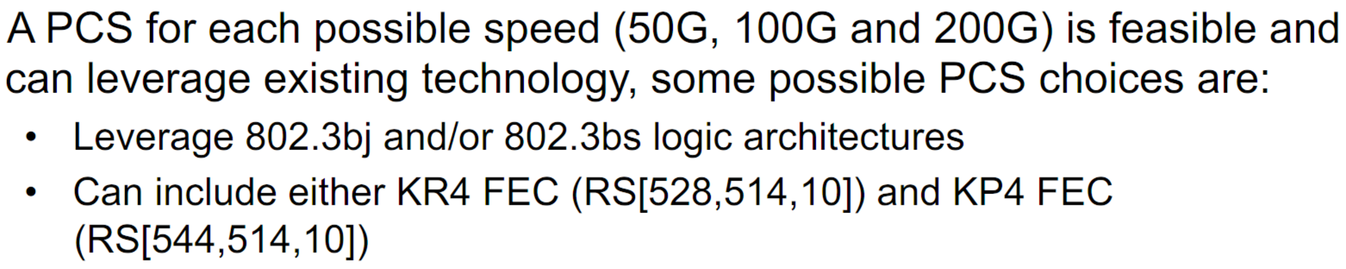
Technical Feasibility of MAC/PCS Layer

Xinyuan Wang, Tongtong Wang

Background and Introduction

- For MAC/PCS technical feasibility in 50 Gb/s Ethernet Over a Single Lane and Next Generation 100 Gb/s & 200 Gb/s Ethernet Call For Interest Consensus Presentation



A PCS for each possible speed (50G, 100G and 200G) is feasible and can leverage existing technology, some possible PCS choices are:

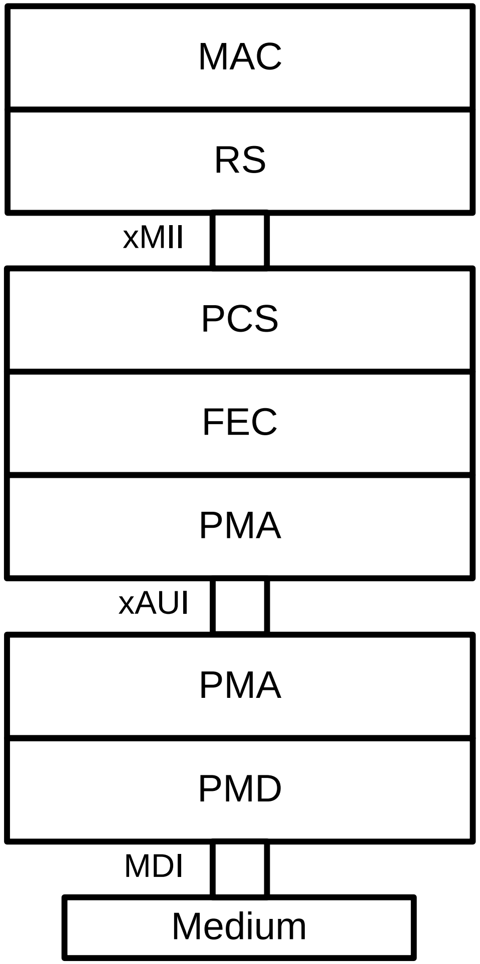
- Leverage 802.3bj and/or 802.3bs logic architectures
- Can include either KR4 FEC (RS[528,514,10]) and KP4 FEC (RS[544,514,10])

- In this contribution, we investigate MAC/PCS approach for 50GbE and NG 100/200GbE from perspective of multi-lanes MAC/PCS architecture and multi-rate implementation

Assumptions for Investigating MAC/PCS

- **PMDs in 50GbE and NG 100/200GbE**
  - Backplane/Twinax Cable/MMF/SMF based on 56Gbps PAM4 technology
- **Single lane vs Multi Lanes**
  - 25Gbps & 50Gbps per lane for electrical interface
  - At least 50Gbps per lane for physical link of PMDs
  - So, 25Gbps per PCS lanes, in line with 802.3bs 400GbE
  - Multi lanes needed for PCS architecture
- **FEC**
  - KR4 FEC: RS(528,514) and/or KP4 FEC: RS(544,514)
- **Enable maximum reuse in Multi rate implementation**
- **50GbE architecture analysis as start point in this contribution**

