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)

- PCS/MLD
- PMA/PMD
- Link
- CAUI/XLAUI
- CGMII/XLGMI

**TX**
- 1
- 2
- ... 4 or 10

**Electrical TX skew**
- PCS TX skew
- PMD/PMA TX skew
- Transmission skew
- PMD/PMA RX skew

**Electrical RX skew**
- PCS RX skew
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
Example – 4 x 10G receiver PCS

When link is established

Lane markers read out of each lane at the same time

Initial skew

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

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 \neq n)

Example – 4 x 25G receiver gearbox

When link is established:

- Lane 0
- Lane 1
- Lane 2
- Lane 3

All buffers half full

Some time later:

- Lane 0
- Lane 1
- Lane 2
- Lane 3

Dynamic skew

Takes no account of the lane markers

\[= \text{first bit of the lane markers}\]
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
• 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
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).
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

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

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

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

<table>
<thead>
<tr>
<th>PMD</th>
<th>‘Standard’</th>
<th>Transmission medium</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>100GBASE-ER4</td>
<td>100GE 40 km</td>
<td>373ps (9.6UI)</td>
<td></td>
</tr>
<tr>
<td>100GBASE-LR4</td>
<td>100GE 10 km</td>
<td>93ps (2.4UI)</td>
<td></td>
</tr>
<tr>
<td>100GBASE-SR10</td>
<td>100GE 100 m</td>
<td>676ps (7UI)</td>
<td></td>
</tr>
<tr>
<td>100GBASE-SR10+</td>
<td>100GE 300 m</td>
<td>2.0ns (21UI)</td>
<td>Speculative</td>
</tr>
<tr>
<td>100GBASE-CR10</td>
<td>100GE Copper 10m</td>
<td>50ps (0.5UI)</td>
<td></td>
</tr>
<tr>
<td>100GBASE-CR10+</td>
<td>100GE Copper 30m</td>
<td>150ps (1.5UI)</td>
<td>Speculative</td>
</tr>
<tr>
<td>40GBASE-LR4</td>
<td>40GE 10 km</td>
<td>0 or 766ps (8UI)</td>
<td></td>
</tr>
<tr>
<td>40GBASE-SR4</td>
<td>40GE 100 m</td>
<td>676ps (7UI)</td>
<td></td>
</tr>
<tr>
<td>40GBASE-SR4+</td>
<td>40GE 300 m</td>
<td>2.0ns (21UI)</td>
<td>Speculative</td>
</tr>
<tr>
<td>40GBASE-CR4</td>
<td>40GE Copper 10m</td>
<td>50ps (0.5UI)</td>
<td></td>
</tr>
<tr>
<td>40GBASE-CR4+</td>
<td>40GE Copper 30m</td>
<td>150ps (1.5UI)</td>
<td>Speculative</td>
</tr>
</tbody>
</table>
### Recommended dynamic skew contributions

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

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 \( m \neq n \), 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
Thank you!
• 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