Proposal for changes in the start-up protocol to enable segment-by-segment training – Part 1: Architecture

Adee Ran, Cisco

Kent Lusted, Intel

Leon Bruckman, Huawei

Anil Mehta, Broadcom

Mike Dudek, Marvell

Ali Ghiasi, Ghiasi Quantum/Marvell

Matt Brown, Alphawave Semi

Zvi Rechtman, NVIDIA

Intro

- In-band training for AUIs as well as links between electrical PMDs was adopted in the January 2024 meeting.
- Further discussions took place in the electrical ad hoc:
  - Challenges with training on segmented links were reviewed in ran_3dj_elec_02_240208. The idea of adding a “Ready to send” (RTS) bit to the protocol, which is propagated from a PCS through segments that are ready to go to data mode, was presented as a solution.
  - Details of the proposed solution were laid out in ran_3dj_elec_01a_240229.
- We present a start-up function for PMAs defined in clause 174 that:
  - Is based on the existing, well-known training protocol of clause 136
  - Creates a well-defined segment-by-segment start-up procedure
  - Enables training on AUIs and PMDs, independently.
- The second part of this presentation includes details for implementation of this proposal in a draft standard.

High level overview – start-up function

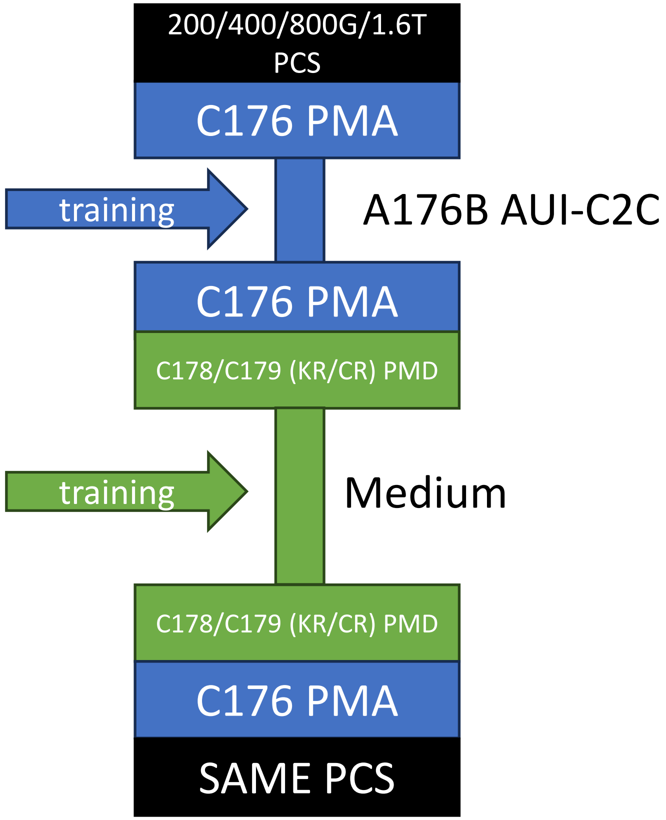
- We propose adding a (mandatory) **PMA start-up function** to the new PMAs defined in P802.3dj (Clause 176). This function may be defined in Annex 176A.
  - The start-up function is separate for each interface of the PMA (see next slides).
- The PMA start-up function includes a **training protocol**, which is required **conditionally** (depending on the PMA's interfaces – some interfaces may not support training).
  - Where it is required (e.g., electrical interfaces), implementation of the training protocol is mandatory, but it can be disabled through management.
- If the sublayer below the PMA is a PMD, the start-up function interacts with the PMD through the PMD signal indication and the PMD transmitter disable functions.
- The PMA start-up function is the same for all PMAs, both within a PHY and within an xMII extender.

