Further consideration of optical auto-negotiation (OAN)

Matt Brown, Alphawave Semi
Kent Lusted, Intel
Gary Nicholl, Cisco

Contributors

Ali Ghiasi, Ghiasi Quantum

Supporters

Ali Ghiasi, Ghiasi Quantum

Introduction

Two previous presentations on OAN

- https://www.ieee802.org/3/dj/public/23_1128/brown_3dj_01a_2311.pdf
- https://www.ieee802.org/3/dj/public/24_01/brown_3dj_02_2401.pdf

Some questions and comments regarding OAN expressed at the January 2024 Interim meeting as follows:

- Provide guidance on how it would work for break-out scenarios.
- Coordination with host.
- Consider using same signal timing as defined in Clause 73.
- Providing alternate advertisement and priority tables to suit other organizations.

Dealing with breakout and lane for AN signaling

Lane assignments for electrical interfaces

- For CR and KR links, IEEE 802.3 defines explicitly specifies that each transmitter lane i on one the local device (SL_i) connects to the corresponding receiver lane i (DL_i) on the link partner.
- For CR, mapping multiple PMD/MDI to a single connector (i.e., break-out) is also explicitly defined in the same way.
- See related slides in Appendix A of this presentation.
- Thus for CR and KR electrical links, specifying AN signaling on lane 0 is sufficient.
  - Only need to transmit on defined lane 0 and receive on defined lane 0, where there is single PMD/MDI on a connector or multiple.

Lane assignments for optical interfaces

- For optical links, IEEE 802.3 does not make explicit mapping of lanes to optical connectors and do not consider breakout cases.
  - See next slide
- The QSFP-DD and CMIS specifications provide explicit mapping of groups of lanes per PMD/MDI to TX fibers and from RX fibers.
  - See related slides in Appendix B.
- However, it is not clear that the fiber plant between will always attach a TX lane to its corresponding RX lane.
  - In other words, lane 0 on TX may not attach to lane 0 on RX.

802.3 optical mapping to lanes

124.6 Lane assignments

There are no lane assignments (within a group of transmit or receive lanes) for 400GBASE-DR4. While it is expected that a PMD will map electrical lane *i* to optical lane *i* and vice versa, there is no need to define the physical ordering of the lanes, as the PCS sublayer is capable of receiving the lanes in any arrangement. The positioning of transmit and receive lanes at the MDI is specified in 124.11.3.1.

Source: 802.3-2022

IEEE 802.3 optical PMD specifications do explicitly not map logical signals to optical connector “contacts”.

Instead, it is assumed that the order does not matter since the PCS will reorder as necessary at the receiver.

Therefore, there is no explicit definition of which fiber is lane 0 (or any other lane #).

Nor is there any consideration for break-out.

124.11.3 Medium Dependent Interface (MDI)

The 400GBASE-DR4 PMD is coupled to the fiber optic cabling at the MDI. The MDI is the interface between the PMD and the “fiber optic cabling” (as shown in Figure 124–5). The 400GBASE-DR4 PMD is coupled to the fiber optic cabling through one connector plug into the MDI optical receptacle as shown in Figure 124–7. Example constructions of the MDI include the following:

a) PMD with a connectorized fiber pigtail plugged into an adapter

b) PMD receptacle

124.11.3.1 Optical lane assignments

The four transmit and four receive optical lanes of 400GBASE-DR4 shall occupy the positions depicted in Figure 124–6 when looking into the MDI receptacle with the connector keyway feature on top. The interface contains eight active lanes within twelve total positions. The transmit optical lanes occupy the left-most four positions. The receive optical lanes occupy the right-most four positions. The four center positions are unused.

