100GE/40GE skew budget for MLD

IEEE 802.3ba
Denver, July 2008

Mark Gustlin – Cisco
Pete Anslow – Nortel
Dimitrios Giannakopoulos - AMCC

Supporters

- **Gary Nicholl – Cisco**
- **Farhad Shafai – Sarance**
- **Francesco Caggioni, Brad Booth – AMCC**
- **Chris Cole - Finisar**

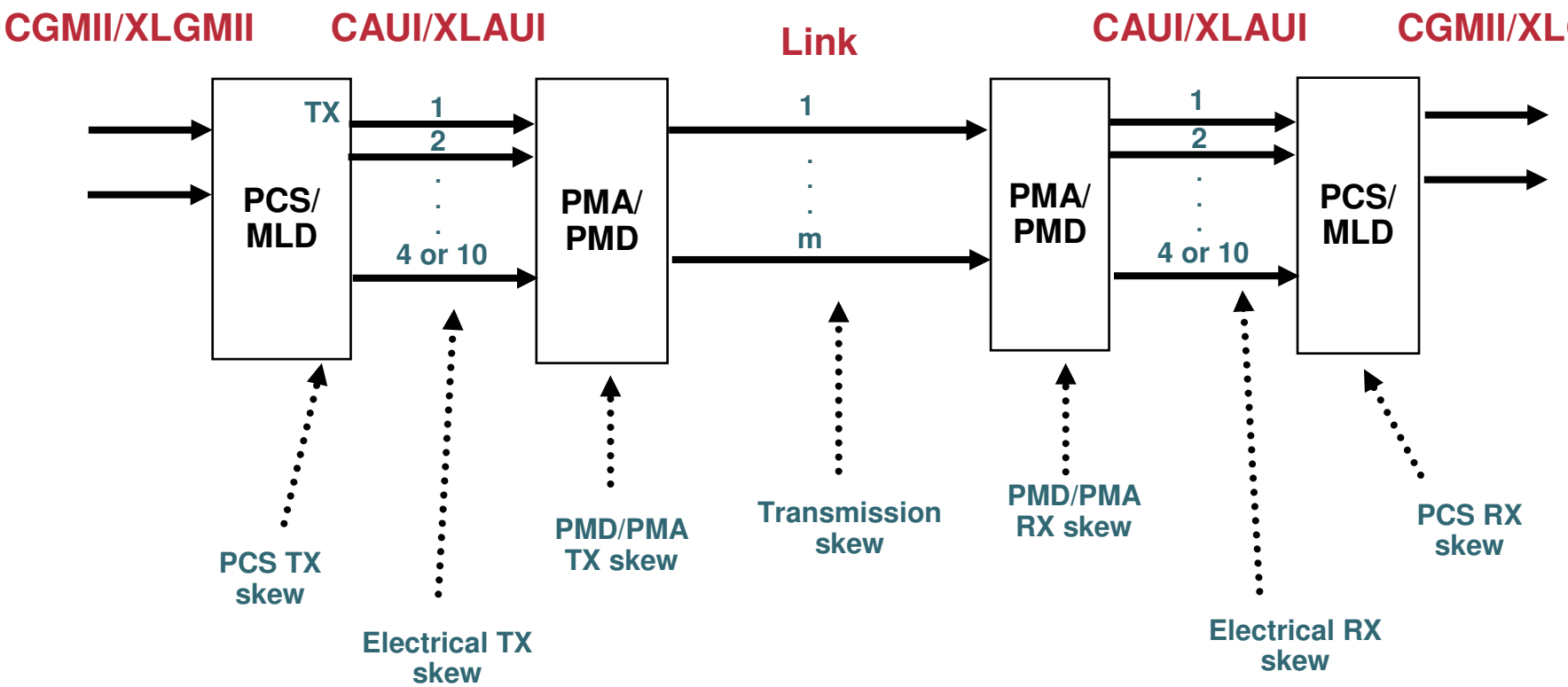
Skew Definition

- **In 40GE and 100GE, information will be transmitted in parallel links (lambdas, fibers, copper cables), typically not serially**
- **Since different paths can have different delays, skew will be introduced between them**
- **Source information needs to be reconstructed at the remote end, therefore de-skewing is needed at appropriate points**
- **Need to identify skew contributors and where we must compensate for skew**
- **Skew considered in this presentation is lane-to-lane skew, not the skew between the positive and negative parts of a differential pair**

