Requirements for a PMD to Achieve 100m over OM3 Fiber

Petar Pepeljugoski, Daniel Kuchta - IBM

John Petrilla, Piers Dawe - Avago Technologies

Outline

- **Desired elements for success**
  - For Modules (specs, form factors, electrical interfaces)
  - For Cable plant (length, connectors)
  - For Testing
- **Example Specs, Power Budgets**

802.3ba objectives over OM3 fiber

- **Support MAC data rates of 100 Gb/s and 40Gb/s**
- **Reach 100m**
- **Achieve better than or equal to 1E-12 BER**

Elements for Success for I/O in High Performance Computing Environment

- **Cost – less than ten/four 10 GbE Solutions**
- **Power consumption – less than ten/four 10 GbE Solutions**
- **High module density per Gb/s– higher than 10 GbE solutions**
- **Cable plant: 100 m OM3 & ≥ 4 connectors**
- **Reliability – better than ten/four 10 GbE Solutions**

Cost and Power Consumption in Parallel Links

- **Module Costs impacted by testing, calibration, programming and yield**
  - **Especially true for a multi-lane solution**
  - Targets requiring narrow operating ranges forces over temperature testing and calibration
  - Relaxation of rise/fall times and jitter specs would improve power consumption*
  - VCSEL array yield can be improved by relaxing wavelength specs*
  - Moving to Class 1M from Class 1 relaxes narrow operating range and reduces module cost by saving on testing
- **Transmitter power consumption is driven by rise/fall times, jitter and input equalization requirements**
  - **Need to allocate more jitter to electrical I/O than 802.3ae did: learn from SFP+ experience**
  - Multilane interface adds crosstalk, increasing TX jitter contribution and reducing RX sensitivity leading to a need for more jitter allocation
  - Higher minimum input amplitude to the TX would also reduce power consumption by simplifying the driver
- **Receiver power consumption is driven by output signal level amplitude, bandwidth and input dynamic range requirements**
  - including equalization, adding a CDR and/or providing linear outputs will increase power consumption

*When compared with 10GBASE-SR specs

Module Density

- **High density brings challenges with heat transfer (power dissipation) and signal integrity (inter-lane crosstalk)**
- **Form Factor Design Points and Projections**
  - SFP(+) Solution
    - Module + Cage Dimensions: 15 mm W x 12 mm H x 59 mm D
    - Horizontal Port Pitch: 16.25 mm
    - Power Level 1 max: 1.0 W
    - Optical Connector: Dual LC
    - Electrical Connector: 20 pin double-sided single edge
  - QSFP Solution
    - Module + Cage Dimensions: 19 mm W x 14 mm H x 79 mm D
    - Horizontal Port Pitch: 21.0 mm
    - Power Level 2 max: 2.0 W
    - Optical Connector: 1 x 12 MPO
    - Electrical Connector: 38 pin double-sided single edge
  - 10SFP Solution
    - Module + Cage Dimensions: ~ 22 mm W x 17 mm H x 79 mm D
    - Horizontal Port Pitch: ~ 24 mm
    - Power Level max: ~ 5.0 W
    - Optical Connector: 2 x 12 MPO
    - Electrical Connector: ~ 84 pin double-sided stacked edges
- **Multi-lane MMF is the lowest power, most compact of any PMD**

Multilane Link Design Point is Crucial

- **User expectation is to see 10x performance at 3x cost**
  - 10GBASE-SR is not a satisfactory basis for multi-lane specification
- **Unless test requirements are changed, test costs not anticipated to scale expectation above**
  - Test costs can easily dominate – need to relax key parameters to simplify testing and improve yield
  - Low costs are not achieved by setting specifications based on best-of-breed but based on industry-wide capabilities.
- **Power/bit not expected to improve unless specifications relaxed**
  - 10GbE, 40GbE & 100GbE solutions will use the same semiconductor technology - reduction in power consumption seems linked to constraining link specifications
  - Use of multiple power supply voltages, e.g. 1.5 V in addition to 3.3 V, will be beneficial
- **Inter-lane crosstalk can adversely impact signal quality and appearance**
  - Impact on link cost and power consumption can be avoided with suitable allocations in signal budgets and proper attention in design
- **Operating life and/or reliability can be adversely impacted by device temperature**
  - Specifications that permit low power consumption will also benefit operating life.
- **Multilane links can adversely impact port density.**
  - Port density is often limited by IO count.
- **Multilane links will have lower max output optical power limits than single lane variants by ~ 1.4 dB/lane**
  - Eye safety standard requirements

