Update and Further Study on FEC Architecture Proposal

Xinyuan Wang, Tongtong Wang

HUAWEI TECHNOLOGIES CO., LTD.

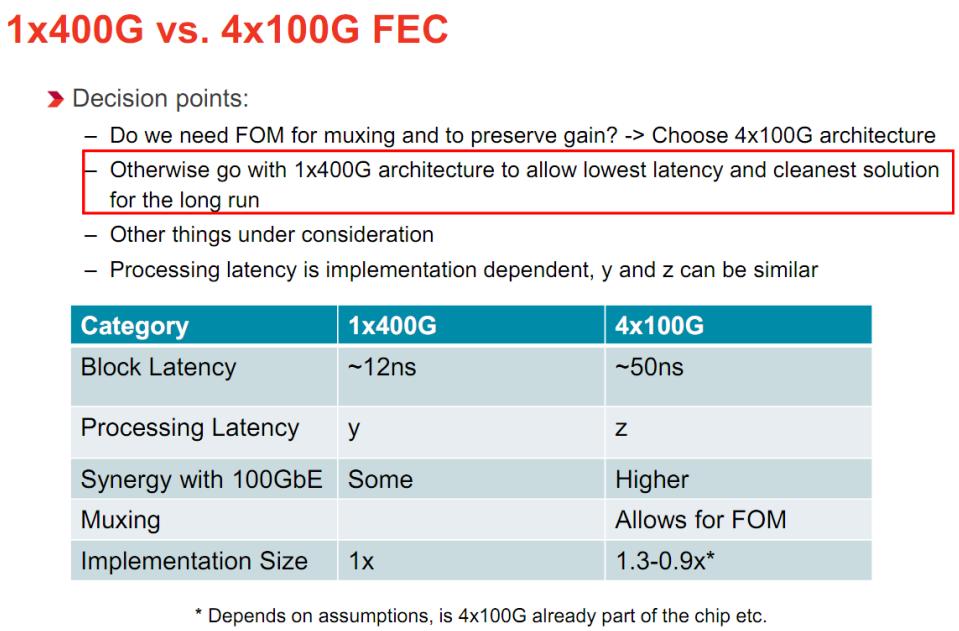
IEEE 802.3bs 400 GbE Task Force

Introduction and Motivation

- This presentation further investigates the FEC architecture for 1X400Gbps versus 4X100Gbps implementation based on KP4 RS FEC
- After Pittsburgh Meeting, how to stripe ingress data flow to FEC instance is still key item to be investigated for moving 400GbE standard forward
- The following FEC architecture open topics in RED related are investigated in this contribution
  - Technical feasibility on 1X400Gbps and 4X100Gbps
  - FEC performance on different bit mux scheme
  - Long run evolution
  - Enable breakout
  - Flexible Ethernet

Long Run Evolution of FEC Architecture?

- In Slides 13 of joint contribution in "gustlin_3bs_02a_0315":



1x400G vs. 4x100G FEC

- Decision points:
  - Do we need FOM for muxing and to preserve gain? -> Choose 4x100G architecture
  - Otherwise go with 1x400G architecture to allow lowest latency and cleanest solution for the long run
  - Other things under consideration
  - Processing latency is implementation dependent, y and z can be similar

Category	1x400G	4x100G
Block Latency	~12ns	~50ns
Processing Latency	y	z
Synergy with 100GbE	Some	Higher
Muxing		Allows for FOM
Implementation Size	1x	1.3-0.9x*

* Depends on assumptions, is 4x100G already part of the chip etc.

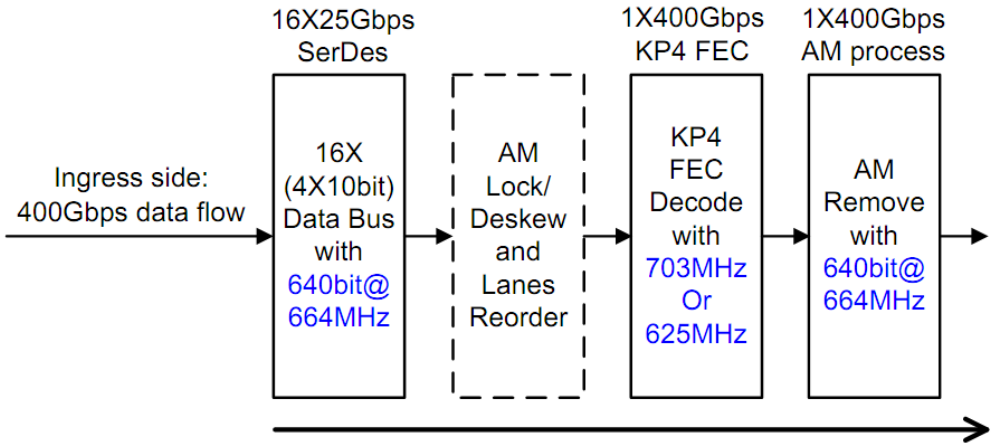
- In "anslow_3bs_03_0515":