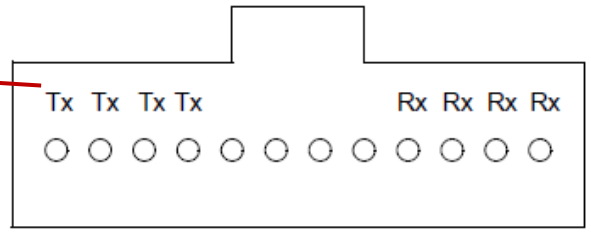


Figure 124–6—400GBASE-DR4 optical lane assignments

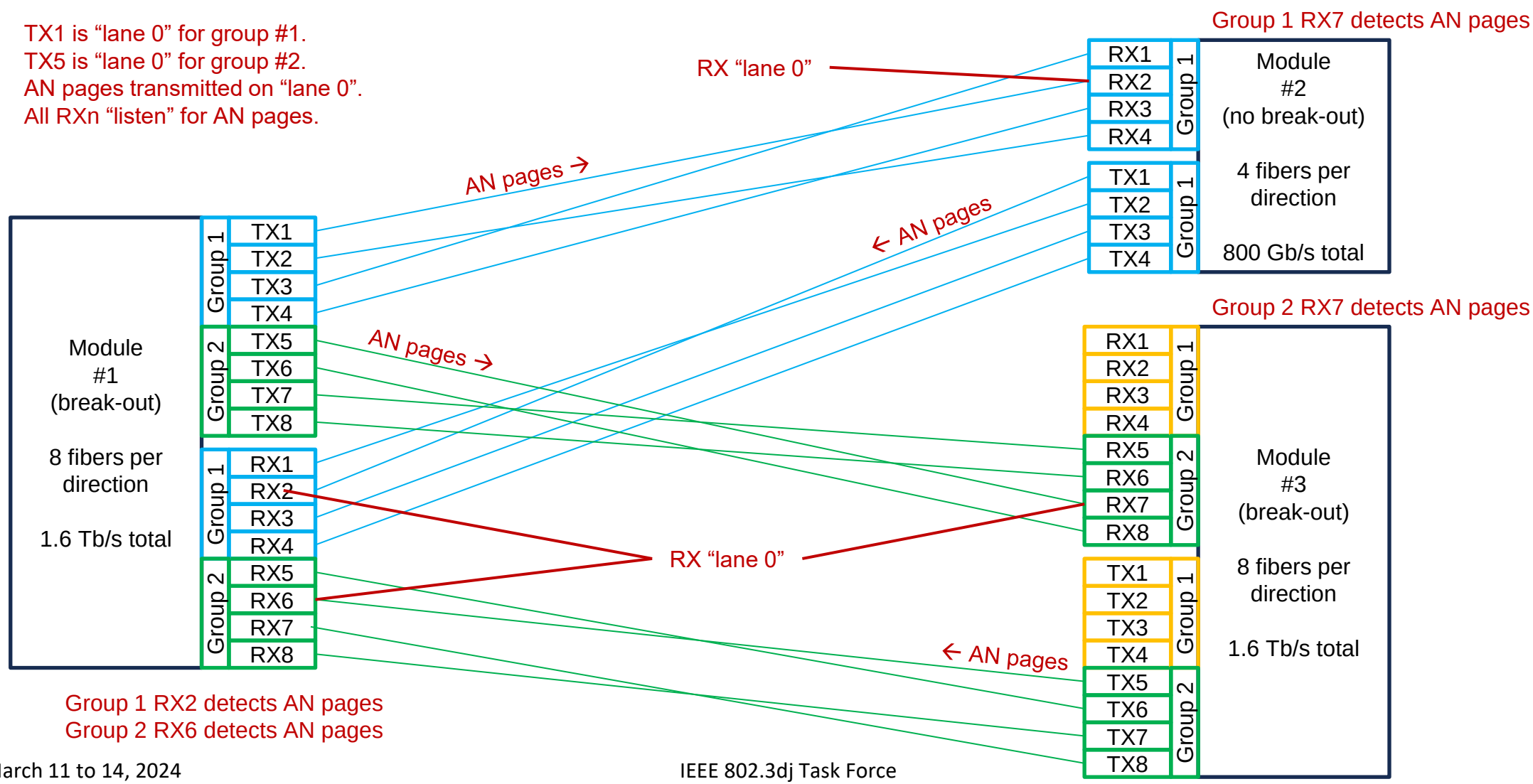
AN over optical links including breakout – lane 0

- The previous slides show that there is well controlled mapping of groups of Ethernet C2M electrical lanes to particular set of optical fibers (each group is a PMD/MDI) on a module.
- To allow for break-out, we can establish a convention that that the lowest optical fiber lane number is “lane 0” from an OAN perspective.
- It is not clear however that optical TX “lane 0” on one module will be connected to optical RX “lane 0” on the link partner module.
  - Might be shuffled by the fiber plant.
- We can nevertheless work around that.

AN over DR (parallel) optical links including breakout – method

- For each transmit group, the lowest optical lane number is designated “lane 0”.
- AN signal (AN pages) are transmitted over “lane 0” for each transmit group.
- For each receive group, all receivers “listen” for AN pages.
- When a receiver on one lane detects AN pages, it becomes “lane 0” from an OAN perspective for that group.
- AN continues with AN pages transmitted and received on “lane 0”, determined as prescribed above.

Illustration of AN messaging for 800GBASE-DR4/DR4-2 lines, with and without break-out



OAN coordination between host and module

Coordination with host – introduction

- This topic deserves a whole presentation on it own. This provides a high-level summary.
- The previous topic looked at how AN might be advertised on a fiber in each direction.
- Contribution brown_3dj_02_2401 proposed a procedure for OAN on the optical side of the module.
- However, part of that procedure requires coordination between the host and module.

Coordination with host – example method

- Assume module management as defined by CMIS.
  - But likely requires new capabilities to be defined.
- Host reads module capabilities, which might be a superset beyond Ethernet PHY types defined in 802.3.
- Host decides which PHY type or types are acceptable and configures a fiber (media lanes) group on the module and configures which PHY types to advertise in OAN.
- The module performs OAN and resolves to a highest common denominator (HCD) PHY type.
- The host completes configuration based on the resolved HCD PHY type.

Dealing with PHY types beyond IEEE

Background for extending OAN to technology types beyond IEEE

- Contribution brown_3dj_02_2401 proposed tables for:
  - AN page bits used for technology advertisement
  - Priority resolution to determine HCD
- Concern was expressed that this table listed only Ethernet PMD types, not others that are listed in CMIS
- Since this is an Ethernet standard, only relevant Ethernet PMD types are specified.
- A different list may be created using extended features of AN
  - AN Selector Field Encoding
  - AN Extended Next Page

Previously proposed advertisement and priority tables

Link codeword advertisement field A[0:27]

Source: https://www.ieee802.org/3/dj/public/24_01/brown_3dj_02_2401.pdf

- Allocate A[0:27] to technologies as shown in the table to the right.
- A value of 1 indicates that the PHY is advertising the corresponding technology.
- It is expected that if advertising a technology, the implementation can support that PHY type.
- All technologies supported by the implementation need not be advertised.
  - e.g., A particular PHY type can be forced by advertising only that technology.

