Proposed Optical Link Training “OLT”

Ali Ghiasi – Ghiasi Quantum/Marvell
Vasu Parthasarathy – Broadcom
Roberto Rodes – Coherent
Chris Cole – Quintessent
Mike Dudek – Marvell
Whay Lee – Marvell

IEEE 802.3dj Task Force
Interim Meeting
Annapolis, Maryland

May 13, 2024

List of Contributors

- Matt Brown – Alphawave
- Adam Healey – Broadcom
- Vasu Parthasarathy – Broadcom
- Adee Ran – Cisco
- Roberto Rodes – Coherent
- Ulf Parkholm - Ericsson
- Ali Ghiasi – Ghiasi Quantum/ Marvell
- Kent Lusted – Intel
- Mike Dudek – Marvell
- Whay Lee – Marvell
- Chris Cole – Quintessent
- Frank Chang – Source Photonics

List of Supporters

- Matt Brown – Alphawave
- David Cassan – Alphawave
- Henry Wong – Alphawave
- Cedric Lam – Google
- Xiang Zhou – Google
- Arash Farhoodfar – Marvell
- Lenin Patra – Marvell
- Drew Guckenberger – Maxlinear
- Sridhar Ramesh – Maxlinear
- Zvi Rechtman – Nvidia.

Overview

- End user's feedback
- Response to some of the questions from Jan-24 interim
- Why optical link up is challenging
- Improving optical link up with DME and RTS
- Benefit of optical link training (OLT)
- OLT PMD requirement
- Proposed OLT
- OLT control and status pages
- OLT flow
- 200G MMF
- Summary.

End Users Feedback on OAN/OLT

- **Traditional DC/Enterprise**
  - With 11 IMDD PMDs with 5 based on FECo and 6 based on FECi OAN can facilitate deployment and configuration
  - Dual mode module can be configured based on remote module capability
  - OAN enables zero-touch in fronthaul networks parkholm_3dj_02_2311
- **Hyperscale feedback**
  - Would not use OAN as deployment is based on build-replicate
  - Reliable link up in timely manner with OLT and RTS
    - Based on past experience reliable optical link may take much longer than 802.3ck AN+LT link up timers (Make it clear)
    - Expecting reliable optical link up at 200G will be more challenging
  - OMA control with LT is most interesting
    - Mitigate overload
    - Mitigate compression
    - Energy efficiency
- **OIF EEI (Energy Efficient Interfaces) project has interest in collaborating and potentially leveraging OLT**
  - See liaison received from OIF OIF_liaison_IEEE_EEI_Project_21Jan24_Redacted.

Respond to Questions from Jan/March-2024

- **What are the key benefit of OLT**
  - Provide coordinated segment by segment reliable link bring up with RTS
  - Pre-coder control
  - Reliable link up starting with known NRZ pattern then transitioning to PAM4 for training
- **Link training should be end-end**
  - OLT build on the ran_3dj_elec_01_240229 multi-segments training for AUI, CR, KR links
  - A key element of Ran proposal for reliable-predictable link up is sending RTS (Ready to Send) across multiple link segments
  - OLT will be compatible with Ran proposal and will pass the RTS across the optical link for more robust-predictable end-end link bring up and training
- **Compatibility of OLT with TDECQ**
  - The current OLT proposal defines only pre-coder control and passing the RTS, so doesn't affect TDECQ test methodology
  - DJ or future IEEE task force may enable additional OLT features that may require some adjustment to TDECQ test methodology but that is not part of this baseline proposal.

Why Optical Link Up Challenging

- **Today's optical SerDes must calibrate and adapt to blind data without even knowing if there is good data**
- **Link training facilitates multi-segment coordinated bring up resulting in more reliable and timelier link up than just relying on blind link up**
- **Before any link up and timing recovery**
  - 1st the SAR(Successive Approximation Register) ADCs must be calibrated
    - Offset calibration between all 16/32/64 sub-SAR ADCs
    - Gain calibration between all 16/32/64 sub-SAR ADCs
    - Timing skew calibration between all 16/32/64 sub-SAR ADCs
    - AGC adjust input signal to fully utilize ADC resolution
  - CDR loop, Mueller-Muller or other type of clock recovery
  - DSP equalizer and adaptation
- **Today's optical DSP SerDes operating with blind start up may require several reset by the module uC while waiting and in hope of getting better SNR**
  - OLT will improve ADC calibration, improve timing recovery, and adaptation!