> To start with, assume a 1 x 400G FEC architecture to allow lowest latency and cleanest solution for the long run (gustlin_3bs_02a_0315 page 13).

- Question:
  - From now and near future technology, 1X400G FEC is still a lowest latency and cleanest solution?
  - What is sweet point of 1X400G FEC with long run evolution?

Issus for Arch A (1X400G RS(544,514)) over 16 lanes

- In "wang_x_3bs_01_0515", 1X400G FEC is not a simple and clean implementation by current process technology.



- A simple issue is that 544/64 (= 8.5) is not an integer*
  - That means, physically, current KP4 FEC design need to be re-considered to work with offset, more cost(area/latency) needed to adopting current SerDes interface
    - Option 1: running at 680bit@625MHz data bus in 8 cycle to complete one FEC codeword encode/decode
    - Option 2: running at 640bit@703MHz data bus in 9 cycle to complete one FEC codeword (over clocking)
    - Option 3: running at 640bit@625MHz data bus in 8 cycle, more logic inside RSFEC block to process the offset
- AM header must be distributed and restored traversing 16 Lanes, thus 160bit granularity is mandatory and higher complexity in option 1 due to data bus width mismatch in function block

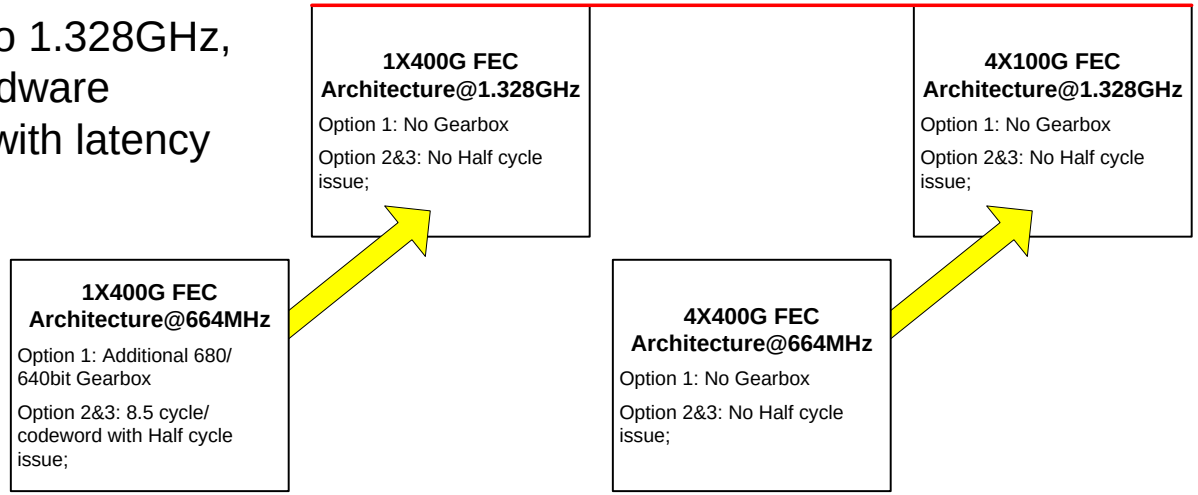
**For 100G .bj FEC over 4 lanes, 544/16 = 34*

Long Run Evolution of FEC Architecture

- Assuming FEC@1.328GHz for long run with comparing to 664MHz right now and regardless of possible power consumption issue for running at 1.328GHz

