100/40GE Max skew budget for MLD

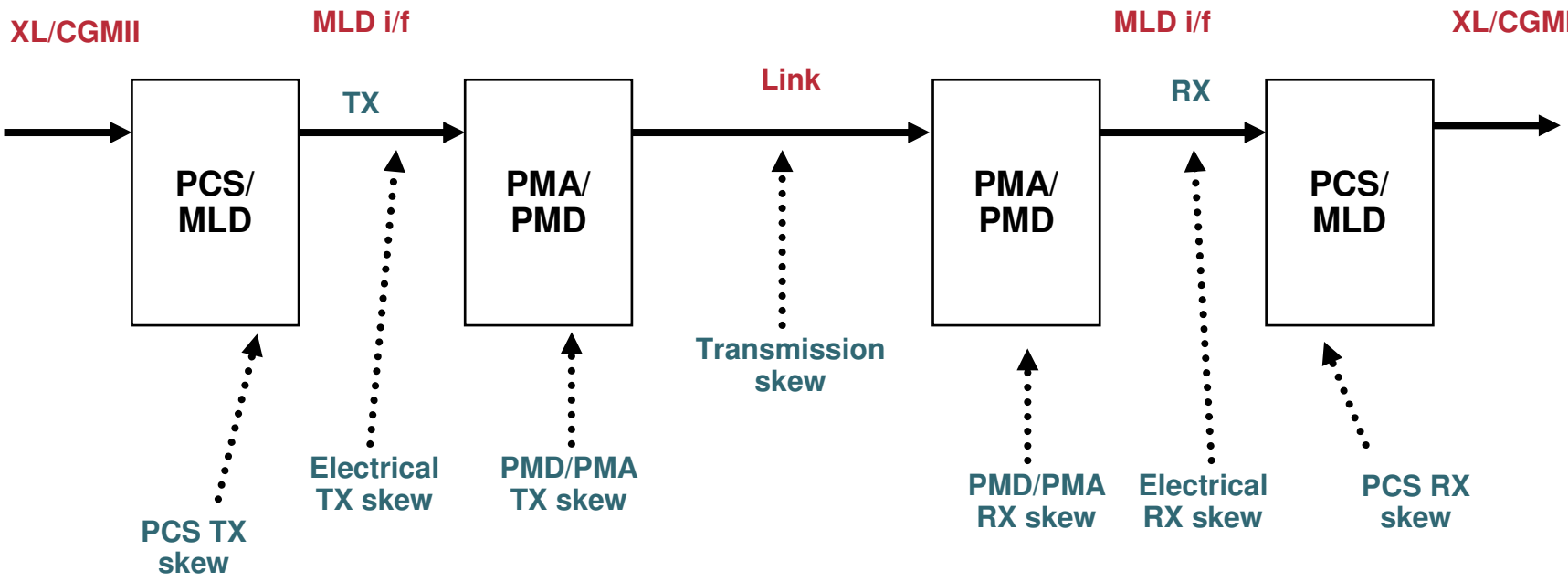
IEEE P802.3ba
Munich, May 2008

Brad Booth, Keith Conroy, Francesco Caggioni,
Dimitrios Giannakopoulos – AMCC

Supporters

- Mark Gustlin, Gary Nichols - Cisco
- Pete Anslow – Nortel
- Farhad Shafai – Sarance
- Craig Hornbuckle, Song Shang - SMI

System architecture (one direction shown)



