16 ‘v’ 80 PCS Lanes for 400GbE
An implementer's Perspective

Cedrik Begin, Gary Nicholl – Cisco Systems

IEEE 802.3 400 Gb/s Ethernet Study Group

IEEE 802.3 September 2013 Interim

York, UK

Topics

- Background
- Contribution Recap
- 400GE & 4x100GE Implementation Options
- Summary

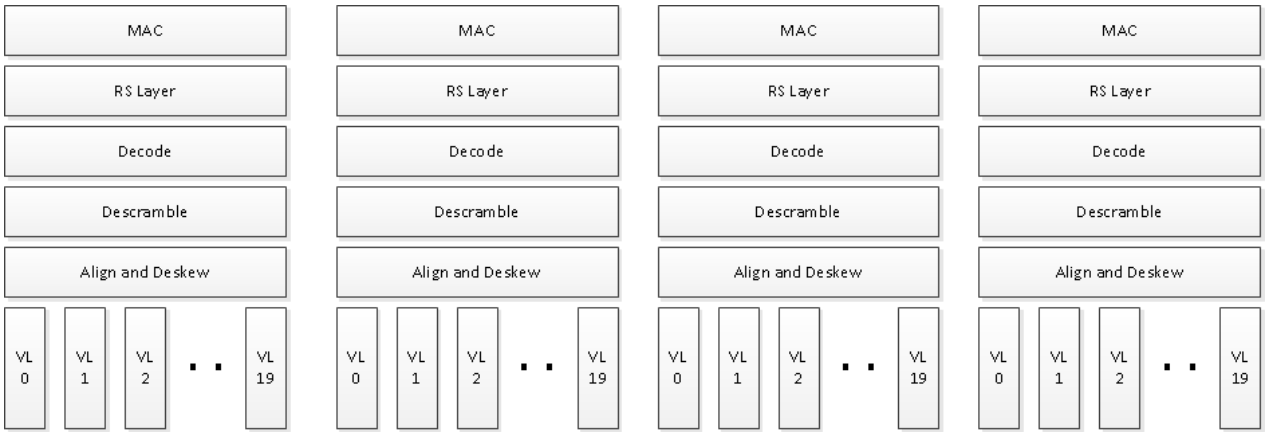
Background

- One point of discussion that has come up during previous meetings, is whether the 400GbE PCS should be based on 16 or 80 PCS lanes
- The primary argument for 16 PCS lanes is that it is simple and forward looking (less is more !)
- The primary argument for 80 PCS lanes is that it makes more reuse of existing 100G technology, and therefore enables a more efficient implementation of a dual rate 1x400G and 4x100G MAC chip
  - Also supports 40Gb/s lanes (i.e. 10x40G)
- This presentation investigates the technical feasibility of both from an implementation perspective.

Contribution Recap

- Here are some previous contributions on the topic.
- gustlin_400_01b_0513 [Slide 7]
- ghiasi_400_01a_0513 [Slide 9]
- gustlin_400_02_0713 [Slides 7 & 8]
- wang_400_01_0713
- ghiasi_400_01_0713 [Slides 7-9]

400GE – 4 x 100G Reuse

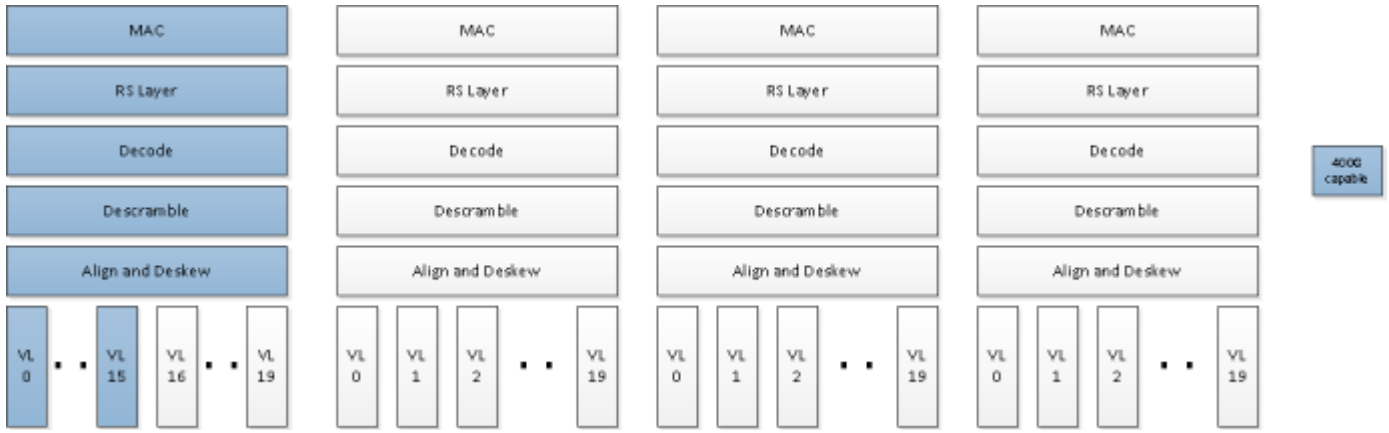


- Start with 4x100GE MACs+PCS.
- Each 100GE MAC/PCS based on 20 PCS Lanes (@ 5.15 Gb/s)
- Note that for simplicity's sake only the Rx path is shown.

