400GbE Logic Challenges

IEEE

May 2013   Victoria

Mark Gustlin   - Xilinx

John D'Ambrosia - Dell

Gary Nicholl - Cisco

Agenda

- Review of past logic related objectives
- Lessons learned from the 802.3ba PCS
- Possible 400 Gb/s PCS issues to investigate

Logic Related Objectives from Past Projects

- Review of past logic related objectives
- There are many common logic related objectives that are applicable to the 400 Gb/s project
- The PMD objectives also can greatly impact the logic functions, number of lanes for instance impact the PCS architecture

Objective	.3ba	.3bg	.3bj	.3bm
Support full-duplex operation only	✓	✓	✓	✓
„Preserve the 802.3 / Ethernet frame format utilizing the 802.3 MAC	✓	✓	✓	✓
„Preserve minimum and maximum FrameSize of current 802.3 standard	✓	✓	✓	✓
„Support a BER better than or equal to 10^{-12} at the MAC/PLS service interface	✓	✓	✓	✓
„Provide appropriate support for OTN	✓			✓
„Support a MAC data rate of x Gb/s	✓	✓		
To define optional Energy-Efficient Ethernet operation for xxx PMD or interface	*	*	✓	✓

* Optional EEE support added by 802.3bj and .3bm for these PMDs

Possible 400GbE Logic Related Objectives

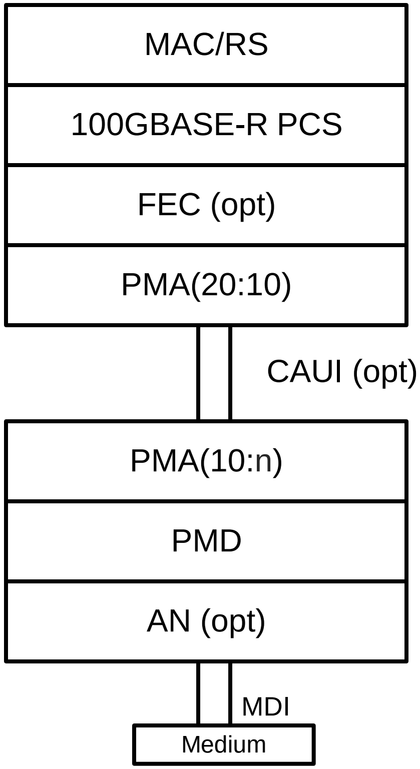
- Support a MAC data rate of 400 Gb/s
- Support full-duplex operation only
- Preserve the 802.3 / Ethernet frame format utilizing the 802.3 MAC
- Preserve minimum and maximum FrameSize of current 802.3 standard
- „Provide appropriate support for OTN
- To define optional Energy-Efficient Ethernet operation for xxx PMD or interface
  - PMD type likely to define what mode(s) are supported, deep sleep vs. fast wake

- Support a BER better than or equal to 10^{-12} at the MAC/PLS service interface
  - Likely to be a lot of discussion around this objective, some want a better BER target than this and FER at the FEC service interface is more appropriate for PHYs with FEC

100GbE Architecture in Review

- Based on a 20 Lane PCS with 64B/66B encoding (5 Gb/s per PCS Lane)
- Data is striped to PCS lanes 66-bit blocks at a time
- Alignment Markers are periodically added to all PCS lanes to enable alignment in the RX PCS
- PMAs do simple bit multiplexing to change lane widths
- Physical lane widths of 20, 10, 5, 4, 2, 1 can all be supported
- Optional KR based FEC is supported
- Initially no support for EEE, now being added by 802.3bj and 802.3bm

- P802.3bj is adding strong FEC to the architecture (below the PCS), but then we lose the simplicity of changing lane widths by bit multiplexing



Lessons Learned from the 100GbE Architecture

- Pros of the architecture:
  - Simple PMA bit multiplexing allows for simple PMAs and flexibility in lane counts
    - Easily supports future signaling or modulation technology
  - Skew and alignment is only performed at the receive PCS
  - One PCS encoding end to end
  - Definition of the service interface allows for flexible placement of electrical interfaces (CAUI)
  - BIP error detection allows quick BER calculations and error isolation per lane
- Negatives
  - No strong or low latency FEC was included, so it had to be added after the PCS in 802.3bj
    - Only 802.3ap FEC was defined, but it was too high in latency for many applications
  - With the 802.3bj FEC you cannot bit multiplex to change lane widths
  - With bit multiplexing, and with electrical DFE which causes burst errors, the protocol is susceptible to MTTFPA issues
  - 20 PCS lanes is a relatively large number and each is low bandwidth, especially now that most PMDs are moving to 4 physical lanes

How Many PCS Lanes to Support?