High Level Skew View Point

- **What are the causes of maximum skew**
  - **Fixed path length differences**
    - **Copper traces**
    - **Cables**
    - **Fibers etc.**
  - **Parallel path FIFOs not synchronized**
  - **Propagation differences between media**
    - **Caused by wavelength differences**
    - **Stress in fibers, etc.**
- **What are the causes of dynamic skew (a subset of max skew)**
  - **Group delay: variable due to laser wavelength shift with temperature and wavelength drift over time**
  - **Fiber stress variation**
  - **DMD: variable due to launch & coupling variation**
  - **Electrical functions**
    - **Temperature variation causing variable gate delay**

System Architecture (one direction shown)



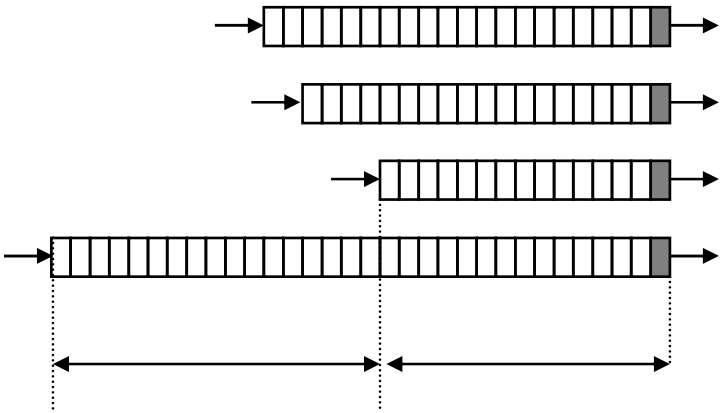
Skew budget definition in the PCS

- **PCS distributes data (with Lane Markers) into 20 (100 GE) or 4 (40 GE) virtual lanes using 66b block distribution**
- **PMAs distribute data by bit multiplexing (when needed)**
- **Skew will be presented as:**
  - **The maximum skew**
  - **The dynamic skew**
    - **Skew change over time due to environmental and/or other conditions**
    - **A subset of the maximum skew**
- **Maximum skew will be compensated for in the Rx PCS**
  - **Determines minimum FIFO size required at PCS sink**
- **Dynamic skew needs to be tolerated at each appropriate sink point, examples are:**
  - **PCS Rx sink**
  - **Tx PMA**
  - **Rx PMA**

De-skew buffer

Example – 4 x 10G receiver PCS

When link is established

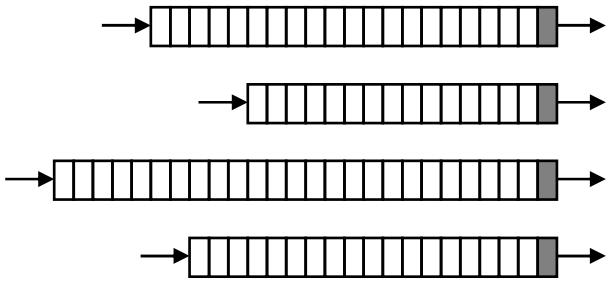


Initial skew

All buffers must contain at least as much as the max dynamic skew

Lane markers read out of each lane at the same time

Some time later



Dynamic skew has changed the fill level of the buffers

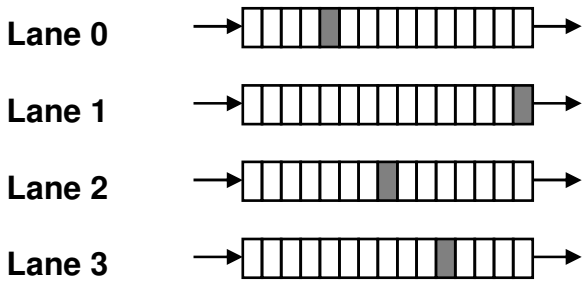
The maximum skew is the largest difference in the fill level of the buffers at any time