High level overview – training

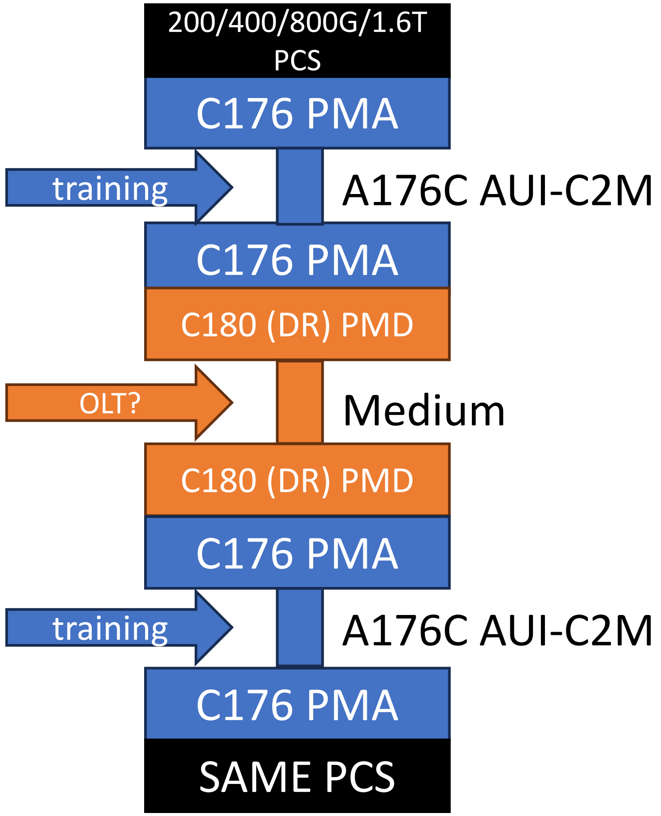
- For electrical PMDs, training becomes part of the PMA start-up function instead of a PMD function.
    - The details of the training protocol, e.g. number of Tx equalizer taps, can be different depending on the interface – PMD or AUI.
- An optical PMD may have a training function in the sublayer above it that is not a PMA, e.g. inner FEC. It can be the same function, or a different one.
    - The details of the training protocol, e.g. fields and bits, may be different from electrical PMDs.
- **The subsequent slides refer only to the PMA and its interfaces.**

Examples of segmented links with P802.3dj sublayers

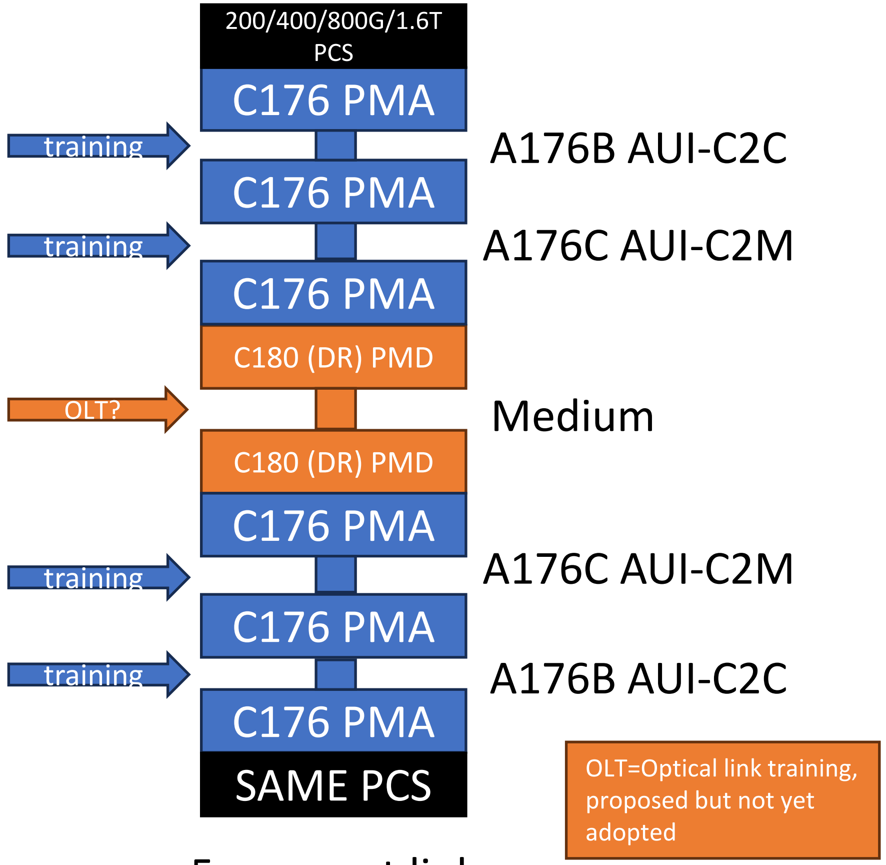
(200G/lane everywhere)



