:

Implementation considerations about FEC and MII extenders for “tight sync”

September 22nd, 2024

Authors (in alphabetical order)

Antonio Tartaglia[1]
François Fredricx[2]

(1) Ericsson, (2) Nokia

:

1. Executive summary

Precision Time Protocol (PTP) is widely used in the Ethernet world. A technology-agnostic approach to characterization and classification of optical pluggables has been proposed by MOPA in the White Paper on Tight Synchronization [MOPA Tight Sync]. This aims to achieve accurate (or "tight") synchronization when relying on PTP, by allowing the host to compensate for the pluggable's static latency, and to estimate the remaining latency inaccuracy.

The presence in pluggables of digital functionalities manipulating Ethernet frames has introduced new potential sources of time synchronization inaccuracy. In this paper we take a look at implementation specific caveats and possible solution paths for achieving tight sync. In particular, the misalignment of the position of alignment markers between Tx side and Rx side would cause extra inaccuracies. Two possible cases where this can occur are MII extenders and segmented FEC. This is particularly relevant for coherent modules. Over the years, coherent optical interfaces have been optimized for shorter reach and lower power consumption, ultimately becoming available as pluggable optics. Standardization of coherent interfaces at IEEE802.3 is bringing coherent into the Ethernet world.

In this paper we review the associated risks and, referencing the ongoing standardization discourse, we highlight some principles that can be used to manage such risks during device design, with the objective to enable (coherent) pluggable optics to meet the more stringent pluggable accuracy classes defined in [MOPA Tight Sync].

:

Contents

:

2. Coherent pluggables

Digital coherent optics (DCO) with coherent ASIC in the pluggable has become the most successful paradigm, as opposed to analog coherent optics (ACO) where the coherent ASIC remains on the host and the host-to-pluggable interface is analog. Success of DCO in the Ethernet context has been determined by the wish of the end users of the host devices to have "universal" ports, in which different types of pluggables can be used. In this document we will only discuss DCO.

2.1. Functions of a coherent pluggable in an Ethernet context

Figure 1 reports, as an example, a simplified block diagram of a digital coherent pluggable connected to an Ethernet host card. While it illustrates a duplex fiber pluggable, it is also valid for the single fiber cases described in [MOPA Coherent Lite].

The host provides Ethernet standard Reed-Solomon (RS) FEC and connects to the pluggable using an electrical "chip to module" (C2M) Attachment Unit Interface (AUI).

Logic in the pluggable is responsible for handling host FEC information, creating a time division multiplexing (TDM) frame structure, mapping Ethernet frames into the TDM frames, adding a stronger FEC required for coherent optical transmission performance.

Optionally, the host-to-pluggable logic interface can also implement a *media independent interface (MII) extender,* allowing when needed to move the MII close to the physical coding sublayer (PCS) via a physical instantiation ("extender sublayer", XS).[1] It is typically used to remove the existing PCS/FEC coding, exposing back the MII so a new PCS/coding tailored for the optical channel can be added.

[1] See for example clause 152 of IEEE802.3-2022, describing 200GMII/400GMII extenders.

:

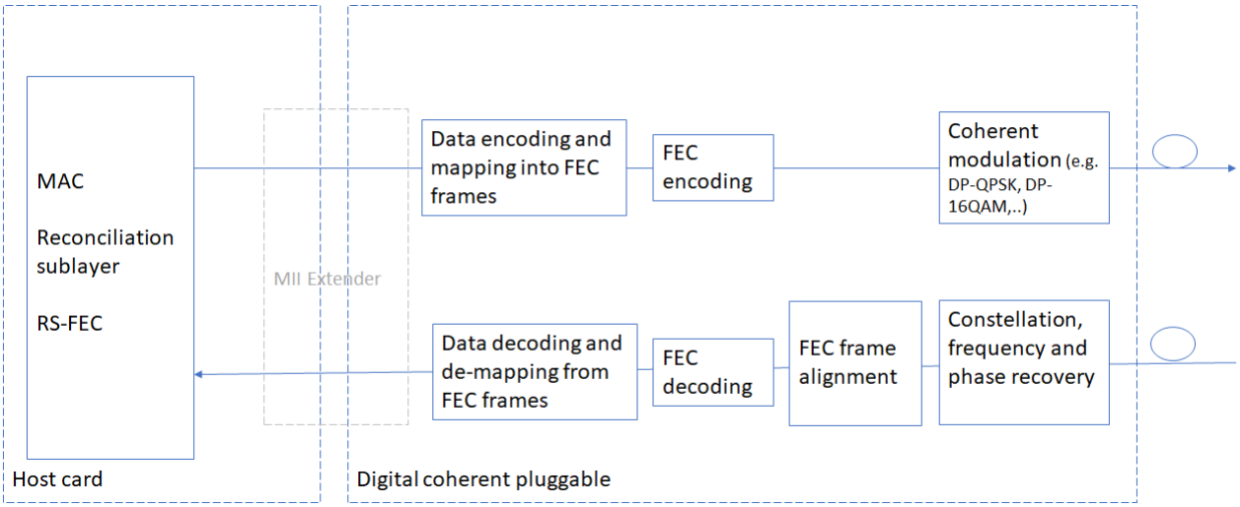

Figure 1: high level description of features in a digital coherent pluggable.

While the list is not exhaustive, the bigger opportunities to introduce time synchronization errors are represented by encoding and mapping of Ethernet frames into the strong FEC frames and by the use of MII extenders.

3. FEC in Ethernet standards

The original formulation of the 100GBASE-R PCS in IEEE802.3 Clause 82 happened when 10Gb/s per lane technology was dominating, and it did not mandate FEC on Ethernet hosts, even if "firecode" FEC has been present in IEEE802.3 Clause 74 for a long time as an optional layer to be placed in between the PCS and PMA sublayers.

With the rise of 25Gb/s per lane technology and above, the increasing transmission challenges brought to the introduction of mandatory Reed-Solomon (RS) FEC described in IEEE802.3 Clause 91.

From the 25Gb/s per lane to the 100Gb/s per lane technology generation the approach has always been to use RS-FEC, integrated in the PCS[2] to cover "end-to-end", both the electrical and optical portions of the physical link, indicating how the FEC correcting power should be shared between the two:

[2] See for example IEEE802.3 Clause 119, describing 400GBASE-R PCS

:

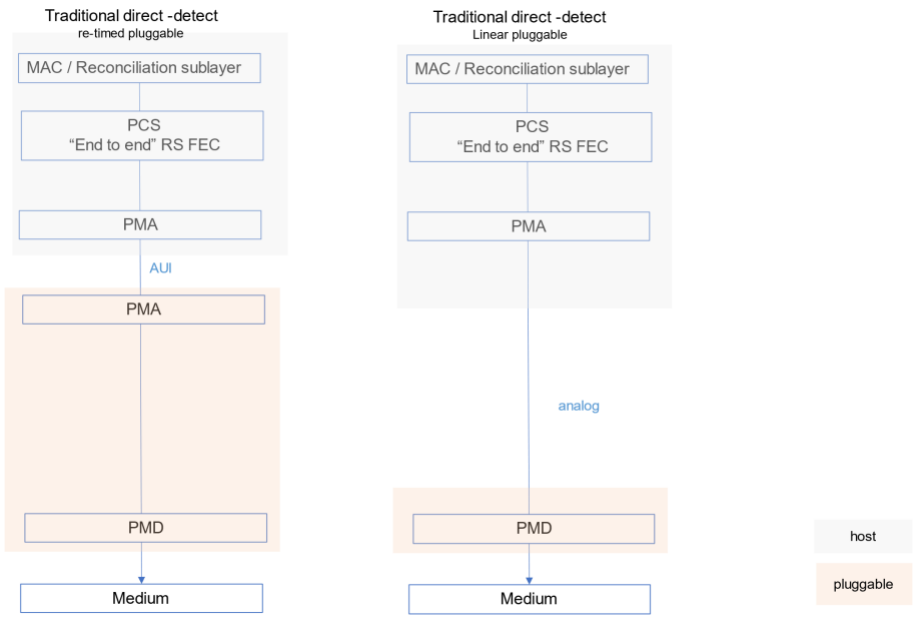

Figure 2: "end-to-end" FEC generated by the host.

The challenges of 200Gb/s per lane technology are bringing the IEEE P802.3dj Task Force to change this approach, adding more powerful FEC schemes. Anyway, this has to be done without requiring extra features to the host port, which could make implementation cumbersome and/or limit the capability of the port to accept the widest possible range of pluggable architectures.