400GE – 4 x 100G Reuse Option #1

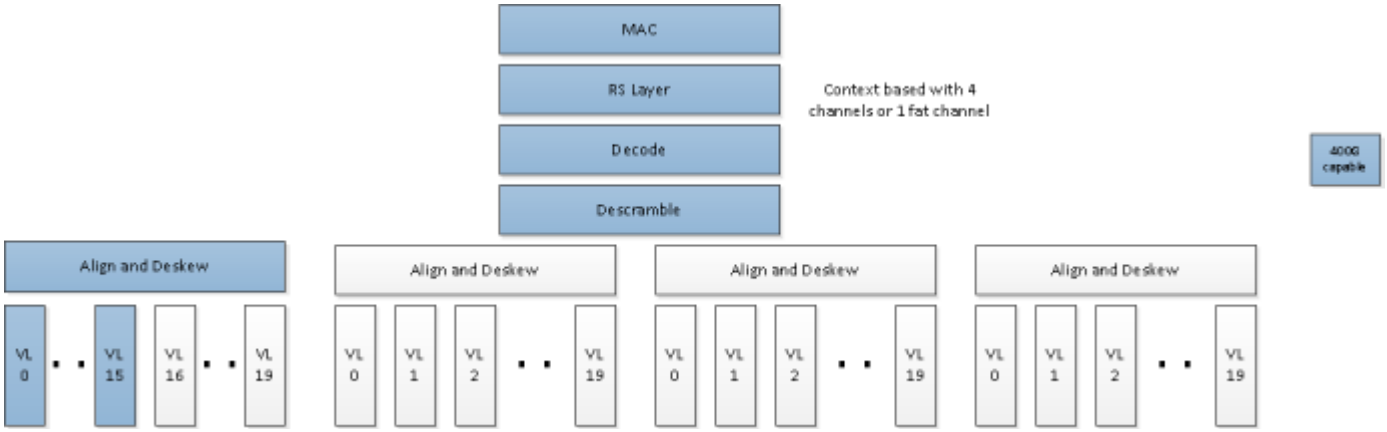
- Add 400G capable data pipe of MAC, RS layer, decode and descrambling (this pipe can also operate at 100G for 4x100G operation)
- Align, deskew, reorder done over 80 VLs (challenging).
  - 4x100G 2:1 mux count = 10032
  - 1x400G 2:1 mux count = 41712
- Need 80 different alignment markers.

400GE – 4 x 100G Reuse Option #2



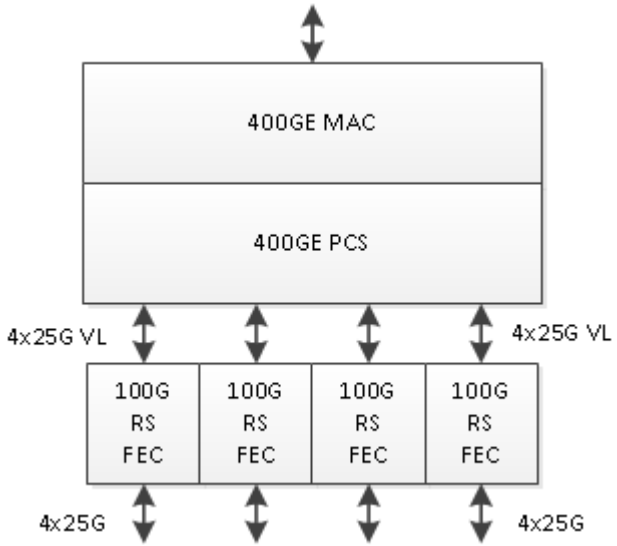
- Like option 1, Add 400G capable data pipe of MAC,RS layer, decode and descrambling (this pipe can also operate at 100G for 4x100G operation)
- 16 of the 80 VLs are 25G capable.
- Deskew, align, reorder much simpler.
  - 1x400G 2:1 Mux count = 8448
  - 3x100G 2:1 Mux count = 7524

400GE – 4x100GE Reuse Option #3



- Like option 2, except that the MAC,RS layer, Decode and descramble operates in either 1x400G or 4x100G modes.
- 16 of the 80 VLs are 25G capable.
- Deskew and demux much simpler.

100G RS-FEC Reuse Proposal



- Send 4x25G VL to each 100G RS FEC.
- PCS sends and expects Alignment Markers every 16399 data blocks on a given VL. (such that alignment markers align to beginning of each 820th 5280-bit word).
- Transcode function: in alignment marker removal/insertion combines 4 66-bit alignment markers into one 257-bit block and vice versa. Ensure that AM0 travels on FEC lane 0, AM1 travels on FEC lane 1…
- FEC alignment lock and deskew will look for Alignment markers every 820th codeword

Summary

- Both 16 and 80 lane PCS options are technically feasible
- However our analysis shows that a 16 lane PCS solution provides the most efficient implementation for both a single rate 400GE MAC and a dual rate 400G/4x100G MAC.
- The main challenge is in the PCS lane reorder block. PCS lane reorder over 80 lanes (400G mode), is significantly more complex than PCS lane reorder over 4 x 20 lanes (4x100G mode).
  - 4x20 lane reorder = ~10,000 2:1 mux eqv.
  - 1x80 lane reorder = ~42,000 2:1 mux eqv.
  - 1x16 lane reorder = ~ 8,500 2:1 mux eqv

4 x 20 $\neq$ 1 x 80 !!