100GE and 40GE PCS (MLD) Proposal

IEEE 802.3ba May 2008 Munich

Contributors and Supporters

David Law – 3com
Steve Trowbridge - Alcatel-Lucent
Jesse Simsarian - Alcatel-Lucent
Brad Booth – AMCC
Dimitrios Giannakopoulos – AMCC
Francesco Caggioni – AMCC
Keith Conroy – AMCC
Piers Dawe – Avago
Rita Horner – Avago
Howard Frazier - Broadcom
Arthur Marris – Cadence
Mike Shahine - Ciena
Mark Nowell - Cisco
Gary Nicholl - Cisco
Hugh Barrass - Cisco
Steve Swanson - Corning
Med Belhadj – Cortina
Chris Cole - Finisar
Krishnamurthy Subramanian – Force10

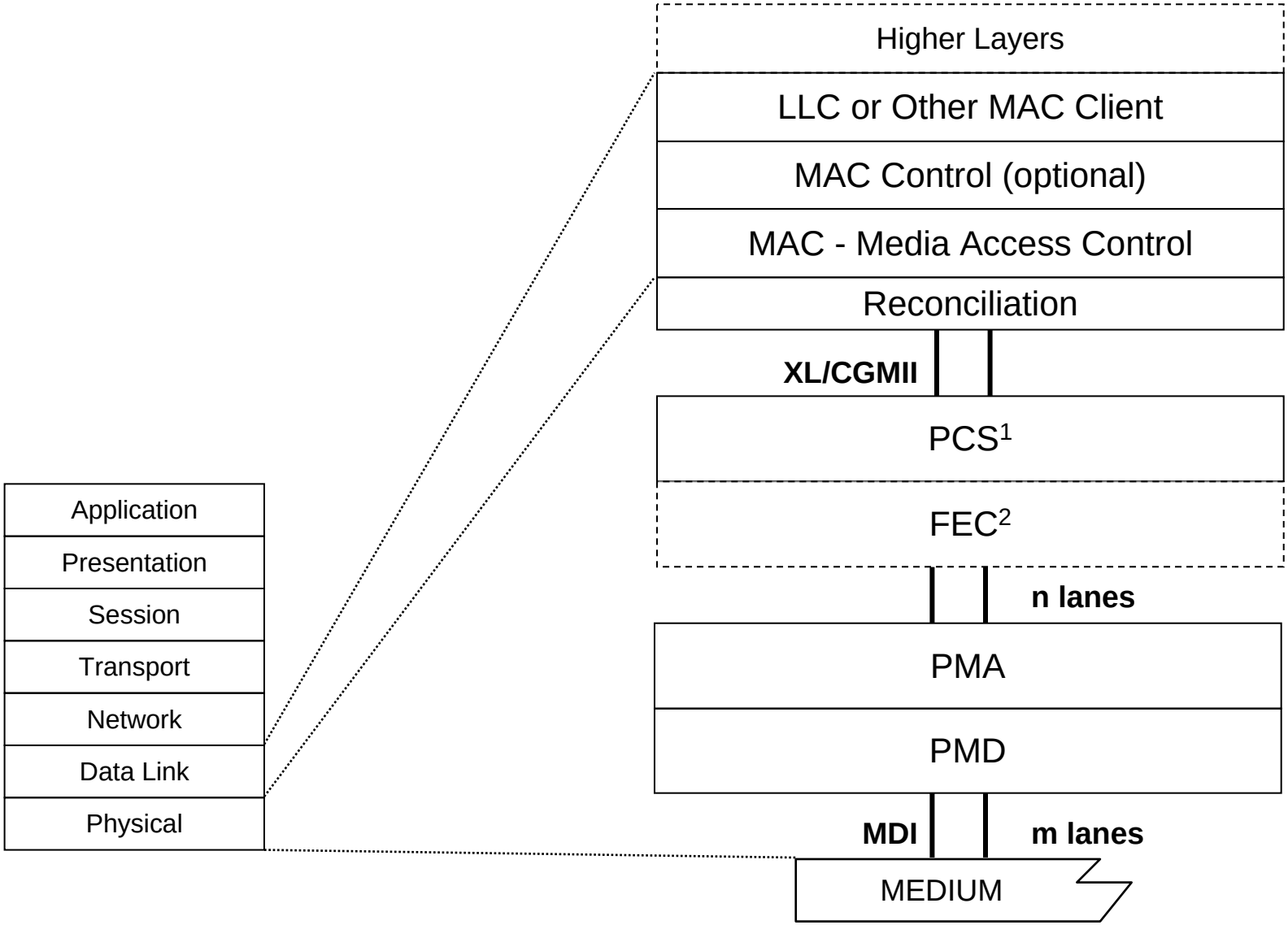
Aris Wong – Foundry Networks
Shashi Patel – Foundry Networks
Bill Ryan – Foundry Networks
Ryan Latchman - Gennum
Justin Abbott - Gennum
Hong Liu – Google
Ashby Armistead – Google
Shinji Nishimura - Hitachi Ltd
Hidehiro Toyoda - Hitachi Ltd
Dan Dove – HP
Petar Pepeljugoski – IBM
John Jaeger - Infinera
Andy Moorwood - Infinera
Drew Perkins - Infinera
Jerry Pepper - Ixia
Thananya Baldwin - Ixia
Faisal Dada - JDSU
Jack Jewell - JDSU
Mike Dudek - JDSU