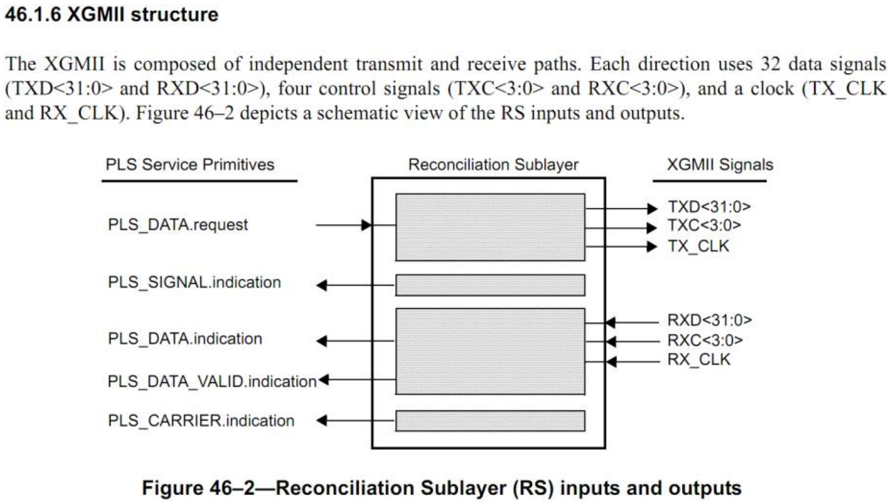
Observation on Logic Architecture



- MAC: Similar as previous 10/100/400GbE
- RS/MII: CL81/117 with 64bit data width or CL46/106 with 32bit data width
- PCS:
  - ✓ Encoding: CL82/119 or CL49/108
  - ✓ AM Structure
  - ✓ RS FEC: Algorithm, Architecture and Implementation
  - ✓ Scramble
- PMA: Bit mux or Block mux

RS and MII

- For 50/NG 100/200GbE, either CL46 in 10GbE or CL81 in 100GbE is feasible, both could use scalable parallel implementation to support higher rate Ethernet

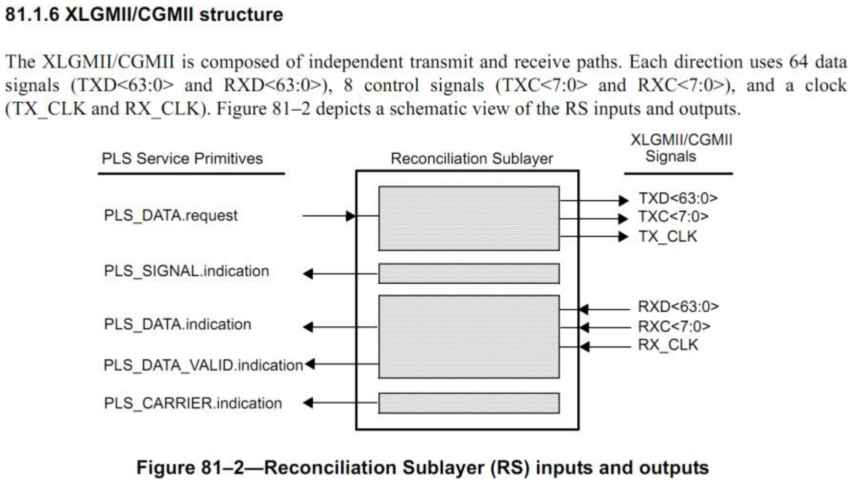


46.1.6 XGMII structure

The XGMII is composed of independent transmit and receive paths. Each direction uses 32 data signals (TXD<31:0> and RXD<31:0>), four control signals (TXC<3:0> and RXC<3:0>), and a clock (TX_CLK and RX_CLK). Figure 46–2 depicts a schematic view of the RS inputs and outputs.

PLS Service Primitives
Reconciliation Sublayer
XGMII Signals
PLS_DATA.request
PLS_SIGNAL.indication
PLS_DATA.indication
PLS_DATA_VALID.indication
PLS_CARRIER.indication
TXD<31:0>
TXC<3:0>
TX_CLK
RXD<31:0>
RXC<3:0>
RX_CLK

Figure 46–2—Reconciliation Sublayer (RS) inputs and outputs



81.1.6 XLGMII/CGMII structure

The XLGMII/CGMII is composed of independent transmit and receive paths. Each direction uses 64 data signals (TXD<63:0> and RXD<63:0>), 8 control signals (TXC<7:0> and RXC<7:0>), and a clock (TX_CLK and RX_CLK). Figure 81–2 depicts a schematic view of the RS inputs and outputs.