- PCS lanes = Virtual lanes from the MLD architecture
- The number of PCS lanes determines the flexibility of the PCS to traverse different electrical and optical lane widths
- The number of PCS lanes should equal the Least Common Multiple of all lane count combinations that you want to support
- Electrical and optical lane counts depend on the technology that is used for the PHY/PMD solutions
- Will 25 Gb/s be the base technology for first generation 400G PHYs?
  - If yes, then 16 PCS lanes makes a lot of sense
  - It can support 50 Gb/s lanes by muxing 2:1, 100 Gb/s lanes by muxing 4:1 etc.
- Is 16 PCS lanes too many for long term?
  - The fewer the PCS lanes the better, less stats, less processing especially as technology advances
  - Should we optimize for 2nd generation 400GbE which is likely to use 50G lanes?
  - You could define 8 PCS lanes and allow them to be further split for 1st generation PMDs
- Do we need to support other lane speeds such as 40 Gb/s?
  - This can impact the number of PCS lanes or we can devise methods to allow multiplexing to 10 lanes

FEC or not?

- Should we define a low latency and strong FEC as part of the PCS?
- As soon as we defined a low latency relatively strong FEC in 802.3bj there is a lot of desire to use it
  - 100GBASE-SR4 is proposed to use it, the PSM4 and PAM8 proposals use it
  - It is used to extend reach and/or simplify link budgets which leads to more cost effective solutions
- Will any of the to be proposed PMDs for 400GbE want to or need to use FEC to achieve the desired cost points and reaches?
  - I assume yes, but time will tell
- Adding FEC to the PCS will mean that we can’t bit mux, rather we would want to multiplex on FEC symbol boundaries
- It is likely desirable to define FEC as part of the PCS from the beginning so we don’t have to bolt on FEC later
  - A related issue is the encoding decision, 64B/66B vs. 256B/257B or something else
  - If it is part of the PCS, do we need to be able to bypass FEC if it is not needed? Or always send it?
  - It is likely that some PMDs will need a stronger FEC on top of whatever might be defined in the PCS?

Other Decision on the MAC/PCS Architecture?

- MAC seems straightforward, run at 400 Gb/s, no other changes?
- MII – same scalable MII as used for 40/100 Gb/s?
- How to change lane widths, assuming FEC, you need to block mux, not bit mux?
- If FEC is needed, stick with an RS code or something else?
  - 400G goes 4x100G rate, so we could have a larger block size to get additional gain and stay < 100ns of added latency
  - We could even re-use the 100GbE FEC as is, use 4x, makes it easier to support 4x100GbE and 1x400GbE in the same device
- How to efficiently align multiple lanes, what flexibility is allowed in lane ordering
- How to scramble the data, scramble across the whole payload or per lane?
- Rules on IPG sizing (deficit idle counter)
- Include BIP error detection and possibly more advanced monitoring?
  - In band Signaling channel, signal fail/degrade alarms, are possible examples
- What is the desired total achievable latency?
  - FEC might have the largest impact on this
- Specify a time synchronization protocol (1588 etc.) reference point for accurate timestamping?

Summary

- A MAC/PCS for 400GbE with or without FEC is feasible today in either FPGA, ASIC or ASSP technology

- There are many possible solutions for a 400GbE PCS
- One simple option is scaling up the 802.3ba PCS
- If there are interfaces that will require FEC, and low latency is important, then a PCS could be defined that incorporates a low latency FEC from the start
  - This applies to both electrical and optical interfaces
- In addition there are many enhancements to the PCS that we can explore to make the PCS more robust and future proof

Thanks!