The two schemes adopted, for different PMDs, by 802.3dj are depicted in figure 3 and describe in detail in [802.3dj logic baseline].

:

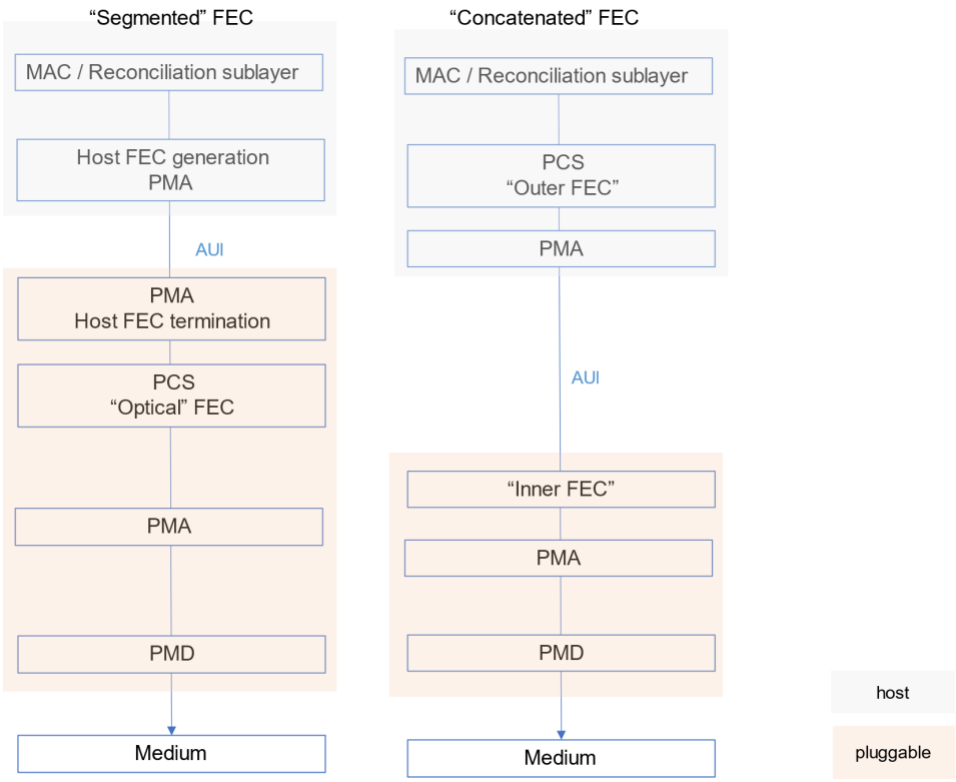

Figure 3: Segmented vs. concatenated FEC schemes.

In a "segmented" scheme, RS-FEC of the host is terminated and a new PCS and the strong FEC is created. RS-FEC coding gain affects the electrical part of the link, while the strong FEC coding gain affects the optical part of the link. This approach can be seen as functionally equivalent to using an "MII Extender". This scheme has been selected by 802.3dj for 800GBASE-ER1 (800G 40km single carrier coherent in C-band) and 800GBASE-ER1-20 (reduced 20km reach variant).

In a "concatenated" scheme, a strong FEC is added to the host FEC, but the two are nested. The RS-FEC from the host, or "outer FEC" (FECo) covers both the electrical and optical part of the link; the stronger FEC that is added, or "inner FEC" (FECi) only provides extra gain for the optical part of the link. This has been selected by 802.3dj for 800GBASE-LR1 (800G single carrier 10km coherent in O-band, with BCH as an inner FEC) and for direct-detect 100 GBaud PAM-4 new interfaces, with a Hamming code as inner FEC (for the 500 m and 2 km interfaces it will be possible to disable the Hamming code and operate in "end-to-end" mode with RS-FEC only).

While in principle nothing prevents implementations that integrate these stronger FECs on the host, in both cases the desire to keep the hosts as simple and universal as possible

:

demands the stronger FEC to be featured by the pluggable optics, as outlined in [FECi bypass].