PLS Service Primitives
Reconciliation Sublayer
XLGMII/CGMII Signals
PLS_DATA.request
PLS_SIGNAL.indication
PLS_DATA.indication
PLS_DATA_VALID.indication
PLS_CARRIER.indication
TXD<63:0>
TXC<7:0>
TX_CLK
RXD<63:0>
RXC<7:0>
RX_CLK

Figure 81–2—Reconciliation Sublayer (RS) inputs and outputs

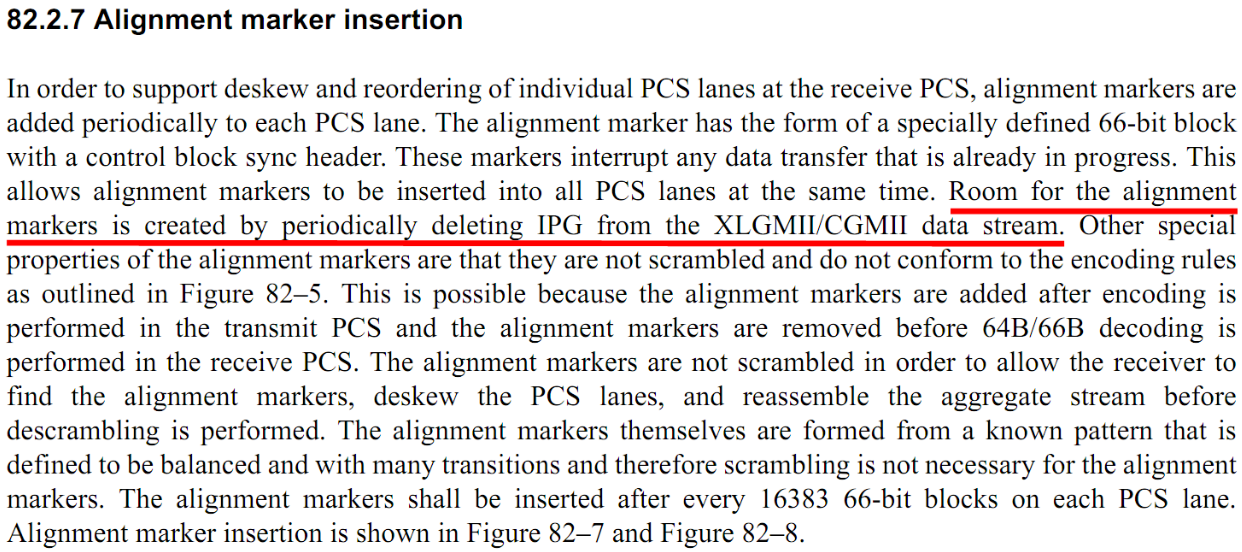
- Further extend RS&MII data bus to be larger than 64bits, e.g. 128b, is not doable, because it will violate Deficit Idle Counter mechanism and minimum IPG requirement, thus compromise line rate transmission in Ethernet

64/66Bit Encode/Decode

- 64/66Bit Encode/Decode as in CL49 and 82 is effective to lower physical line rate
- Block type difference between CL49 in 10GE and CL81 in 40/100GE may cause interoperation problem

Alignment Marker

- From 802.3ba 40/100GE, PCS layer utilizes AM mechanism to support multi PCS lanes



82.2.7 Alignment marker insertion

In order to support deskew and reordering of individual PCS lanes at the receive PCS, alignment markers are added periodically to each PCS lane. The alignment marker has the form of a specially defined 66-bit block with a control block sync header. These markers interrupt any data transfer that is already in progress. This allows alignment markers to be inserted into all PCS lanes at the same time. Room for the alignment markers is created by periodically deleting IPG from the XLGMII/CGMII data stream. Other special properties of the alignment markers are that they are not scrambled and do not conform to the encoding rules as outlined in Figure 82–5. This is possible because the alignment markers are added after encoding is performed in the transmit PCS and the alignment markers are removed before 64B/66B decoding is performed in the receive PCS. The alignment markers are not scrambled in order to allow the receiver to find the alignment markers, deskew the PCS lanes, and reassemble the aggregate stream before descrambling is performed. The alignment markers themselves are formed from a known pattern that is defined to be balanced and with many transitions and therefore scrambling is not necessary for the alignment markers. The alignment markers shall be inserted after every 16383 66-bit blocks on each PCS lane. Alignment marker insertion is shown in Figure 82–7 and Figure 82–8.

