50GbE and NG 100GbE Logic Baseline Proposal

Gary Nicholl - Cisco
Mark Gustlin - Xilinx
David Ofelt - Juniper

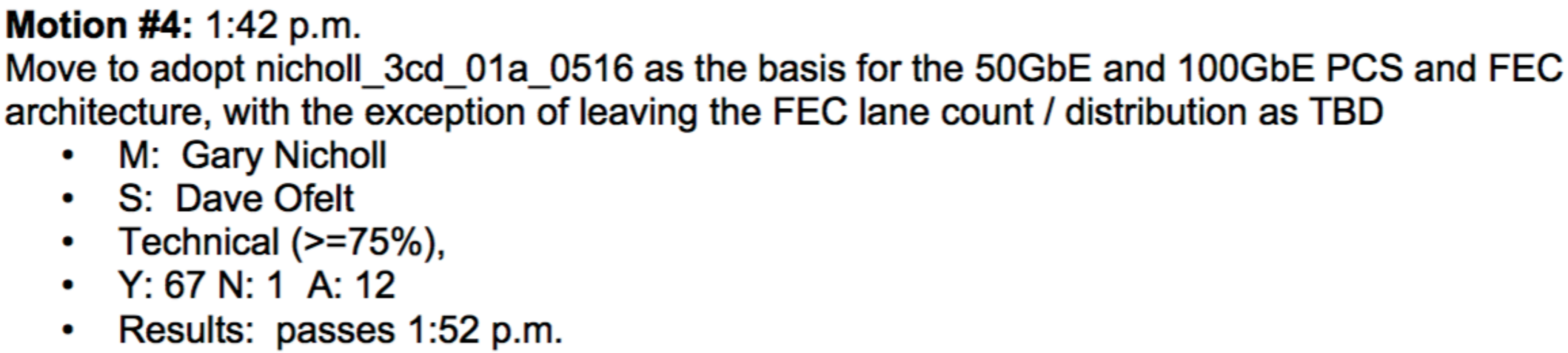
IEEE 802.3cd Task Force, July 25-28 2016, San Diego

Supporters

- Jonathan King - Finisar
- Chris Cole - Finisar
- Tom Palkert - Molex
- Bharat Tailor - Semtech
- Hanan Leizerovich - Multiphy
- Kent Lusted - Intel
- Pirooz Tooyserkani - Cisco
- Cedrik Begin - Cisco
- Mike Li - Intel
- David Lewis - Lumentum
- Kohichi Tamura - Oclaro
- Ryan Latchman - Macom
- Rob Stone - Broadcom
- Eric Baden - Broadcom
- Matt Brown - APM
- Paul Brooks - Viavi Solutions
- David Estes - Spirent
- Jerry Pepper - Ixia
- Thananya Baldwin - Ixia
- Peter Anslow - Ciena
- Adee Ran - Intel
- Arthur Marris - Cadence
- Faisal Dada- Xilinx
- Scott Irwin - Mosys
- Don Cober - Comira
- Jeff Twombly – Credo
- Phil Sun - Credo
- Tom Issenhuth - Microsoft
- Brad Booth - Microsoft
- Ali Ghiasi - Ghiasi Quantum LLC

Background

- At the 802.3cd meeting in Whistler it was agreed to adopt nicholl_3cd_01a_0516 as the basis for the 50GbE and 100GbE PCS and FEC architecture, with the exception of leaving the FEC lane count / distribution as TBD.
- http://www.ieee802.org/3/cd/public/May16/nicholl_3cd_01a_0516.pdf



Motion #4: 1:42 p.m.
Move to adopt nicholl_3cd_01a_0516 as the basis for the 50GbE and 100GbE PCS and FEC architecture, with the exception of leaving the FEC lane count / distribution as TBD

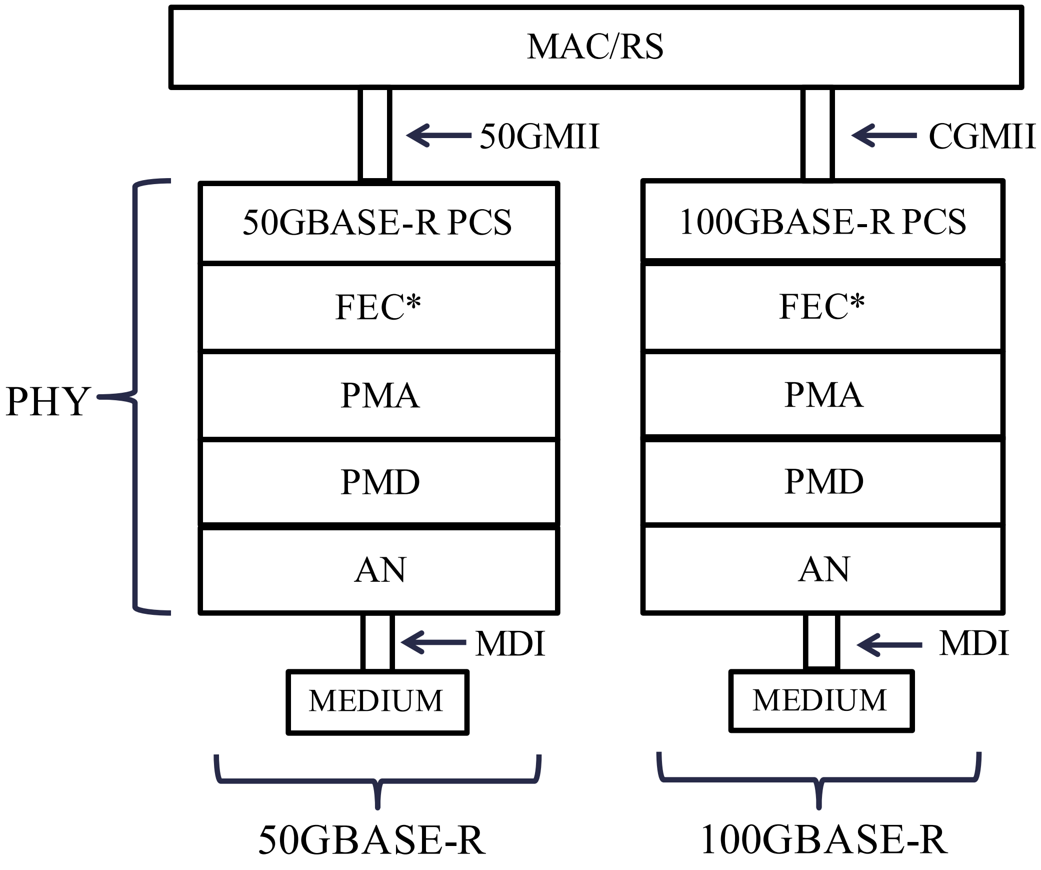
- M: Gary Nicholl
- S: Dave Ofelt
- Technical (>=75%),
- Y: 67 N: 1 A: 12
- Results: passes 1:52 p.m.

Introduction

- This presentation builds upon nicholl_3cd_01a_0516 and provides baseline proposals for the 50GbE and NG 100GbE logic layers
- The primary change for this presentation is related to the FEC lane count:
  - 50GbE: 2x26G FEC lanes; 100GbE 4x26G FEC lanes
  - A bit muxing PMA maps FEC lanes to 53Gb/s PAM4 lanes
  - This change was made to enable a broader range of implementation choices
- Additional changes include:
  - Update to AM to FEC lane mapping to reflect the adoption of 26G FEC lanes
  - PAM4 precoding added to mitigate impact of burst errors (on links with dominant first DFE tap). Optional to enable
  - RS/MII baseline proposal
  - PMA baseline proposal

Architecture Overview

- Based on 802.3ba system architecture
- Separate PCS and FEC sub layers
- PCS and FEC can be separated by an optional AUI (not shown)
- FEC is mandatory for all 802.3cd PHY types
- FEC can be carried over a 25Gb/s or 50Gb/s per lane optional AUI (not shown)
- PMA is based on blind bit muxing



