Feasibility of a 400GbE PCS

IEEE HSE Consensus

September 2012 Geneva

Mark Gustlin - Xilinx

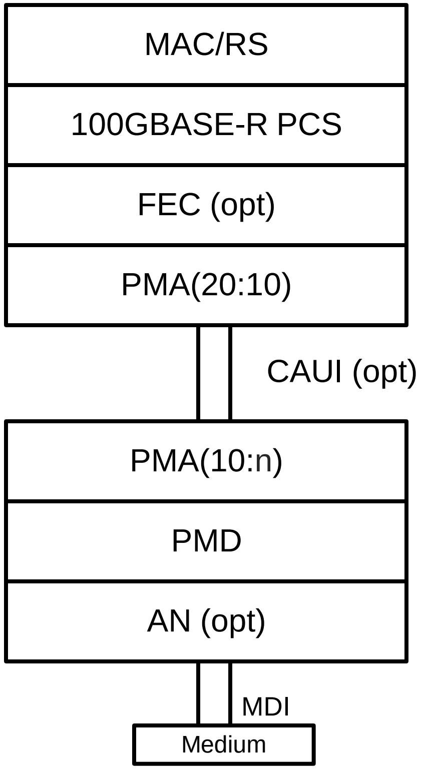
Introduction

- The following slides explore the feasibility of a 400GbE PCS
- A couple of feasible PCS architecture options are shown at 400GbE, building on the 802.3ba PCS and the work that has been done within P802.3bj so far

100GbE Architecture in Review

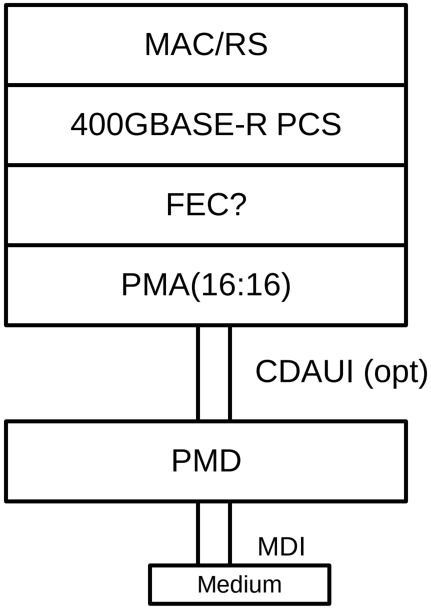
- Based on a 20 Lane PCS with 64B/66B encoding (5 Gb/s per PCS Lane)
- Data is striped to PCS lanes 66-bit blocks at a time
- Alignment Markers are periodically added to all PCS lanes to enable alignment in the RX PCS
- PMAs do simple bit multiplexing to change lane widths
- Lane widths of 20, 10, 5, 4, 2, 1 can all be supported
- Optional KR based FEC is supported

- Pros of this architecture
  - Very flexible, can support future lane widths without a PCS change
  - Most of the complexity is in the PCS, PMAs are very simple bit multiplexers
- Cons of this architecture
  - KR based FEC is not very strong and has high latency
  - With 25G SerDes and DFE we are seeing a higher probability of correlated errors which can cause MTTFPA issues with bit interleaved PCS lanes
- P802.3bj is adding strong FEC to the architecture (below the PCS), but then we lose the flexibility of changing lane widths by bit multiplexing



400GbE Possible Architecture #1

- Base on a 16 Lane PCS with 64B/66B encoding (25 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
- Lane widths of 16, 8, 4, 2, 1 can all be supported
- FEC support??
- Optional 16 lane CDAUI interface
- Pros of this architecture
  - Very flexible, can support future lane widths without a PCS change
  - Most of the complexity is in the PCS, PMAs are very simple bit multiplexers
  - Root for PCS lanes is 25G, which is becoming mainstream
- Cons of this architecture
  - What to do about FEC?
- This architecture for 400GbE would be feasible to implement in current ASIC or FPGA technology