Bit	Technology	Capability
A0	1.6TBASE-DR8-2	1.6 Tb/s, 2 km, parallel
A1	1.6TBASE-DR8	1.6 Tb/s, 500 m, parallel
A2	800GBASE-LR4	800 Gb/s, 10 km, duplex
A3	800GBASE-FR4	800 Gb/s, 2 km, d
A4	800GBASE-DR4-2	800 Gb/s, 2 km, p
A5	800GBASE-FR4-500m	800 Gb/s, 500 m,
A6	800GBASE-DR4	800 Gb/s, 500 m
A7	400GBASE-DR2-2	400 Gb/s, 2 km, p
A8	400GBASE-DR2	400 Gb/s, 500 m
A9	200GBASE-FR1	200 Gb/s, 2 km, p
A10	200GBASE-DR1	200 Gb/s, 500 m
A11	Reserved	
…		
A27	Reserved	

Priority resolution table

- If the two devices, one at each end of a fiber, advertise multiple common technologies, then a mechanism to reconcile is required.
- A prioritization table is defined to determine which of a set of common technologies to select.
- A priority table, to the right, is proposed as a means to determine the highest priority common technology.
- The proposed order of priority based upon the following criteria, in order of prioritization:
  - Ethernet rate: higher rate = higher priority
  - Reach: longer reach = higher priority
  - Data rate per lane: higher data rate = higher priority
  - Number of fibers (inverse): fewer fibers = higher priority
    - e.g., FR4 higher priority than DR4-2

Priority	Technology	Capability
1 highest	1.6TBASE-DR8-2	1.6 Tb/s, 2 km, parallel
2	1.6TBASE-DR8	1.6 Tb/s, 500 m, parallel
3	800GBASE-LR4	800 Gb/s, 10 km, duplex
4	800GBASE-FR4	800 Gb/s, 2 km, duplex
5	800GBASE-DR4-2	800 Gb/s, 2 km, parallel
6	800GBASE-FR4-500m	800 Gb/s, 500 m, duplex
7	800GBASE-DR4	800 Gb/s, 500 m
8	400GBASE-DR2-2	400 Gb/s, 2 km, parallel
9	400GBASE-DR2	400 Gb/s, 500 m
10	200GBASE-FR1	200 Gb/s, 2 km, parallel
11 lowest	200GBASE-DR1	200 Gb/s, 500 m

AN Selector Field Encoding

- The selector field in an Auto-negotiation base page identifies the types of message being sent by AN
  - The selector field in Cl 73 currently has a single value (00001b = IEEE Std 802.3)
- For OAN, new values for the selector field could be defined to enable other organizations to use the optical AN
  - Different AN page interpretations and priority tables can be defined for each organization

D 0	D 1	D 2	D 3	D 4	D 5	D 6	D 7	D 8	D 9	D 10	D 11	D 12	D 13	D 14	D 15
S 0	S 1	S 2	S 3	S 4	E 0	E 1	E 2	E 3	E 4	C 0	C 1	C 2	RF	Ack	NP

D 16	D 17	D 18	D 19	D 20	D 21	D 22	D 23	D 24	D 25	D 26	D 43	D 44	D 45	D 46	D 47
T 0	T 1	T 2	T 3	T 4	A 0	A 1	A 2	A 3	A 4	A 5	A 22	F 2	F 3	F 0	F 1

Figure 73–6—Link codeword Base Page

Table 73–3—Selector Field Encoding

S4	S3	S2	S1	S0	Selector description
0	0	0	0	1	IEEE Std 802.3

Table 28A–1—Selector Field value mappings

S4	S3	S2	S1	S0	Selector description
0	0	0	0	0	Reserved for future Auto-Negotiation development
0	0	0	0	1	IEEE Std 802.3
0	0	0	1	0	IEEE Std 802.9a-1995 (withdrawn)
0	0	0	1	1	IEEE Std 802.5v-2001 (withdrawn)
0	0	1	0	0	IEEE Std 1394
0	0	1	0	1	INCITS
0	0	1	1	X	Reserved for future Auto-Negotiation development
0	1	X	X	X	Reserved for future Auto-Negotiation development
1	X	X	X	X	Reserved for future Auto-Negotiation development

OAN Extended Next Page

- Additional information beyond the Link Codeword Base Page can be exchanged with a partner by using a Next Page function
  - Example: 73.7.7 for Next Page function
- Two Next Page types could be defined: Formatted and Unformatted
- The Message Code 5 type (e.g. OUI) could be created to identify the meaning and format of sub-sequent Next Pages for the exchange of information defined elsewhere
  - Retains the AN transmission and acknowledge protocol

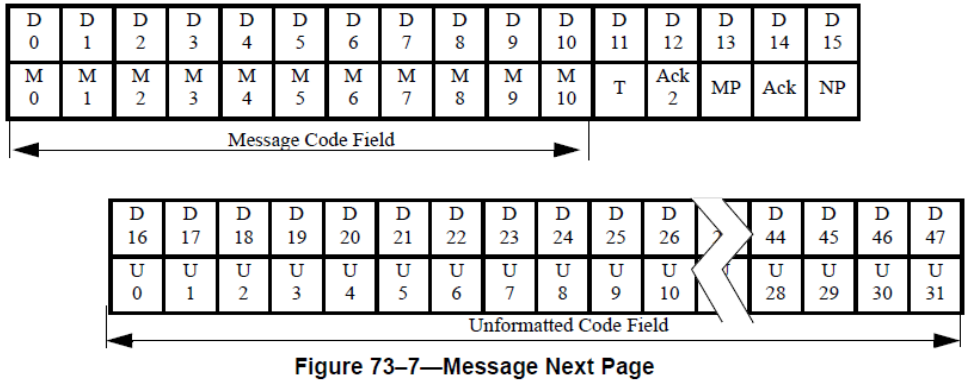


Figure 73–7—Message Next Page

OAN signal timing

Background for OAN signal timing

- Contribution brown_3dj_02_2401 proposed to change the AN signal timing parameters for OAN.
- However, it was suggested that it should be technically okay to use the same timing as defined in Clause 73.
  - This would make OAN RX/TX implementation almost identical to electrical AN implementation allowing reuse.
- Agree that making the timing the same would be the best starting point.
  - If we find that this is not compatible with typical optical implementations, then we can revisit.

Previously proposed OAN signal timing…

Source:
https://www.ieee802.org/3/dj/public/24_01/brown_3dj_02_2401.pdf

Let's forego this proposal and revisit later if necessary.

OAN Timing Characteristics

- For OAN, the signal will be defined by optical rather than electrical characteristics.
- The table below is proposed as a starting point.
- Typical transition distance based on u[illegible]nterval fo[illegible].125 GBd / 4 = 13.28125 GBd and with minimum and maximum values scaled as in Table [illegible].