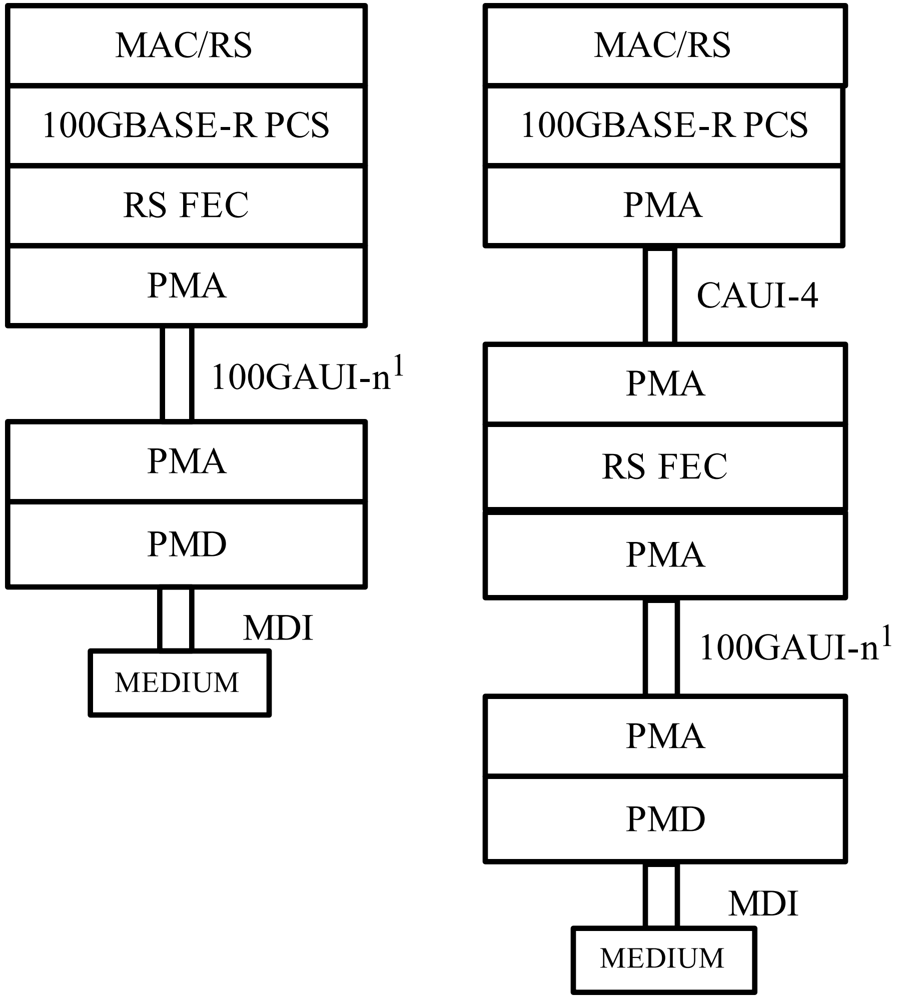
* FEC mandatory for all 802.3cd PHY types

RS/MII Baseline

- RS adapts the bit serial protocols of the MAC to the parallel format of the PCS
- MII provides an optional logical interfaces between the MAC/RS sublayers and the Physical Layer (PHY).
  - not physically instantiated
  - defines a common logical interface for all PHY types
  - often used for connecting RTL blocks within ASICs/FPGAs
- 100G RS and MII are already defined in Clause 81
- 50G RS and MII to be based on Clause 81

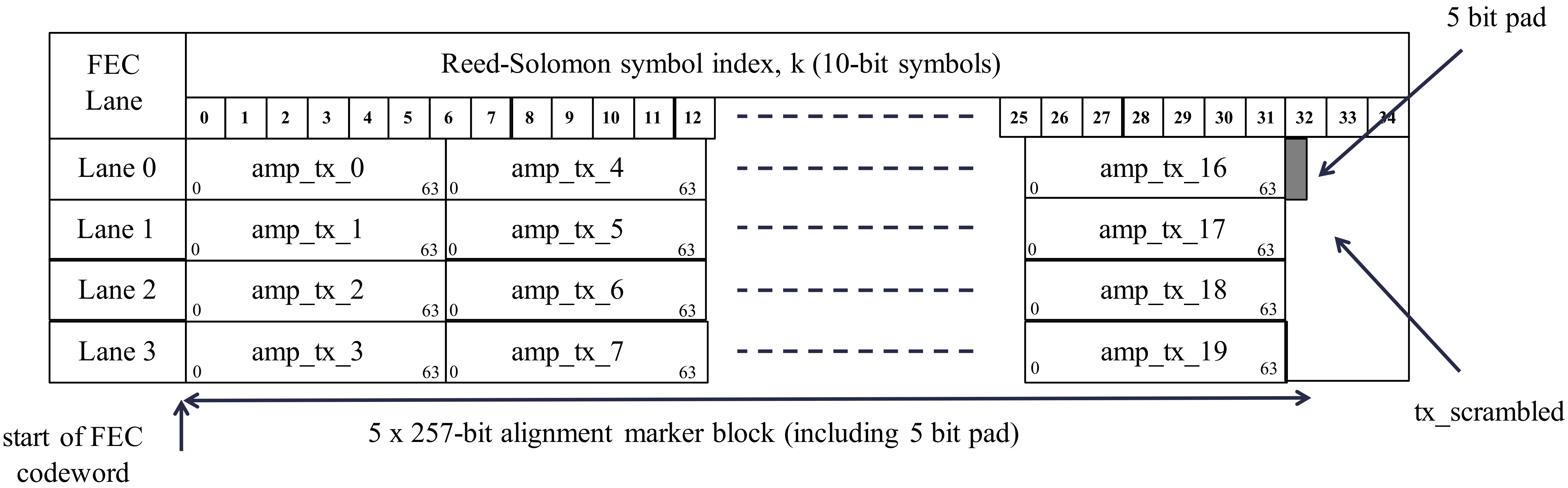
NG 100GbE PCS & FEC Overview

- PCS
  - Re-use existing 100GbE (Clause 82) PCS
  - No changes proposed
- FEC
  - Based on 802.3bj RS(544,514) FEC (Clause 91)
  - Need to modify AM mapping to enable bit muxing
- 100GAUI-n name is a placeholder
  - CAUI-4: 25.78125G per lane (Annex 83 D/E)
  - 100GAUI-4: 26.5625G per lane (based on Annex 120 B/C)
  - 100GAUI-2: 53.125G per lane (based on Annex 120 D/E)



Note 1: n = 2 or 4 lanes

NG 100GbE - Alignment Marker mapping to FEC lanes



5 bit pad

FEC Lane

Reed-Solomon symbol index, k (10-bit symbols)