Optical Link Up Process Outdated

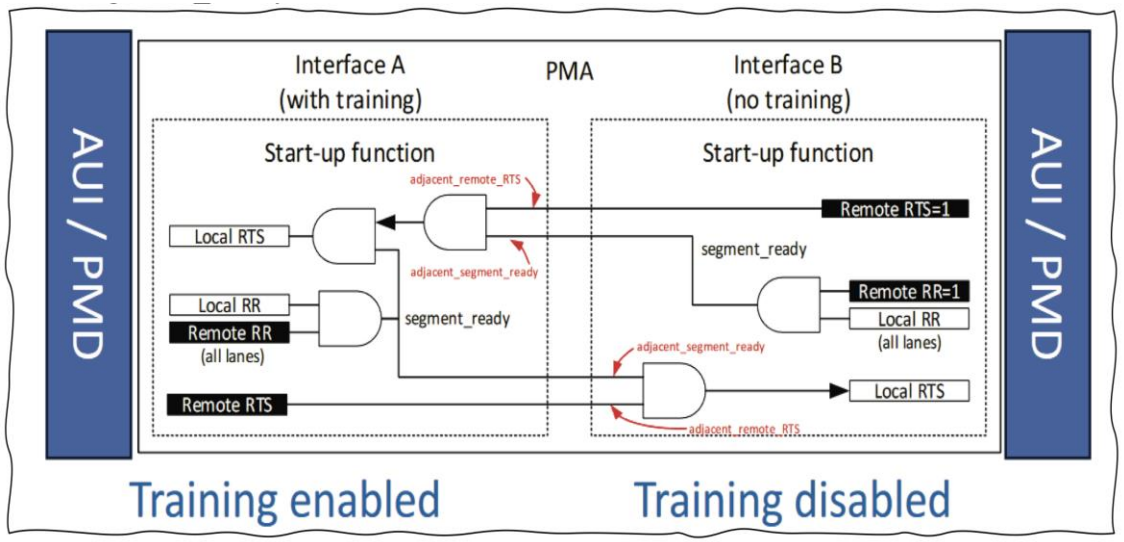
- **Current IMDD link up goes back to early days of FC**
  - Early optics were unretimed without any equalization on AUIs/PPI or optics with PCS relying on SD (Signal Detect)
- **What is making the current optical link up cumbersome and complex**
  - CMIS initializes module and data paths
  - CMIS configures AUI SI parameters, then module enables it's optical TX
  - SD and CDR locks are used as indicator of good optical signal
  - As the module CDR/DSP is calibrating its ADC, recovering, equalizing, and propagating data to host there is no guarantee good data is being transmitted to the optical receiver
  - Downstream electrical link may simultaneously be trying to adapt/equalize to "not guaranteed good data" and may end up adapting to invalid data
  - Optical links are multi-segments and complex, and inherently unpredictable to host-module and PMD-PMD interactions
  - By relying on SD/CDR lock to unsquelch module out may result sending invalid data to the host that require host keep resetting the module Egress DSP and hoping for the best!

Improving Optical Link Up Process with OLT

- **OLT start up DME facility with known NRZ PRBS then switching to PAM4 improve optical receiver/DSP calibration and adaptation**
- **OLT passing RTS, ran_3dj_elec_01_240229, will improve reliable link up on both AUIs segments and optical link**
  - Optical SD/CDR lock are not always a reliable indicator of signal goodness and may falsely send RTS onto local AUI segment
  - OLT DME facility starting with known NRZ PRBS then switching to PAM4 offers improved optical SerDes receiver start up
  - But real benefit of OLT is RTS propagation from AUI segments across optical link
- **Reliable and predictive optical link up is required for 200G optics considering 100G optics pain points**
  - OLT with RTS will improve 200G optics to have more reliable and consistent link up in a timely manner!