	Parameter	Min.	Typ.	Max.	Units
T1	Transition position spacing (period)	[illegible].2941 – 0.01%	75.2941	75.2941 + 0.01%	ps
T2	Clock transition to clock transition	[illegible]45.8823	150.5882	155.2941	ps
T3	Clock transition to data transition (d[illegible] 1)	[illegible]5882	75.2941	80	ps
T4	Transitions in a DME page	51	—	100	—
T5	DME page width	7482.3529	7981.1765	8480	ps
T6	DME Manchester violation delimiter width	282.3529	301.1765	320	ps

AN signal timing per Clause 73

Source:
https://www.ieee802.org/3/dj/public/24_01/brown_3dj_02_2401.pdf

Let's make this the proposed timing for OAN.

Clause 73 – AN signaling (redux)

Table 73–1—DME electrical characteristics

Parameter	Value	Units
Transmit differential peak-to-peak output voltage	600 to 1200	mV
Receive differential peak-to-peak input voltage	200 to 1200	mV

Table 73–2— DME page timing summary

	Parameter	Min.	Typ.	Max.	Units
T1	Transition position spacing (period)	3.2 −0.01%	3.2	3.2 +0.01%	ns
T2	Clock transition to clock transition	6.2	6.4	6.6	ns
T3	Clock transition to data transition (data = 1)	3.0	3.2	3.4	ns
T4	Transitions in a DME page	51	—	100	—
T5	DME page width	338.8	339.2	339.6	ns
T6	DME Manchester violation delimiter width	12.6	12.8	13.0	ns

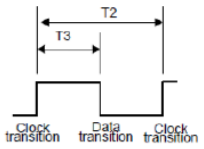


Figure 73–4—DME page transition timing

Given 3.2 ns width (T3) this is equivalent to NRZ signaling rate 312.5 MBd

The encoding of data using DME bits in an DME page is illustrated in Figure 73–3.

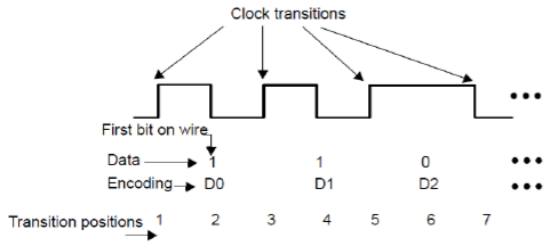


Figure 73–3—Data bit encoding within DME pages

73.5.3.1 Manchester violation delimiter

A violation is signaled as shown in Figure 73–5.