Numer of Symbols	Data Bus Width Per Lanes(Bit)	Clock Rate	Data Bus Width Per 400Gbps FEC(Bit)	Number of clock cycle for 400Gbps FEC	Data Bus Width Per 100Gbps FEC(Bit)	Number of clock cycle for 100Gbps FEC
1	10	2.65625GHz	160	34	40	136
2	20	1.328GHz	320	**17**	80	**68**
3	30	885MHz	480	11.333	120	45.333
4	**40**	**664MHz**	**640**	**8.5**	160	**34**
5	50	531MHz	800	6.8	200	27.2
6	60	443MHz	960	5.667	240	22.667
7	70	379MHz	1120	4.857	280	19.429
8	**80**	**332MHz**	**1280**	**4.25**	320	**17**
9	90	295MHz	1440	3.778	360	15.111
10	100	265MHz	1600	3.4	400	13.6

- Even for long run evolution to 1.328GHz, 1X400G FEC get similar hardware complexity to 4X100G FEC with latency advantage
- 400GbE standard lifecycle VS long run, how far away approached for 1.328GHz?



Why Enable Breakout in 400GbE?

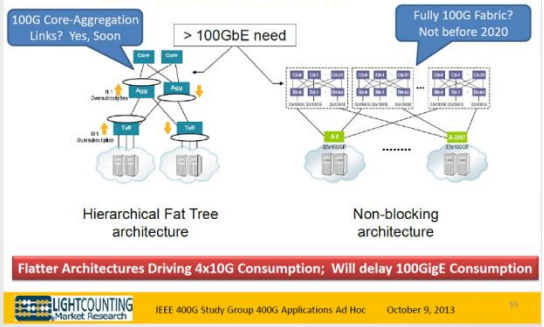
- November 2013, Copy and paste from Joint contribution "Breakout Functionality" by John D'Ambrosia and David Law

Looking to the Future

Future Breakout Scenarios

400G Call for Interest Slide

Data Center Architectures



Flatter Architectures Driving 4x10G Consumption; Will delay 100GigE Consumption

LIGHTCOUNTING Market Research IEEE 400G Study Group 400G Applications Ad Hoc October 9, 2013

Source: Dale Murray, LightCounting, http://www.ieee802.org/3/400GSG/public/adhoc/app/murray_app_01a_1013.pdf

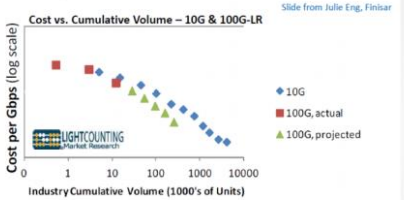
Observations for 400GbE

- Reasonable assumption that 40G/100G will ship in greater volumes than 400G.
- Multiple higher density 40G/100G scenarios envisioned by 400GbE time frame.
- Multiple scenarios can be envisioned where 400GbE ports could support higher density / lower rate 40GbE and or 100 GbE PMDs. Some include:
  - 400 GbE based on 16 x 25 Gb/s
    - Could be divided into 4 ports of 100G @ 4 x25Gb/s
  - 400 GbE based on 8 x 50 Gb/s
    - Run 50Gb/s at 40 Gb/s for 8 ports of 40GbE
    - Divide into 4 ports of 100G @ 2 x 50Gb/s
  - 400 GbE based on 4x 100Gb/s (assuming modulation)
    - Divide into 4 ports of 100G @ 1 x 100Gb/s
    - Change modulation to support 40G and support 4 ports @ 1 x 40 Gb/s

Leveraging Lower Speeds

100GigE has to follow the same curve

Slide from Julie Eng, Finisar



- 100G falls on the same curve as 10G for the volume shipped
- Cost projected to erode more quickly than 10G
- Best way to reduce cost of 100G components: Bring on the Volume!

LIGHTCOUNTING Market Research IEEE 400G Study Group 400G Applications Ad Hoc October 9, 2013

Cost Reductions
- Integration via higher port density
- Volume

400 GbE implementations with breakout can drive lower costs via higher density lower speeds

Shared volumes can drive lower cost for 400 GbE

Source: Dale Murray, LightCounting, http://www.ieee802.org/3/400GSG/public/adhoc/app/murray_app_01a_1013.pdf