Jeffery J. Maki - Juniper Networks
David Ofelt - Juniper Networks
Brad Turner - Juniper Networks
Adam Healey - LSI
Martin White – Marvell
Andy Weitzner – Marvell
Pete Anslow – Nortel
David W. Martin – Nortel
Osamu Ishida - NTT
Shoukei Kobayashi - NTT
Matt Traverso – Opnext
Farhad Shafai - Sarance Technologies
Farzin Firoozmand – SMI
Craig Hornbuckle – SMI
Song Shang - SMI
Ted Seely - Sprint
Kengo Matsumoto - Sumitomo Electric
Shimon Muller - Sun
Andre Szczepanek – TI
Martin Carroll - Verizon
Frank Chang - Vitesse

Agenda

- 40GE/100GE Architecture
- PCS and MLD layer details
- Possible XL/CGMII Interface
- Alignment details
- Alignment performance metrics
- Clocking example
- Skew
- Summary

40GE/100GE Generic Architecture



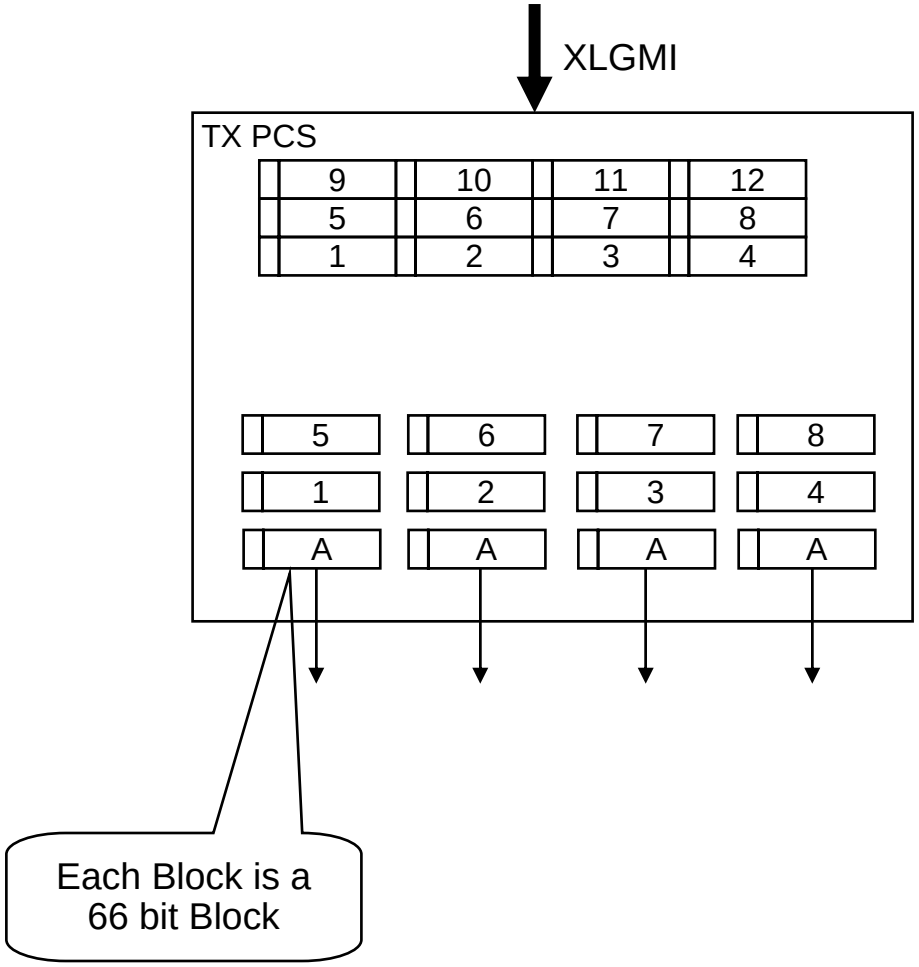
1: Includes MLD functionality 2: For 40GE Backplane

Proposed 100GE/40GE PCS

- 10GBASE-R 64B/66B based PCS
  - Run at 100Gbps or 40Gbps serial rate
  - Includes 66 bit block encoding and scrambling
- Multi-Lane Distribution
  - Data is distributed across n virtual lanes 66 bit blocks at a time
  - Round robin distribution
  - Periodic alignment blocks are added to each virtual lane to allow deskew in the rx PCS
- PMA maps n lanes to m lanes
  - PMA is simple bit level muxing
  - Does not know or care about PCS coding
- Alignment and static skew compensation is done in the Rx PCS only