Segment Ready and RTS Propagation

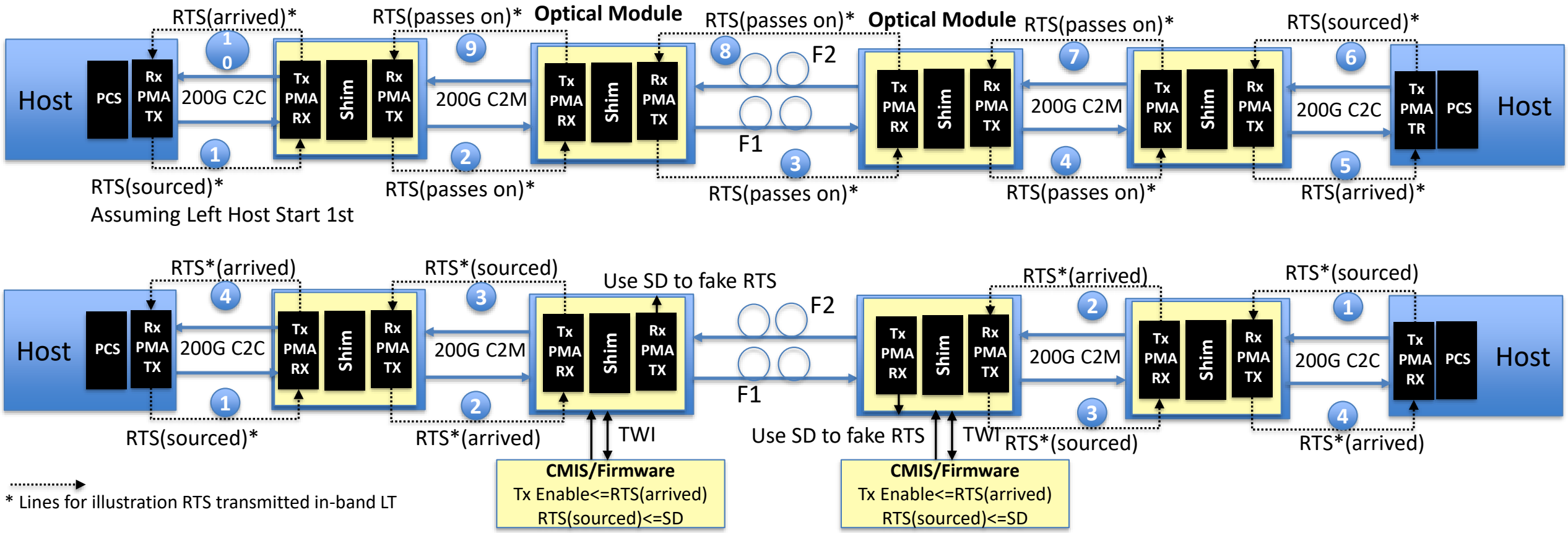
- **If training is disabled or next segment (optical PMD) doesn't support training then Remote RTS=1 and RR=1 permanently with local AUI not taking advantage of propagated RTS**
  - Local AUI instead of relaying on remote RTS instead relies on PMA unsquelching (Local RR) driven by the optical PMD OMA level or CDR lock without knowledge of Remote RR
    - OMA/CDR lock indicator are not a reliable link up indicator
  - Even if OMA/CDR lock were reliable link up indicator there is no guarantee when the remote PMD/PMA will unsquelch
  - Multi-segment link with optical segment are most problematic regarding reliable and predictable link up and will benefit the most from RTS.



See [24_0404/ran_3dj_elec_01_240404]