2-segment link



3-segment link

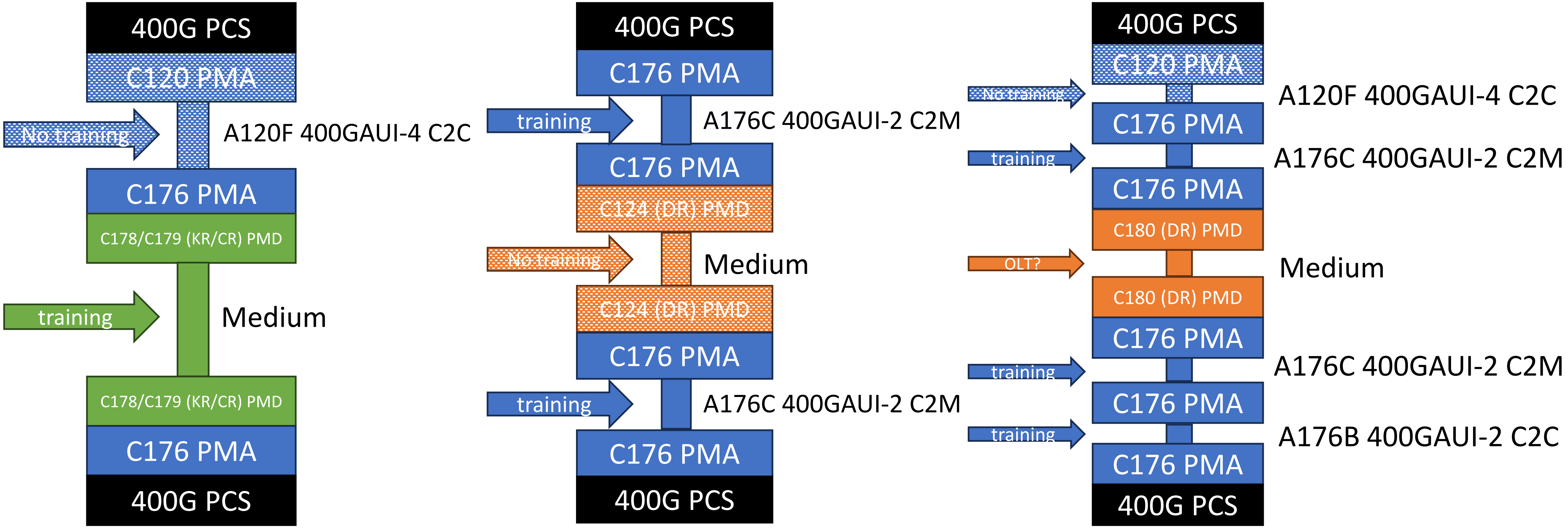


OLT=Optical link training, proposed but not yet adopted

5-segment link

Examples of segmented links with pre-P802.3dj sublayers

(100G/lane on some segments)



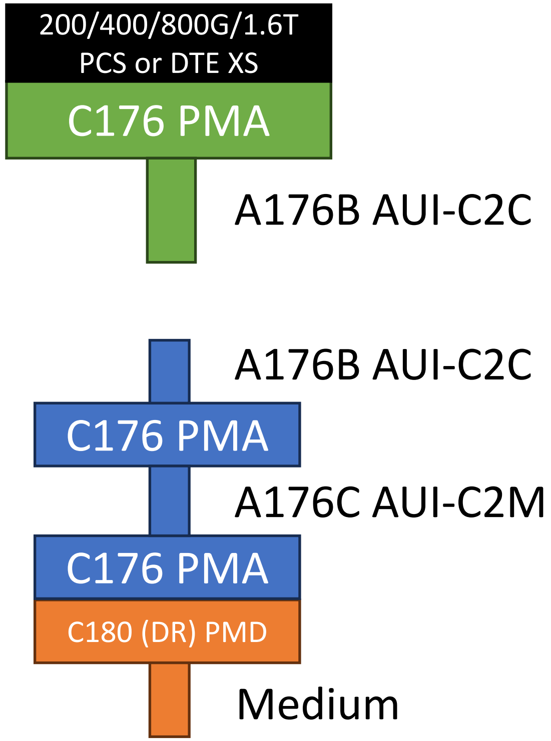
2-segment link

3-segment link

5-segment link

PMA interfaces

- A PMA adjacent to a PCS or a DTE XS has one interface that can potentially use training. The interface with the PCS or XS is never exposed.
- Other PMAs (referred to as retimers) have two interfaces that can potentially use training.
  - Each interface can be an AUI or a PMD.
  - Training may be performed in an intermediate sublayer such as inner FEC.
- A PCS-to-PCS link without retimers has one segment. Retimers separate the link into segments.