Options for 100/40 Gb/s Parallel PHY - Distances up to 100m

- Parallel (over OM3 multimode fiber)
  - N (10/4) channels x ~10 Gb/s (preferred)
    - Easiest to implement, cable cost high for longer lengths
    - ISI and DJ low, relaxed jitter budget
    - High Density, low power consumption
    - Transceiver commonality with other protocols (Quad Data Rate Infiniband)
    - Component availability: singlets available now at 10 Gb/s, lowest risk
  - 4 channels x ~25 Gb/s
    - Highest potential savings on cable cost
    - Commercially available direct modulated OE devices do not exist today
    - Cable plant compatible with likely 40Gb/s solutions
    - Highest ISI penalty, very tight jitter budget
      - energy efficiency needs to be examined
    - Highest module density
    - Can we reach 100m without equalizers?
  - 6 channels x 17-20 Gb/s
    - Saves on cable cost, but requires high speed devices that may not be available
    - Higher ISI penalty, tighter jitter budget, tighter device specs (BW, rise times, laser rms linewidth, etc.)
    - High Density, low power consumption
    - Compatibility with 17Gb/s FC devices if suitable laser available

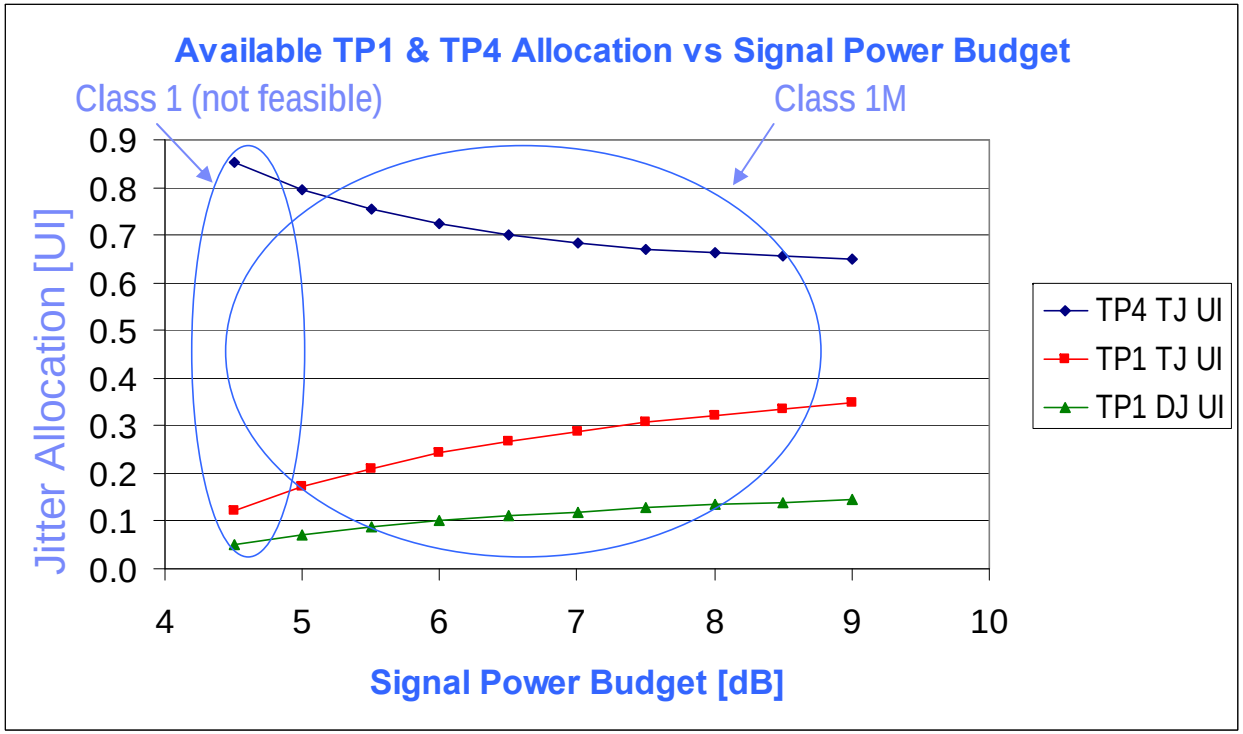
Transmitter and Receiver Solution Space

- **Link Budget Items with example values (10.3125 Gb/s)**
  - Transmitter:
    - Max P_average for Eye Safety Class 1: ~ -2.5dBm
      Max P_average for Eye Safety Class 1M: ~ 3.5dBm
    - Min OMA for Eye Safety Class 1: ~ -6.3dBm