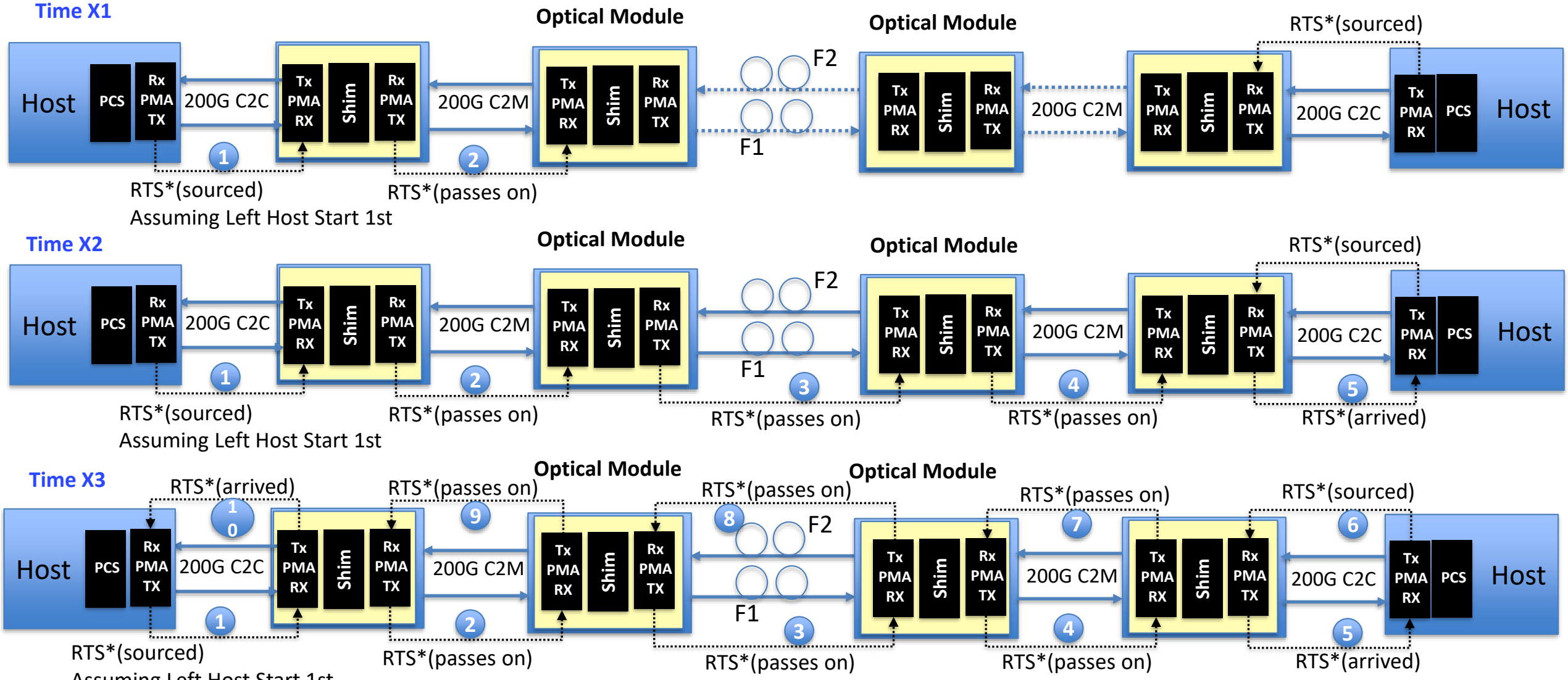
RTS Propagation

- **RTS propagated across optical link with OLT offers reliable-consistent link up in timely manner**
  - RTS terminated in the optical module doesn't address optical link up challenges
  - Generating RTS based on not always reliable SD/CDR lock can exasperate optical link up!



* Lines for illustration RTS transmitted in-band LT

Example of RTS Propagation



Why pre-coder on/off necessary for DJ optical PMDs

- **Largely due to severe BW limitation from TIA-VGA-ADC cascaded bandwidths requiring the DFE/MLSE to work very hard**
  - TDECQ is measured with BW=Baudrate (53.125 GHz FECo/56.72 GHz FECi) and pre-emphasis adjusted on that basis
  - Initial 200G optics front end TIA-VGA-ADC BW expect to have an aggregate BW ~34 GHz but over time aggregate BW expected to increase ~50 GHz
    - On top these BW there will be ~±15% BW variation due to components variations (early on some low BW TIA-DSP may fail sensitivity due to low aggregate BW)
- **Some receiver may have unacceptable burst errors without a pre-coder enabled**
  - Pre-coder may be required on any link from DR to LR4
  - Enabling pre-coder on every optical link requires optical PMDs to have better pre-FEC BER
  - Controlling pre-coder on/off with the method of CL136/162 LT is the best proven method.

Optical OLT PMD Requirements

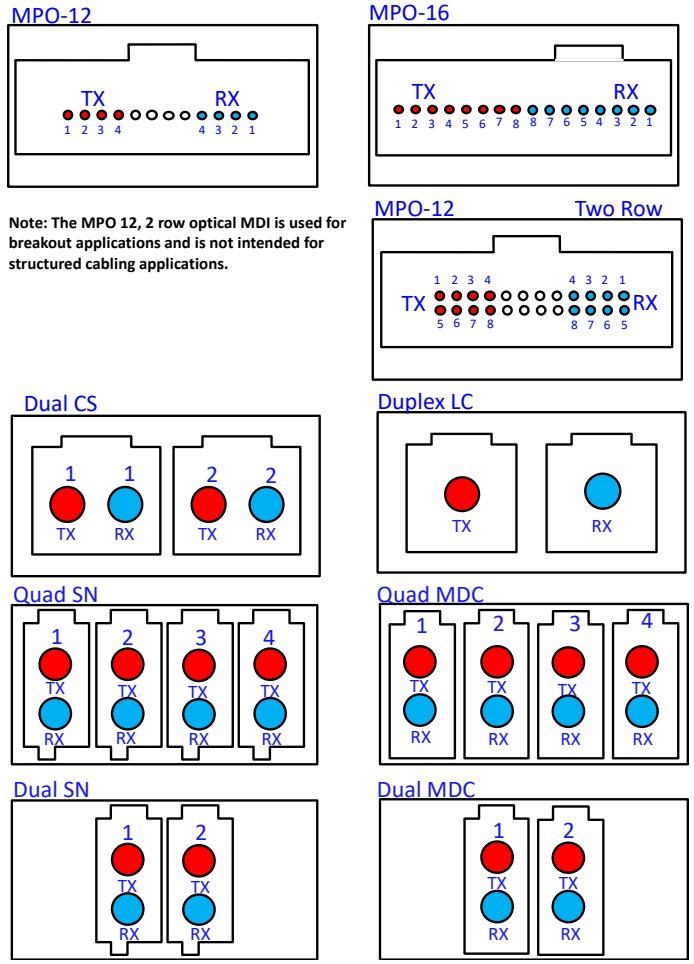
- **Transmit and receive must be grouped in duplex pairs PMDs to support optical loop back and breakout applications**
  - Current definition of optical PMDs
    - TX TX TX TX - RX RX RX RX
    - Was defined for connivence of routing any TX to RX lanes
  - We would need to define the optical PMDs as following
    - TX1 TX2 TX3 TX4 - RX4 RX3 RX2 RX1
    - Support breakout and optical loopback
- **In IEEE we have not grouped the TX/RX into duplex pairs but actual products are based on duplex pair grouping**
  - QSFP-DD MSA and OSFP MSA specify TX/RX lane # on the MDI
  - CMIS associate PMA lanes to optical PMD lanes – otherwise you wouldn't know which transmitter was turned on
- **Structure cable plants compliant to ANSI/TIA-568.3-E 2022 preserve duplex pairs and lane grouping.**

Both CMIS and Optical Modules Enforce Tx/Rx Pairs

- **CMIS data-path pairs electrical Tx/Rx lanes to optical Tx/Rx lanes**
  - Table and figure below from QSFP-DD MSA Rev. 7.0 illustrates the concept.

