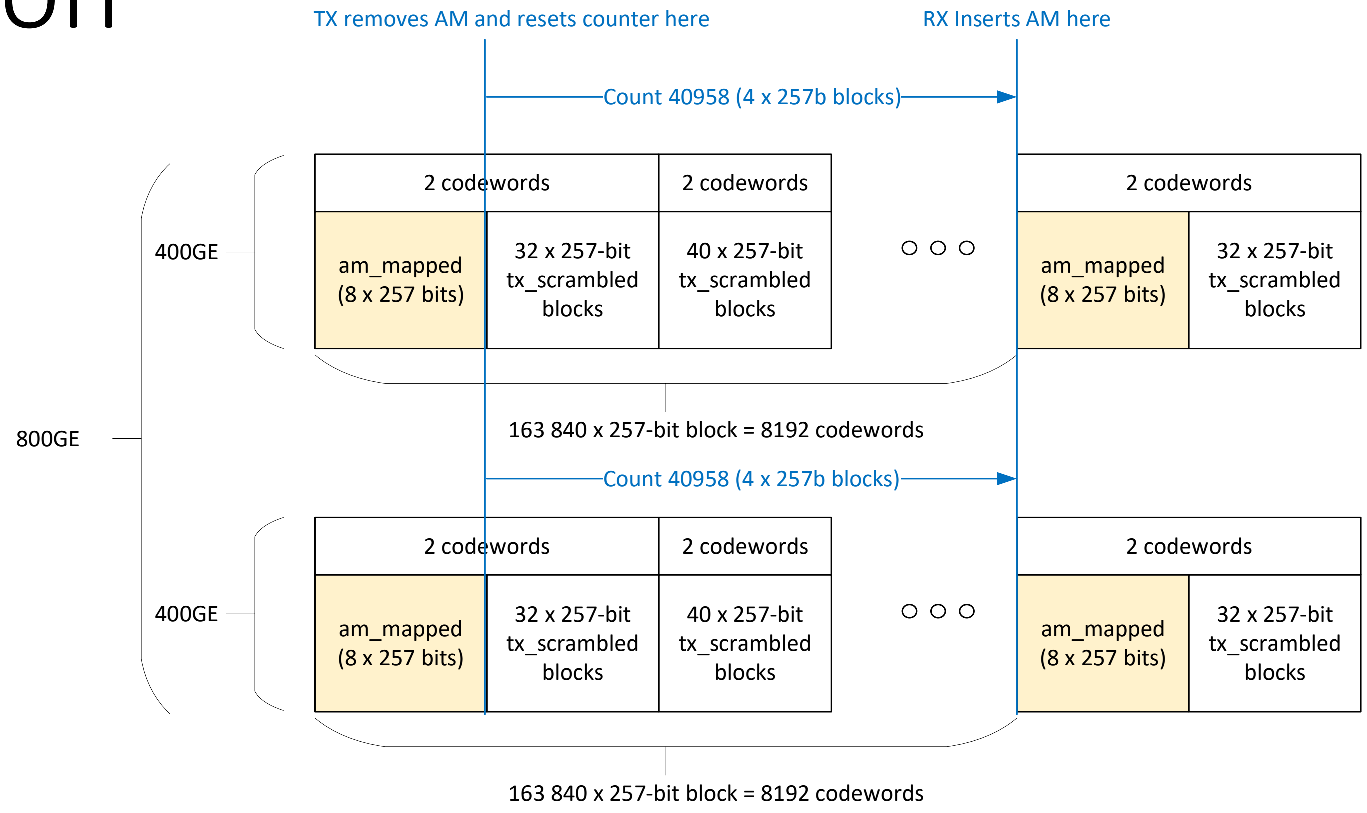
* Logical Equivalent ImplementaionT he rate matching functions shall be disabled to from the400 GMII interface in both directionsNo Idle deletion/ insertion shall be performed

Solution concept

- The 800GBASE-R FEC frame (after serializing the two flows) consists of AMs (16×257b) and payload blocks (327664×257b)
  - The AMs are removed prior to mapping into the GMP frame
- This frame floats within the GMP frame used by the 800GBASE-ER1 PCS
  - The GMP frame has overhead and a payload that is a mix of data and stuff blocks
  - The payload is organized as groups of four 257b blocks that are either data or stuff
- The position of the 800GBASE-R AMs can be conveyed from the transmitter to the receiver by encoding their relative position to the mapped data in the client payload overhead block of the GMP frame
  - The receiver can use this to determine where AMs need to be inserted