- Reuse mature technology in CL82/81 enable multi PCS lanes requirement in 50/NG 100/200GbE
  - ✓ Further analysis in TF is needed for reusing from 802.3bj or 802.3bs

KR4 or KP4 RS FEC

- For easy analysis, assuming most FEC coding gain is to cover physical link of PMDs and BER Objective is 1E-12 in 50GbE/NG 100GbE

RS FEC(n,k,t,m)	CG	NCG*	BERin	Overhead	SerDes Rate	Block Time	Latency**	Area Ratio
RS(528,514,7,10)	5.39	5.28	5.30E-05	0%	25.78125	102.4ns	~175ns	1X
RS(544,514,15,10)	6.64	6.39	3.60E-04	3.03%	26.5625	102.4ns	~197ns	2.9X

*: Block time and latency based on 50GbE

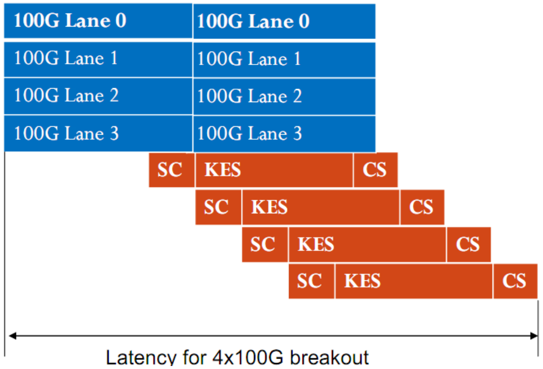
*: Refer to wang_x_3bs_01a_0115

- Either KR4 RS FEC and KP4 RS FEC is feasible with minor difference in latency for 50GbE
- FEC performance and architecture selection depend on physical link of PMDs requirement

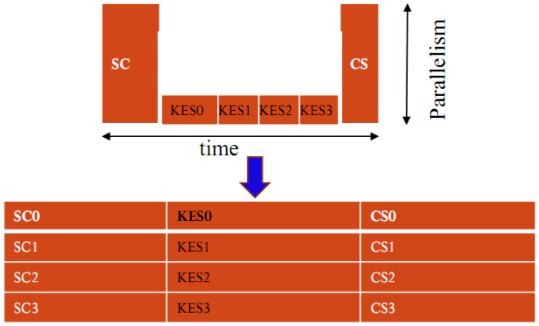
RS FEC with Breakout

- For RS FEC, Supporting Breakout is similar as implementing Multi-rate Ethernet with 50G/NG 100G/200GbE in one ASIC
- As in “sun_3bs_01_0715” , breakout can be achieved by logic or time sharing