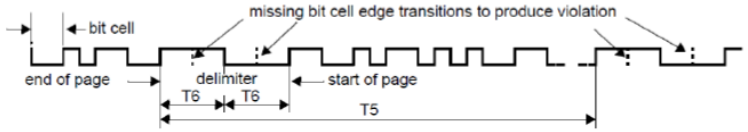


Figure 73–5—Manchester violation

Appendix A
CR and KR lane mapping

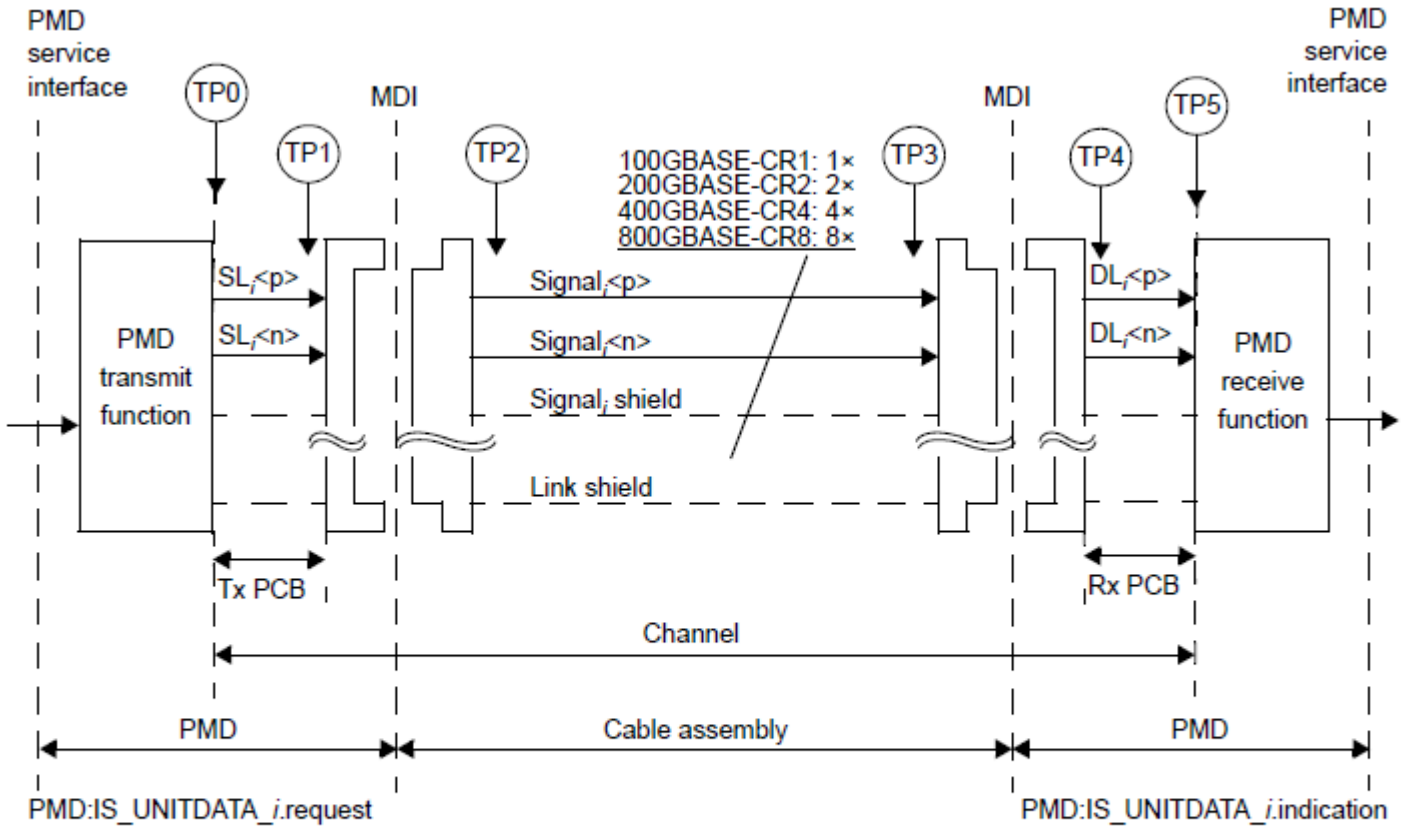
800GBASE-CR8 signaling

Source:
IEEE Std 802.3df Draft 3.2

Mapping of lanes from transmitter (SL_i) to receiver (DL_i) is explicit.

Lane 0 at transmitter is always connect to lane 0 at receiver.

AN signaling on Lane 0.



NOTE—One direction is illustrated.

Figure 162–2—100GBASE-CR1, 200GBASE-CR2, ~~or~~ 400GBASE-CR4, or 800GBASE-CR8 link

CR – 1:1 mapping logical PMD/MDI lanes to to connector contacts

Source: IEEE Std 802.3ck-2022

Mapping of logical PMD lanes to physical connector lanes is explicitly defined for non-breakout and breakout cases. Therefore, for each Ethernet group, lane 0 is explicit.

Table 162C–2—PMD to connector signal assignments

PMD signal <PMD number>:<PMD signal>				Connector signal
100GBASE-CR1	200GBASE-CR2	400GBASE-CR4	800GBASE-CR8	
1-lane 0:DL0n	0:DL0n	0:DL0n	0:DL0n	DL0n
1-lane 0:DL0p	2-lane 0:DL0p	0:DL0p	0:DL0p	DL0p
1-lane 1:DL0n	2-lane 0:DL1n	0:DL1n	0:DL1n	DL1n
1-lane 1:DL0p	0:DL1p	4-lane 0:DL1p	0:DL1p	DL1p
1-lane 2:DL0n	1:DL0n	0:DL2n	0:DL2n	DL2n
1-lane 2:DL0p	2-lane 1:DL0p	0:DL2p	0:DL2p	DL2p
1-lane 3:DL0n	1:DL1n	0:DL3n	0:DL3n	DL3n
1-lane 3:DL0p	1:DL1p	0:DL3p	8-lane 0:DL3p	DL3p
1-lane 4:DL0n	2:DL0n	1:DL0n	0:DL4n	DL4n
1-lane 4:DL0p	2-lane 2:DL0p	1:DL0p	0:DL4p	DL4p
1-lane 5:DL0n	2:DL1n	1:DL1n	0:DL5n	DL5n
1-lane 5:DL0p	2:DL1p	4-lane 1:DL1p	0:DL5p	DL5p
1-lane 6:DL0n	3:DL0n	1:DL2n	0:DL6n	DL6n
1-lane 6:DL0p	2-lane 3:DL0p	1:DL2p	0:DL6p	DL6p
1-lane 7:DL0n	2-lane 3:DL1n	1:DL3n	0:DL7n	DL7n
1-lane 7:DL0p	3:DL1p	1:DL3p	0:DL7p	DL7p
0:SL0n	0:SL0n	0:SL0n	0:SL0n	SL0n
0:SL0p	0:SL0p	0:SL0p	0:SL0p	SL0p
1:SL0n	0:SL1n	0:SL1n	0:SL1n	SL1n

Table 162C–4—MDI connector contact mapping for QSFP112 and QSFP-DD800

QSFP112	QSFP-DD800	Connector signal name	Description
1	1	GND	Ground
2	2	SL1n	Transmitter inverted data input
3	3	SL1p	Transmitter non-inverted data input
4	4	GND	Ground
5	5	SL3n	Transmitter inverted data input
6	6	SL3p	Transmitter non-inverted data input
7	7	GND	Ground
13	13	GND	Ground
14	14	DL2p	Receiver non-inverted data output
15	15	DL2n	Receiver inverted data output
16	16	GND	Ground
17	17	DL0p	Receiver non-inverted data output
18	18	DL0n	Receiver inverted data output
19	19	GND	Ground
20	20	GND	Ground
21	21	DL1n	Receiver inverted data output
22	22	DL1p	Receiver non-inverted data output