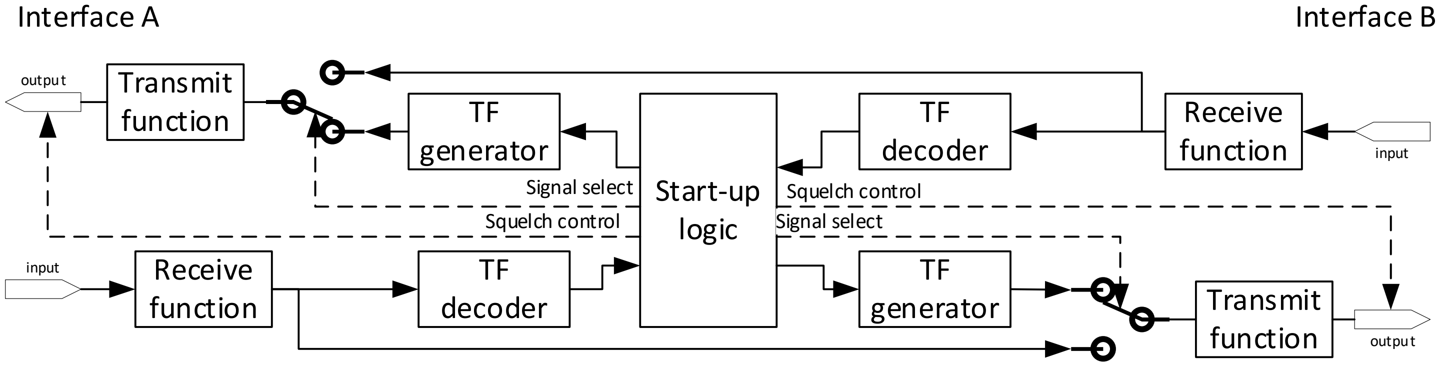


Start-up concept

- In links with one or more segments:
  - local_RTS propagates in the transmit direction from the local PCS toward the remote PCS.
  - remote_RTS propagates similarly and independently in the receive direction from the remote PCS.
  - RTS is propagated only across a segment that is ready to send data.
  - When a PMA interface both **sends** and **receives** RTS in both directions, it means all the segments are ready and it can switch to data mode. When all PMAs are in data mode, the link between the PCSs is established.
- RTS is communicated between two Clause 176 PMAs during start-up using one of two methods:
  A. Training frames (using a protocol defined in Annex 176A), if available and enabled
  B. The **transmit disable** and **PMA signal detect** (receiver ready) functions, otherwise
- Interaction with earlier PMAs (e.g., those defined in Clause 120 or Clause 173), and across optical links that don’t have training, is performed only using the second method.

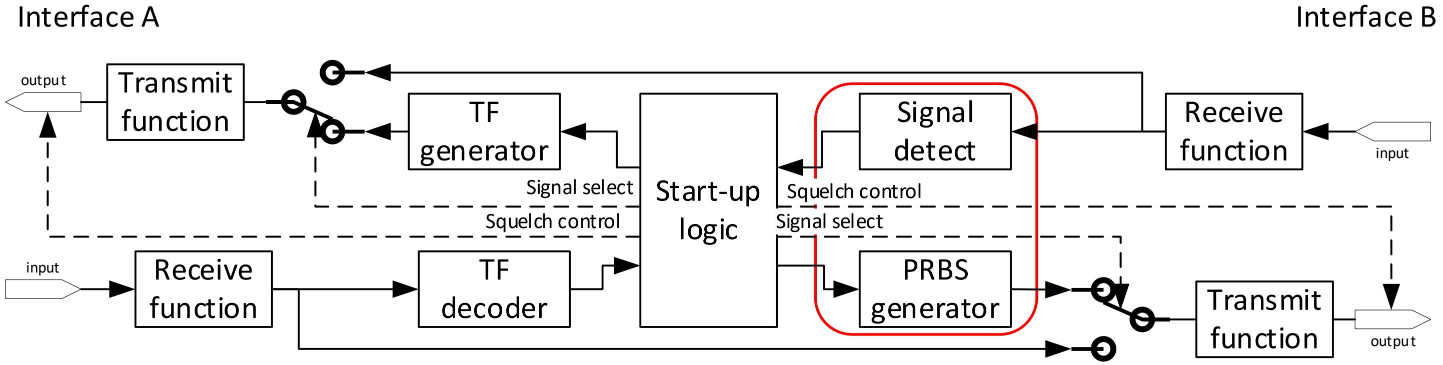
Further illustration of this concept can be found in the November 2023 OIF contribution https://www.oiforum.com/get/53958.

Block diagram:
A retimer that has training on both interfaces (A and B)



Represents start-up functions of both interfaces and their interaction.