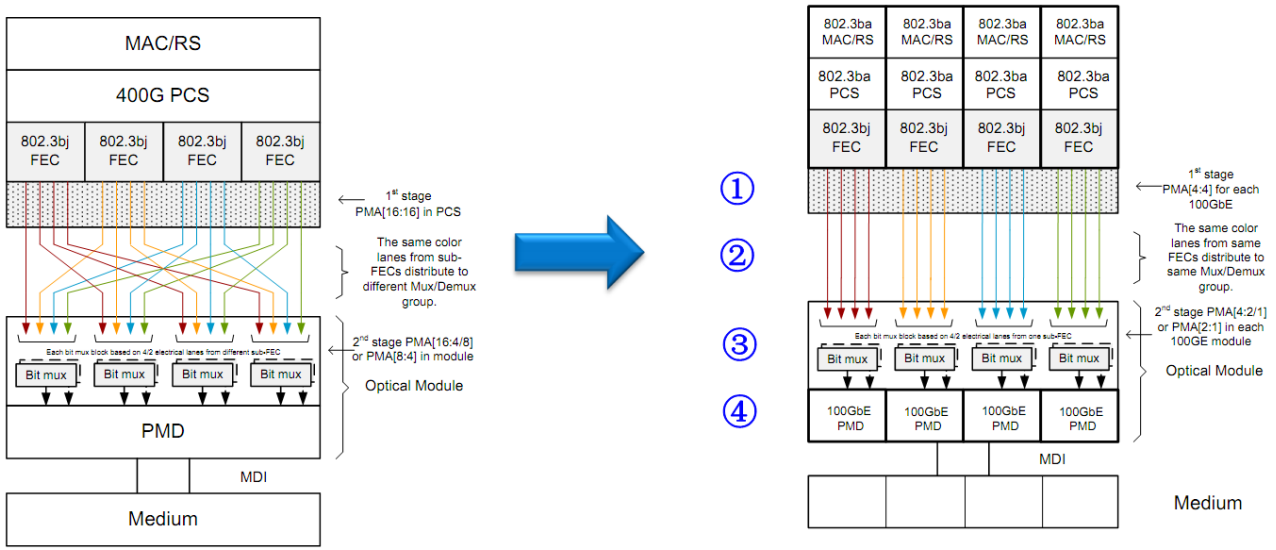
Conclusions

- The market is adopting this "breakout functionality" with 10GbE / 40GbE
  - Breakout functionality – the ability to use a port in a lower rate / higher density mode of operation
- Providing an upgrade path forward could further improve this scenario for lower speeds
- "Breakout functionality" will enhance broad market potential of 400GbE by enabling adoption to support higher density / lower rate lower speeds to enable lower 400GbE cost.
- Proposed objective–
  - Provide appropriate support for breakout functionality

How to Enable Breakout in 400GbE?

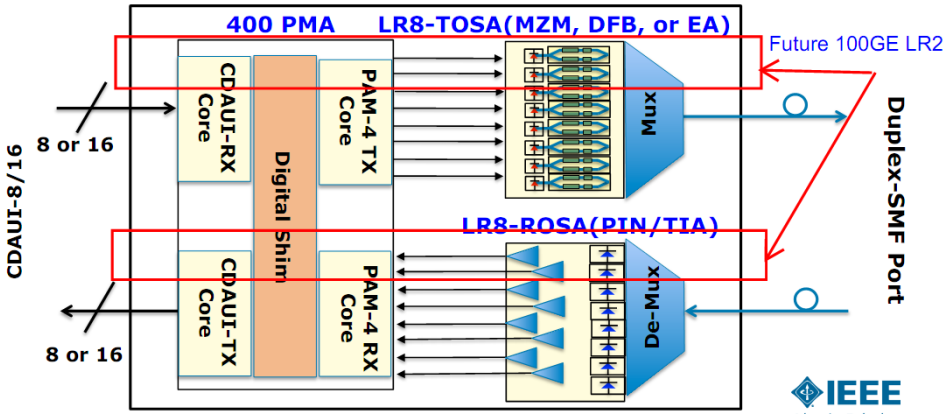
- From general observation, breakout is from PMD perspective. However, it is a essentially systemic scheme in 400GbE. It should help increase broad market potential without more additional cost to 400GbE
- July 2014, in "wang_x_3bs_01_0714", breakout in 400GbE was investigated

  - Breakout should be implemented in the following four sub-layer.
    - ① In host ASIC side, for lower silicon cost purpose;
    - ② One unified CDAUI interface layout for both implementation;
    - ③ Share one gearbox solution for 1X400GbE/4X100GbE;
    - ④ PMD breakout for reuse;