Striping Mechanism

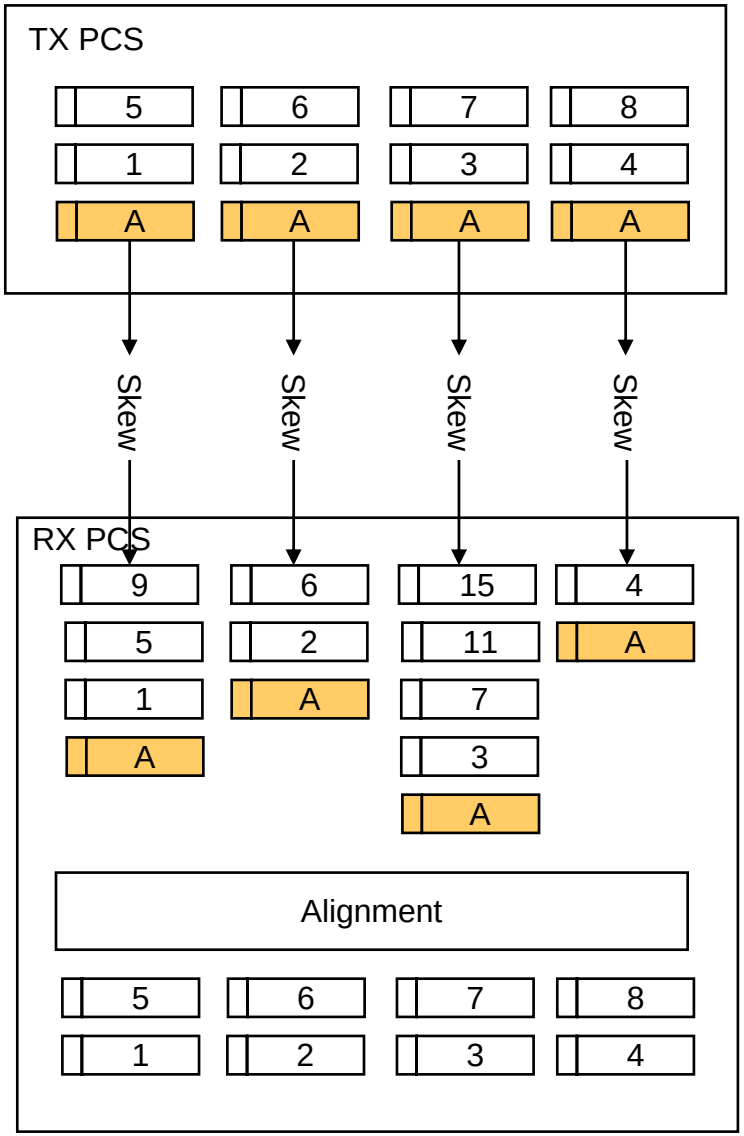
This example is 40GE with 4 electrical and 4 optical lanes



PCS Functions:

- 66 bit encoding
- Scrambling
- Periodic alignment block addition
- Round robin block distribution

Alignment Mechanism – 40GE Example



RX PCS Functions:

- Re-Align 66 bit blocks
- Remove the Alignment blocks
- Then descramble and decode