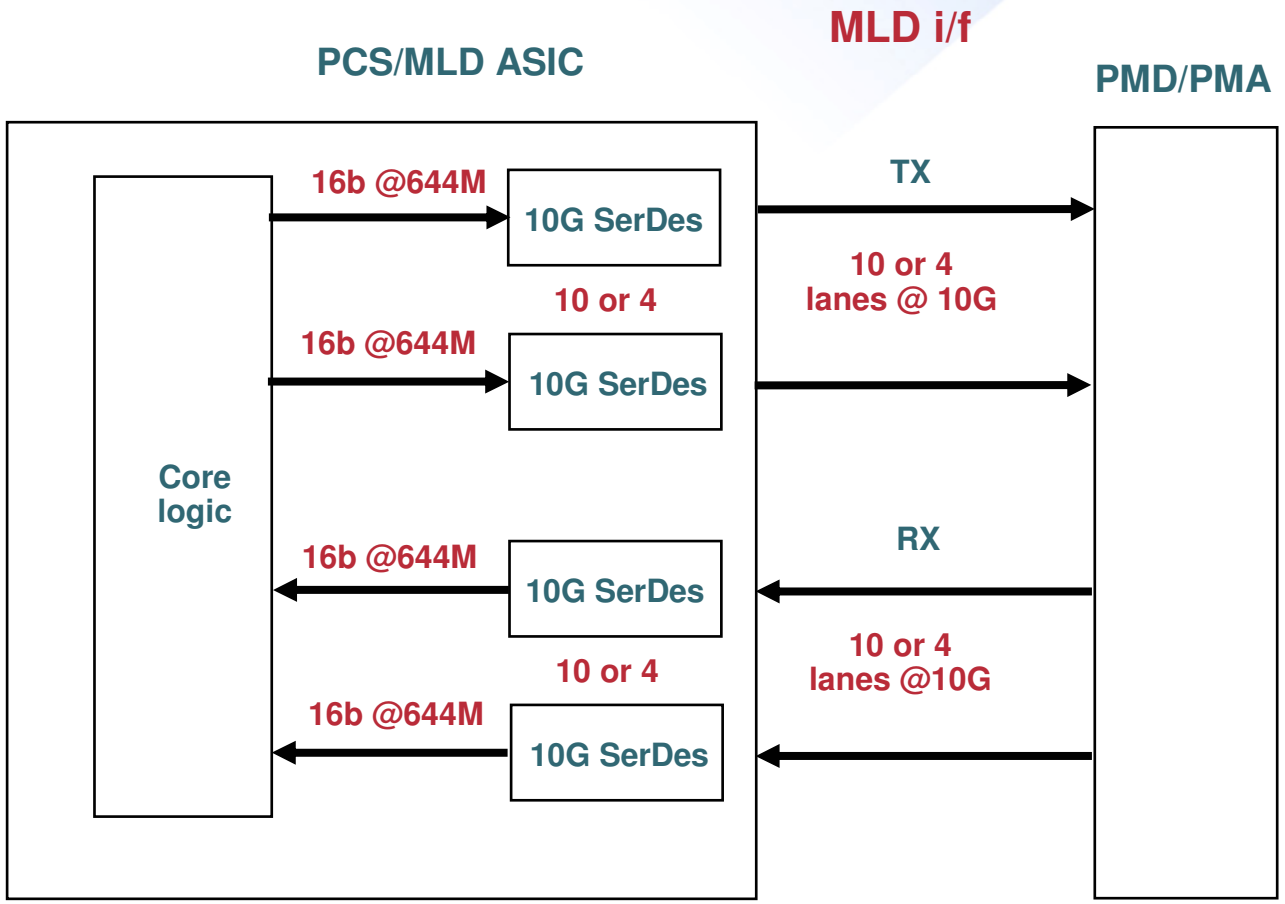
Total Skew introduced end-to-end

- Skew considered in this presentation is max lane-to-lane skew, not P-N skew in a differential pair
- We need to add up all of the skew to see how much total skew must be compensated for at the receiver
- Skew contributors are PCS/MLD, Electrical interface, PMD/PMA and transmission skew, can be broken further into:
  - TX PCS/MLD
  - TX Electrical (PCS to PMA)
  - TX PMD/PMA
  - Transmission (medium)
  - RX PMD/PMA
  - RX Electrical (PMA to PCS)
  - RX PCS/MLD

PCS/MLD skew (TX and RX)