Breakout in 400GbE from Optical Module perspective

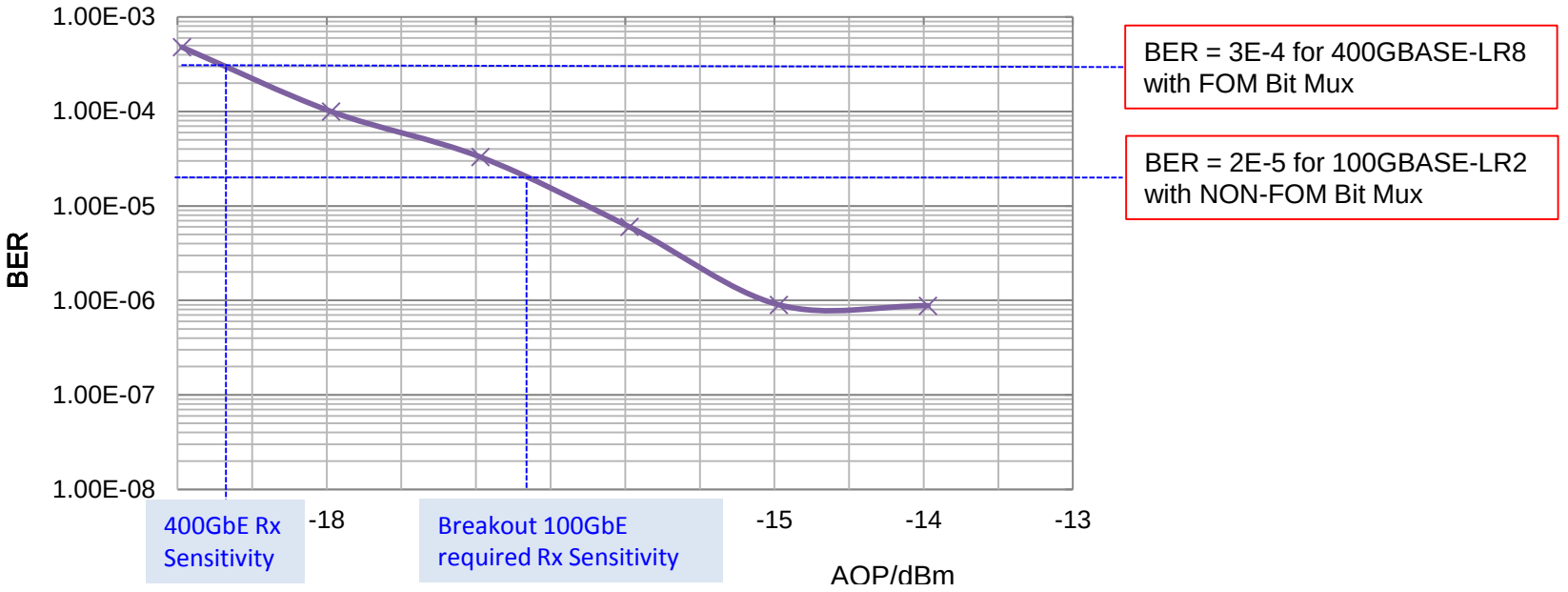
- For 100m over MMF and 500m over SMF objective with parallel fiber, solution and benefits of breakout is easy to understand
- For 2/10km over SMF with duplex fiber:
  - Share one optical solution/platform for 400GE and breakout 100GE
    - Take 400GBASE-LR8 from "ghiasi_3bs_01b_0515" as example, it can used to breakout into possible 4X(100GBASE-LR2) independent new 100GE module



- 400GBASE-LR8 use 8:1 optical Mux/DeMux@3.5dB, while 100GbE use 2:1 optical Mux/Demux@1.5dB. This difference should be included in Breakout 100GE PMDs. Similar result as future 4X100G PAM4 in 2km.

Physical Link Margin for Breakout 400GBASE-LR8 into 4X100GE

- For 4dB additional loss decrease by 2:1 optical Mux/DeMux, it is possible to allocate 2dB for extra optical link margin, the other 2dB for improve Receiver sensitivity.
- So in the following test result, the corresponding optical link BER is 2E-5 in breakout100GE under the similar solution with 400GBASE-LR8.
- 100G FEC should improve BER from 2E-5 to 1E-12, as defined in 100GbE.
- Error floor in FEC performance is still an issue in Breakout 100G PMD if with NON-FOM Bit Mux and CDAUI-8.