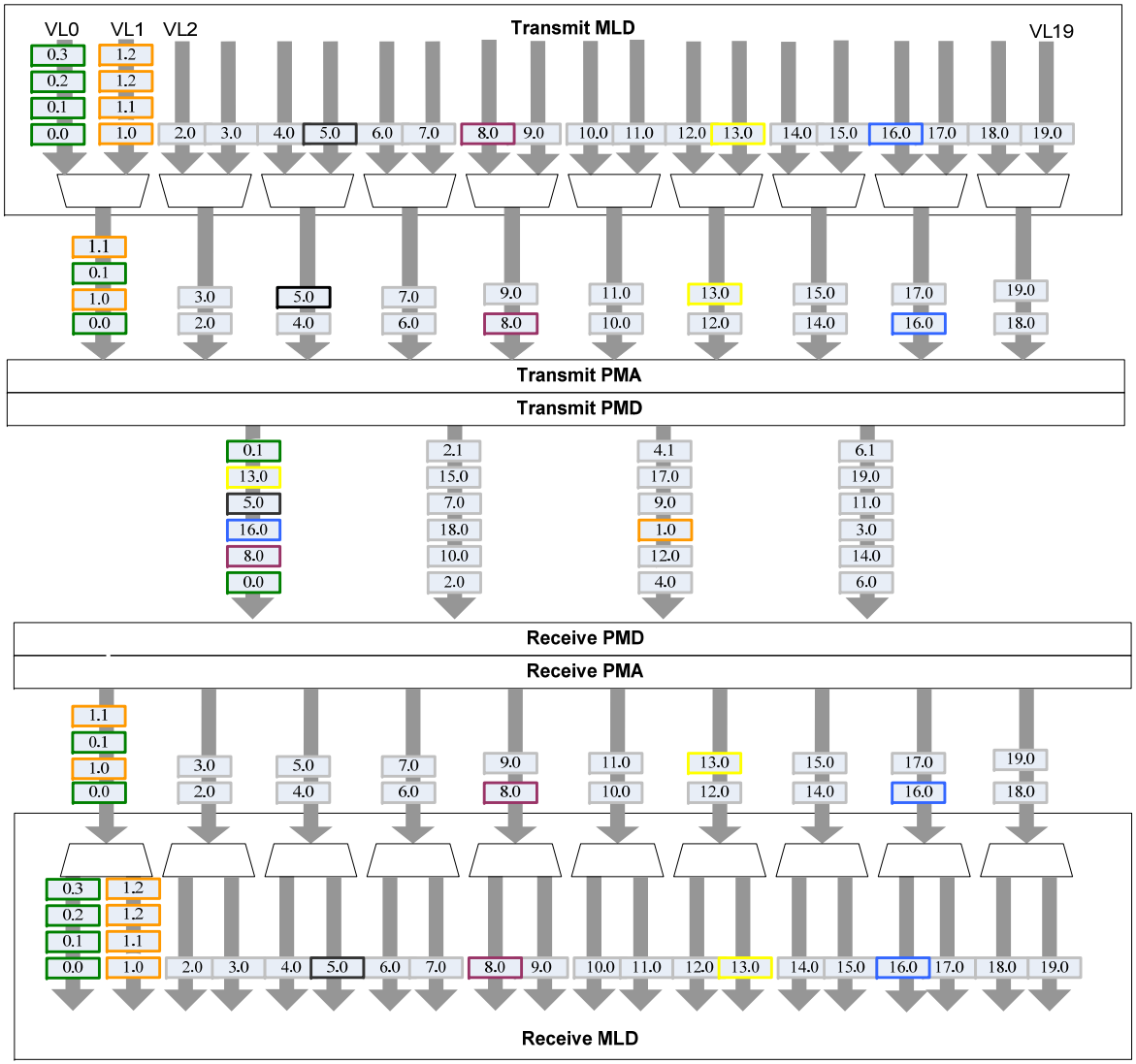
Key Concept – Virtual Lanes

- Virtual lanes may or may not correspond to physical lanes
- Virtual lanes are created by distributing PCS encoded data in a round robin fashion, on a 66 bit block basis
- The number of virtual lanes generated is scaled to the Least Common Multiple (LCM) of the n lane electrical interface and the m lane PMD

  This allows all data (bits) from one virtual lane to be transmitted over the same electrical and optical lane combination

  This ensures that the data from a virtual lane is always received with the correct bit order at the Rx MLD
- The alignment markers allow the Rx PCS to perform skew compensation, realign all the virtual lanes, and reassemble a single 100G or 40G aggregate stream (with all the 64B/66B blocks in the correct order)

Bit Flow Through – 100GE 4 lane PMD



- 20 VLs
- 10 Electrical lanes
- 4 Optical lanes
- With Skew, VLs move around
- RX MLD puts things back in order

How Many Virtual Lanes are Needed?

- **4 VLs For 40GE, this covers all of the possible combinations of lanes:**

Electrical Lane Widths	PMD Lane Widths	Virtual Lanes Needed
4, 2, 1	4, 2, 1	4

- **20 VLs For 100GE, this covers all of the possible combinations of lanes:**

Electrical Lane Widths	PMD Lane Widths	Virtual Lanes Needed
10, 5, 4, 2, 1	10, 5, 4, 2, 1	20

PCS Encoding

- Same 10GBASE-R PCS (Clause 49) encoding

Input Data	Sync	Block Payload							
Bit Position: / Data Block Format:	0 1	2							65
$D_0 D_1 D_2 D_3/D_4 D_5 D_6 D_7$	01	D_0	D_1	D_2	D_3	D_4	D_5	D_6	D_7
Control Block Formats:		Block Type Field							
$C_0 C_1 C_2 C_3/C_4 C_5 C_6 C_7$	10	0x1e	C_0	C_1	C_2	C_3	C_4	C_5	C_6	C_7
~~$C_0 C_1 C_2 C_3/C_4 D_5 D_6 D_7$~~	~~10~~	~~0x2d~~	~~C_0~~	~~C_1~~	~~C_2~~	~~C_3~~	~~C_4~~	~~D_5~~	~~D_6~~	~~D_7~~
~~$C_0 C_1 C_2 C_3/D_4 D_5 D_6 D_7$~~	~~10~~	~~0x33~~	~~C_0~~	~~C_1~~	~~C_2~~	~~C_3~~		~~D_5~~	~~D_6~~	~~D_7~~
~~$O_0 D_1 D_2 D_3/D_4 D_5 D_6 D_7$~~	~~10~~	~~0x66~~	~~D_1~~	~~D_2~~	~~D_3~~	~~O_0~~		~~D_5~~	~~D_6~~	~~D_7~~
~~$O_0 D_1 D_2 D_3/O_4 D_5 D_6 D_7$~~	~~10~~	~~0x55~~	~~D_1~~	~~D_2~~	~~D_3~~	~~O_0~~	~~O_4~~	~~D_5~~	~~D_6~~	~~D_7~~
$S_0 D_1 D_2 D_3/D_4 D_5 D_6 D_7$	10	0x78	D_1	D_2	D_3	D_4		D_5	D_6	D_7
$O_0 D_1 D_2 D_3/C_4 C_5 C_6 C_7$	10	0x4b	D_1	D_2	D_3	O_0	C_4	C_5	C_6	C_7
$T_0 C_1 C_2 C_3/C_4 C_5 C_6 C_7$	10	0x87		C_1	C_2	C_3	C_4	C_5	C_6	C_7
$D_0 T_1 C_2 C_3/C_4 C_5 C_6 C_7$	10	0x99	D_0		C_2	C_3	C_4	C_5	C_6	C_7
$D_0 D_1 T_2 C_3/C_4 C_5 C_6 C_7$	10	0xaa	D_0	D_1		C_3	C_4	C_5	C_6	C_7
$D_0 D_1 D_2 T_3/C_4 C_5 C_6 C_7$	10	0xb4	D_0	D_1	D_2		C_4	C_5	C_6	C_7
$D_0 D_1 D_2 D_3/T_4 C_5 C_6 C_7$	10	0xcc	D_0	D_1	D_2	D_3		C_5	C_6	C_7
$D_0 D_1 D_2 D_3/D_4 T_5 C_6 C_7$	10	0xd2	D_0	D_1	D_2	D_3	D_4		C_6	C_7
$D_0 D_1 D_2 D_3/D_4 D_5 T_6 C_7$	10	0xe1	D_0	D_1	D_2	D_3	D_4	D_5		C_7
$D_0 D_1 D_2 D_3/D_4 D_5 D_6 T_7$	10	0xff	D_0	D_1	D_2	D_3	D_4	D_5	D_6	

