Multi-rate support in 400GbE logic layer?

Xinyuan Wang, Tongtong Wang,
Wenbin Yang, Suping Zhai

July 2013

HUAWEI

Background

- In the first SG meeting at Victoria, We get the following strawpoll result:

Strawpoll #3 Made by the Chair
Are you interested in multi-rate support (backward compatibility from 400GE to 100GE and/or 40GE):

Results

Yes	50
No	10

Strawpoll #4 Made by Mark Gustlin

I believe that FEC should be an integral part of the 400GbE architecture

Results

Yes	44
No	1
Undecided	27

http://www.ieee802.org/3/400GSG/public/13_05/minutes_400_01_0513_unapproved.pdf

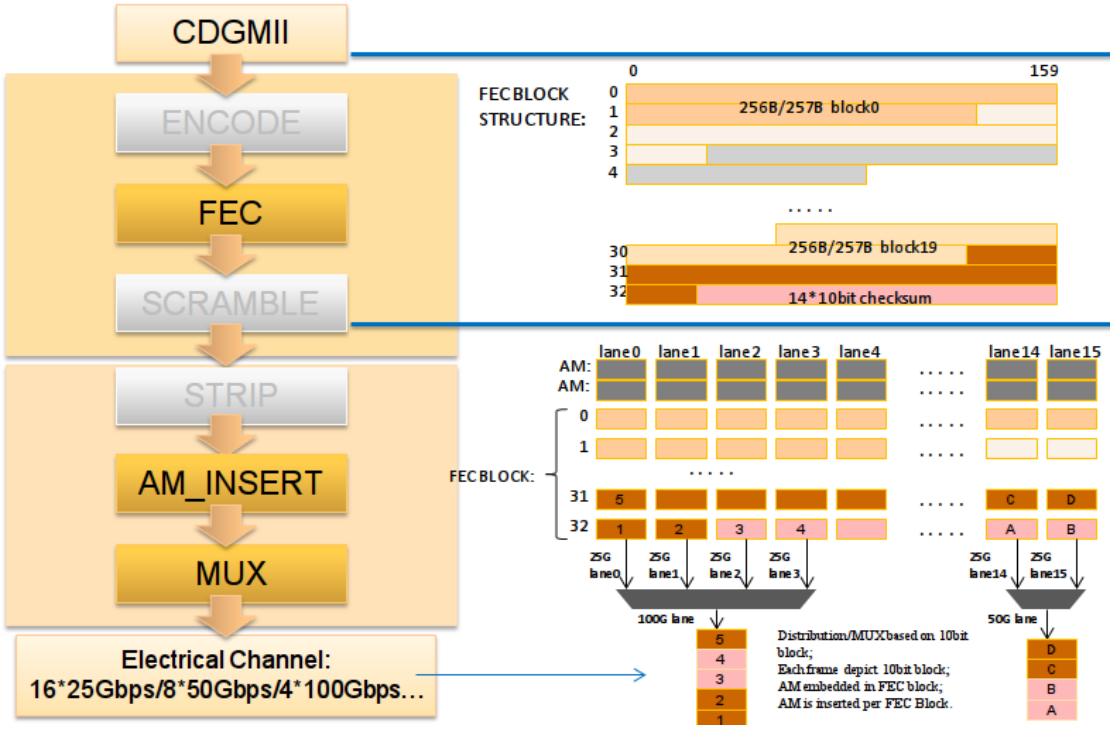
- Is there any technical/economic feasibility to support 100GbE and/or 40GbE in 400GbE logic architecture?

Background(Cont'd)

- 100GbE/40GbE Architecture: 802.3ba
  - Use 20/4 PCSLs(Virtual Lanes) with 64/66B coding (5.15625/ 10.3125 per PCSL) .
  - High latency & low gain FEC(2112,2080) implemented per PCSLs(Virtual Lanes) .
- 100GbE Architecture: 802.3bj
  - Use 20 PCSLs(Virtual Lanes) with 64/66B coding (5.15625 per PCSL) .
  - RS(528, 514, 7, 10)/RS(544, 514, 15, 10) for 100GBASE-KR4/KP4.
  - The architecture is based on 64/66B to 256/257B transcode, 64/66b scrambler, RS-FEC and symbol distribution to 4-25Gbps Physical Lanes.
- Possible 400GbE without FEC Architecture
  - Upgrade from 802.3ba architecture.
  - Use 16/80 PCSLs(Virtual Lanes) with 64/66B coding (25.78125/5.15625 per PCSL) .

Possible 400GbE with FEC Architecture

- Encode/decode by either 64/66B or 256/257B;
- Use RS(528, 514, 7, 10)/RS(544, 514, 15, 10) for most PMD, such as 16X25GB NRZ for 10km Duplex SMF or 8X25GB PAM4 for 2km Duplex SMF;
- FEC Symbol based distribution in PCS/PMA;



Scenarios of Multi-rate support in 400GbE

- What is the essential requirement of multi-rate support? And is it feasible?
  - Scenario #1: 400GbE Architecture without FEC interoperate with 802.3ba 100 and/or 40GbE.
  - Scenario #2: 400GbE Architecture with FEC interoperate with 802.3ba 100 and/or 40GbE.
  - Scenario #3: 400GbE Architecture with FEC interoperate with 802.3bj 100 and/or 40GbE.
  - Scenario #4: 400bGE Architecture will support 100Gbps and/or 40Gbps, with no interoperating with 802.3ba and/or 802.3bj.
- All above scenarios are divergent and have different requirements on 400GbE architecture. Which one shall we choose?