0 1 2 3 4 5 6 7 8 9 10 11 12 ------------ 25 26 27 28 29 30 31 32 33 34

Lane 0: 0 amp_tx_0 63 | 0 amp_tx_4 63 | ------------ | 0 amp_tx_16 63

Lane 1: 0 amp_tx_1 63 | 0 amp_tx_5 63 | ------------ | 0 amp_tx_17 63

Lane 2: 0 amp_tx_2 63 | 0 amp_tx_6 63 | ------------ | 0 amp_tx_18 63

Lane 3: 0 amp_tx_3 63 | 0 amp_tx_7 63 | ------------ | 0 amp_tx_19 63

tx_scrambled

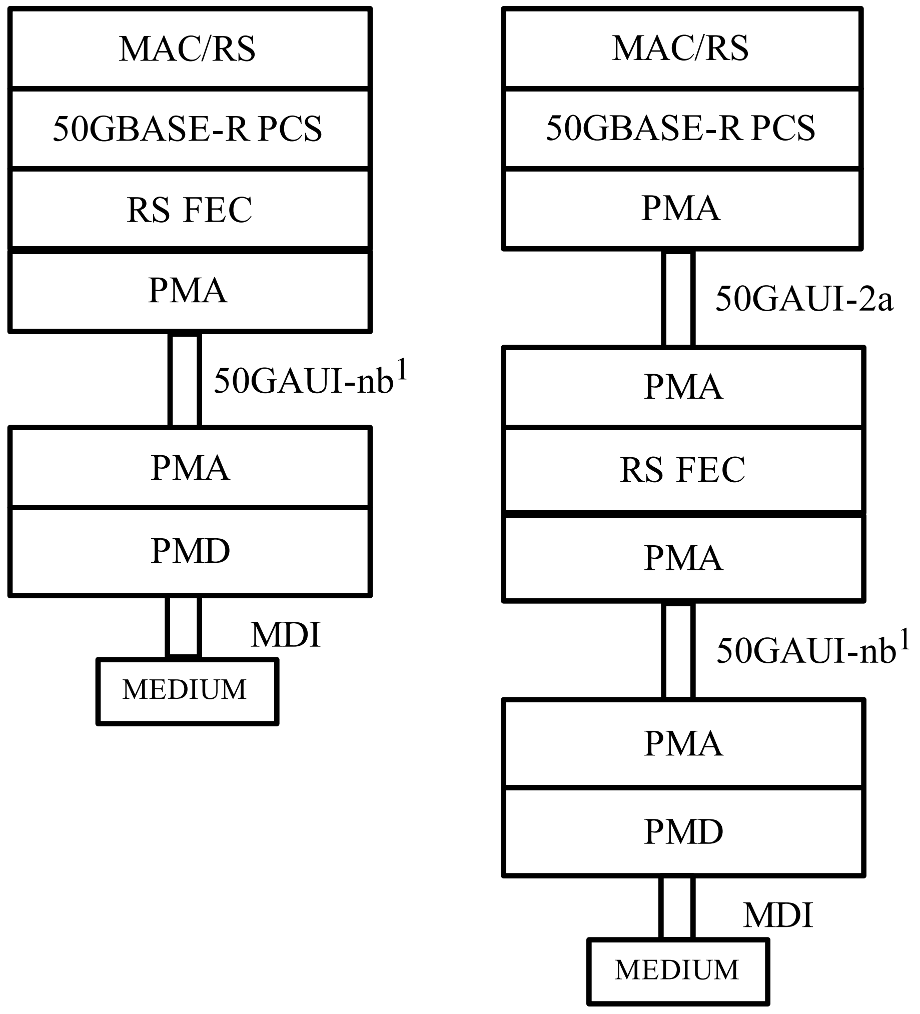
start of FEC codeword

5 x 257-bit alignment marker block (including 5 bit pad)

- Based on Clause 91. Exact AM mapping still TBD
- Initial analysis indicates that Clause 91 AM mapping needs to be modified to avoid clock content issues with repeating AM0 and AM16 patterns when bit muxing FEC lanes.

50GbE PCS & FEC Overview

- PCS
  - Based on overclocked 40GbE PCS (Clause 82)
  - 4 x PCS lanes running at 12.890625 Gb/s
  - AM spacing changed from 16k to 20k* to better support FEC sublayer
- FEC
  - Based on 802.3bj RS(544,514) FEC (Clause 91)
  - FEC symbols distributed to 2 x 26G FEC lanes
  - Modified AM mapping format (4 PCS lanes, 2 FEC lanes, 10b alignment and to enable bit muxing)
- AUI names are placeholders
  - 50GAUI-2a: 25.78125G per lane (based on Annex 83 D/E)
  - 50GAUI-2b: 26.5625G per lane (based on Annex 120 B/C)
  - 50GAUI-1b: 53.125G per lane (based on Annex 120 D/E)

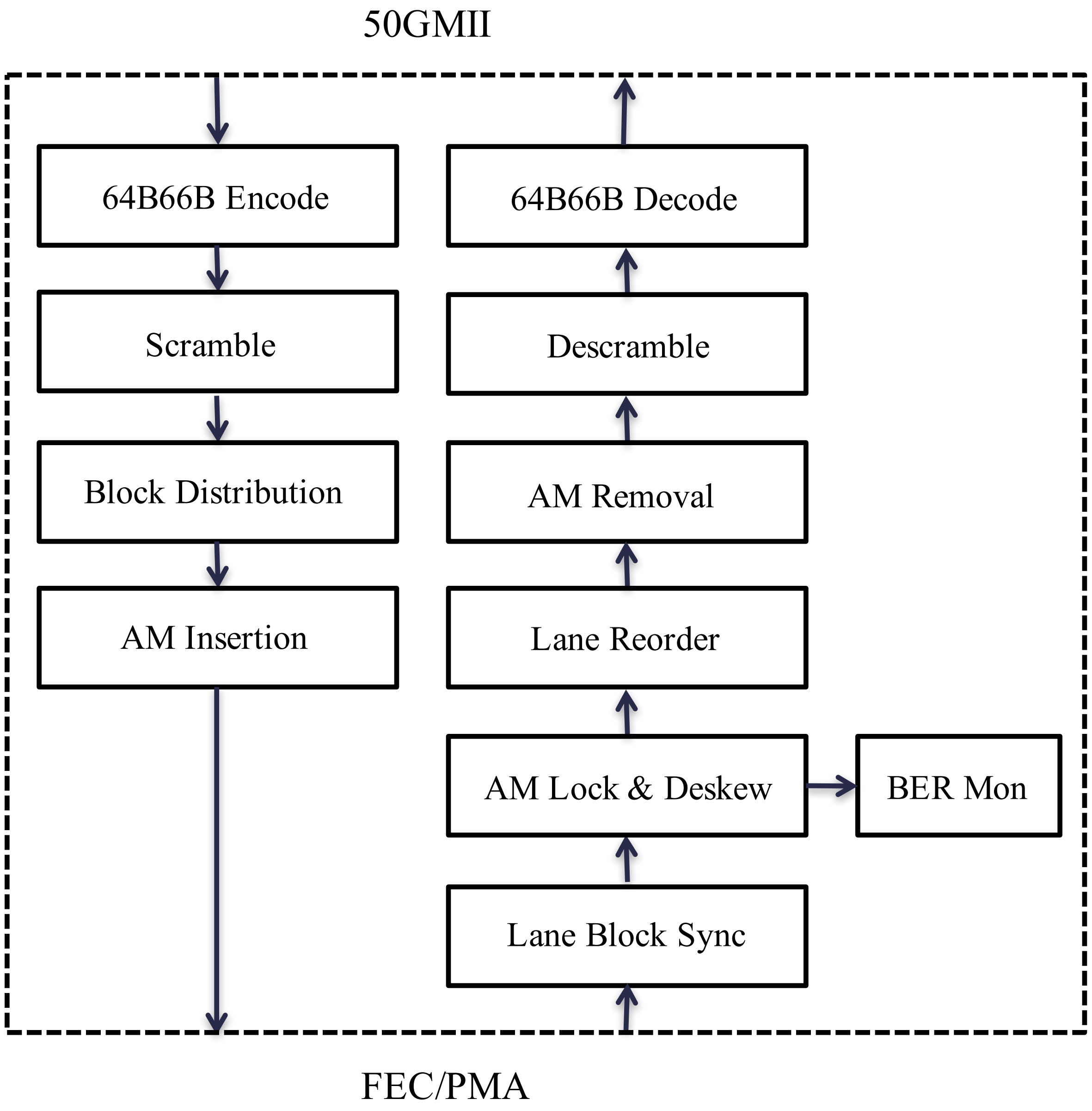


Note 1: n = 1 or 2 lanes

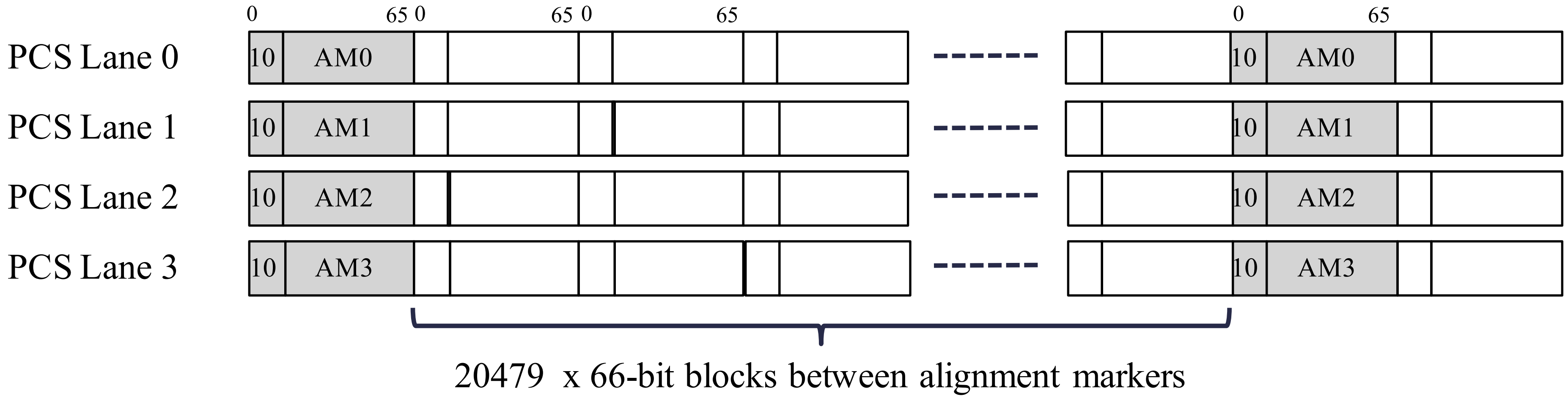
* 20k spacing is consistent with other 50G FEC implementations.