AM Synchronization

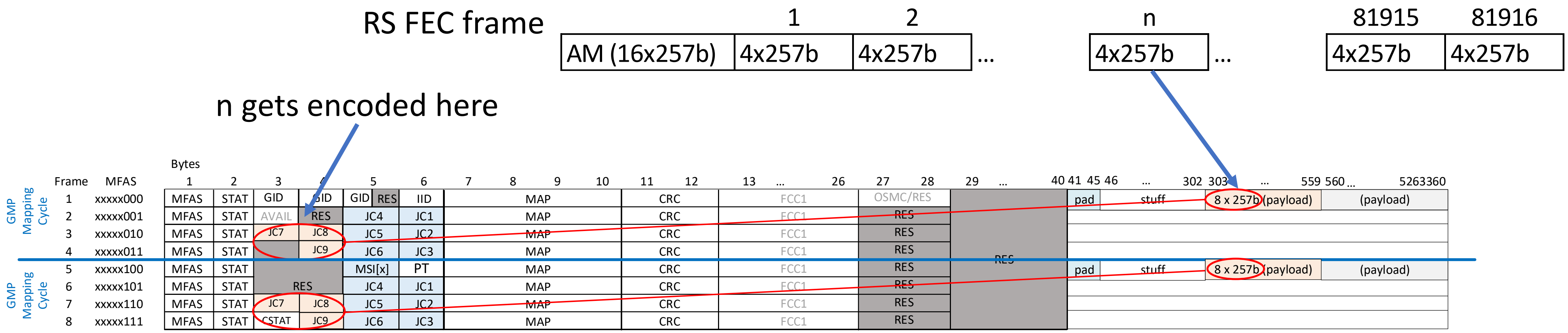
The purpose of the Counter is to convey the TX_START/END (Point when AM are removed in the TX mapper) to RX_START/END (Point where AM insertion is required in the RX de-mapper)


TX removes AM and resets counter here
RX Inserts AM here
Count 40958 (4 x 257b blocks)
800GE
400GE
2 codewords
2 codewords
am_mapped (8 x 257 bits)
32 x 257-bit tx_scrambled blocks
40 x 257-bit tx_scrambled blocks
○ ○ ○
2 codewords
am_mapped (8 x 257 bits)
32 x 257-bit tx_scrambled blocks
163 840 x 257-bit block = 8192 codewords
Count 40958 (4 x 257b blocks)
400GE
2 codewords
2 codewords
am_mapped (8 x 257 bits)
32 x 257-bit tx_scrambled blocks
40 x 257-bit tx_scrambled blocks
○ ○ ○
2 codewords
am_mapped (8 x 257 bits)
32 x 257-bit tx_scrambled blocks
163 840 x 257-bit block = 8192 codewords

Host Rate	Code Words/AM Block	257-bit blocks/AM	800GE 4 x 257b Counter (modulo)
800G	8192	163840	81916

Proposal – Mark AM position using GMP extender JC9,JC7,JC8 (MSB…LSB) OH bytes.

- The GMP Mapping block Includes a counter which increments for every 4 x 257-bit block that includes data in the 800ZR transmission frame. The counter is reset at the start of every 800GBASE-RS AM delineated block.
- The count value is encoded in JC9, JC7, JC8 (MSB..LSB) OH fields and indicates the offset position of the first 4 x 257b blocks carrying data relative to the AM framed 800GBASE-RS FEC.