Multi-rate support scenario #1: Compatible with 802.3ba w/o FEC

- If logic resource is not reusable between 400GbE and 100/40GbE, it makes no sense for 400GE to support 802.3ba.
- MAC/RS sub-layer can share between 400GbE and 100/40GbE.
- 400G PCS implementation with 16 VL is rather different from 20/4 VL scheme of 802.3ba, sharing few common logics; while 80 VL PCS could be compatible with 4x100GbE and 400GbE, bearing significantly more cost.
- Consider the following statistic data, which is from FPGA implementation of a general 400GbE receive end without FEC:

Sub-layer	Clock rate	Logic resource			
		LUT#	REG	Percent（LUT）	Percent（REG）
MAC/RS	312MHz	60k	61k	34%	28%
PCS	312/156MHz	104k	144K	60%	67%
PMA	161MHz	10k	10K	6%	5%

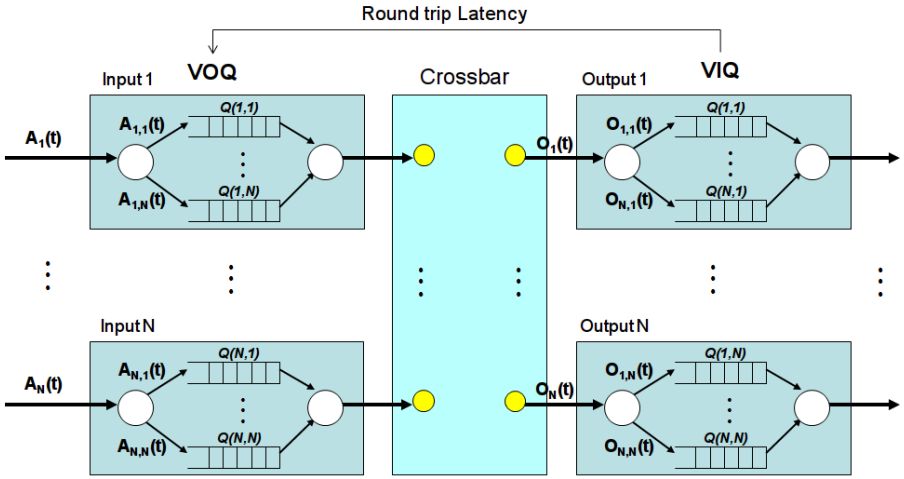
- PCS sub-layer is around 60-70% of total resource required. If PCS share no common resources, we would rather to have different PCS specification for 400GbE and 100/40GbE,each one for its own best performance and cost.

Multi-rate support scenario #2: Compatible with 802.3ba with FEC

- IEEE 802.3ba define FEC(2112,2080) on each PCSL(Virtual Lanes) and bit multiplex on physical lanes.
- If 400GE integrate RS-FEC, it will implement it in PCS and distribute symbols on 16/8/4/2/1 Physical lanes.
- Since the FEC algorithm is different in 100/40GbE(firecode) and 400GbE(RS-FEC), bit flow in physical lanes is entirely different. It is meaningless to define a 400GbE standard in this scenario with multi-rate support as 100GbE/40GbE in 802.3ba.
- 400GE logic implementation in this scenario is different to 802.3ba 100/40GbE, and only limited logic in MAC/RS sub-layer is in common.

Multi-rate support scenario #3: Compatible with 802.3bj

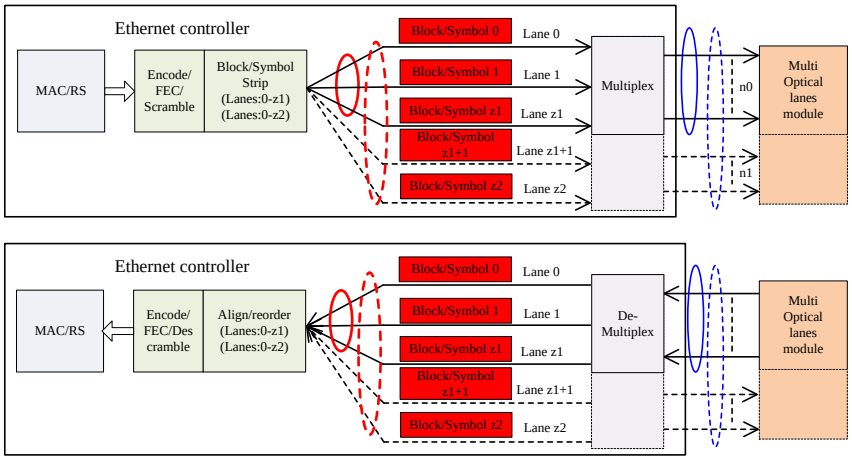
- In 802.3bj for 100GBASE-KR4/KP4, FEC symbol distribute to 4 Physical Lanes(25GB NRZ/12.5GB PAM4 SerDes).
- The suggested 400GbE architecture can interoperate with 802.3bj by 20/80 lanes (1X100GbE/4X100GbE) exactly as 802.3bj requires, keeping 256/257b transcoding.
- The main difference between 4x100G 802.3bj and 400GbE is RS-FEC implementation and its latency.
  - If reuse some of the 802.3bj RS-FEC in 400GbE, higher latency would be induced than doing RS-FEC across all lanes, which requires 4 times larger fabric buffer to balance latency in backplane interconnect application.