- Time sharing



- Logic sharing

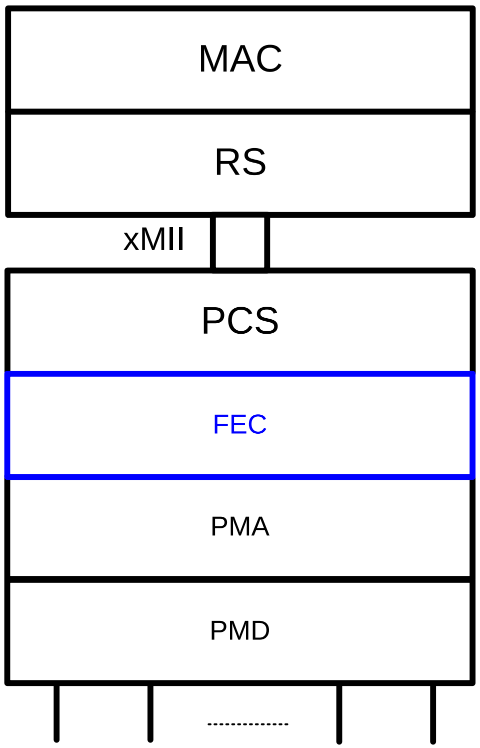


Breakout example:1x400G FEC to 4x100G KP4 FEC

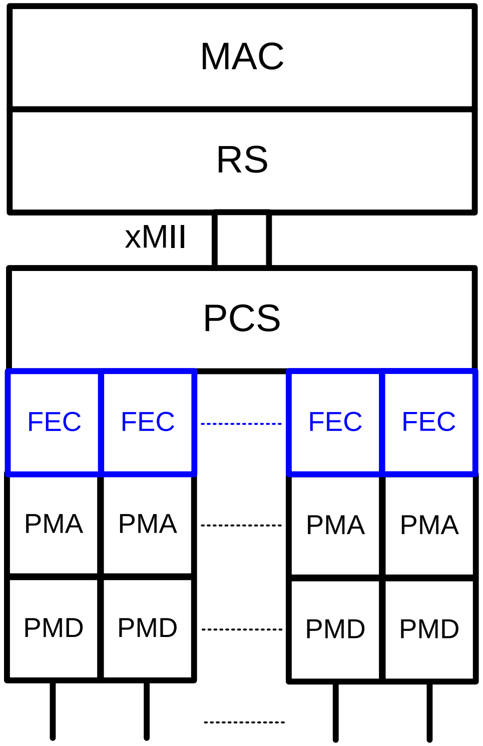
- Additional logic resource and latency for logic or time sharing to support Breakout

RS FEC: Per PHY or Per Lane Architecture

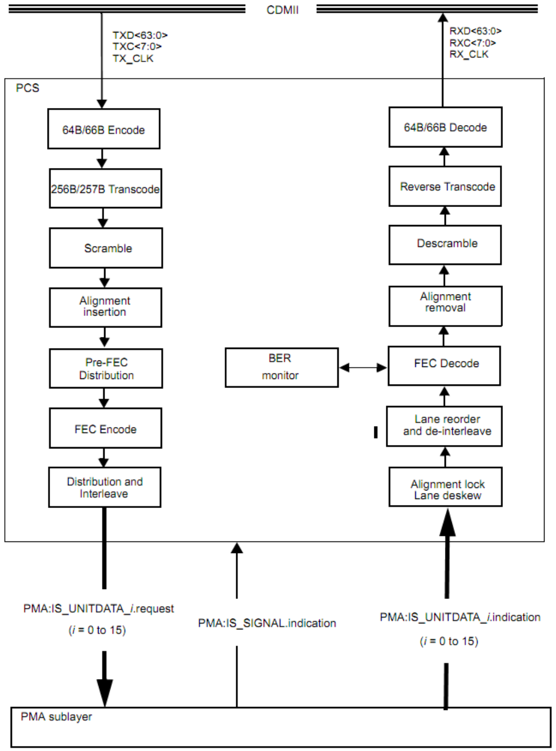
- FEC per PHY



- FEC per lane



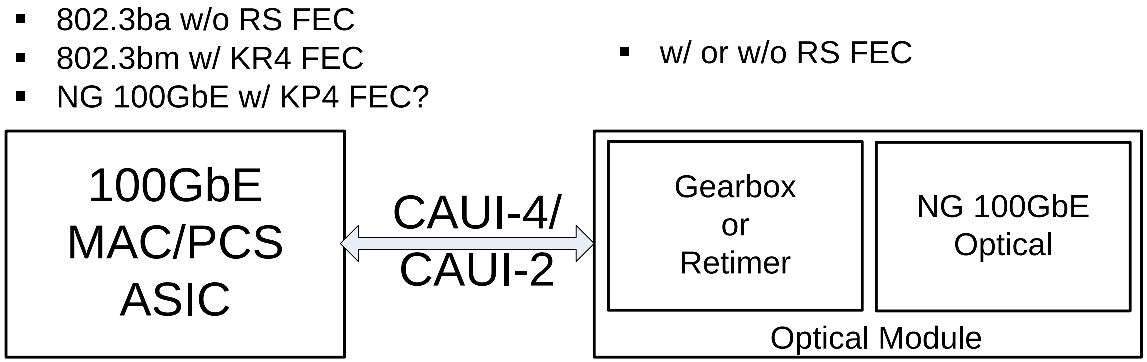
- 802.3bs 400GbE



- FEC per PHY is better than per lane with advantage of low latency
- FEC per Lane give higher FEC coding gain in multiplexing for facing burst error
- Tradeoff in latency and FEC performance as FEC architecture of 802.3bs