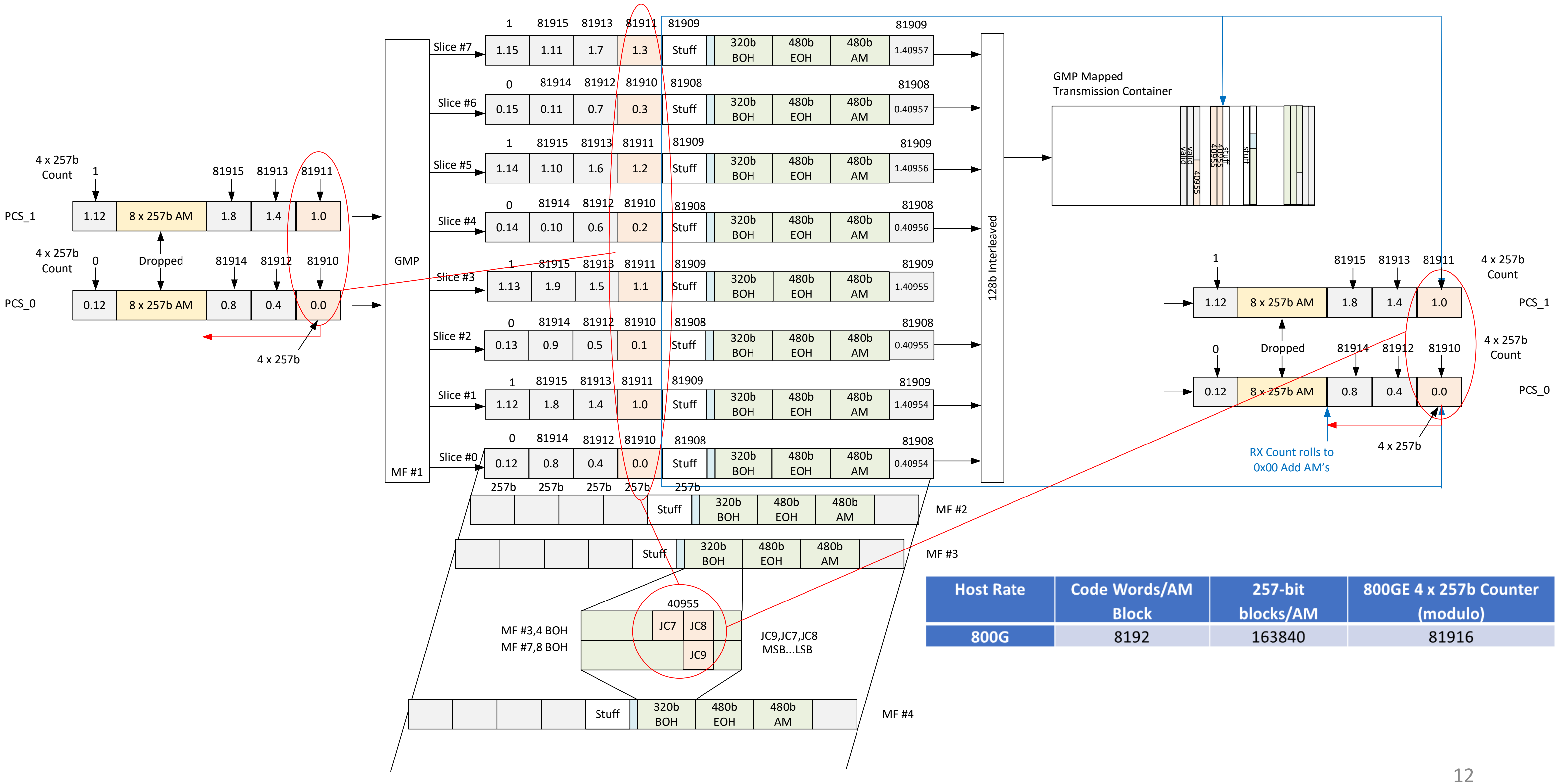
GMP Mapping Cycle	Frame	MFAS	Bytes 1	2	3	4	5	6	7 8 9 10	11 12	13 … 26	27 28	29 … 40	41 45	46 … 302	303 … 559	560 … 5263360
GMP Mapping Cycle	1	xxxxx000	MFAS	STAT	GID	GID	GID RES	IID	MAP	CRC	FCC1	OSMC/RES	RES	pad	stuff	8 x 257b (payload)	(payload)
	2	xxxxx001	MFAS	STAT	AVAIL	RES	JC4	JC1	MAP	CRC	FCC1	RES					
	3	xxxxx010	MFAS	STAT	JC7	JC8	JC5	JC2	MAP	CRC	FCC1	RES					
	4	xxxxx011	MFAS	STAT		JC9	JC6	JC3	MAP	CRC	FCC1	RES					
GMP Mapping Cycle	5	xxxxx100	MFAS	STAT	RES		MSI[x]	PT	MAP	CRC	FCC1	RES		pad	stuff	8 x 257b (payload)	(payload)
	6	xxxxx101	MFAS	STAT			JC4	JC1	MAP	CRC	FCC1	RES					
	7	xxxxx110	MFAS	STAT	JC7	JC8	JC5	JC2	MAP	CRC	FCC1	RES					
	8	xxxxx111	MFAS	STAT	CSTAT	JC9	JC6	JC3	MAP	CRC	FCC1	RES					