400GbE Possible Architecture #1 cont

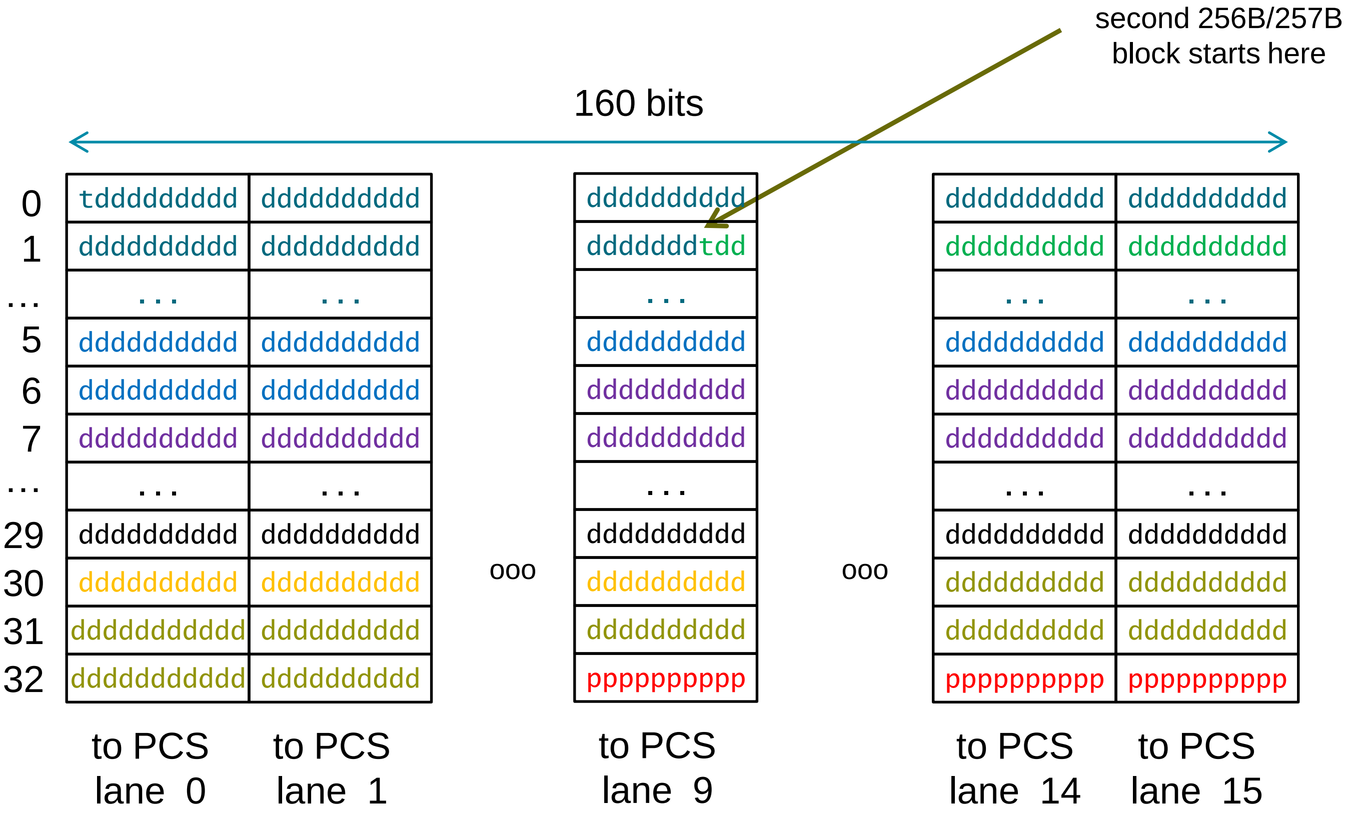
- Why 16 PCS Lanes?
  - 25G SerDes are becoming mainstream, so a 16x25G CDAUI interface is technically feasible and seems like a likely first thing to standardize
    - A CFP form factor module could support 16x25G lanes with a 0.6mm pitch connector
  - Can be used with the optical technology developed for 100 Gb/s
    - 100GBASE-LR4 has 25G lanes
    - 100GBASE-SR4 has 25G lanes
    - If an optical interface is based on advanced modulation was developed, then that most likely looks like a 100G lane
    - All of these work well with 16 PCS lanes at 25 Gb/s each (aggregate of 400 Gb/s)
  - Less is better, the more PCS lanes the more logic for AM lock SMs, more registers to keep etc.
- Another option is 80 PCS lanes (5 Gb/s, just 4 x 20 PCS lanes from 100GE)
  - This could make designing a single PCS that supports 4x100GbE and 1x400GbE simpler?
    - Depends on FEC and other things so not necessarily true
  - Downside is this is a lot of logic for 400GbE