- **Skew could be introduced due to 10G Tx SerDes FIFOs not aligned, difference in FIFO fill translates into skew**
- **Basically, each SerDes has its own FIFO, all 10 (100 GE) or 4 (40 GE) FIFOs should be aligned if possible to reduce skew – essentially "reset" the pointers within some tolerance**
- **Another contributor can be the high speed serializer or deserializer stage in the SerDes**
- **2 case studies: ASIC or FPGA solution**

PCS/MLD skew – ASIC case

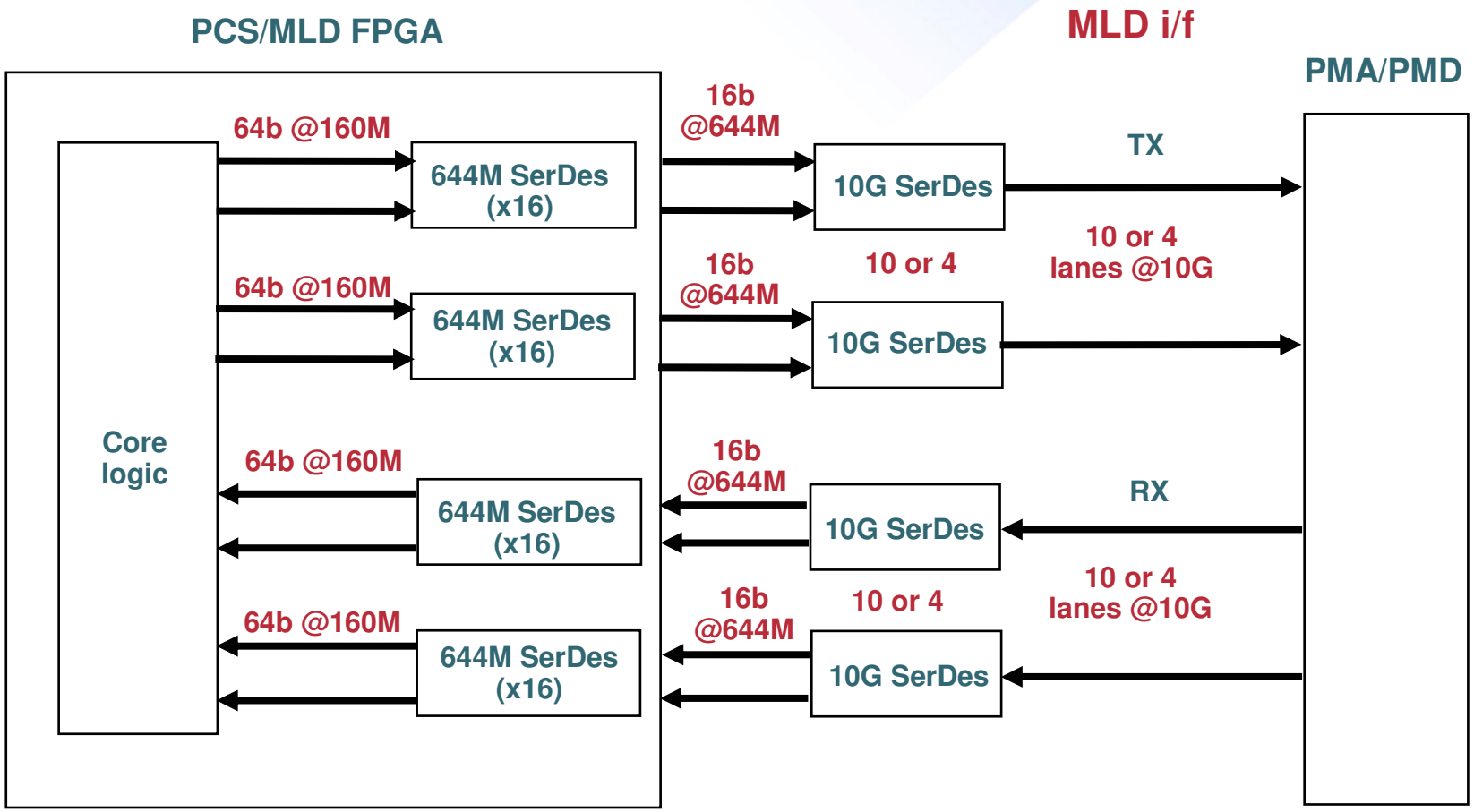


MLD i/f: 10 lanes for 100 GE or 4 lanes for 40GE @10G

PCS/MLD skew – ASIC case (cont.)