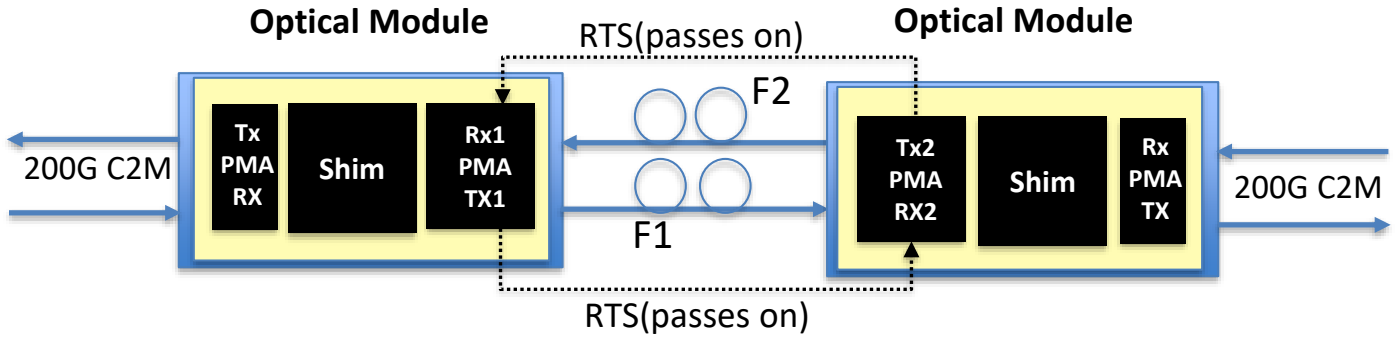
Electrical data input/output	Optical port mapping (see Figure 15)				
	Duplex LC, CS, SN, or MDC	MPO-12, Dual (CS, SN, MDC, Duplex LC, or MPO-12)	MPO-12, Quad (SN or MDC)	MPO-12 (two row), MPO-16, or Dual MPO-12	MPO-12, SN, MDC (BiDi)
	1 TX fiber 1 RX fiber [1]	2 TX fibers 2 RX fibers [1]	4 TX fibers 4 RX fibers [1]	8 TX fibers 8 RX fibers [1,3]	8 Tx (Rx) fibers [2,3]
Tx1	TX-1	TX-1	TX-1	TX-1	TR1
Tx2				TX-2	RT1
Tx3			TX-2	TX-3	TR2
Tx4				TX-4	RT2
Tx5		TX-2	TX-3	TX-5	TR3
Tx6				TX-6	RT3
Tx7			TX-4	TX-7	TR4
Tx8				TX-8	RT4
Rx1	RX-1	RX-1	RX-1	RX-1	RT1
Rx2				RX-2	TR1
Rx3			RX-2	RX-3	RT2
Rx4				RX-4	TR2
Rx5		RX-2	RX-3	RX-5	RT3
Rx6				RX-6	TR3
Rx7			RX-4	RX-7	RT4
Rx8				RX-8	TR4

Notes:
1. TX-n or RX-n where n is the optical port number as defined Figure 15.
2. TRn or RTn where n is the optical port number as defined Figure 15.
3. Some QSFP-DD/QSFP-DD800/QSFP-DD1600 modules may require fewer CS, SN, or MDC connectors. In such cases, Port #1 is always the left-most port. Successive ports then follow sequentially from left-to-right as shown in Figure 15.



Proposed Optical OLT

- **Leverage CL136/162/176A to enable pre-coder on/off based on receiver requests**
  - DME frame (request/status) with training frame are transmitted F1 and F2
    - Rx1 or Rx2 receiver may need pre-coder to be enabled on the transmitter
    - RX1 or Rx2 may respectively request precoder to be enabled on Tx2 and Tx1
- **Propagate the RTS from the 200G AUI link across the 200G optical links**
  - In case of 200G optical PMDs operating with 100G AUI RTS not utilized
  - In case of 200G optical PMDs with 100G AUI on one end and 200G AUI on the other end, RTS is sourced and terminated in the 200G module PMA attached to 100G AUI
- **Illustration F1/F2 is for one duplex link, OLT can be supported across any 1 to N lanes PMDs as long as Tx/Rx pairing are preserved.**



Leverage CL 136/162/176A Control/Status Field Structure for OLT

- **Use Table 176A-2 and 176A-3 with number of fields changed to reserved and include control for RTS**
  - Number of DME pages proposed to be 4 instead of 2 (with all fields in page 3 and 4 reserved).