= first bit of the lane markers

Dynamic skew buffer (m != n)

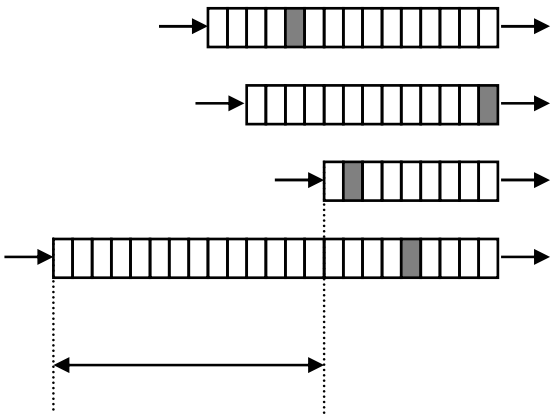
Example – 4 x 25G receiver gearbox

When link is established



All buffers half full

Some time later



Dynamic skew

Takes no account of the lane markers

= first bit of the lane markers

Maximum Skew

- **We need to add up all of the skew to see how much total skew must be compensated for at the Rx PCS**
  - **Maximum skew includes a dynamic component**

- **Max skew contributors are:**
  - **TX PCS**
  - **TX Electrical (CAUI/XLAUI)**
  - **TX PMD/PMA**
  - **Transmission (medium, electrical or optical)**
  - **RX PMD/PMA**
  - **RX Electrical (CAUI/XLAUI)**
  - **RX PCS**

PCS maximum skew (TX and RX)

- **Skew can be introduced due to 10G Tx SerDes FIFOs not being aligned, differences in FIFO fill levels translates into skew**
- **Another contributor can be the high speed serializer or deserializer stage in the SerDes**
- **2 case studies: ASIC or FPGA solution**

  ASIC case: TX= 2 ns, RX= 2 ns

  FPGA case: TX = 25.5 ns, RX = 14.3 ns
- **More detailed analysis in giannakopoulos_01_0508**

CAUI/XLAUI maximum skew (TX and RX)

- **A good starting point would be the CAUI/XLAUI interface for a chip to chip interconnect**
  - **Current proposal allows for up to 8-12” on the host board, and there would be 1” or so on the module**
- **Propose a generous 4” of trace length difference allowance, equates to 0.88 ns per direction (TX/RX)**

PMD/PMA max skew (TX and RX)

- **The most complicated PMA is a simple bit MUX/DeMUX**

  internal skew should be less than 0.4 ns (per chip, per direction), including analog and digital skew

- **PMA to PMD connection**

  Traces should in any case be carefully laid out

  Propose 1” (per direction), which is 0.22 ns (TX/RX)

- **Total = 0.4 ns + 0.22 ns = 0.62 ns per direction**

Maximum Transmission skew of planned PMDs

PMD	Description	Max Skew Budget ns (UI@10G)	Notes
100GBASE-ZR4??	100GE 80 km	33.2 ns (332UI)	Speculative
100GBASE-ER4	100GE 40 km	1.3ns (13UI)	
100GBASE-LR4	100GE 10 km	0.3ns (3UI)	
100GBASE-SR10	100GE 100m	4.5ns (45UI)	
100GBASE-SR10+	100GE 300m	13.6ns (136UI)	Speculative
100GBASE-CR10	100GE Copper 10m	0.5ns (5UI)	
100GBASE-CR10+	100GE Copper 30m	1.5ns (15UI)	Speculative
40GBASE-LR4 or 40GBASE-LR	40GE 10 km	1.7ns (17UI) or 0ns (0UI)	Depends on solution that is chosen
40GBASE-SR4	40GE 100m	4.5ns (45UI)	
40GBASE-SR4+	40GE 300m	13.6ns (136UI)	Speculative
40GBASE-CR4	40GE Copper 10m	0.5ns (5UI)	
40GBASE-CR4+	40GE Copper 30m	1.5ns (15UI)	Speculative