Breakout from ASIC perspective

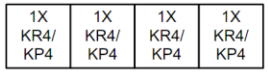
- In "wang_x_3bs_01_0315" of Berlin meeting, from one ASIC to implement 1X400GbE and 4X100GE perspective, areas are estimated for the following most popular scenarios

Area Estimate of 1x400 & 4x100GbE Compatible FEC

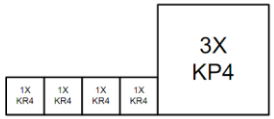
- 4x100Gbps *VS* 1x400Gbps+4x100Gbps

Scenario 2: KR4 FEC in 100GbE and KP4 FEC in 400GbE

Area of 1X KR4 FEC= a
Area of 1X KR4/KP4 FEC= b =2.9a



- FEC architecture Option 1:
  4X100Gbps KR4/KP4 FEC:
  4X=4X2.9a=11.6a



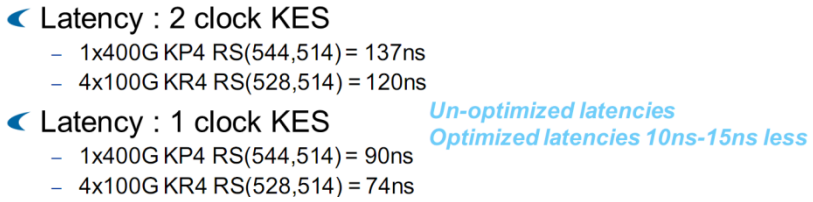
- FEC architecture Option 2:
  4x100Gbps KR4 + 1X400Gbps KP4 FEC:
  4a+3X(2.9a)=12.7a

- Using 4x100Gbps FEC(Option 1) for 400GbE is more area efficient in compatible FEC design
- If scale up from more realistic 100G FEC* , the area for Option 2 is enlarged to 14.15a

- If only KR4 FEC is required in 100GbE, both 4X100G and 1X400G FEC architecture have similar area cost.
- If KP4 FEC is required in 100GbE, 4X100G FEC architecture have low area advantage and better FEC performance.
- From ASIC flexible perspective, 4X100Gbps FEC architecture in 400GbE is more reasonable

What is Difference to Enable Breakout in 1X400G and 4X100G FEC Architecture?

- In "langhammer_02_0615_logic" , logic sharing solution in 1X400G KP4 FEC can enable breakout into 4X100G KR4 FEC
- 1X400G FEC already faces more difficulties in fitting before breakout, adding in more logic to enable 4x100G FEC breakout, additional timing closure problem is further accumulated and increase 1X400G FEC latency in "langhammer_02_0615_logic".



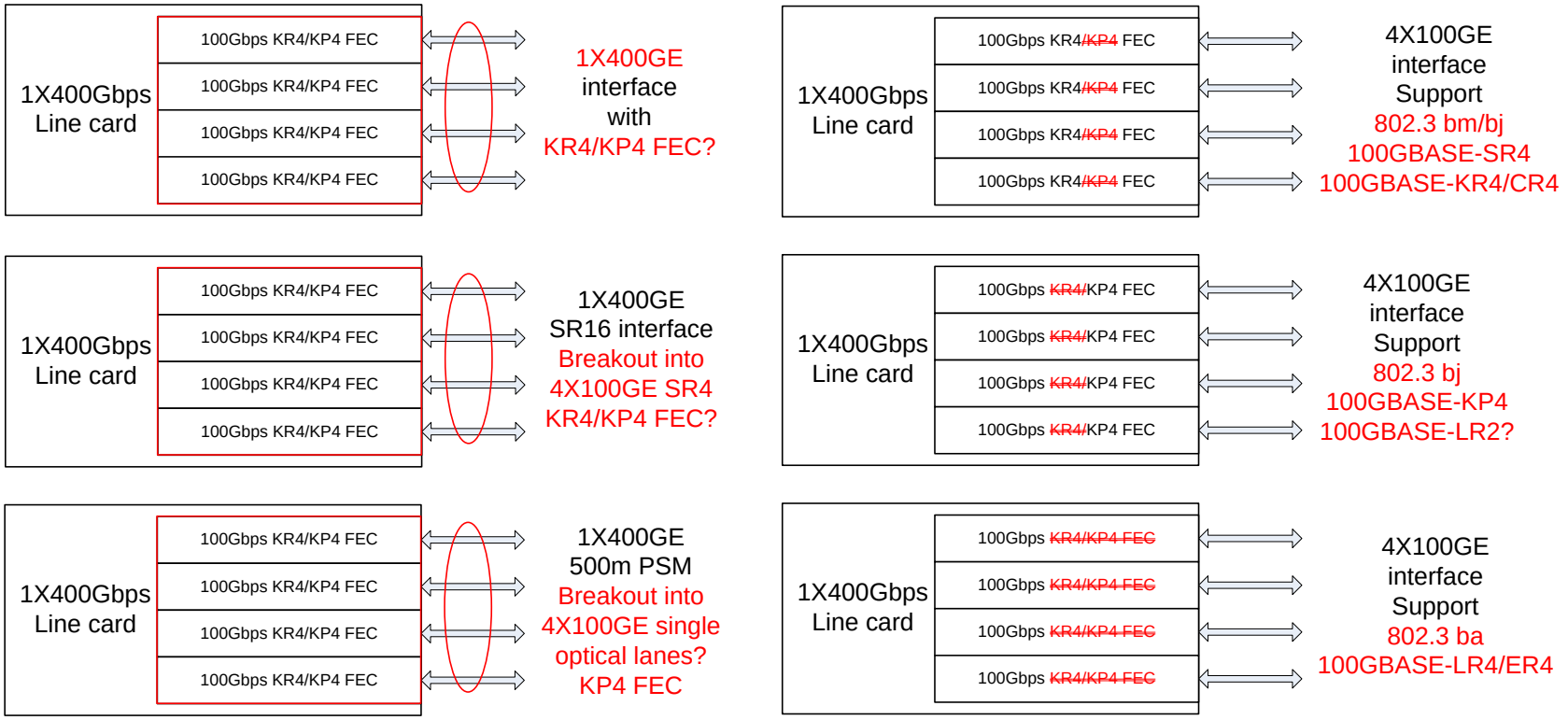
Additional Error marking latency ~12ns should be included as in "wang_x_3bs_01a_0115"