Table 176A-2 Control field structure with modification highlighted

Bit(s)	Name	Description
15:14	Reserved	Transmit as 0, ignore on receipt
13:10	Reserved	Transmit as 0, ignore on receipt
9:7	Modulation and precoding request	9 8 7 1 1 1 = PAM4 free-running PRBS31 with precoding 1 0 1 = Reserved 1 1 0 = PAM4 free-running PRBS13 with precoding 1 0 0 = PAM4 PRBS13 0 1 1 = PAM4 free-running PRBS31 0 1 0 = PAM4 free-running PRBS13 0 0 1 = PAM2 free-running PRBS31 0 0 0 = PAM2 PRBS13
6:5	Reserved	Transmit as 0, ignore on receipt
4:2	Reserved	Transmit as 0, ignore on receipt
1:0	Reserved	Transmit as 0, ignore on receipt

Table 176A-3 Status field structure with modification highlighted

Bit(s)	Name	Description
15	Receiver ready	1 = Training is complete and the receiver is ready for data 0 = Request for training to continue
14	One	Transmit as 1
13	Reserved	Transmit as 0, ignore on receipt
12:10	Modulation and precoding status	9 8 7 1 1 1 = PAM4 free-running PRBS31 with precoding 1 0 1 = Reserved 1 1 0 = PAM4 free-running PRBS13 with precoding 1 0 0 = PAM4 PRBS13 0 1 1 = PAM4 free-running PRBS31 0 1 0 = PAM4 free-running PRBS13 0 0 1 = PAM2 free-running PRBS31 0 0 0 = PAM2 PRBS13
9	Receiver frame lock	1 = Frame boundaries identified 0 = Frame boundaries not identified
8	Initial condition status	1 = Updated 0 = Not updated
7	Parity	Even parity bit
6	Extended Training (RTS)	1 = No data is available, continue training 0 = Switch to data when training is completed
5:3	Reserved	Transmit as 0, ignore on receipt
2:0	Reserved	Transmit as 0, ignore on receipt

C2C/C2M CL176A Multi-Segment LT with RTS

- **OLT will use AUI multi segments link training with RTS based on [lusted_3dj_02_2401] and [ran_3dj_elec_01_240229] proposals now in CL176A.**