Copper break-out cables

Break out 400 Gb/s host port to 4x 100GBASE-CR1 links

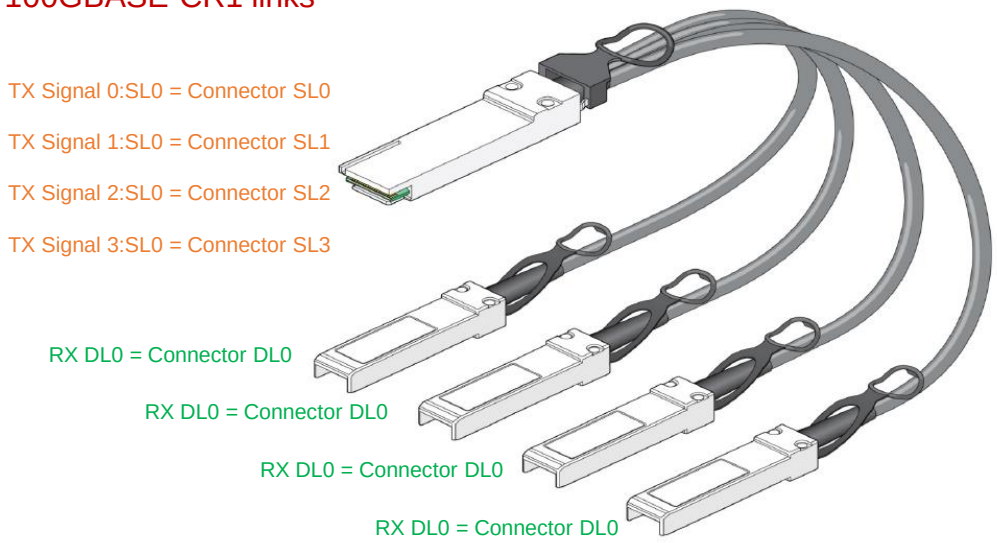


Figure 136D–3— QSFP to 4x SFP **cable assembly**

Break out 800 Gb/s host port to 4x 200GBASE-CR2 links

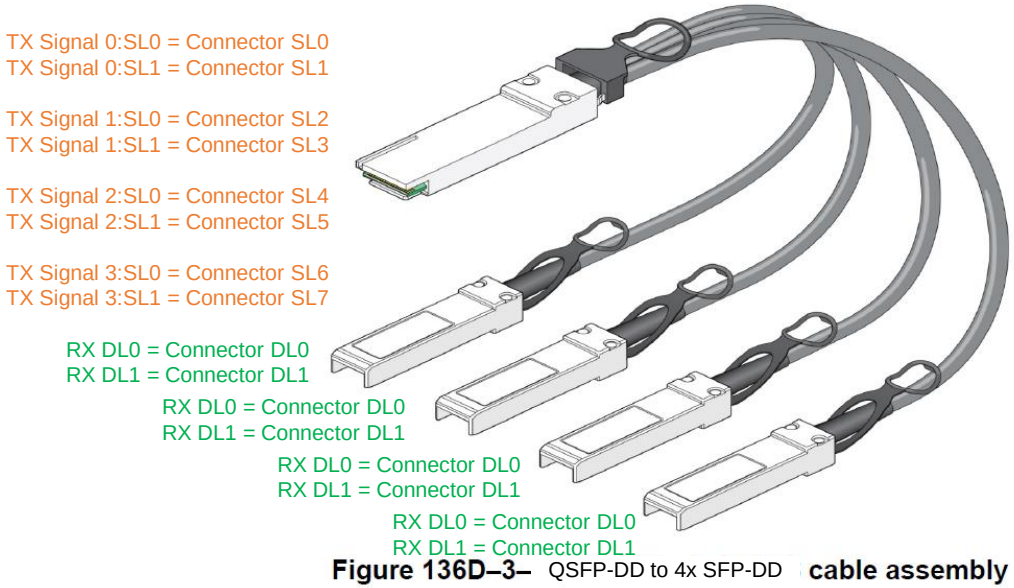


Figure 136D–3– QSFP-DD to 4x SFP-DD **cable assembly**

CMIS electrical mapping

- The CMIS specification also explicitly defines for copper cables the mapping of groups to particular media lanes.
  - See section 8.3.8.
  - https://www.oiforum.com/wp-content/uploads/OIF-CMIS-05.2.pdf
- It includes various breakout configurations.
- But it is does not explicitly identify lanes numbers within a group
- We might assume that the lowest “host lane number” in a group is lane 0 for that group.

Backplane signaling

163.11 MDI specifications

The MDI for the 100GBASE-KR1, 200GBASE-KR2, ~~and~~ 400GBASE-KR4, and 800GBASE-KR8 PHYs is an implementation-dependent direct electrical connection between the PMD and the medium. The MDI is composed of $2 \times n$ differential pairs, one pair for the transmit function and one pair for the receive function on each lane, marked by TP0 and TP5 in Figure 163–2.