  - Low latency is also required in DC switch with ethernet in backplane application.
  - Multi-rate support in 400GE backplane is about balance of RS-FEC implementation and latency. We should have a different RS-FEC realization in 400GbE for its very low latency.

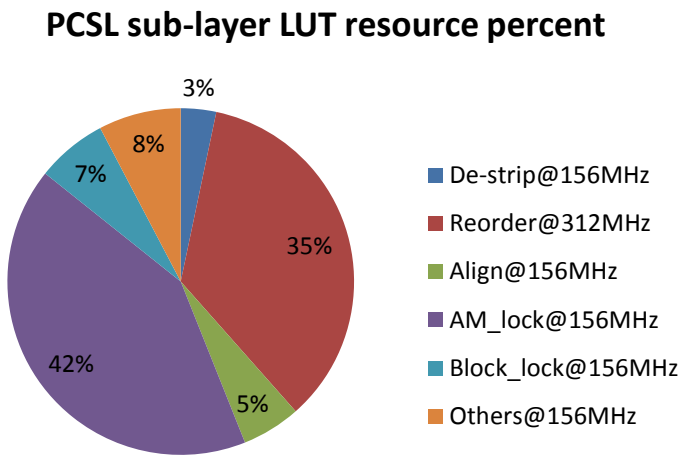
Multi-rate support scenarios #4: No compatible with 802.3bj, but with arbitrary sub-interfaces

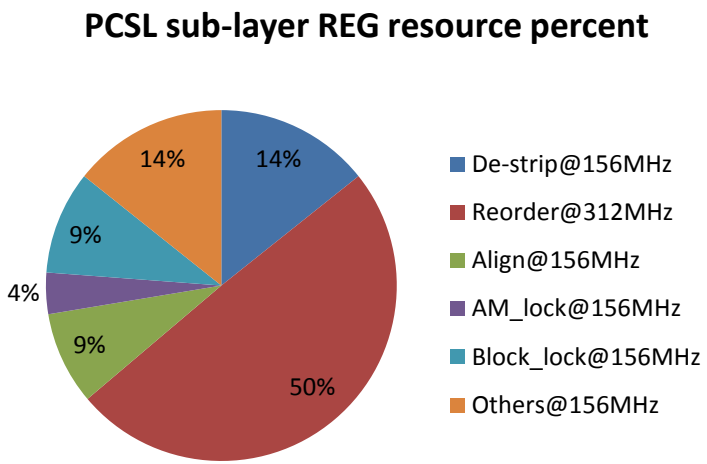
- What about having multi-rate support in 400GbE is just to have 100Gbps and/or 40Gbps sub-interfaces but non-compatible with 802.3 ba/bj, will it be acceptable?
- Yes? !
  - We could define a 400GbE with new 100Gbps/40Gbps sub-interface. This idea will base on minimum bandwidth unit in electrical/optical lanes and bundle any arbitrary numbers of physical lanes in 400GbE. 25Gbps will be an ideal candidate granularity.
  - We will define a flexible/scalable MAC/PCS/FEC architecture for this purpose. More work need to be done in future if this idea is interesting to the industry.



Non Multi-rate support and compatible: PCSL number?

- PCSL number is the key factor in 400GbE logic layer. The difference in PCSL number will result in different data width & clock rate per PCSLs;
- In PCS layer, the PCSLs reorder and AM lock process make up most of total area consumption and these functions can not share in different PCSLs number architectures;
- Consider the following example, which is an implementation of the most general/complex receive end for 400GbE with 80 lanes on FPGA without FEC:





- The complexity of reorder block directly depends on the number of PCSLs, which also affect total area consumption.
- 16VLs structure only use 20% data width of the scheme of 80VLs and thus requires less area. It can also scale down to 8, 4, 2, or 1 physical lane(s) in the future 400GbE PMD.

Summary

- Use 400GbE to support multi-rate and 100GbE/40GbE compatibility is not a good architecture in logic layer.
- 16 PCSLs in 400GbE is less resource/area requiring in FPGA/ASIC and is an ideal choice with no support to multi-rate 400GbE.
- Is it worthwhile to make a 400GbE architecture with multi-rate sub-interfaces yet 100GbE/40GbE non-compatible?
  - If Yes, the following item should be focus:
    - Sub-rate granularity;
    - FEC algorithm and integrate in MAC/PCS architecture is key factor.
    - Need to define a flexible logic layer for 400GbE,

Thank you