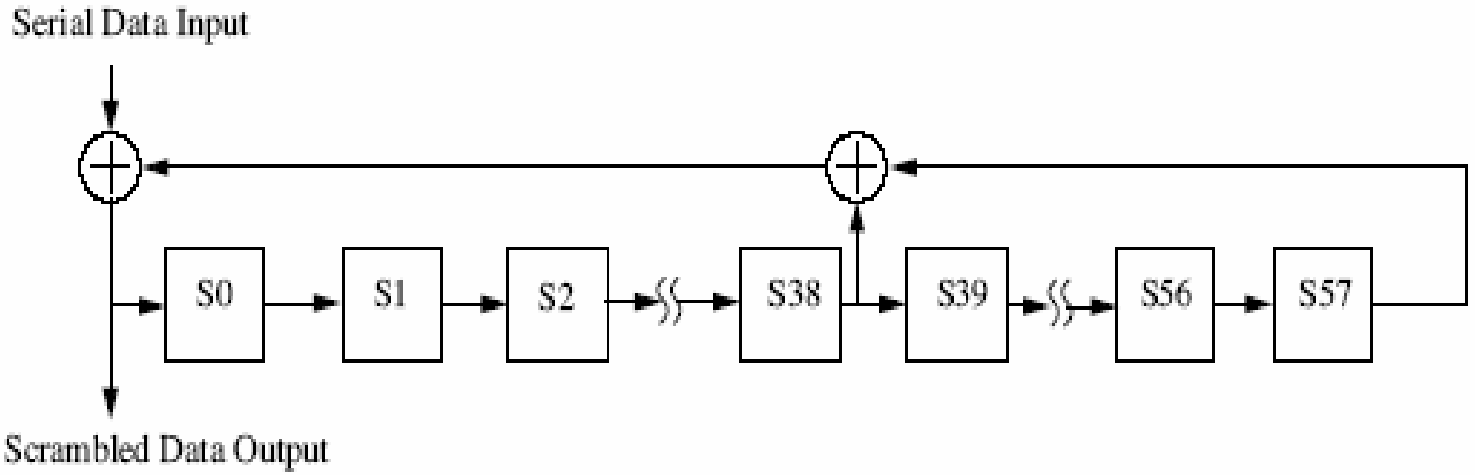
Not used since we have 8B alignment

Only block type used for ordered sets

PCS Scrambling

- Identical 10GBASE-R PCS (Clause 49) scrambler

  Runs at 40Gbps or 100Gbps now



PCS Idle Deletion/Insertion rules

- Straight from 802.3ae (except for highlighted text):

  Idle insertion or deletion occurs in groups of eight Idle characters

  Idle characters are added following idle or ordered_sets

  Idle characters are not added while data is being received

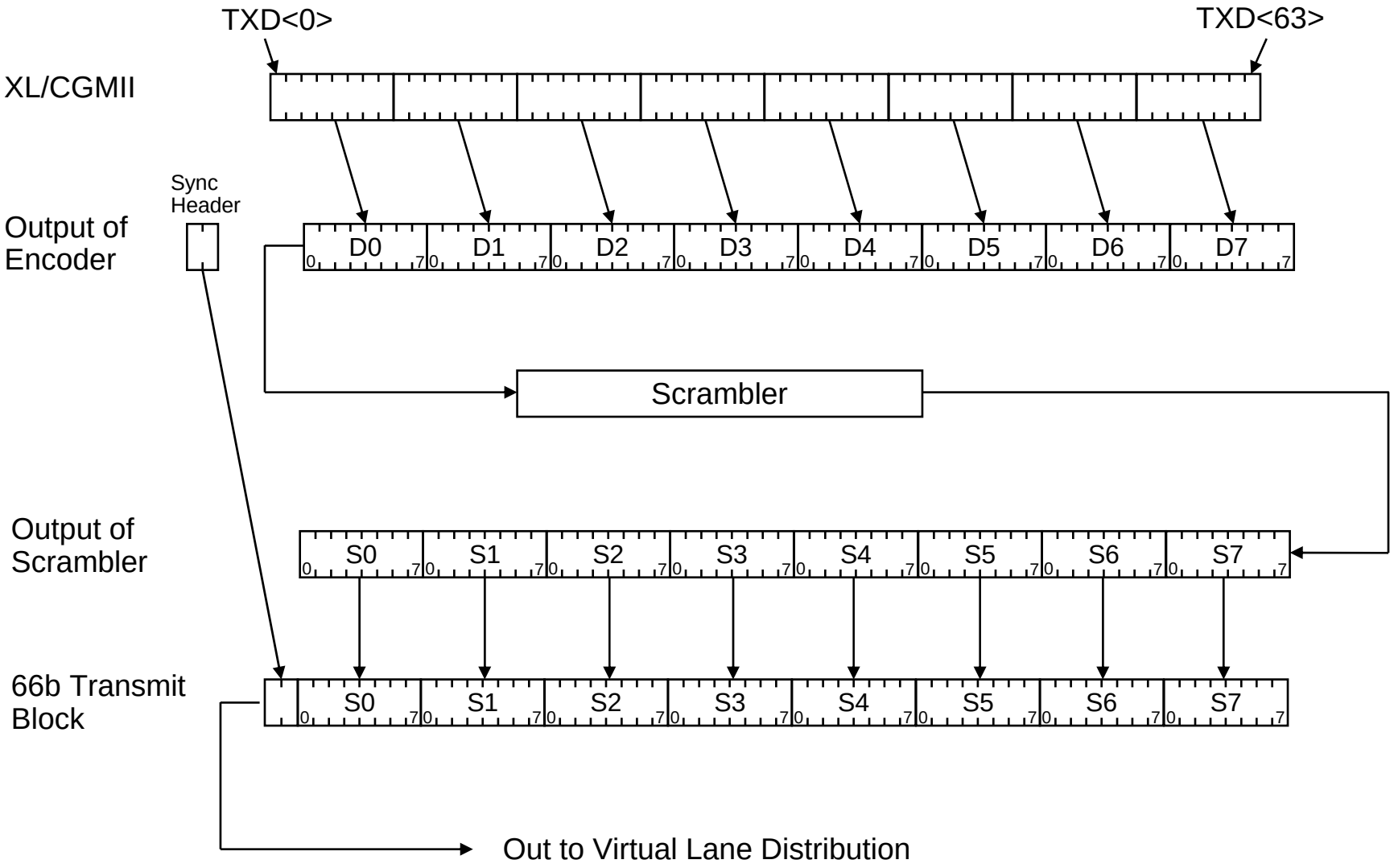
  When deleting idles, the minimum IPG of one character is maintained