3.1. Segmented FEC for coherent: is it really new?

There are existing standards covering coherent optical interfaces and, while Ethernet clients only cover a part of the use cases, when dealing with Ethernet all the standards adopt a segmented FEC scheme and leverage a well-known mechanism standardized by OTN G.709 to accommodate constant bit rate clients, such as the generic mapping procedure (GMP) (see figure 4)

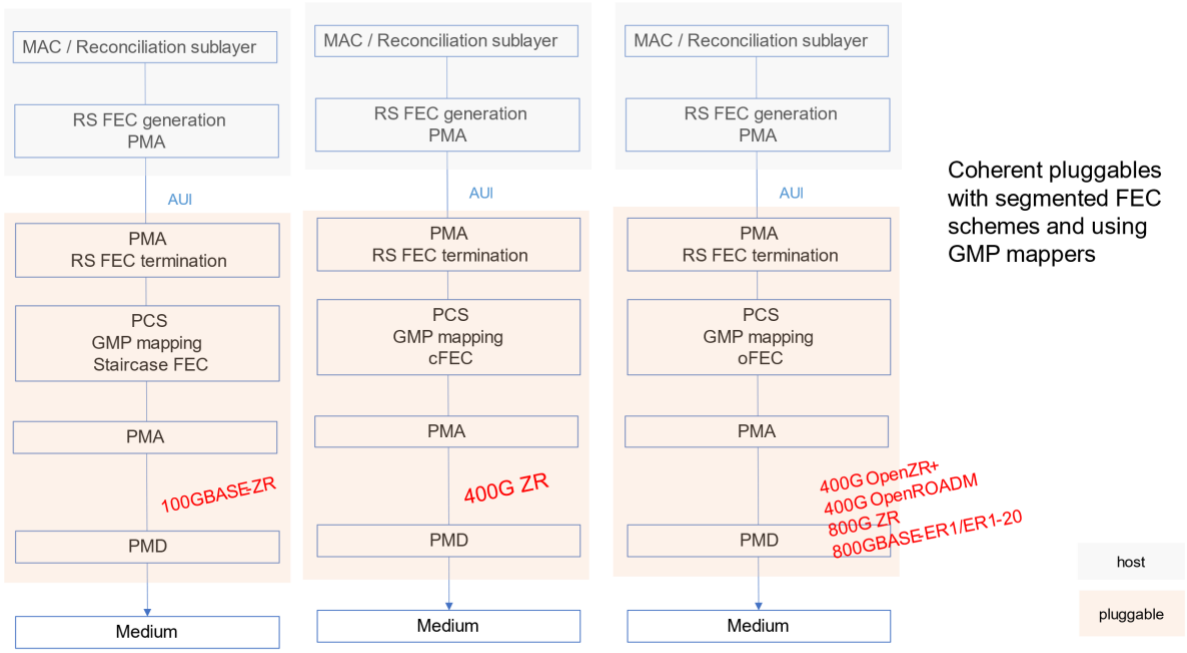

Figure 4: simplified view of 100GBASE-ZR, OIF's 400G ZR and coherent standards using oFEC.

The exception to this rule will be the new 800GBASE-LR1 being standardized by 802.3dj, mirroring OIF's 800G LR, which is based on a concatenated FEC scheme and uses a direct mapping of Ethernet into FEC frames, generating them synchronously with the client Ethernet clock.

:

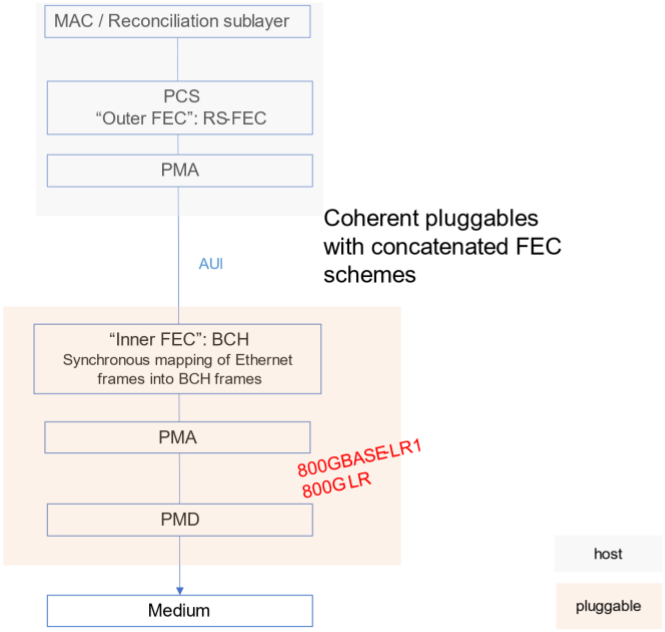

Figure 5: Simplified view of 800BASE-LR1.