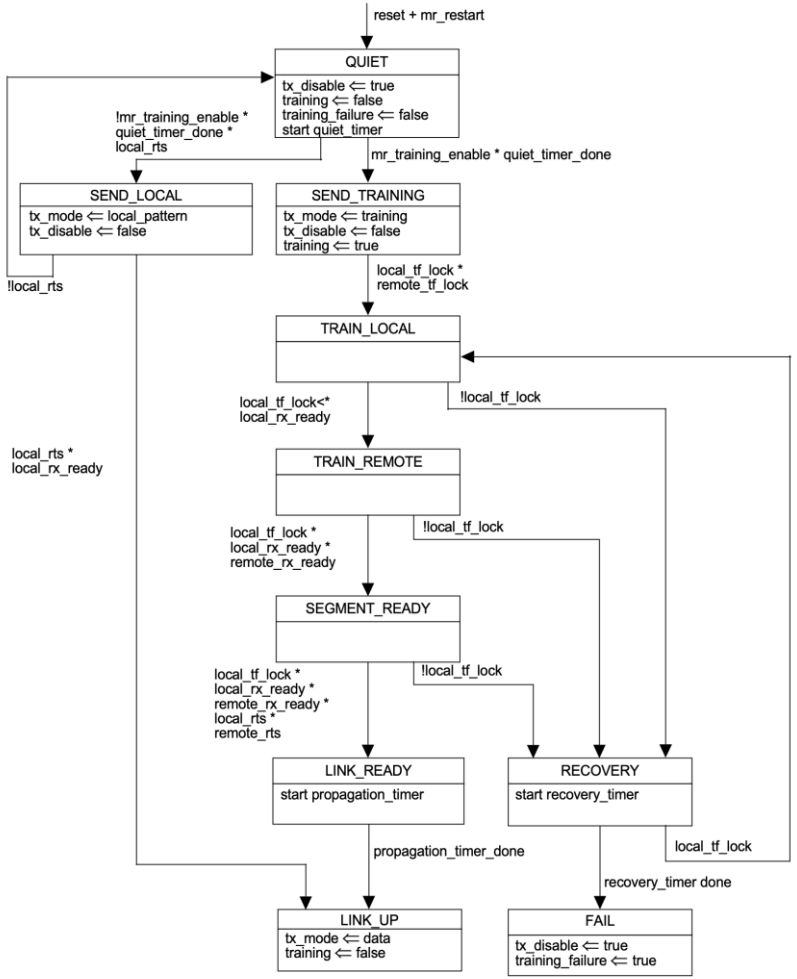


Figure 176A–6—Interface control state diagram

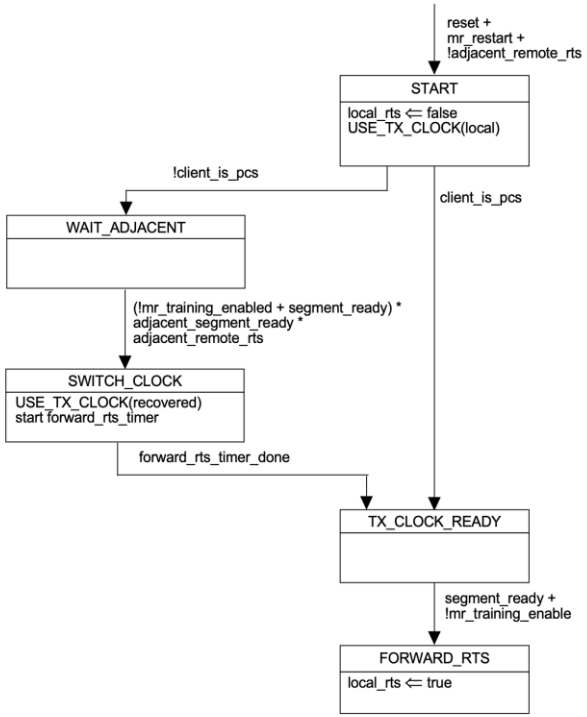


Figure 176A–7—RTS update state diagram

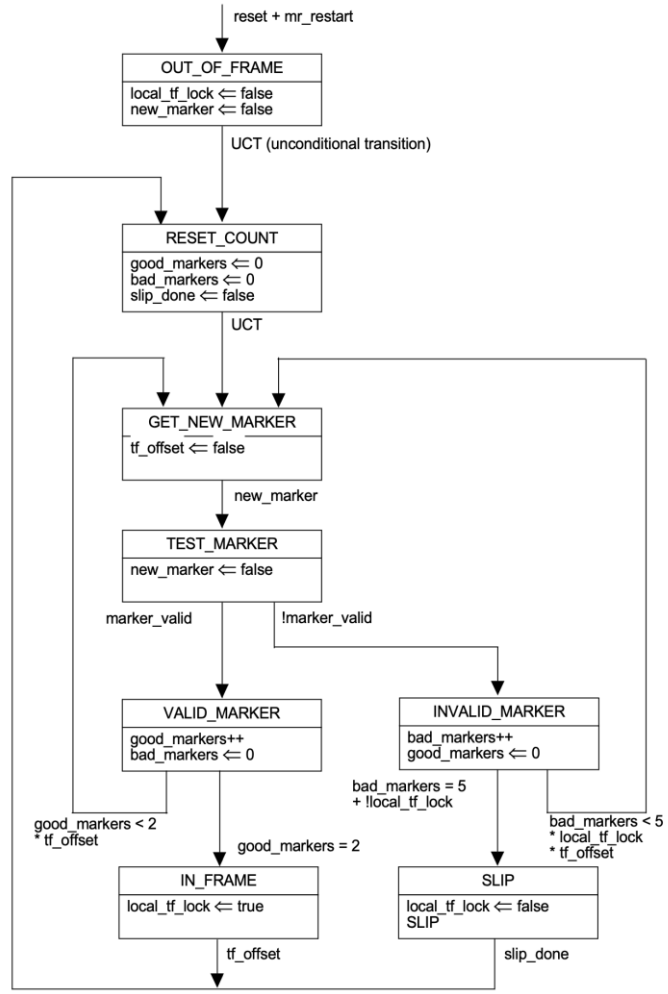
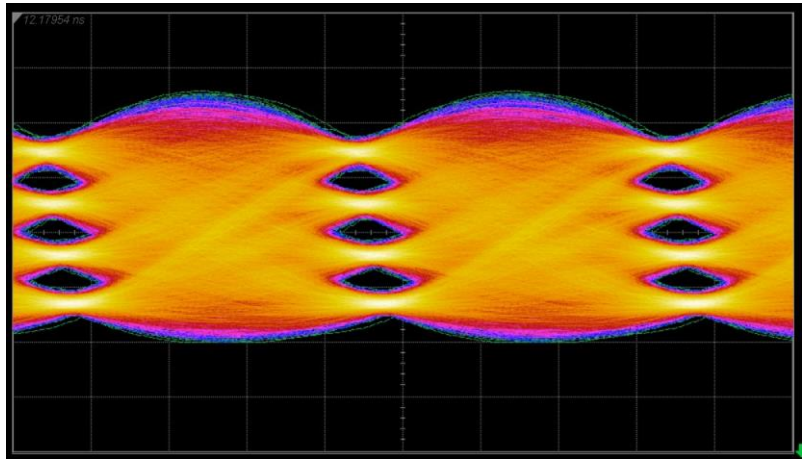