- **ASIC skew**
  - **ASICs can integrate 10G SerDes technology. This is assuming all 10 (100 GE) or 4 (40 GE) SerDes in same chip**
  - **This should potentially introduce only a small amount of skew quantified by the uncertainty on the FIFO reset mechanism and de/serializer-related skew**
  - **There should be no skew in the stages feeding the Tx FIFOs**
  - **Total amount of TX skew should be in the order of 1.8 ns for reset/serializer uncertainty and another 0.2 ns for skew induced from driver/package, total of upto 2 ns**
  - **Numbers for RX side are also in the order of upto 2 ns**

PCS/MLD skew – FPGA case



MLD i/f: 10 lanes for 100 GE or 4 lanes for 40GE @10G

PCS/MLD skew – FPGA case

- **FPGA solution with external 10G SerDes devices**
  - **Since current FPGA technology does not support 10G links (SerDes), board level solutions need to connect a MAC/PCS FPGA to 10 (100 GE) or 4 (40 GE) 10G SerDes devices**
  - **Assuming a 16-bit i/f between FPGA and external 10G SerDes: still need internal FPGA SerDes to convert from a wide databus to serial 644 Mb/s (16 of them per 10G) – a 4-bit or 2-bit i/f are alternative options**
  - **Stages feeding the internal SerDes can introduce upto 12.8 ns (2x64 bits) of skew**
  - **Number of pins and internal FPGA SerDes required might prohibit single-FPGA device implementation for 100GE, unless narrower external busses are used (RXAUI for example)**

PCS/MLD skew (TX and RX), cont. - FPGA

- **FPGA solution with external 10G SerDes devices (cont.)**
  - **External SerDes devices are difficult to synchronize, so skew can be introduced by different fill levels in their FIFOs – with a FIFO size of 8 by 16 bits, could have a skew of 7*1.6 ns = 11.2 ns (applies to Tx, no FIFO in Rx)**
  - **Add in upto 1.5 ns max for serializer skew**
  - **In Rx, no FIFO is needed, deserializer skew can be upto 1.5 ns**
  - **Do not expect electrical skew in a 16-bit clocked i/f @644M between FPGA and external SerDes**

MLD interface Electrical skew (TX and RX)

– Calculation of skew requires modeling using controlled impedance traces on standard FR4 low-cost PCBs

– A good starting point would be the XFI interface for a chip to chip interconnect

- current XFI i/f allows for 9.6dB of loss @5.5GHz
- allows for a range from 1” to 8-10” typically

MLD i/f Electrical skew (TX and RX) (cont.)

- If we assume a 10” range then the upper bound of skew is 10”, which translates into 10*220ps/in = 2.2ns, so 4.4 ns for TX and RX – Could never happen ☺
- A more realistic scenario would be 1-2” of board skew, 2” translates to 0.88 ns
- PCB designers could give more accurate skew data, depends on board size, number of layers etc
- Propose a generous 4” of trace length difference allowance, equates to 1.76 ns for both RX and TX

PMD/PMA skew (TX and RX), 4x25G

- **MLD protocol supports bit muxing at the MLD i/f as well as at the line side**
- **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
  - Should be less than 1” (per direction), which is 0.45 ns (RX and TX)

Transmission skew

- **Dependent on PMD type**

- **For an SMF optical solution, a 10Km range has a max skew @1300 nm of 1.7ns for transmission skew (4x25G or 4x10G case)**
- **For a parallel fiber MMF case @850nm, assuming 44.3ps/m skew accumulation, a 100m range results in 4.43 ns of skew**
- **Objective for copper is 10m, therefore the skew contribution should be smaller than the optical cases**