400GbE PCS Architecture #2 (with FEC)

- It is very likely that multiple 400GbE interfaces will require FEC
  - For instance a speculative SR16 PMD could leverage SR4 which is currently being developed within IEEE, to get to 100m we may need FEC (25G per lane)
  - A PMD based on advanced modulation technology would also very likely need FEC, but not clear if this could be common with other FEC requirements
  - A speculative LR16 might not need FEC
  - A 25G-VSRx16 (CDAUI) electrical interface won't need FEC, even a longer reach 25G electrical interface likely won't need FEC, but a future 50G electrical interface might?
    - With 25G PCS lanes there is no bit interleaving on 25G lanes so we are not that susceptible to MTTFPA burst error degradation
- What if we add FEC to the PCS and when changing widths we re-multiplex based on RS symbol boundaries?
  - This can allow us to be flexible in our widths but still keep the correction properties of the RS intact
- One possibility:
  - Support 16 lanes at 25.78125G for each PCS lane, encode directly into 256B/257B, add in alignment markers per PCS Lane, stripe to each PCS Lane on RS boundaries
    - This allows you to change lane widths (from 16 down to 8, 4, 2, 1)
    - Do you have to deskew at each step – no!
    - Just align to the nearest RS symbol per lane
    - Should be very low latency also
- This architecture for 400GbE would be feasible to implement in current ASIC or FPGA technology

FEC Frame Structure

- Use 256B/257B encoding directly (not transcoding)
- Re-use the RS FEC code from 802.3bj, RS(528,514) but add the FEC into the PCS sublayer
- The last 14 10-bit words are the FEC checksum