Figure 176A–8—Training frame lock state diagram

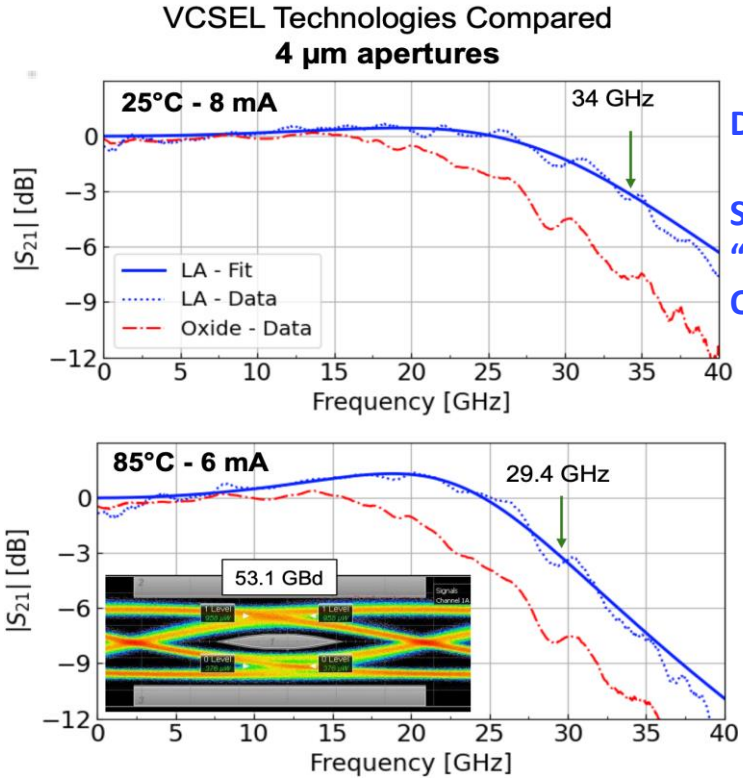
Progress in 200G MMF

- **Two technical papers in OFC 2024 demonstrated initial feasibility and path toward defining 200GBASE-SR class of optics**
  - Current advance VCSELs have BW in the low 30 GHz but with goal to get to ~40 GHz for 200GBASE-SR
  - 200G MMF CFI likely will be sometime in the 1st half of 2025
  - Defining OLT framework in DJ allow future MMF TF to leverage the OLT facility and enable Presets or TX controls.

Data Courtesy of Broadcom
Ramana Murty, et. al.,
"Toward 200G per Lane VCSEL Based Multimode Links"
OFC 2024 M2D.3



Symbol rate:	**106.25 GBd**
Pattern:	**PRBS15Q**
Tx FIR:	**7-tap**
DCA SIRC filter:	**53.1 GHz**
DCA FFE:	**25-tap**
ER:	**2.7 dB**
Temperature:	**Room**
Fiber:	**2 m**



Data Courtesy of Coherent

S. Tirelli, et. al.,
"Datacom VCSELs Beyond 100G"
OFC 2024 M2D-1

Summary

- **Multi-segment link training with RTS is a very important feature that has been adopted in the 802.3dj CL176A for AUIs**
  - Unless we carry RTS signal across optical link from PCS-PCS with OLT the benefit of RTS for the link is negated
- **OLT will be based on multi-segment C2C/C2M link training of CL176A with RTS**
  - CL176A transmitter FFE select/echoes are reserved as we are not enabling any transmitter control that affects TDECQ
  - DJ or future project may choose to enable additional optical transmit controls
- **Proposed OLT will be based on clause 176A and provides the following**
  - Pre-coder control will mitigate problematic burst on optical links from DR to LR4
  - The OLT training facility starting with NRZ then transition to PAM4 provide robust receiver/ADC/DSP start up
  - RTS transmitted across optical links provides graceful mechanisam to transition from training to PCS data in predictable and timely manner
  - Proposed OLT also will address many of the 100G optical link/DSP start up pain-point issues
  - OLT framework facility allow future project to enable preset, OMA control, or transmit FFE control that may be necessary for the future 200G MMF project or OIF EEI
- **Maintaining optical lane assignments within a PMDs already done to support break-out and remote/local loop-backs and doesn't add any burden**
- **802.3dj is the right project to add OAN and OLT to 200G optical Phys, while the PMA's for 200G are being developed**
  - Trying to add OLT facility later in a future PMD project will be more challenging and we may not have all the experts.

BACKUP

Other OLT Potential Capabilities

- **OLT baseline with 4 pages will have sufficient space that future projects may choose to control various type of optical parameters to improve optical link performance and/or reach – not part of current proposed baseline**
  - Preset
  - Adjusting transmit FFE (pre-emphasis)
  - OMA control
  - MZM compression
  - EA modulators asymmetrical compression
  - CD penalty on links > 2km on outer wavelengths L0 and L3
- **Some of the above controls may impact baseline TDECQ which not always produce the best BER, as long as adjustment are not disruptive receiver MMSE will converge to the best link BER.**