50GbE PCS Data Flow

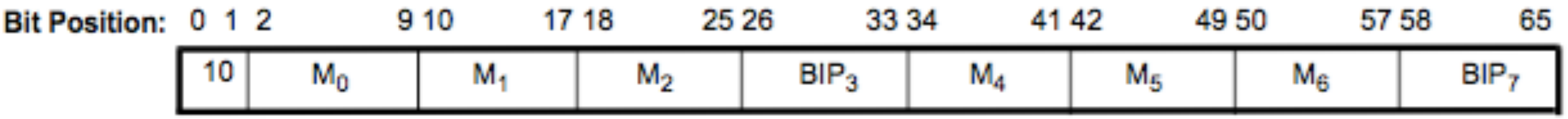
- Identical data flow to 40GbE PCS in Clause 82
- 4 x PCS lanes running at 12.890625 Gb/s (overclocked rate)
- 4 x 66-bit alignment markers (AM), one per PCS lane, inserted periodically
- AM spacing (start of one AM to the start of next AM) modified to 20480 66-bit blocks, to better align with FEC codeword boundaries



50GbE PCS AM Details



50GBASE-R Alignment marker spacing



50GBASE-R Alignment marker format

50GbE PCS AM Details

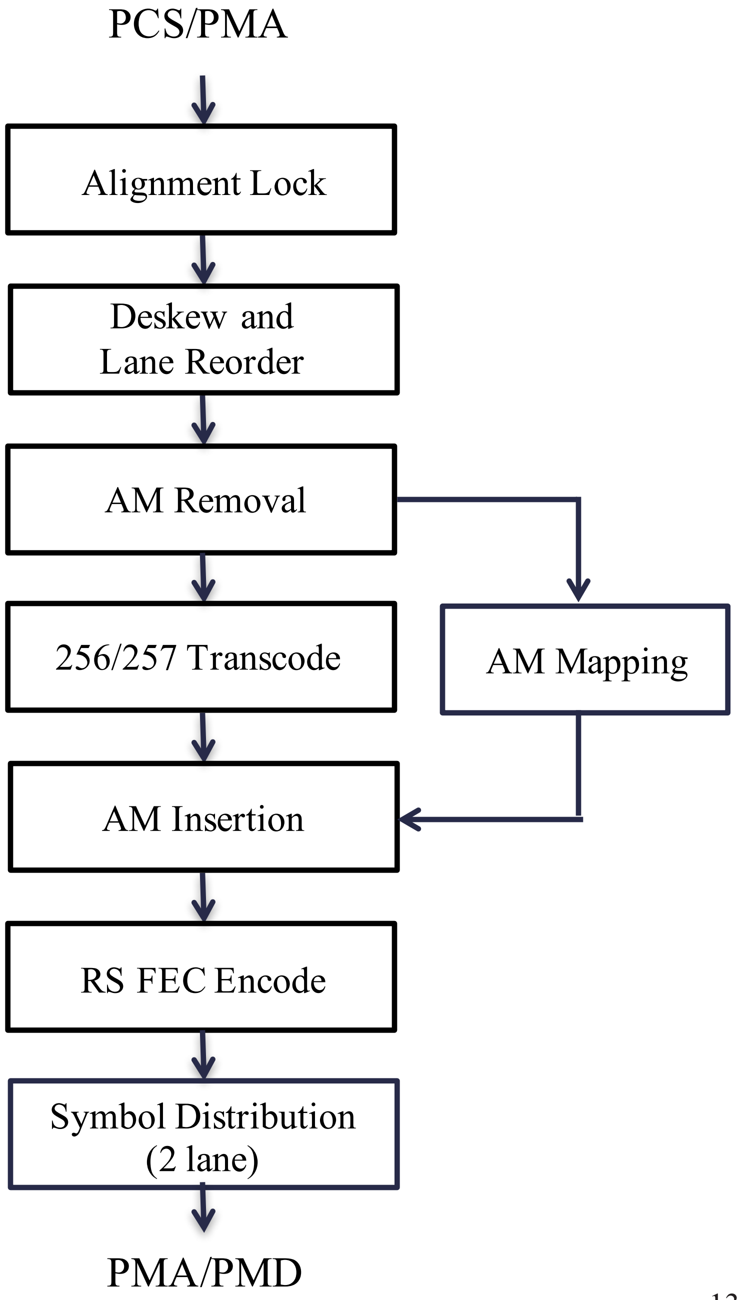
PCS Lane Number	Encoding {M0,M1,M2,BIP3,M4,M5,M6,BIP7}
0	0x90, 0x76, 0x47, BIP3, 0x6F, 0x89, 0xB8, BIP7
1	0xF0, 0xC4, 0xE6, BIP3, 0x0F, 0x3B, 0x19, BIP7
2	0xC5, 0x65, 0x9B, BIP3, 0x3A, 0x9A, 0x64, BIP7
3	0xA2, 0x79, 0x3D, BIP3, 0x5D, 0x86, 0xC2, BIP7

Note: Each octet is transmitted LSB to MSB

50GBASE-R Alignment marker encodings
(identical to 40GBASE-R encodings)

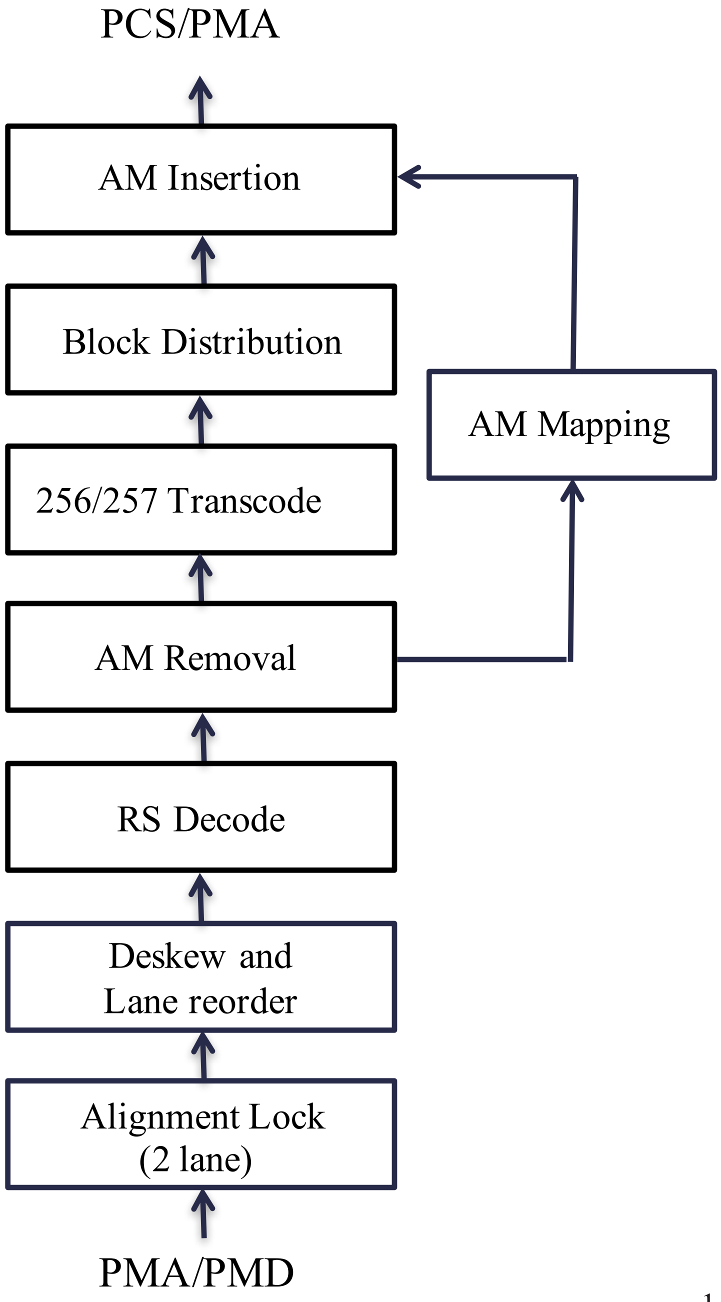
50GbE Tx FEC Data Flow

- Data flow is based on 802.3bj Clause 91
- FEC encoder is RS(544,514) running in a 1x50G configuration
- FEC encoder output is distributed to 2 FEC lanes on a symbol by symbol basis
- A single 257-bit alignment marker (AM) is inserted into the first 257 message bits to be transmitted from every 1024th FEC codeword