:

4. General concerns on MII Extenders

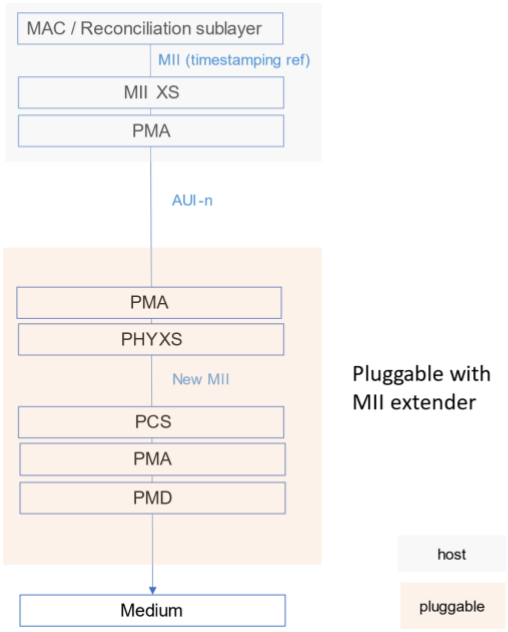

Figure 6: simplified view of MII extenders.

MII extenders are a popular solution in coherent pluggables, as they allow to strip off the features of a port designed to work with end-to-end RS-FEC, expose again the media independent interface (MII) and add the new PCS and FEC demanded by coherent transmission, but it is an option that also direct-detect interfaces can use.

IEEE802.3 specifies that the PTP timestamp must report the time at the Generic Reconciliation Sublayer (gRS), adjusted to account for the delay through the PHY up to the media dependent interface (MDI) called "*path data delay*". 802.3cx-2023 explains how to manage cyclic delay variations through the PHY and uses this to assume that the *path data delay* is constant in time. In 802.3cx-2023, the PCS communicates with the Generic Reconciliation Sublayer in case of Alignment Marker position inserted (Tx) or deleted (Rx).

In case of MII Extenders, the PCS may not be in the same device as the reconciliation sublayer, and this communication is impossible. Moreover, if alignment markers received from the host are eliminated and new alignment markers are generated in Tx, since alignment markers define the position in time of the timestamping packet the effective latency will vary randomly, contributing to time error exactly like a propagation delay

asymmetry. In the case of a 800G MII extender, the resulting timestamping error has been shown to be in the order of 5ns worst case [MII_extenders].

4.1. A practical example: "single 100G lambda" QSFP28 and possible implementations

100G "single lambda", 50Gbaud PAM-4 duplex fiber optical interfaces are defined by IEEE802.3 Clause 140, and new single fiber BiDi interfaces are being defined by IEEE P802.3dk.

In both cases, RS-FEC as per Clause 91 is mandatory and it prescribes use of RS(528,514), also known as "KR FEC", with 25Gb/s per lane interfaces and RS(544,514), also known as "KP FEC" with 50Gb/s and 100Gb/s per lane interfaces, such as for example 100GBASE-DR, 100GBASE-FR1, or 100GBASE-LR1 of Clause 140.

An example is the case of a host designed with 25Gb/s electrical serdes technology that needs to use QSFP28 pluggables. The electrical interface to the pluggable is a C2M (Chip to Module) AUI (Attachment Unit Interface) and the possible implementations are in figure 7.