second 256B/257B block starts here

160 bits

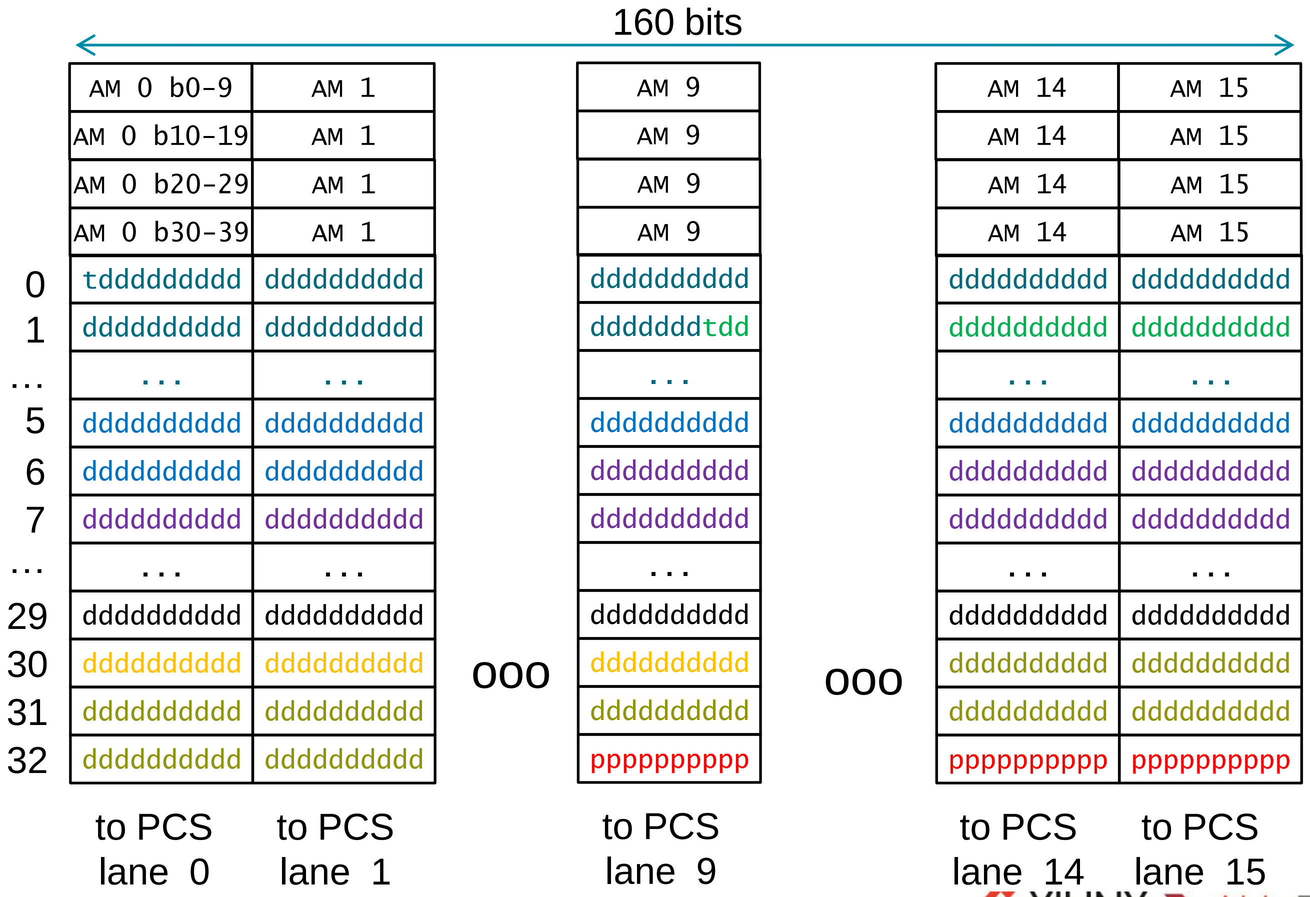
	to PCS lane 0	to PCS lane 1	to PCS lane 9	to PCS lane 14	to PCS lane 15
0	tddddddddd	dddddddddd	dddddddddd	dddddddddd	dddddddddd
1	dddddddddd	dddddddddd	dddddddtdd	dddddddddd	dddddddddd
…	. . .	. . .	. . .	. . .	. . .
5	dddddddddd	dddddddddd	dddddddddd	dddddddddd	dddddddddd
6	dddddddddd	dddddddddd	dddddddddd	dddddddddd	dddddddddd
7	dddddddddd	dddddddddd	dddddddddd	dddddddddd	dddddddddd
…	. . .	. . .	. . .	. . .	. . .
29	dddddddddd	dddddddddd	dddddddddd	dddddddddd	dddddddddd
30	dddddddddd	dddddddddd	dddddddddd	dddddddddd	dddddddddd
31	ddddddddddd	dddddddddd	dddddddddd	dddddddddd	dddddddddd
32	ddddddddddd	dddddddddd	pppppppppp	pppppppppp	pppppppppp

ooo

ooo

Alignment Markers

- Add an alignment marker to each PCS periodically, it does not need to be part of the FEC blocks, and it seems to make it easier if they are not part of the FEC block (so you don’t have Alignment issues)
- Below the AMs are 40 bits each, but this is flexible, just must be nx10-bit