Recommended maximum skew contributions

Contributor	Maximum	Proposed Standard	Cumulative
PCS TX	25.5ns (255UI)	40ns (400UI)	40ns
Electrical CAUI/XLAUI i/f TX	.88ns (8.8UI)	2ns (20UI)	42ns
PMA/PMD TX	.62ns (6.2UI)	2ns (20UI)	44ns
Transmission	33.2ns (332UI)	120ns (1200UI)	164ns
PMA/PMD RX	.62ns (6.2UI)	2ns (20UI)	166ns
Electrical CAUI/XLAUI i/f RX	.88ns (8.8UI)	2ns (20UI)	168ns
PCS RX	14.3ns (143UI)	30ns (300UI)	198ns
TOTAL			198ns (~ 2k bits)

Propose to make all numbers Normative, except from the Rx PCS which can be Informative

Note: All UI are @10G

Standards normally are in bit times??? But at 100 or 40G?

Maximum dynamic skew of planned PMDs

PMD	'Standard'	Transmission medium	Notes
100GBASE-ER4	100GE 40 km	373ps (9.6UI)	
100GBASE-LR4	100GE 10 km	93ps (2.4UI)	
100GBASE-SR10	100GE 100 m	676ps (7UI)	
100GBASE-SR10+	100GE 300 m	2.0ns (21UI)	Speculative
100GBASE-CR10	100GE Copper 10m	50ps (0.5UI)	
100GBASE-CR10+	100GE Copper 30m	150ps (1.5UI)	Speculative
40GBASE-LR4	40GE 10 km	0 or 766ps (8UI)	
40GBASE-SR4	40GE 100 m	676ps (7UI)	
40GBASE-SR4+	40GE 300 m	2.0ns (21UI)	Speculative
40GBASE-CR4	40GE Copper 10m	50ps (0.5UI)	
40GBASE-CR4+	40GE Copper 30m	150ps (1.5UI)	Speculative

Recommended dynamic skew contributions

Contributor	Worst Case	Proposed Standard	Cumulative
PCS TX	194ps (2UI)	194ps (2UI)	194ps (2UI)
Electrical CAUI/XLAUI i/f TX	0	0	0
PMA/PMD TX	194ps (2UI)	194ps (2UI)	388ps (4UI)
Transmission	2.0ns (21UI)	2.91ns (30UI)	3.3ns (34UI)
PMA/PMD RX	194ps (2UI)	194ps (2UI)	3.49ns (36UI)
Electrical CAUI/XLAUI i/f RX	0	0	0
PCS RX	194ps (2UI)	194ps (2UI)	3.69 (38UI)
TOTAL			3.69 (38UI)

Standards normally are in bit times??? But at 100 or 40G?

More on Dynamic Skew

- **Ok, I have the dynamic skew numbers, now what?**
- **For designs with a PMA gearbox (m != n), the gearbox has a dynamic skew buffer per input lane**

  Size is 2x the max dynamic skew for that corresponding path

  Start reading out of the wander buffers when they are half full

- **For designs without a PMA gearbox (m=n), the maximum skew already includes the dynamic skew numbers, your receive PCS input FIFOs/buffers need to be able to track the dynamic skew**

  Note: An increase in maximum skew capability does not impact latency, only buffer depth. An increase in dynamic skew capability does increase latency because you must wait to start reading out data from the receive FIFOs until there is enough data in the least filled FIFO to allow for the maximum dynamic skew variation that we expect for a worst case interface.

- **In addition, depending on the PMA design it might need to track dynamic skew**

  For instance, if you clock all outputs with a common clock

Q & A

Thank you !

Backup - Fiber characteristics tool

- **Fiber characteristics tools (spreadsheets) officially adopted by IEEE and used by P. Anslow to calculate transmission skews are in:**

http://www.ieee802.org/3/ba/public/tools/Fibre_characteristics_V_3_0.xls

http://www.ieee802.org/3/ba/public/may08/kolesar_02_0508.xls