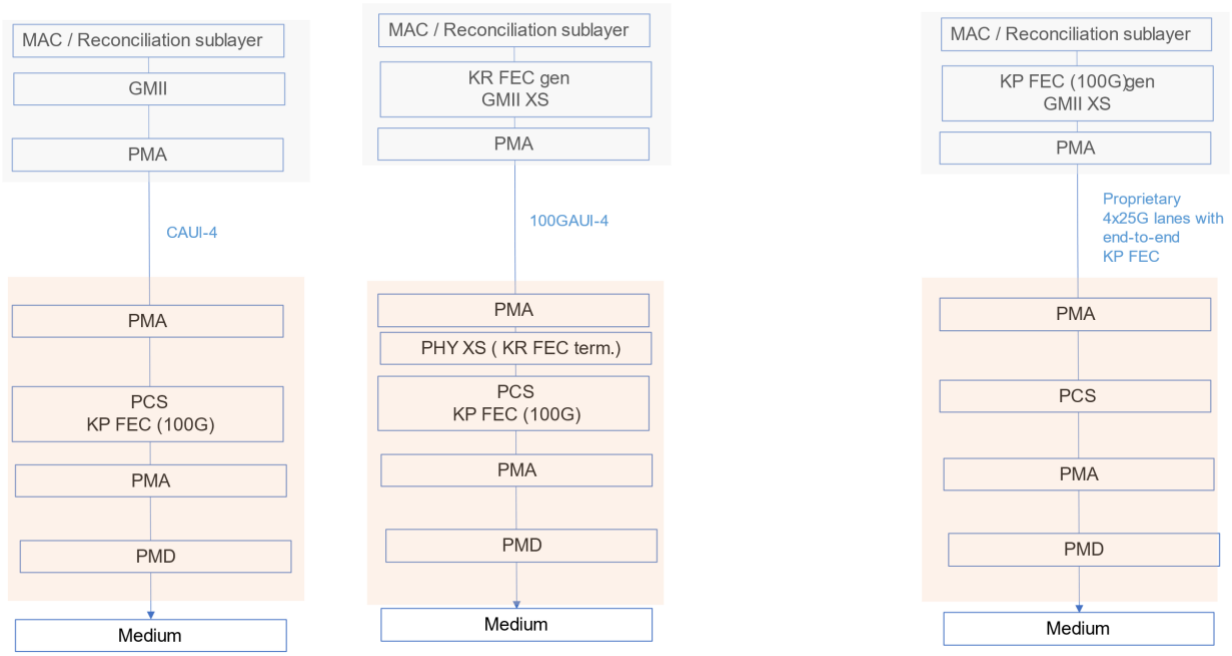

Figure 7 (a, b, c): Simplified view of QSFP28 "single lambda" pluggables.

:

(Fig 7a) The host could have been designed around old standards like 100GBASE-LR4, and therefore not supporting RS-FEC at all: in this case, the electrical CAUI-4 interface needs to work without FEC and the “KP” FEC is added by an integrated circuit in the pluggable.

(Fig 7c) The host could have been designed with 100G “single lambda” operation in mind, incorporating “KP” FEC as per Clause 91 directly in the host, and “KP” FEC would be working end-to-end on the electrical and optical part of the link, even if this operation mode is not contemplated by IEEE802.3.

(Fig 7b) More likely, the host will integrate “KR” FEC to be able to use modern 4x25G lanes optical standards like 100GBASE-SR4. Therefore, it will be necessary to terminate the “KR” FEC coming from the host, only covering the electrical interface, and add “KP” FEC in the pluggable only covering the optical link. In other words, a solution resembling MII extenders and segmented FEC schemes, with all the caveats associated with the unknown time relationship between the KR alignment markers removed and the new KP alignment markers added in the pluggable.

5. Recommended practices and principles for “tight sync” support of coherent pluggables

IEEE802.3cx-2023 has updated the original Clause 90 with prescriptions and considerations on how to make sure the timestamping inaccuracy can be minimized, allowing to meet stringent accuracy targets.; anyway, 802.3 does not tell extensively which features are on the host and which are supposed to be in the optical pluggable, and system partitioning has the potential to break the chain of propagation delay management described by 802.3cx like in the case of use of MII Extenders.

When designing coherent pluggables that are meant to be used in PTP networks with stringent time synchronization requirements, the first recommendation is to be aware of the requirement and to take all the design choices that allow to best support it – ideally, guaranteeing performance by design.

Points of attention to keep under control and exclusively driven by the implementation are pipeline delays, FIFO delays, interconnect delays, serdes delays, clock domain crossings. This normally brings to constant delay values when in operation, but the values change at startup.

The embedded capability of a device to exactly measure delays, report them, possibly equalize them, is another good to have feature.