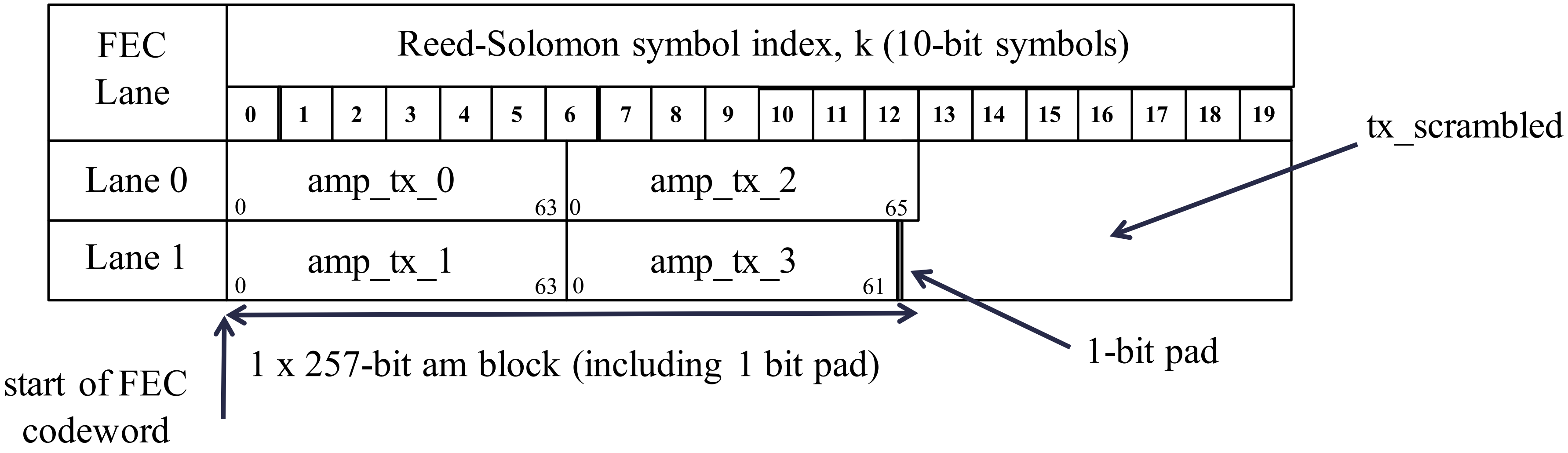


50GbE Rx FEC Data Flow

- Reverse of Tx



50GbE - Alignment Marker mapping to FEC lane



FEC Lane	Reed-Solomon symbol index, k (10-bit symbols)
	0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
Lane 0	0 amp_tx_0 63 | 0 amp_tx_2 65
Lane 1	0 amp_tx_1 63 | 0 amp_tx_3 61

tx_scrambled

1-bit pad

1 x 257-bit am block (including 1 bit pad)

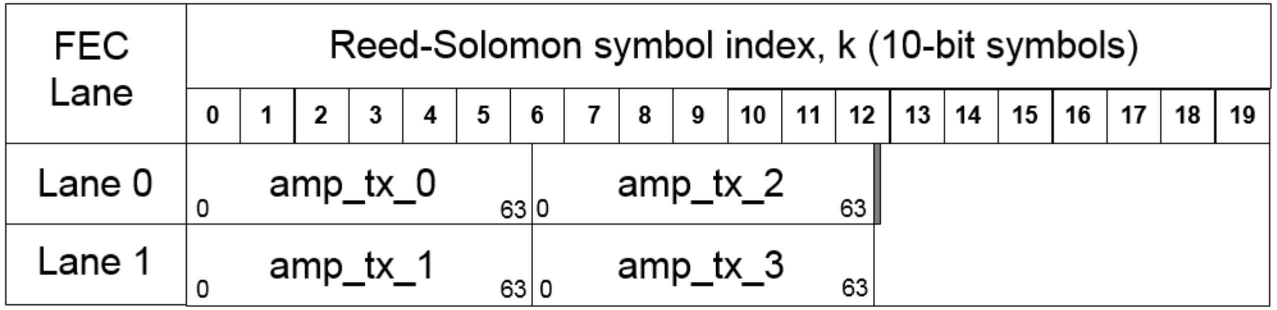
start of FEC codeword

- Based on Clause 91 mapping, but modified to support 4 PCS lanes, 2 FEC lanes, 10b alignment and to enable bit muxing of FEC lanes
- Exact AM mapping still TBD
- Initial proposal for the format of amp_tx_2/3
    - amp_tx_2: am2_m0/am2_m1/am2_m2/am2_bip3/am2_m4/am2_m5/am2_m6/am2_bip7/am3_bip7[0:1]
    - amp_tx_3: am3_m0/am3_m1/am3_m2/am3_bip3/am3_m4/am3_m5/am3_m6/am3_bip7[2:7]

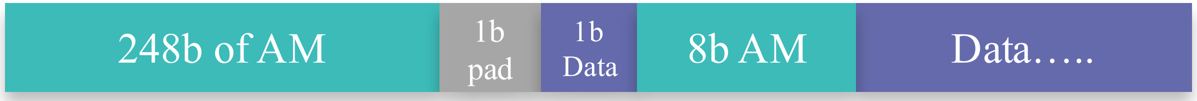
Why 10-bit Alignment is Good on the AMs

- The two figures show the options for how to map AMs into the FEC lanes
- 10-bit alignment simplifies the operation and description

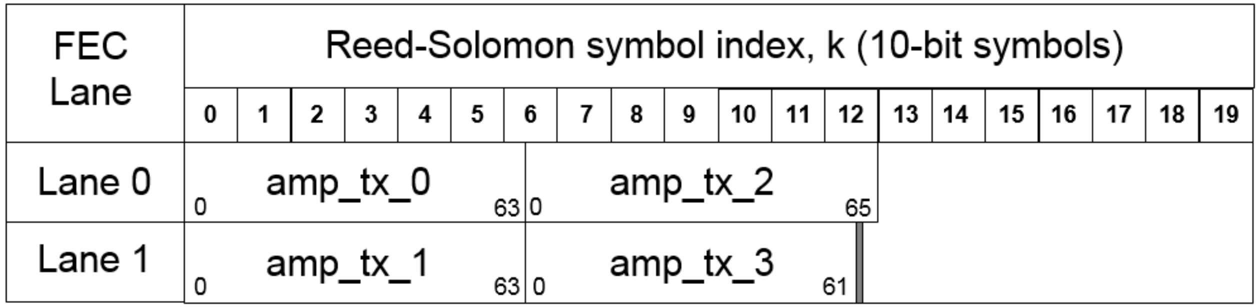
Non 10-bit aligned



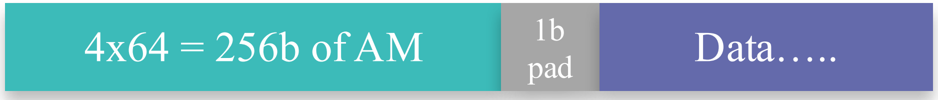
Start of Codeword Data Format



10-bit aligned



Start of Codeword Data Format



PMA Baseline

- Identical PMA functions as described in Clause 120
  - Support for bit muxing
  - Support for Gray coding and PAM4 coding on 50G interfaces
  - Mapping between logical and physical lanes is not defined nor constrained
- PAM4 Precoding
  - Mandatory to implement for Tx on PMAs driving backplane, copper and C2C.
  - Optional to enable (on links with dominant first DFE tap)
  - Precoding is implemented after Gray coding
- With no FEC the per lane signaling rate is 25.78125Gb/s
  - 2x25.78125Gb/s NRZ for 50GbE or 4x25.78125Gb/s NRZ for 100GbE (defined in Clause 83)
- With RS 544 FEC the per lane signaling rate is 26.5625Gb/s or 53.125 Gb/s
  - 2x26.5625Gb/s NRZ for 50GbE or 4x26.5625Gb/s NRZ for 100GbE
  - 1x53.125 Gb/s PAM4 for 50GbE or 2x53.125 Gb/s PAM4 for 100GbE

EEE

- The EEE baseline proposal is being addressed in a separate presentation

Open Issues

- Naming of the AUI interfaces
  - In particular to differentiate between different speeds for the same number of lanes
- Should precoding support be mandatory on Rx?
  - With optional use
- Exact patterns for AMs
  - How much data is repeated
  - Do we keep the same 40GbE AMs for 50GbE?