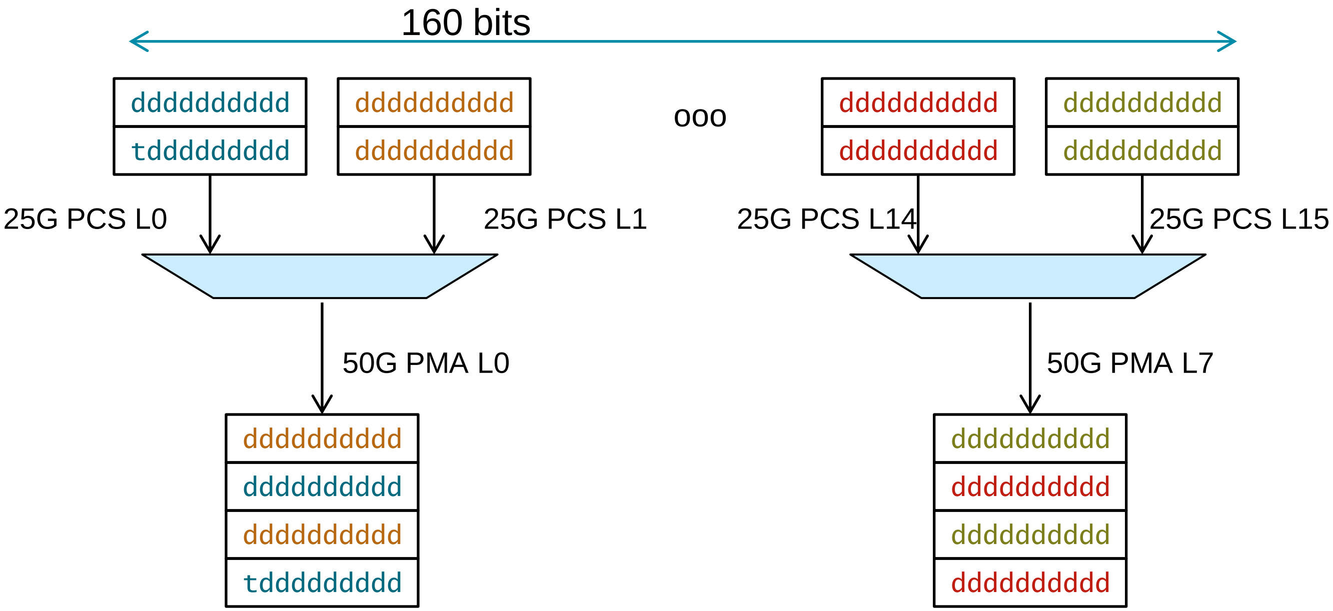


160 bits

	AM 0 b0-9	AM 1		AM 9		AM 14	AM 15
	AM 0 b10-19	AM 1		AM 9		AM 14	AM 15
	AM 0 b20-29	AM 1		AM 9		AM 14	AM 15
	AM 0 b30-39	AM 1		AM 9		AM 14	AM 15
0	tddddddddd	dddddddddd		dddddddddd		dddddddddd	dddddddddd
1	dddddddddd	dddddddddd		dddddddtdd		dddddddddd	dddddddddd
...	...	...		...		...	...
5	dddddddddd	dddddddddd		dddddddddd		dddddddddd	dddddddddd
6	dddddddddd	dddddddddd		dddddddddd		dddddddddd	dddddddddd
7	dddddddddd	dddddddddd		dddddddddd		dddddddddd	dddddddddd
...	...	...		...		...	...
29	dddddddddd	dddddddddd		dddddddddd		dddddddddd	dddddddddd
30	dddddddddd	dddddddddd	ooo	dddddddddd	ooo	dddddddddd	dddddddddd
31	dddddddddd	dddddddddd		dddddddddd		dddddddddd	dddddddddd
32	dddddddddd	dddddddddd		pppppppppp		pppppppppp	pppppppppp
	to PCS lane 0	to PCS lane 1		to PCS lane 9		to PCS lane 14	to PCS lane 15

Multiplexing

- With 16 PCS lanes, you can multiplex down to 8, 4, 2, or 1 lane(s)
- All multiplexing must be on RS boundaries (10-bit in the case shown)
  - To preserve error correction capability in the face of burst errors
- First you must find alignment marker lock to find 10-bit boundaries, then you multiplex on RS boundaries
  - No need to deskew the various lanes
- Below shows muxing from 16 lanes down to 8 lanes



Summary

- There are many possible solutions for a 400GbE PCS, this paper shows a couple of options that are feasible with today’s technology (either ASIC or FPGA)
- One simple option is scaling the 802.3ba PCS up in speed
- But if there will be interfaces that 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

Thanks!