Block diagram:
A retimer that has training only on one interface (A)



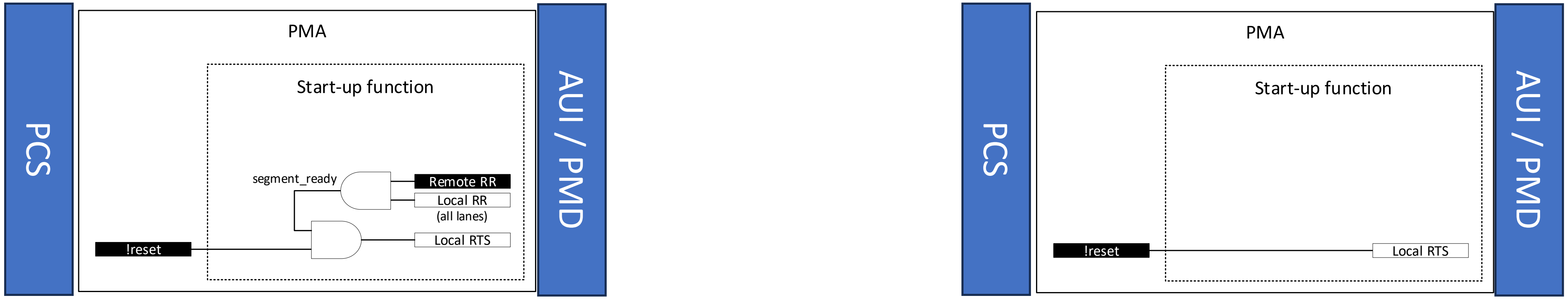
Represents start-up functions of both interfaces and their interaction.

Logical view of the start-up function and RTS

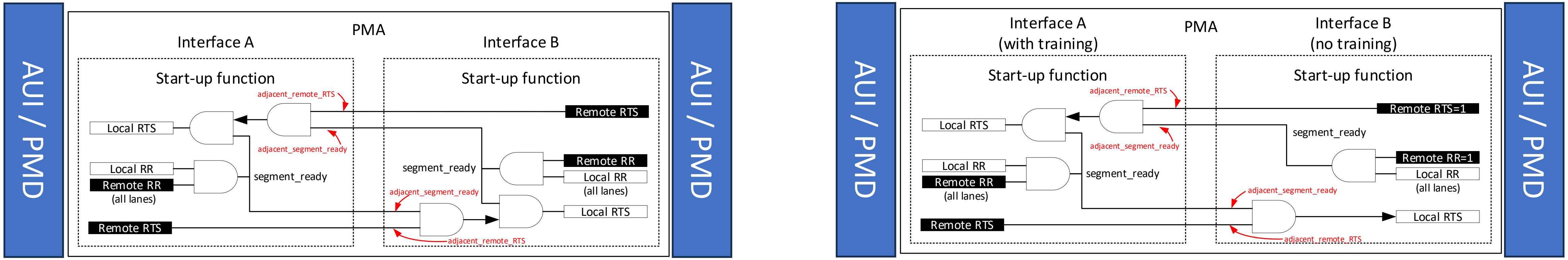
Excluding clocking and timing

Updated version of ran_3dj_elec_01a_240229 slide 16

PMA attached to a PCS



Retimer



Training enabled

Training enabled

Training enabled

Training disabled

Backup

Training in retimers (including modules)

- Training protocol transmission starts with local clock and transitions to recovered clock when available.
- Local_RTS is set to true on the egress interface only after the transmit clock is derived from the local PCS clock; on the ingress interface, only after the transmit clock is derived from the remote PCS clock
  - This is a specified in the **RTS state diagram**
  - The transition between clock sources occurs while sending local_RTS=false. This ensures that the whole link is running with the correct clocks before retimers go to “mission mode”.
- Propagation of RTS across a retimer:
  - Exchanging the RTS between the two PMA interfaces (i.e., copying remote_RTS to adjacent_remote_RTS) may be implemented in various ways. It may be done either autonomously inside the PMA, or using external management (e.g., CMIS).
  - When remote_RTS=1 is received on an interface that sends local_RTS=0, the propagation to the other interface does not need any timing requirements.
  - However, when remote_RTS=1 is received in on an interface that sends local_RTS=1, **it should be propagated to the other interface within a reasonable time** (e.g., 100 ms) to prevent unnecessary delay in bringing up the link (other retimers may have already transitioned to data mode).

xMII Extenders

- Training on a segment within an extender follows the same process, except that RTS is communicated to the PHY XS using its IS_SIGNAL.indication and IS_SIGNAL.request primitives.
- PMAs within an extender can train before or in parallel to the main link, and training signaling will continue until the main link is ready.
  - Same behavior as PMAs within the PHY.