Conclusion

- This presentation provides 50GbE and NG 100GbE baseline proposals for:
  - RS/MII
  - PCS
  - FEC
  - PMA

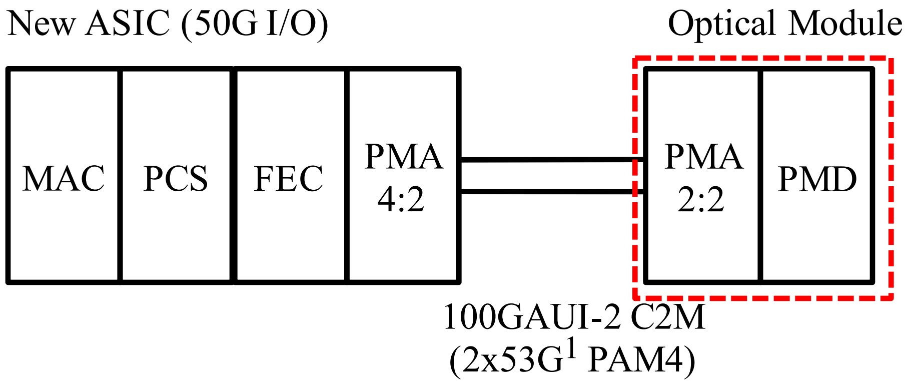
Thanks !!

Backup: Use cases

- The following slides provide examples of use cases that can be supported with the proposed 50GbE and NG 100GbE architecture.

NG 100GbE Use Cases (New Port ASIC)

USE CASE #1



New ASIC (50G I/O)

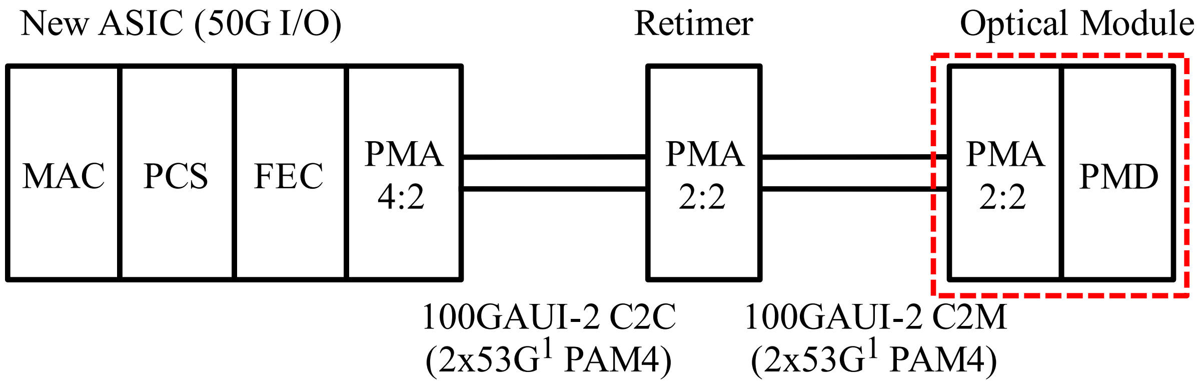
MAC | PCS | FEC | PMA 4:2

Optical Module

PMA 2:2 | PMD

100GAUI-2 C2M
(2x53G^{1} PAM4)

USE CASE #2



New ASIC (50G I/O)

MAC | PCS | FEC | PMA 4:2

Retimer

PMA 2:2

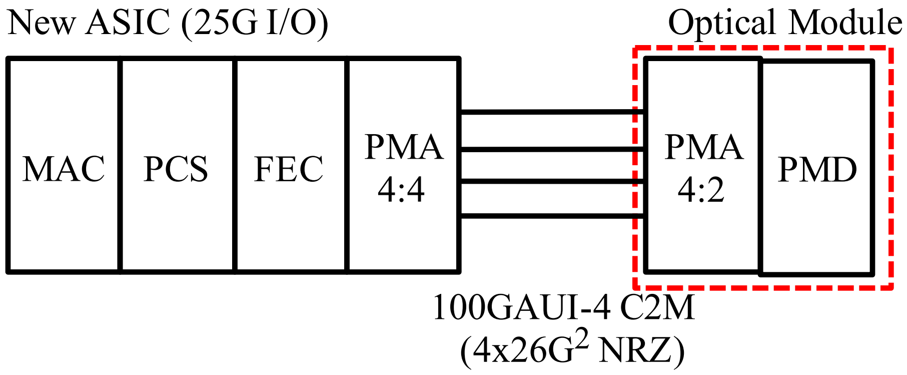
Optical Module

PMA 2:2 | PMD

100GAUI-2 C2C
(2x53G^{1} PAM4)

100GAUI-2 C2M
(2x53G^{1} PAM4)

USE CASE #3



New ASIC (25G I/O)

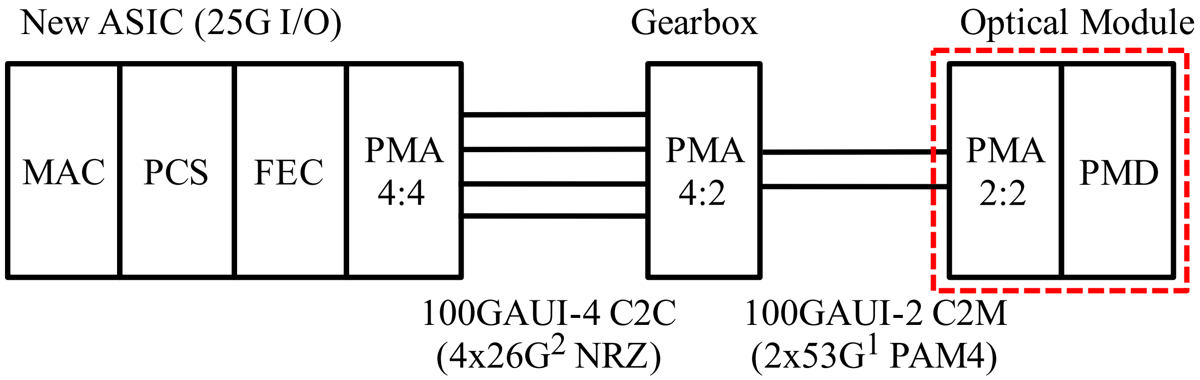
MAC | PCS | FEC | PMA 4:4

Optical Module

PMA 4:2 | PMD

100GAUI-4 C2M
(4x26G^{2} NRZ)

USE CASE #4



New ASIC (25G I/O)

MAC | PCS | FEC | PMA 4:4

Gearbox

PMA 4:2

Optical Module

PMA 2:2 | PMD

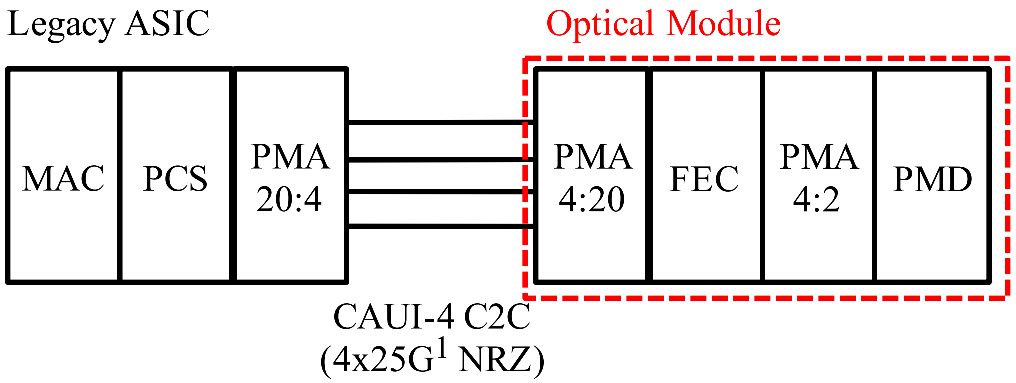
100GAUI-4 C2C
(4x26G^{2} NRZ)

100GAUI-2 C2M
(2x53G^{1} PAM4)

1 = 53.125 Gb/s, 2 = 26.5625Gb/s

NG 100GbE Use Cases (Legacy Port ASIC)