:

802.3cx-2023 also details principles to calculate and allocate cyclical, bit-dependent delays on the host, so that the Tx and Rx propagation delays continue to be cyclical in nature but their sum is a constant value, as summarized in [802.3cx_802.3dj] slide #7. The same approach, abstracted in figure 8, can be used for any digital part inside the pluggable module. The pluggable would still have to report fixed propagation delays towards the host, allowing for accurate calculation of the real position of the timestamping packet in time. A way of doing that is, of course, using the EEPROM register features described in [MOPA Tight Sync].

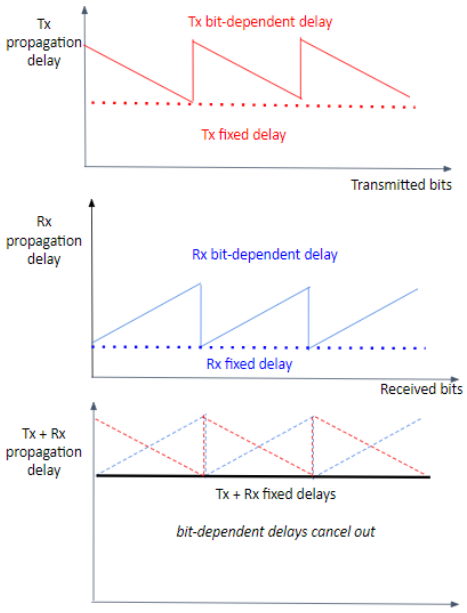

Figure 8: Nulling of symmetric cyclical, bit-dependent propagation delays.

5.1. Removing the alignment marker indetermination due to segmented FEC in coherent pluggables

During discussions of 802.3dj, a proposal has been made [SegmentedFEC and GMP] to fix the issue introduced by segmented FEC in 800GBASE-ER1 and ER1-20. The proposal restores the determinism of timestamping position required by 802.3cx-2023 by taking note of the alignment marker position in the ingress host Ethernet frames and transmitting them to the remote egress host, that can use the information to recreate alignment markers in the right position when generating egress Ethernet frames.

:

The proposed mechanism leverages the features of Generic Mapping Procedure (GMP), specifically uses some of the JC (justification counter) bytes in the GMP overhead to deliver alignment marker information to the other host.

Since it only uses GMP features, the proposed mechanism can work in principle with all coherent implementations based on GMP to map Ethernet into FEC frames, whatever the specific FEC code: in other words, the same fix is potentially applicable also to IEEE802.3ct-2021 100GBASE-ZR, OIF's 400G ZR, OpenZR+ and OpenROADM 400G and of course IEEE P802.3dj's 800GBASE-ER1/ER1-20 and OIF's 800G ZR. The only known coherent interface that cannot use it is 800GBASE-LR1, but it does not need it as it is based on synchronous mapping of Ethernet into BCH FEC frames and on a concatenated FEC scheme that does not touch the original alignment markers.

5.2. Pluggables without GMP

In principle, for pluggables integrating digital parts which manipulate Ethernet frames, changes in Alignment Marker could be sent to the other host using Ethernet-specific features. While it would make the solution fairly independent from the specific pluggable form factor, it calls for further studies.

:

6. References

[MOPA Tight Sync] Technical paper on Optical pluggable performance for tight synchronization v1.0

[MOPA Coherent Lite]
https://mopa-alliance.org/wp-content/uploads/2024/03/MOPA_Coherent_Lite_paper-v1.0.pdf

[IEEE802.3-2022] 802.3-2022 - IEEE Standard for Ethernet

[MII_extenders] Timestamping across an 800GE MII Extender (ieee802.org)

[SegmentedFEC] Timestamp consideration with segmented FEC (ieee802.org)

[802.3cx_802.3dj] Time Synchronization Clarifications for 802.3dj (ieee802.org)

[800GBASE-LR1 baseline]
Logic Baseline proposal for 800G single-wavelength coherent PHY with concatenated FEC (ieee802.org)

[802.3dj logic baseline]
https://www.ieee802.org/3/df/public/22_05/22_0517/gustlin_3df_01a_220517.pdf

[FECi bypass] FEC Inner Code Bypass Options for 200G/L IMDD Optics (ieee.org)

[SegmentedFEC and GMP] https://www.ieee802.org/3/dj/public/24_05/sluyski_3dj_01a_2405.pdf