  Sequence ordered_sets are deleted to adapt between clock rates

  Sequence ordered_set deletion occurs only when two consecutive sequence ordered_sets have been received and deletes only one of the two

  Only idles are inserted for clock compensation

PCS Bit Order



Alignment Proposal

- Send alignment on a fixed time basis
- Alignment word also identifies virtual lanes
- Sent every 16384 66bit blocks on each virtual lane at the same time
  - ~216usec for 20 VLs @ 100G
  - ~108usec for 4 VLs @ 40G
- It temporarily interrupts packets
- Takes only 0.006% (60PPM) of the Bandwidth
- Rate Adjust FIFO will delete enough IPG so that the MAC still runs at 100.000G or 40.000G with the interface running at 10.3125G

Alignment Word Proposal

Requirements:

- Significant transitions and DC balanced – word is not scrambled
- Keep in 66 bit form, but no relation to 10GBASE-R is needed
- But why not keep it close? – Because of the clock wander concerns
- Contains Virtual Lane Identifier

Proposed Alignment Word

10	VL	~VL

- This is DC balanced
- No relationship to the normal 10GBASE-R blocks
- Added after and removed before 64/66 processing
- Alignment block is periodic, no Hamming distance concerns with 64/66 block types

Alignment Word Proposal – 100GE

The encoding of the VL markers is as follows (based on $x^{58} + x^{39} + 1$ scrambler output):

VL Number	32 Bit encoding	VL Number	32 Bit encoding
0	C1,68,21,F4	10	FD, 6C, 99, DE
1	9D, 71, 8E, 17	11	B9, 91, 55, B8
2	59, 4B, E8, B0	12	5C, B9, B2, CD
3	4D, 95, 7B, 10	13	1A, F8, BD, AB
4	F5, 07, 09, 0B	14	83, C7, CA, B5
5	DD, 14, C2, 50	15	35, 36, CD, EB
6	9A, 4A, 26, 15	16	C4, 31, 4C, 30
7	7B, 45, 66, FA	17	AD, D6, B7, 35
8	A0, 24, 76, DF	18	5F, 66, 2A, 6F
9	68, C9, FB, 38	19	C0, F0, E5, E9

Note that data is played out in VL order, 0, 1, 2, …19, 0, 1…

Alignment Word Proposal – 40GE

The encoding of the VL markers is as follows (based on $x^{58} + x^{39} + 1$ scrambler output):

VL Number	32 Bit encoding
0	C1,68,21,F4
1	9D, 71, 8E, 17
2	59, 4B, E8, B0
3	4D, 95, 7B, 10

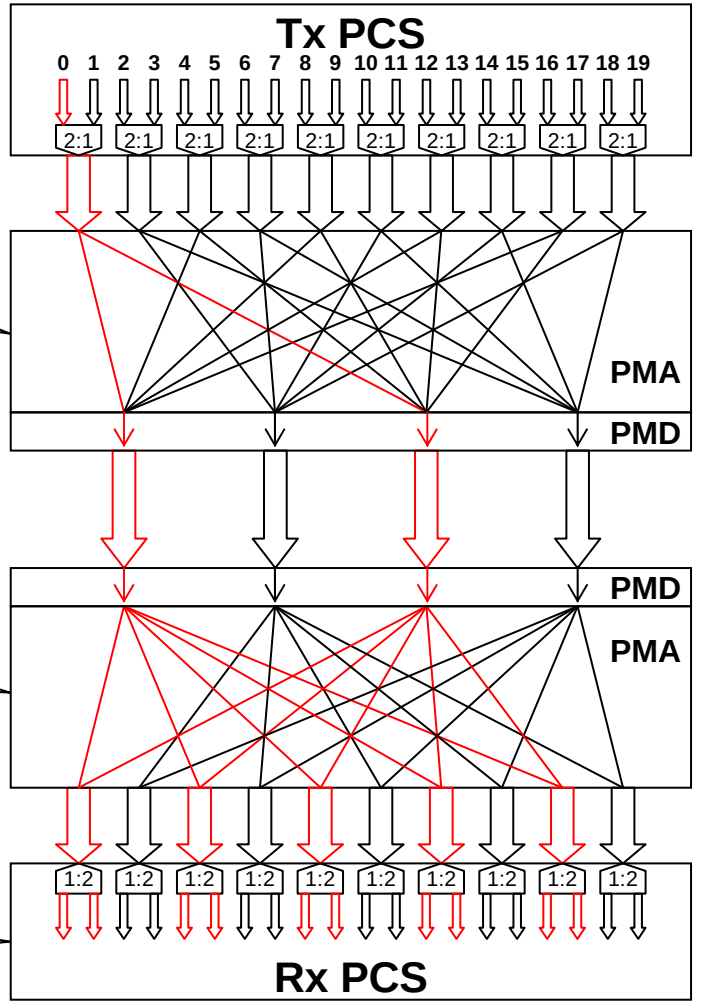
Note that data is played out in VL order, 0, 1, 2, 3, 0…