USE CASE #5



Legacy ASIC

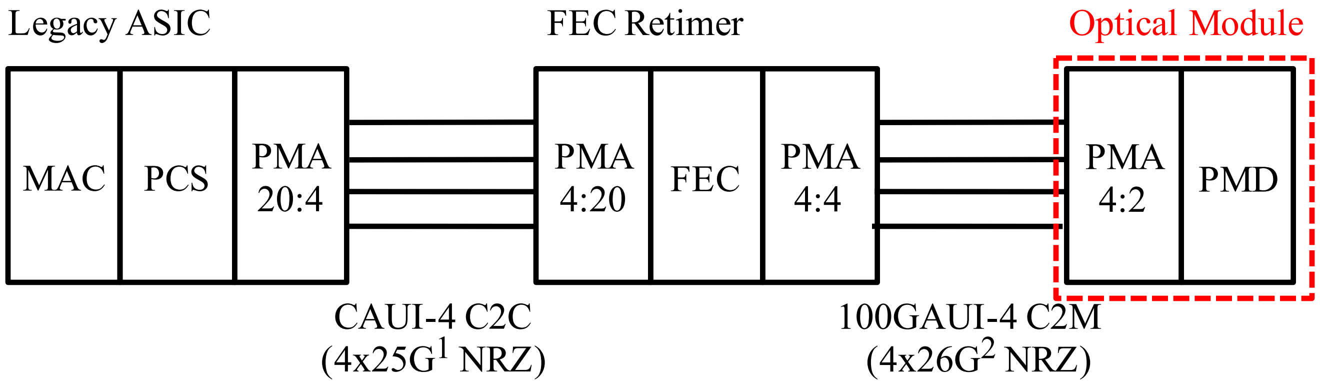
MAC | PCS | PMA 20:4

Optical Module

PMA 4:20 | FEC | PMA 4:2 | PMD

CAUI-4 C2C
(4x25G[1] NRZ)

USE CASE #6



Legacy ASIC

MAC | PCS | PMA 20:4

FEC Retimer

PMA 4:20 | FEC | PMA 4:4

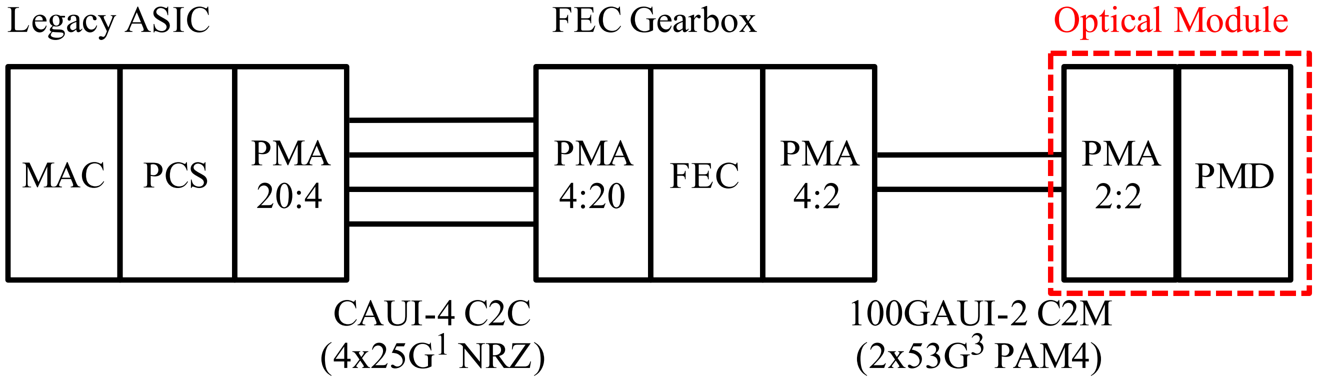
Optical Module

PMA 4:2 | PMD

CAUI-4 C2C
(4x25G[1] NRZ)

100GAUI-4 C2M
(4x26G[2] NRZ)

USE CASE #7



Legacy ASIC

MAC | PCS | PMA 20:4

FEC Gearbox

PMA 4:20 | FEC | PMA 4:2

Optical Module

PMA 2:2 | PMD

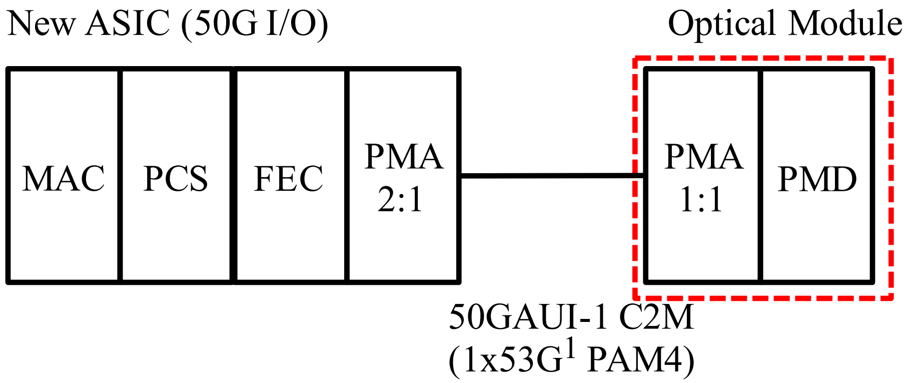
CAUI-4 C2C
(4x25G[1] NRZ)

100GAUI-2 C2M
(2x53G[3] PAM4)

1 = 25.78125Gb/s , 2 = 26.5625Gb/s, 3 = 53.125 Gb/s

50GbE Use Cases (New Port ASIC)

USE CASE #1



New ASIC (50G I/O)

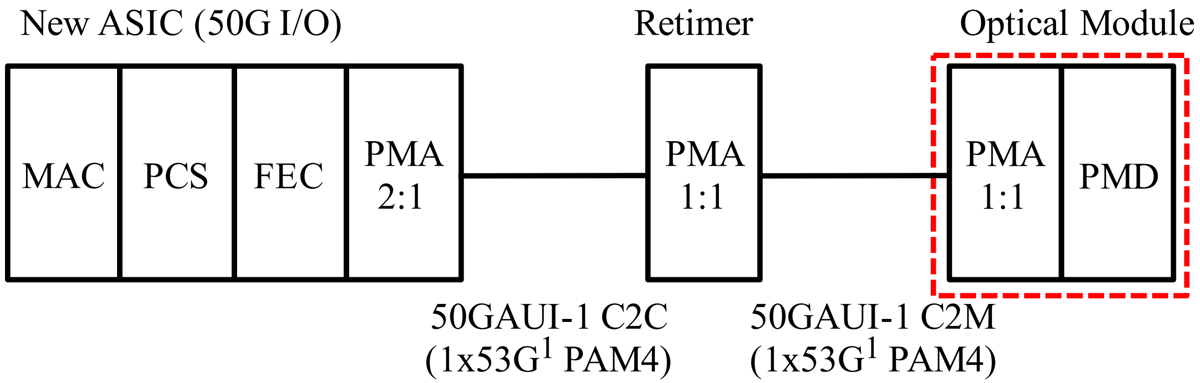
MAC | PCS | FEC | PMA 2:1

Optical Module

PMA 1:1 | PMD

50GAUI-1 C2M
(1x53G^1 PAM4)

USE CASE #2



New ASIC (50G I/O)

MAC | PCS | FEC | PMA 2:1

Retimer

PMA 1:1

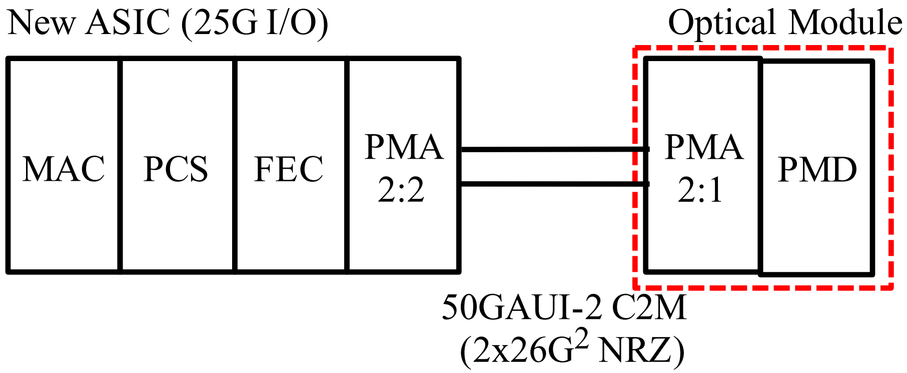
Optical Module

PMA 1:1 | PMD

50GAUI-1 C2C
(1x53G^1 PAM4)

50GAUI-1 C2M
(1x53G^1 PAM4)

USE CASE #3



New ASIC (25G I/O)