▨	Proposed GMP extensions for use in AM synchronization
▨	GMP and Tributary OH fields
□	FlexO-x information Structure OH defined by ITU-T G.709.1 to carry various mapped payloads

AM Synchronization Mechanisms

- JC9-JC7 (MSB-LSB) bytes in the 800GBASE-ER1 transmission container can be used to provide a mechanism (16-counter) to transmit phase information of the Ethernet payload relative to the original FEC Alignment Marker field (AM).
  - Purpose of the counter is to convey the “Start/End” of the 800GBASE-R RS block transfer over the 800GBASE-ER1.
  - For 800GE there are 2 x 400G blocks of 8192 code words (2 x (163840 x 257b blocks)) code words between 16-lane AM marked RS coded blocks (2 x 8 x 257b). Markers are inserted on the same 257b block for both flows (PCS-[0,1]). The two flows are serialized as part of the GMP mapping, with the first block of the serialized flow coming from PCS-0.
  - GMP mapping is done using 4 x 257b words over 8 x 100G slices (each slice is mapped independently but uses the same GMP parameters). The counter is incremented for each valid (non-stuff) 4 x 257b data blocks mapped into the 800GBASE-ER1 transmission container. The counter is reset to zero every AM cycle. Terminal count is therefore 81916 x (4 x 257b) blocks.
  - The value of the counter indicating the 1st 4 x 257b data block from the RS FEC frame that is GMP mapped into JC9,JC7,JC8 (MSB…LSB).
  - On the RX side of the 800GBASE-ER1 link the RX GMP de-mapper initializes its version of an AM sync counter to correspond with the values sent from the receiver. Sync is achieved when the values are consistent over 3 8-frame multi-frames. Sync is lost when the values are inconsistent over 3 8-frame multi-frames.
  - The De-mapper increments for each valid (non-stuff) 4 x 257b clock de-mapped from the 800GBASE-ER1 transmission container. Terminal count for RX counter is also 81916 x (4 x 257b).
  - AM are inserted between the payload block where the counter rolls to 0.

Mapping detail



Host Rate	Code Words/AM Block	257-bit blocks/AM	800GE 4 x 257b Counter (modulo)
800G	8192	163840	81916

Misconfiguration scenarios

- If the transmitter has this optional feature enabled and the receiver does not, the only impact is that the PTP accuracy is not improved (i.e., the receiver is simply not using the information being made available in JC7-JC9)
- If the transmitter has not implemented this feature or if the feature is disabled, the receiver will detect this condition as not available.
  - An all-zero value in two consecutive multi-frames indicates to the receiver this feature is not implemented or disabled.

Mismatches with use of XS

- The use of the XS is not mandatory
- The 800GBASE-ER1 PCS is designed to provide the correct format whether or not an XS is present
- This proposal assumes an XS is present
- If only one end of the link uses an XS, it will be the same as if the end without the XS has disabled the optional feature described in this proposal
- If neither end has an XS, the optional feature cannot be used, but it is also not necessary because AMs are not removed and re-inserted

Summary

- A potential issue was identified related to timestamping accuracy when using an 800GBASE-ER1 PHY with an 800GMII Extender, which is not adequately compensated for by Clause 90.
- This contribution proposes a method to mark the position where the Alignment Markers (AM) are removed in the transmit 800GMII Extender and carry this information over the 800BASE-ER1 link such that the receive 800GMII Extender can re-insert the AMs in the same location from which they were removed.
- This optional mechanism would allow an implementation choice for 800GBASE-ER1 that could comply with the intent of the note in 90.7.2 (referenced in our response to the MPOA liaison letter) and essentially brings 800GBASE-ER1 into line with the other 802.3dj PHYs in this regard.
- Note: the same solution is applicable to the 800GBASE-ER1-20 PHY

Thanks !