Possible Paths Through the Link

Skew on input electrical interface determines which optical lane VL 0 passes through

Skew on optical interface and previous electrical interface determines which output electrical lane VL 0 passes through

VL 0 can appear on any of the red outputs depending on the skew of the electrical and optical interfaces



Note: These possible paths are based on a 10:4 and 4:10 function based on round-robin distribution. Other arrangements which give different paths are possible.

Virtual Lane Location on the Receive Side

- Due to how virtual lanes are multiplexed, and due to skew, and in order to be future proof:
  - All receivers must support receiving a transmitted virtual lane on any received virtual lane
  - This is true for 100GE and 40GE

Finding VL Alignment

- After reception in the rx MLD, you have x VLs, each skewed and transposed
- First you find 66bit alignment on each VL

  Each VL is a stream of 66 bit blocks

  Same mechanism as 10GBASE-R (64 valid 2 bit frame codes in a row)
- Then you hunt for alignment on each VL

  Look for one of the 20 VL patterns repeated and inverted
- Alignment is declared on each VL after finding 2 consecutive non-errored alignment patterns in the expected locations (16k words apart)
- Out of alignment is declared on a VL after finding 4 consecutive errored frame patterns
- Once the alignment pattern is found on all VLs, then the VLs can be aligned

Alignment Performance Parameters – 100GE

- Mean Time To Alignment (MTTA)

  Mean time it takes to gain Alignment on a lane or virtual lane for a given BER

  Nominal time = 314usec

- Mean Time To Loss of Alignment (MTTLA)

  Mean time it takes to lose Alignment on a lane or virtual lane for a given BER

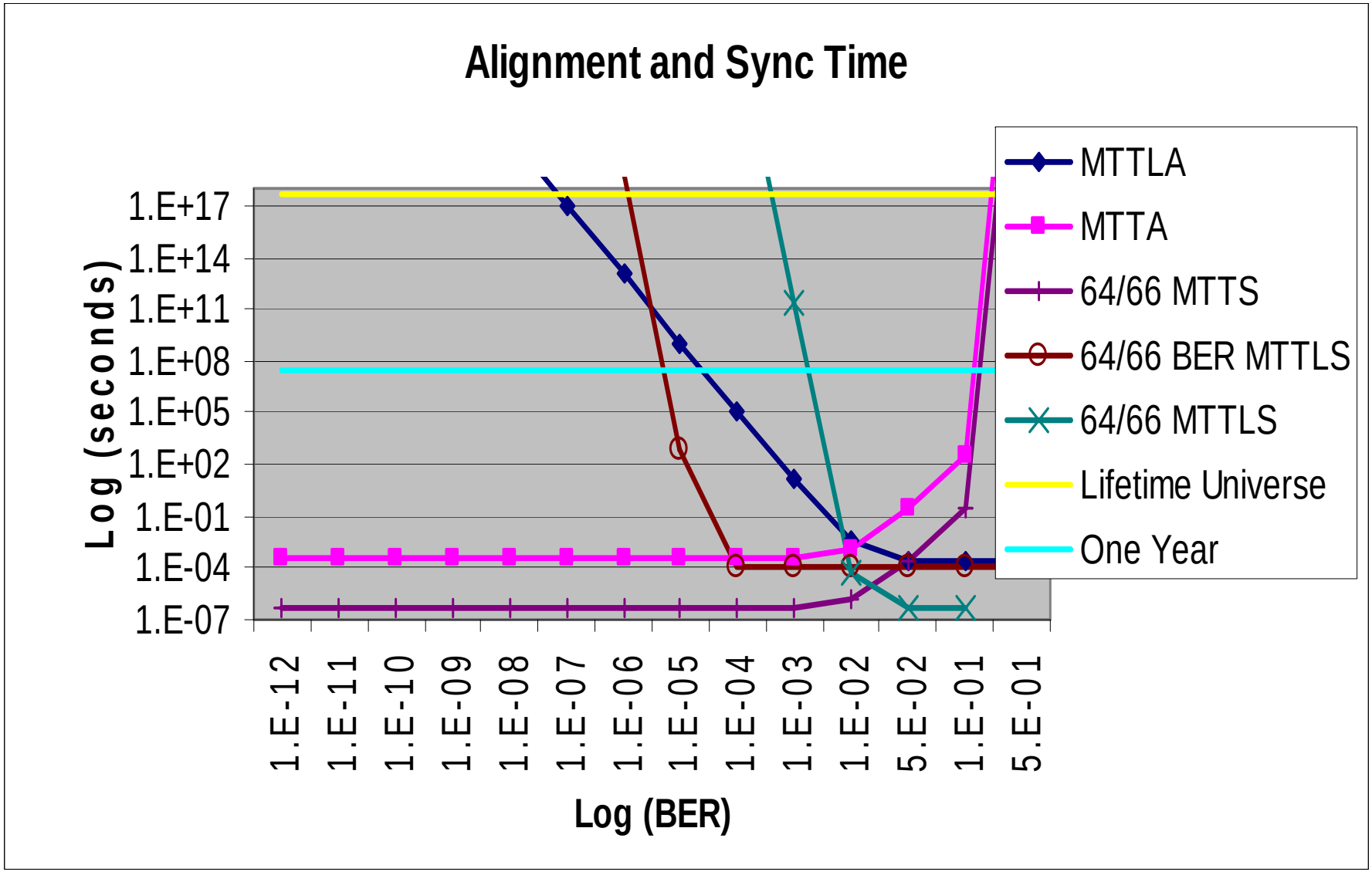
- Probability of False Alignment (PFA) = 3 E-40
- Probability of Rejecting False Alignment (PRFA) = ~1
- Also have 64/66 sync stats on the graph for comparison