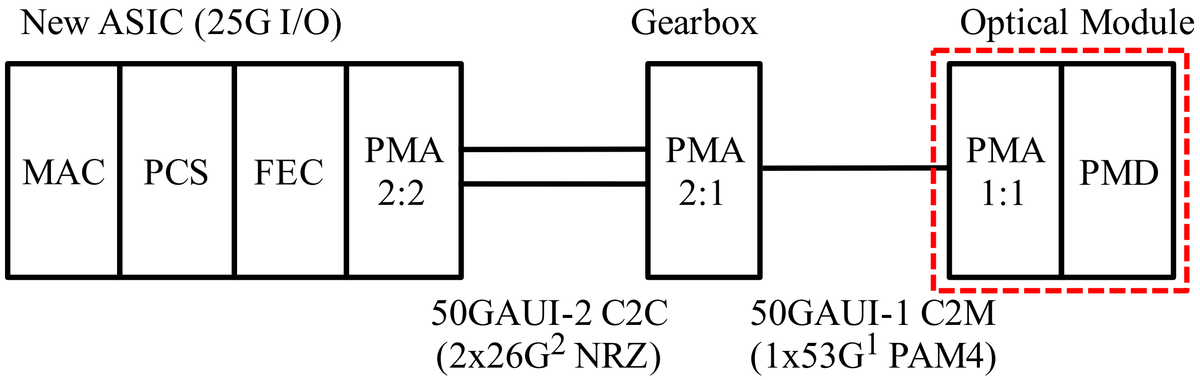
MAC | PCS | FEC | PMA 2:2

Optical Module

PMA 2:1 | PMD

50GAUI-2 C2M
(2x26G^2 NRZ)

USE CASE #4



New ASIC (25G I/O)

MAC | PCS | FEC | PMA 2:2

Gearbox

PMA 2:1

Optical Module

PMA 1:1 | PMD

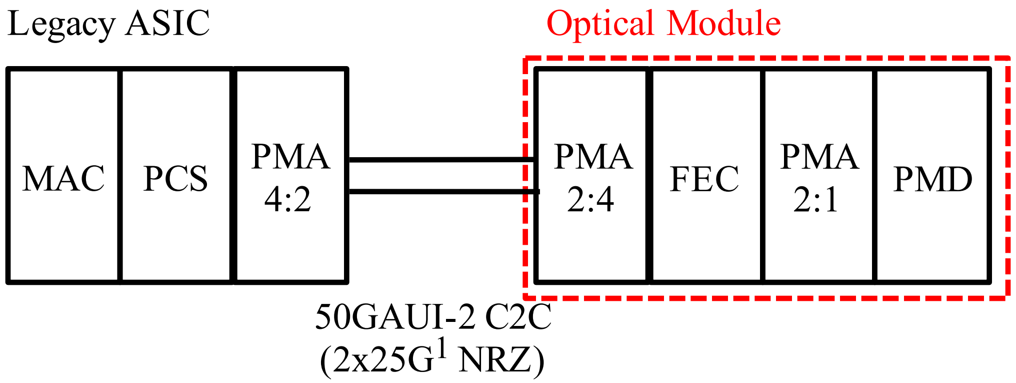
50GAUI-2 C2C
(2x26G^2 NRZ)

50GAUI-1 C2M
(1x53G^1 PAM4)

1 = 53.125 Gb/s, 2 = 26.5625Gb/s

50GbE Use Cases (“Legacy” Port ASIC - PCS only)

USE CASE #5


Legacy ASIC

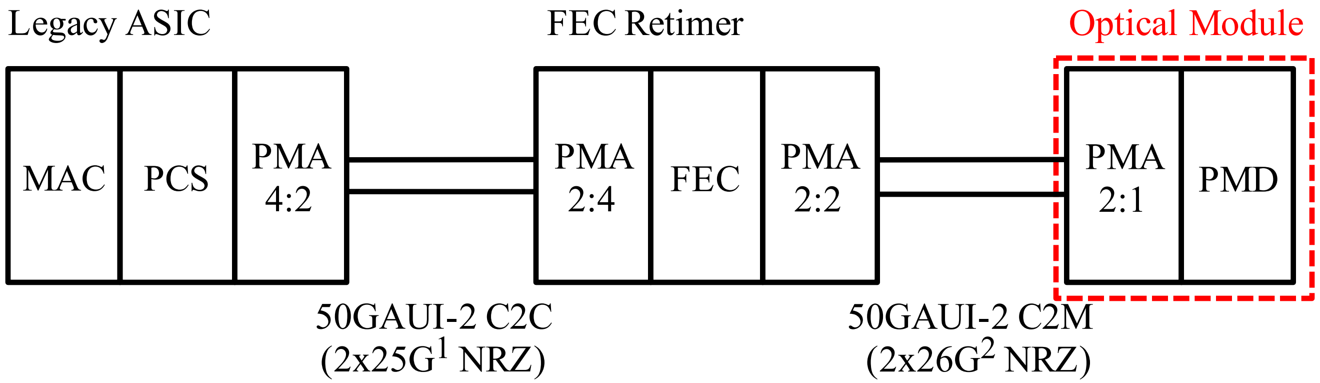
Optical Module

MAC | PCS | PMA 4:2

PMA 2:4 | FEC | PMA 2:1 | PMD

50GAUI-2 C2C
(2x25G[1] NRZ)

USE CASE #6


Legacy ASIC

FEC Retimer

Optical Module

MAC | PCS | PMA 4:2

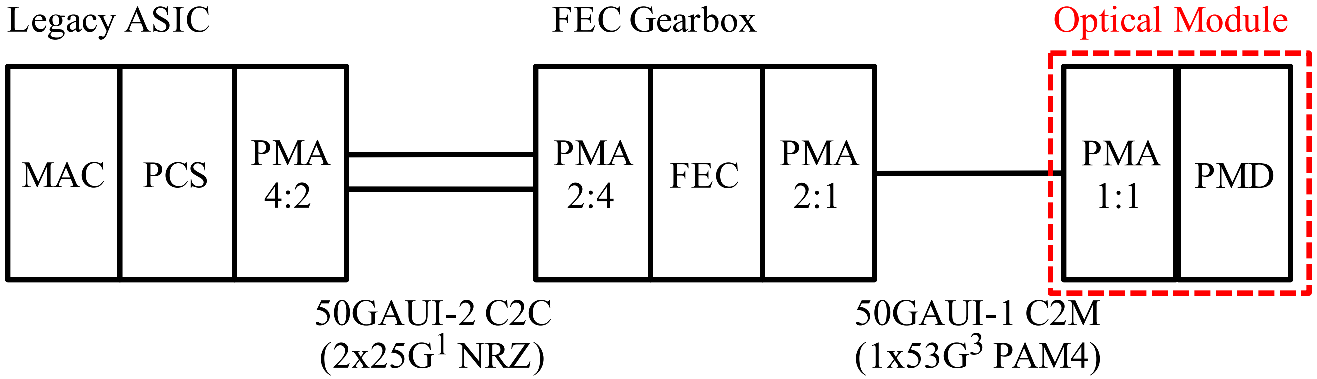
PMA 2:4 | FEC | PMA 2:2

PMA 2:1 | PMD

50GAUI-2 C2C
(2x25G[1] NRZ)

50GAUI-2 C2M
(2x26G[2] NRZ)

USE CASE #7


Legacy ASIC

FEC Gearbox

Optical Module

MAC | PCS | PMA 4:2

PMA 2:4 | FEC | PMA 2:1

PMA 1:1 | PMD

50GAUI-2 C2C
(2x25G[1] NRZ)

50GAUI-1 C2M
(1x53G[3] PAM4)

1 = 25.78125Gb/s , 2 = 26.5625Gb/s, 3 = 53.125 Gb/s

Backup: PMA Examples

50GbE PMA Examples:

PMA (2:2): 2x25G NRZ retimer (for connecting the PCS with the FEC sublayer)
PMA (2:2): 2x26G NRZ retimer (for connecting RS 544 FEC sublayer to a PMA 2:1)
PMA (2:1): 2x26G NRZ to 1x53G PAM4 (to connect RS 544 FEC sublayer to a PMA 1:1 or PMD)
PMA (1:1): 1x53G PAM4 retimer (to go between a PMA 1:1 and a PMD)

100GbE PMA Examples:

PMA (4:4): 4x25G NRZ retimer (for connecting the PCS with the FEC sublayer)
PMA (4:4): 4x26G NRZ retimer (for connecting RS 544 FEC to a PMA 4:2 sublayer)
PMA (4:2): 4x26G NRZ to 2x53G PAM4 (to connect RS 544 FEC sublayer to a PMA 2:2 or a PMD)
PMA (2:2): 2x53G PAM4 retimer (to go between a PMA 2:2 and a PMA 2:2 or a PMD)