- **Not a current objective, but if in the future a 80 km solution is used, then 80 km @ 1550 nm (800 GHz spacing in C band, 4x25G) would result in a 33.2 ns skew**

Max skew budget – 40GE

- **In all modes:**
  - TX electrical (FPGA): 25.5+0.88+0.62 = 27 ns
  - RX electrical (FPGA): 14.3+0.88+0.62 = 15.8 ns
  - TX+RX electrical (FPGA) = 42.8 ns

- **10Km SMF, CWDM mode**
  - Optical interface (1300 nm, 4x10G): 1.7ns
  - TOTAL: 42.8 + 1.7 = **44.5 ns**
- **100m MMF, parallel fibers**
  - Optical interface (850 nm): 4.43ns
  - TOTAL: 42.8 + 4.43 = **47.23 ns**
- **300m MMF, parallel fibers (not an objective)**
  - Optical interface (850 nm): 13.29ns
  - TOTAL: 42.8 + 13.29 = **56.09 ns**

Max skew budget – 100GE

- **In all modes:**
  - **TX+RX electrical (FPGA) = 42.8 ns (same as in 40GE)**

- **10Km SMF, CWDM**
  - **Optical interface (1300 nm, 4x25G): 1.7ns**
  - **TOTAL: 42.8 + 1.7 = 44.5 ns**
- **40Km SMF, DWDM**
  - **Optical interface (800 GHz spacing): 0.72ns**
  - **TOTAL: 42.8 + 0.72 = 43.52 ns**
- **80Km SMF, DWDM (not an objective)**
  - **Optical interface (1550 nm, 800 GHz spacing): 33.14ns**
  - **TOTAL: 42.8 + 33.14 = 75.94 ns**
- **100m MMF, parallel fibers**
  - **Optical interface (850 nm): 4.5ns**
  - **TOTAL: 42.8 + 4.43 = 47.23 ns**
- **300m MMF, parallel fibers (not an objective)**
  - **Optical interface (850 nm): 13.29ns**
  - **TOTAL: 42.8 + 13.29 = 56.09 ns**

Summary Table for max skew

Contributor	Max Skew (ns) for objectives
PCS/MLD TX	2 (ASIC) 12.8+11.2+1.5 = 25.5 (FPGA solution)
Electrical MLD i/f TX	0.88
PMA/PMD TX	0.62
Transmission	1.7 Optical SMF (4x10 or 4x25G @1300nm) 4.43 Parallel Fiber MMF (4x or 10x10G @850nm)
PMA/PMD RX	0.62
Electrical MLD i/f RX	0.88
PCS/MLD RX	2 (ASIC) 12.8+1.5 = 14.3 (FPGA solution)
TOTAL	**11.43** (ASIC) **47.23** (FPGA solution)

Proposals for skew contributors

- Propose a 4” MLD interface board trace skew which results in 1.76 ns skew (total RX+TX)

- Propose a 1” board trace skew for PMA to PMD electrical skew (0.44 ns total RX+TX)

Max Skew budget proposal

- From summary Table:

  max_estimated_skew = 47.23 ns (for objectives)

- Propose allowing for a wide margin
  - for future technology (80 Km reach e.g. ?)
  - even a number much higher seems acceptable since total memory at PCS/MLD Rx end is still small (reduces risk)
  - bigger buffers do not result in more latency (latency is determined by FIFO fill)
- PROPOSAL = 204.8 ns
  - FIFO size needed for deskew at PCS/MLD remote end:
    - 40GE: 2048 bits per VL (4 VLs @ 10G), TOTAL memory of 8192 bits
    - 100 GE: 1024 bits per VL (20 VLs @ 5G), TOTAL memory of 20480 bits

Thank you !

Backup - Fiber characteristics tool

- Fiber characteristics tool (spreadsheet) officially adopted by IEEE and used by P. Anslow to calculate transmission skew is in:

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