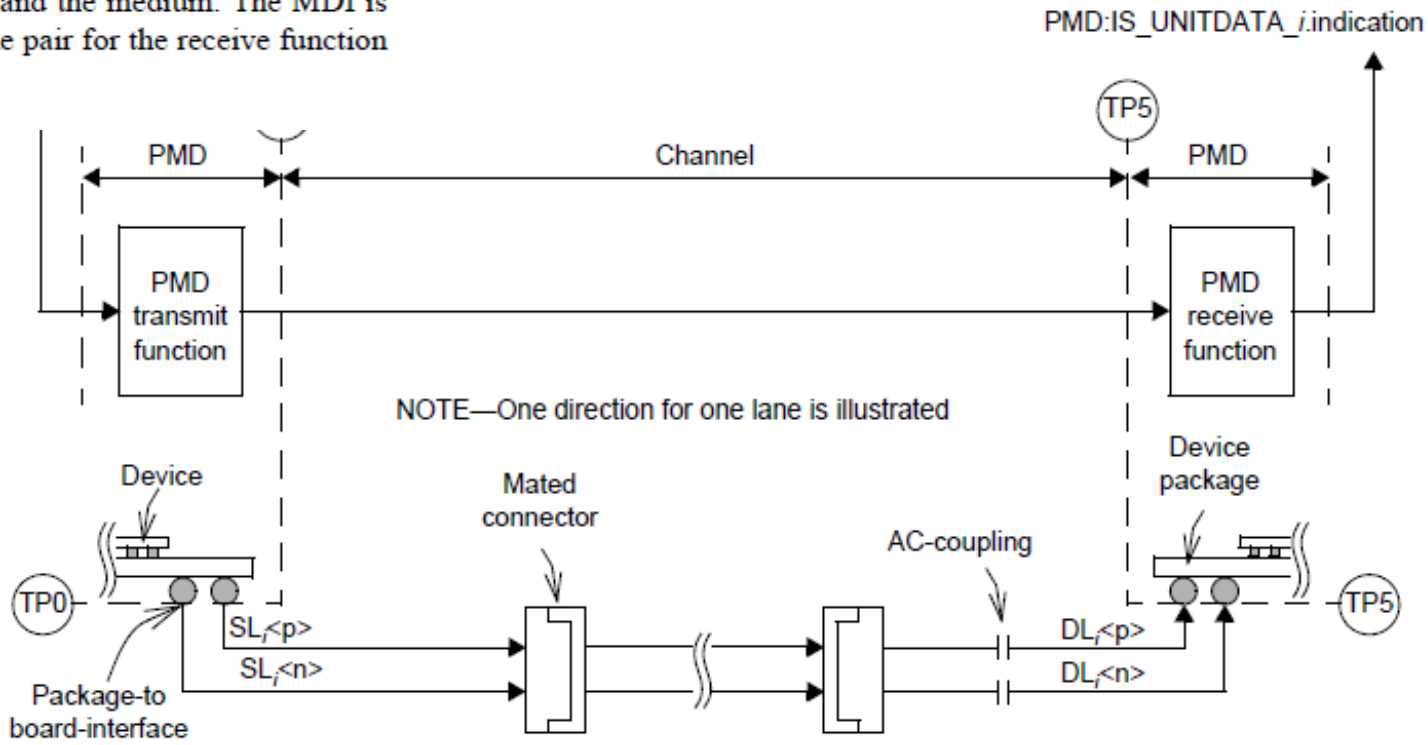


Figure 163–2—100GBASE-KR1, 200GBASE-KR2, ~~or~~ 400GBASE-KR4, or 800GBASE-KR8 link

For backplane, there is an explicit one to one mapping from local transmitter (SL_i) to link partner receiver (DL_i).

Therefore lane 0 is explicit.

Breakout is not a concern for backplane since it is a closed point to point system.

Appendix B
Optical lane mapping

QSFP-DD mapping

Source:
"QSFP-DD/QSFP-DD800/QSFP112 Hardware Specification for QSFP DOUBLE DENSITY 8X AND QSFP 4X PLUGGABLE TRANSCEIVERS", Revision 6.01, May 28, 2021

Module mapping from electrical input to optical fibers is explicit.

However, the grouping of electrical inputs into Ethernet groups is not defined. But it does specify how groups of electrical lanes are mapped to a single fiber (e.g., for electrical or wavelength multiplexing).

It is therefore not clear which groups of TX outputs would be mapped to a single Ethernet rate, nor which we might deem to be "lane 0".

Nor is it defined whether like ports will be connected at each end of the fibers.

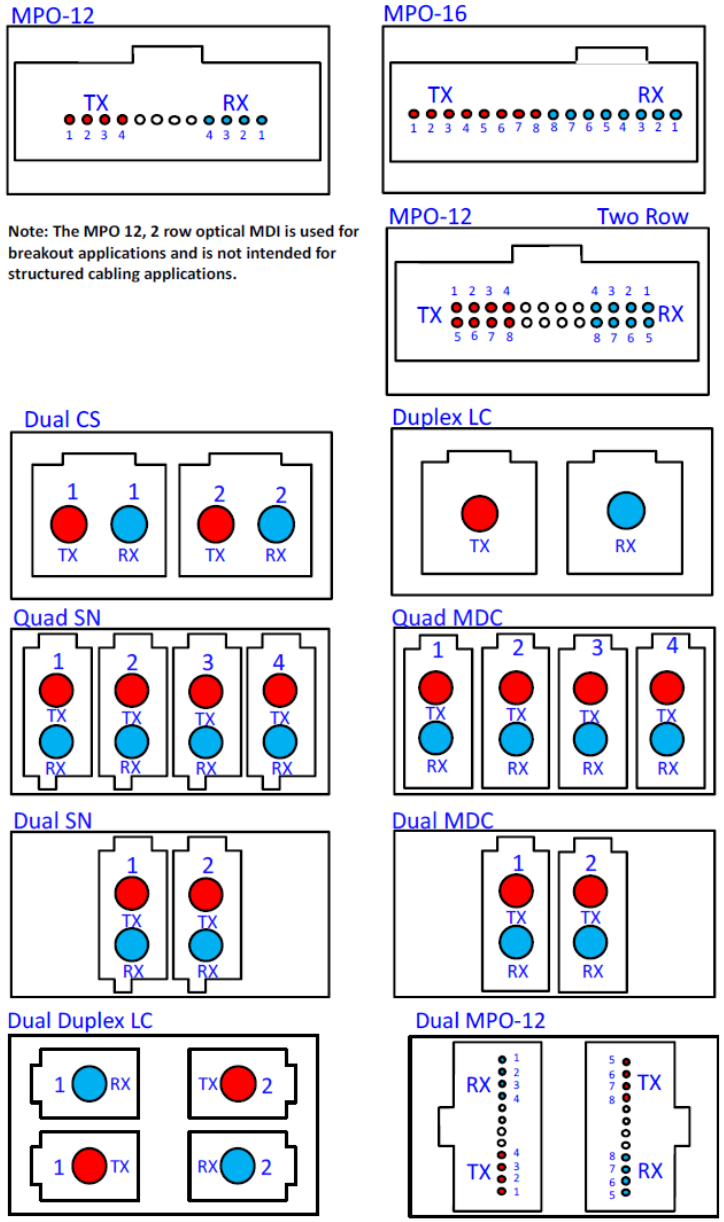
6.1 Electrical data input/output to optical port mapping

Table 22 defines the mapping for QSFP-DD/QSFP-DD800 electrical Tx data inputs and Rx data outputs to optical ports combinations. Note that there is no defined mapping of electrical input/output to optical wavelengths for WDM applications. The QSFP112 with 4 transmit lanes [Tx1-Tx4] and 4 receive lanes [Rx1-Rx4] allows optical port mapping as shown in Table 22, but the Tx/Rx lanes 5-8 should be ignored.