  MTTS – Mean Time To Sync (64 non errored syncs in a row)

  BER MTTLS – With the 125usec BER window, what is the Mean Time To Lose Sync

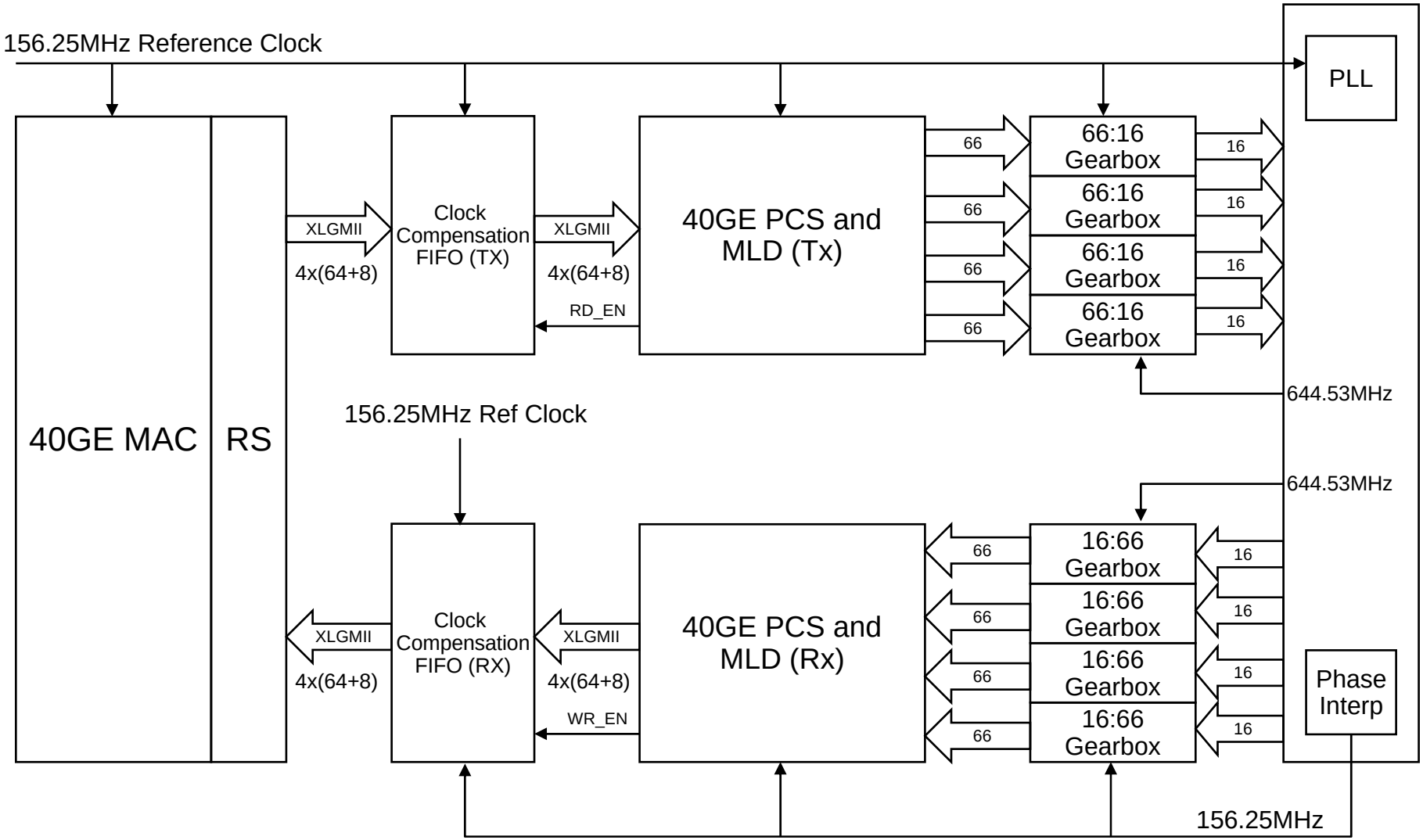
  MTTLS - Mean Time To Lose Sync

Alignment Performance Parameters – 100GE



40GE Alignment Performance will be similar

Clocking Example – 40GE



Skew Handling

- Both dynamic and static skew budgets need to be identified
- See other presentations for details

Summary

- Simple 10GBASE-R based PCS
- MLD layer to support multiple physical lanes/lambdas
- Complexity is low within the MLD layer
  - Simple block data striping
- Complexity in the optical module is low
  - Simple bit muxing even when m != n
- Based on proven 64B/66B framing and scrambling
- Electrical interface is feasible at 10x10G or 4x10G
- Allows for a MAC rate of 100.000G or 40.000G
  - Overhead very low and independent of packet size
- Supports an evolution of optics and electrical interfaces