      Min OMA for Eye Safety Class 1M: ~ -3.3dBm
    - Min Extinction Ratio: 3.0 dB
    - Max RMS Spectral Width: 0.65 nm
    - Center Wavelength Range: 840 to 860 nm
    - Max RIN12OMA: -132 dB/Hz
  - Receiver
    - Max Sensitivity: ~ -11.1 dBm
    - Max Input Power: 0.5 dBm
    - Min Bandwidth: 7500 MHz

Good and bad precedent

- **10GBASE-S**
  - Was designed to achieve a headline distance rather than for low cost
  - It was designed assuming CDRs in the module
  - It has been too expensive and late to volume
  - Trying to move the CDRs out of the module (SFP+) has been painful
- **800-Mx-SN-S (8GFC limiting)**
  - Learned from 10GBASE-S's mistakes and 7 years of experience
  - Shorter headline distance gives much better cost structure, addresses most links
  - Better allocation of jitter to the electrical I/O
- **Any new MMF spec should learn from 8GFC**

Boundary Condition: Class 1 or Class 1M Eye safety



Class 1 requires extremely tight jitter specs for TP1, TP4 - hard for CDR

Class 1M achieves better jitter balancing

- **The above chart shows the tradeoff expected between possible jitter allocations at TP1 and TP4 and the available link signal power budget. The available allocation is arbitrarilly split between TP1 and TP4**
- **Link attributes assumed 100 m OM3, 1.5 dB connector loss, 0.3 dB Pmn*, 0 BLW**, 10 ps Tx DJ, 0.5ps fiber DJ & 10 ps Rx DJ**

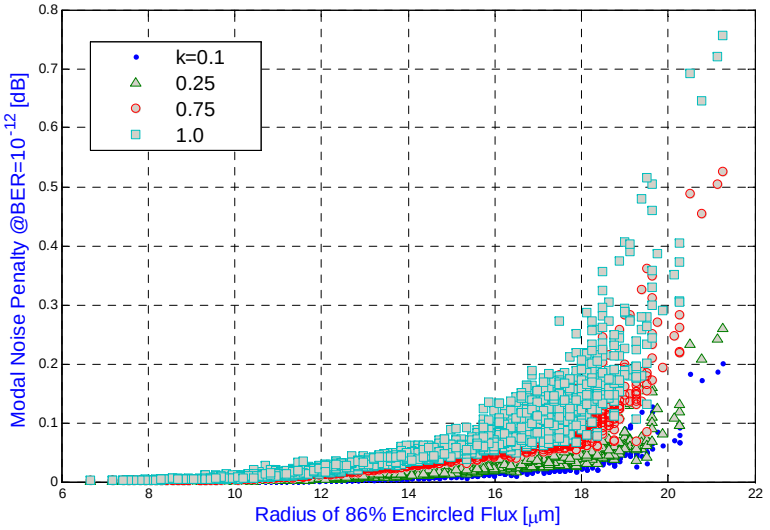
*Pmn: modal noise penalty **BLW: baseline wander