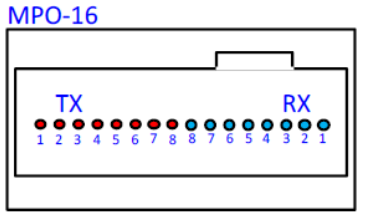
Table 22- Electrical Signal to Optical Port Mapping

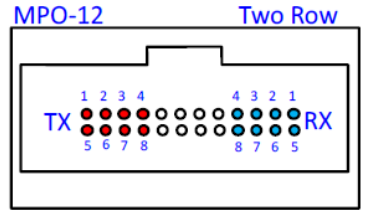
Electrical data input/output	Optical port mapping (see Figure 28)				
	Duplex LC, CS, SN, or MDC	MPO-12, Dual (CS, SN, MDC, Duplex LC, or MPO-12)	MPO-12, Quad (SN or MDC)	MPO-12 (two row), MPO-16, or Dual MPO-12	MPO-12, SN, MDC (BiDi)
	1 TX fiber 1 RX fiber [1]	2 TX fibers 2 RX fibers [1]	4 TX fibers 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 28.
2. TRn or RTn where n is the optical port number as defined Figure 28.
3. Some QSFP-DD/QSFP-DD800 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 28.



QSFP-DD mapping (8 fiber module)





Explore auto-negotiation for one of the case where the optical connector supports 8 TX and 8 RX fibers.

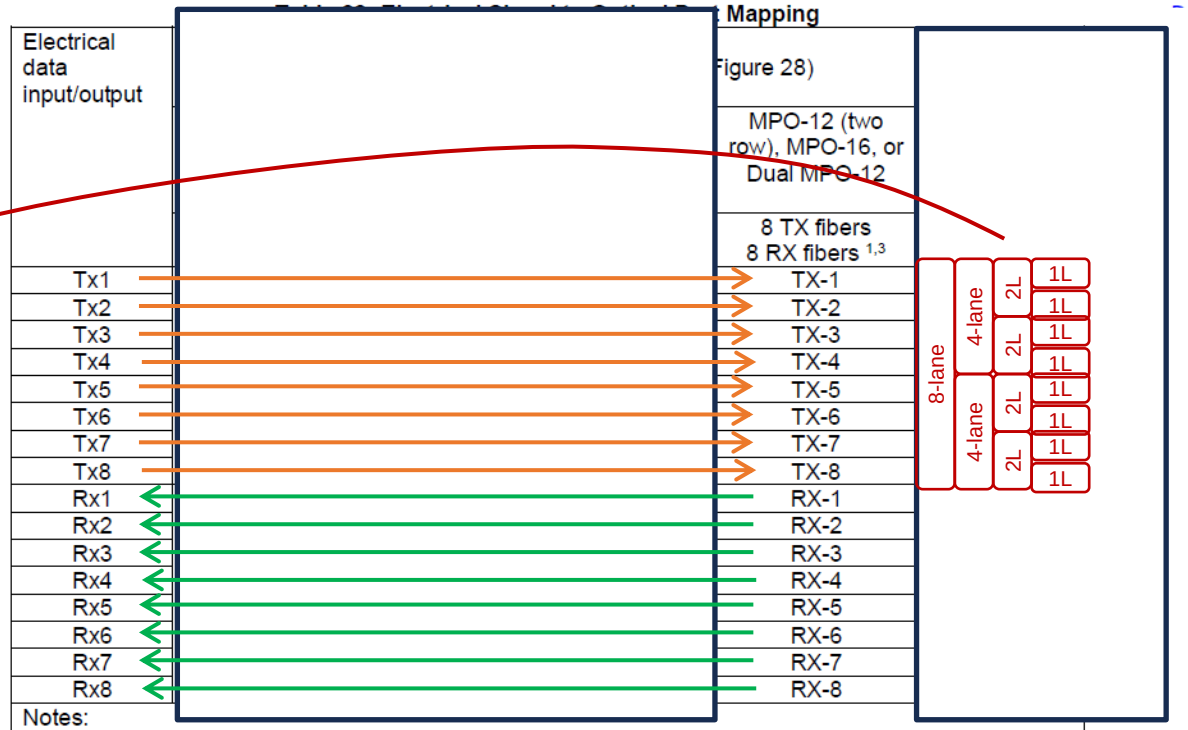
Each electrical input maps directly to a specific, corresponding optical fiber.
e.g., electrical Tx1 maps to fiber TX-1

We might also assume that Ethernet groups (for breakout) are grouped as shown. This is similar to how grouping of multiple electrical lanes to a single fiber are specified and similar to the way breakout is specified for CR. We might also assume that the lowest index is lane 0.

However, it is not clear that each optical fiber maps directly to a specific, corresponding optical fiber on the link partner:
e.g., node 1 Tx1 maps to node 2 Rx1.

6.1 Electrical data input/output to optical port mapping

Table 22 defines the mapping for QSFP-DD/QSFP-DD800 electrical Tx data inputs and Rx data outputs to optical ports combinations. Note that there is no defined mapping of electrical input/output to optical wavelengths for WDM applications. The QSFP112 with 4 transmit lanes [Tx1-Tx4] and 4 receive lanes [Rx1-Rx4] allows optical port mapping as shown in Table 22, but the Tx/Rx lanes 5-8 should be ignored.



Notes:
1. TX-n or RX-n where n is the optical port number as defined Figure 28.
2. TRn or RTn where n is the optical port number as defined Figure 28.
3. Some QSFP-DD/QSFP-DD800 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 28.

CMIS optical mapping

- The CMIS 5.2 specification also explicitly defines the mapping of groups of electrical lanes to particular fibers, but does not specifically identify lanes numbers within a fiber group.
  - See section 6.2.1.3.
  - https://www.oiforum.com/wp-content/uploads/OIF-CMIS-05.2.pdf
- However, we might assume that the lowest “module media lane” number in a group is “lane 0” for that group.

Thanks