RS FEC: Backward Compatibility

- 50/200GbE by 50Gbps PAM4 physical lane has no backward compatibility requirement for it will be a new chip and optical module, while NG 100GE need to consider it



- 802.3ba w/o RS FEC
- 802.3bm w/ KR4 FEC
- NG 100GbE w/ KP4 FEC?

- w/ or w/o RS FEC

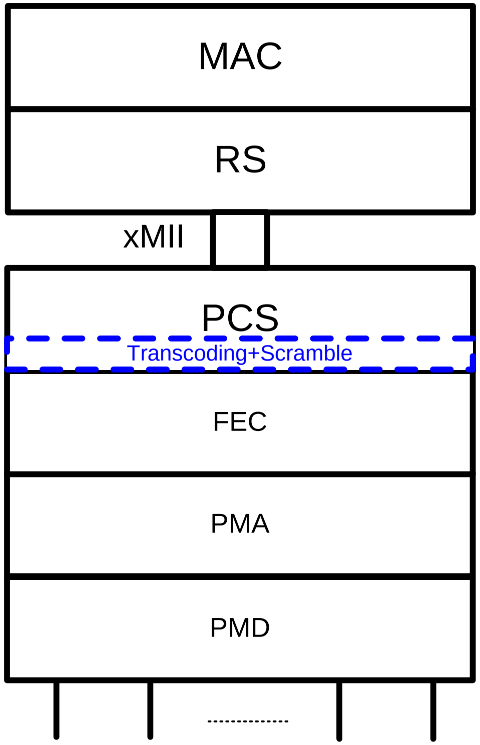
- Backward compatible requirement is only in host ASIC with complying 802.3bm and CAUI-4 interface

PMA: Bit mux or Block mux

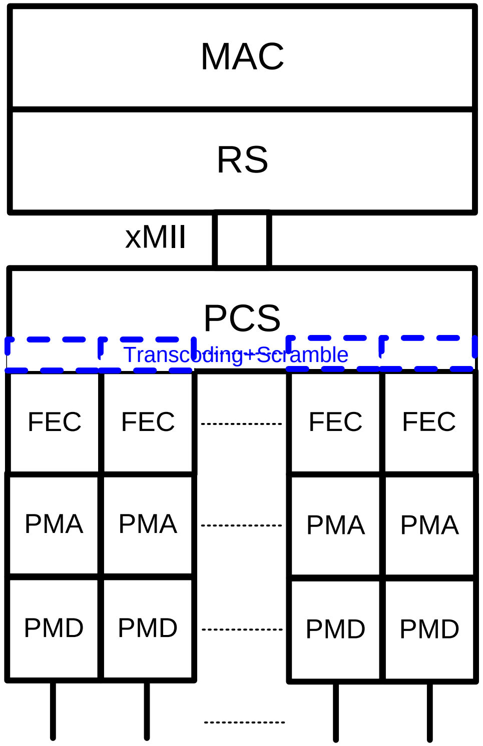
- Bit Mux is preferred,
  - Enable protocol agnostic optical module and reuse in OTN/Fiber channel/Inifiniband
  - 802.3ba w/o RS FEC or 802.3bs w/ KP4 RS FEC
- Block mux:
  - Only protocol aware optical module
  - 802.3bj w/ KR4/KP4 RS FEC
  - Better performance in FEC for facing burst error than Bit Mux

Scramble: Per PHY or Per Lane

- Scramble in per PHY



- Scramble in per lane



- Reuse "Transcoding+Scramble" scheme in 802.3bs
- Scramble with Per PHY is much better than per lanes as no baseline/clock wander issue if same seed for each scrambler

Summary

- For 50/NG 100G/200GbE, MAC/PCS layer with RS FEC, either RS(528,514) or RS(544, 514), is technical feasible
- Most portion of 802.3bs logic layer can be reused in this new project
- Further work is needed to clear FEC performance requirement from optical and electrical links

Thank You