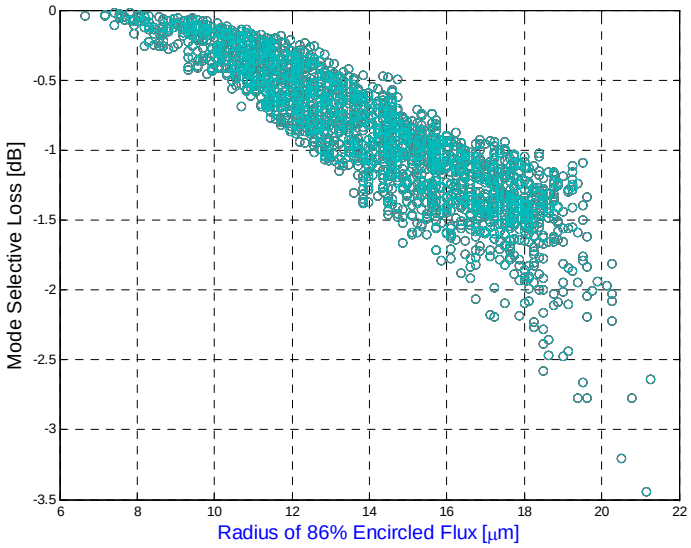
Cable Plant Trade-offs, Connectors and Length

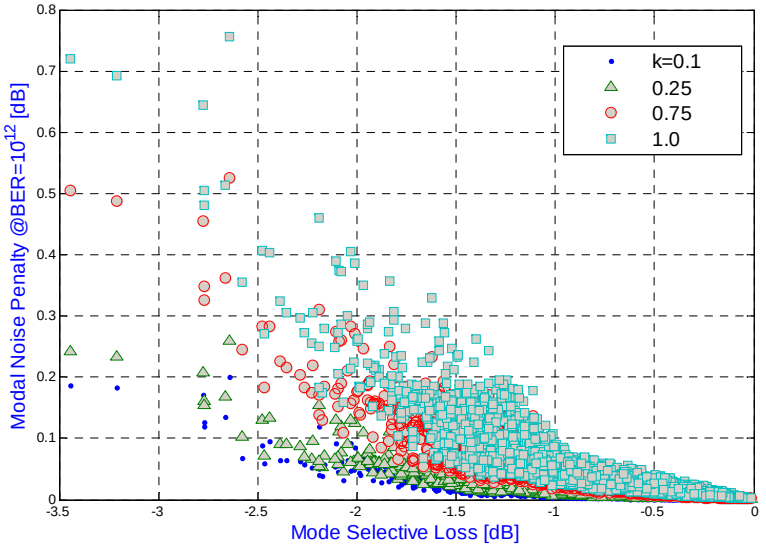
- **100m is more than sufficient to cover all distances in HPC environment**
  - OM3 is ideal (minimizes the cost)
  - Any longer distances should come by using better fibers (i.e. OM4), no tighter specs on the modules or use another PMD over SMF
- **Number of intermediate connector pairs 4 or more**
  - Connectors interfacing to modules not counted
  - Ethernet assumes 4 connectors, we need to address the discrepancy between assumptions when specs were adopted and practice
    - Users routinely trade distance for loss in extra connectors
    - This may be the most challenging issue for any MMF solution
  - Parallel connectors have higher loss than single connectors
  - Need to limit aggregate loss and individual loss
    - Impact on modal noise and link performance

Why do we need to limit connector loss?

- **Larger encircled flux radius leads to higher modal noise penalty**
  - Need EF to manage modal noise penalty
- **Larger number of connectors increases the total mode selective loss**
  - Need to limit individual and aggregate connector loss to manage modal noise penalty
- **Larger number of connectors also decreases effective modal bandwidth**
  - Encircled flux important even for 100m links







Conclusion/Recommendations

- **Multi-lane MMF the lowest power, most compact solution for HPC and other short-reach applications**
- **10GBASE-S specifications not well optimized –need better focus on all aspects of cost**
- **With sufficient power budget reasonable jitter allocations can be accommodated**
- **Moving from Class 1 to Class 1M eye safety limits enables larger power budgets by increasing the TXmax limits on optical power and should be considered**
- **Other alternatives that may improve the power budget (FEC for example) should be examined**
- **The Task Force should address the number of connectors in the link; understanding their performance and accommodating their impact on signal quality is essential**