With no low latency advantage of 1x400G FEC nor easy implementing advantage of 4x100G FEC

- Comparing to that, 4x100G FEC architecture has latency at ~110ns at either 1X400GbE or 4X100GbE to enable breakout in "sun_01_0615_logic", and easy to implement, faster to market.

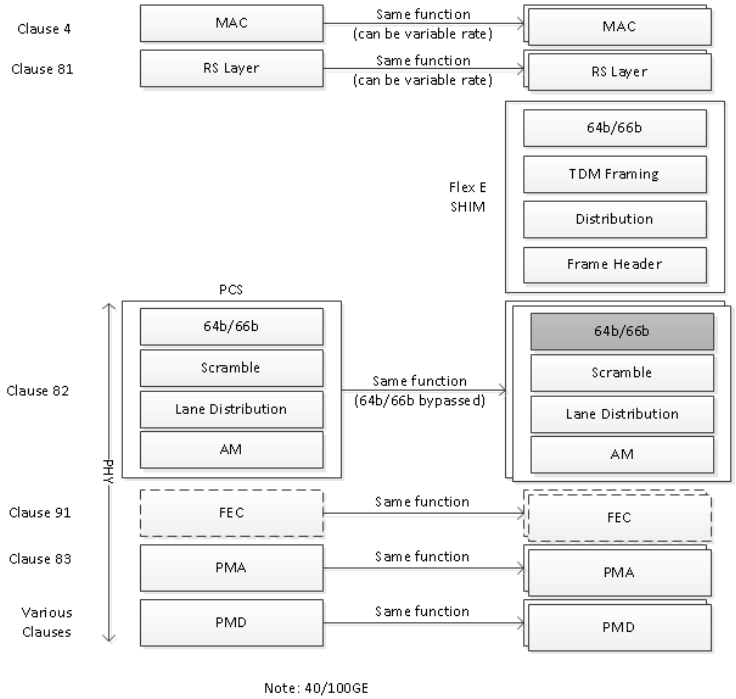
Breakout from System Perspective: 400GbE with 4X100Gbps FEC Architecture

- In order to support 400GbE and breakout into 4X100GbE, based on 4X100Gbps KR4/KP4 FEC(802.3bj) architecture, a unified host line card implementation can be realized to lower investment and achieve more robust system



Flex Ethernet and Relationship with 400GbE

- A Initial Text Proposal “oif2015.127.01” for FlexE was adopted at the Q215 OIF meeting in Lisbon
  - FlexE will mainly used in Router to Transport connection as in the initial proposal, similar as 2/10km objective in IEEE 400GbE project
  - Proposed that the first version of the Implementation Agreement specify bonding of 100GBASE-R PHYs only and up to 4 PHYs bonding is probably common

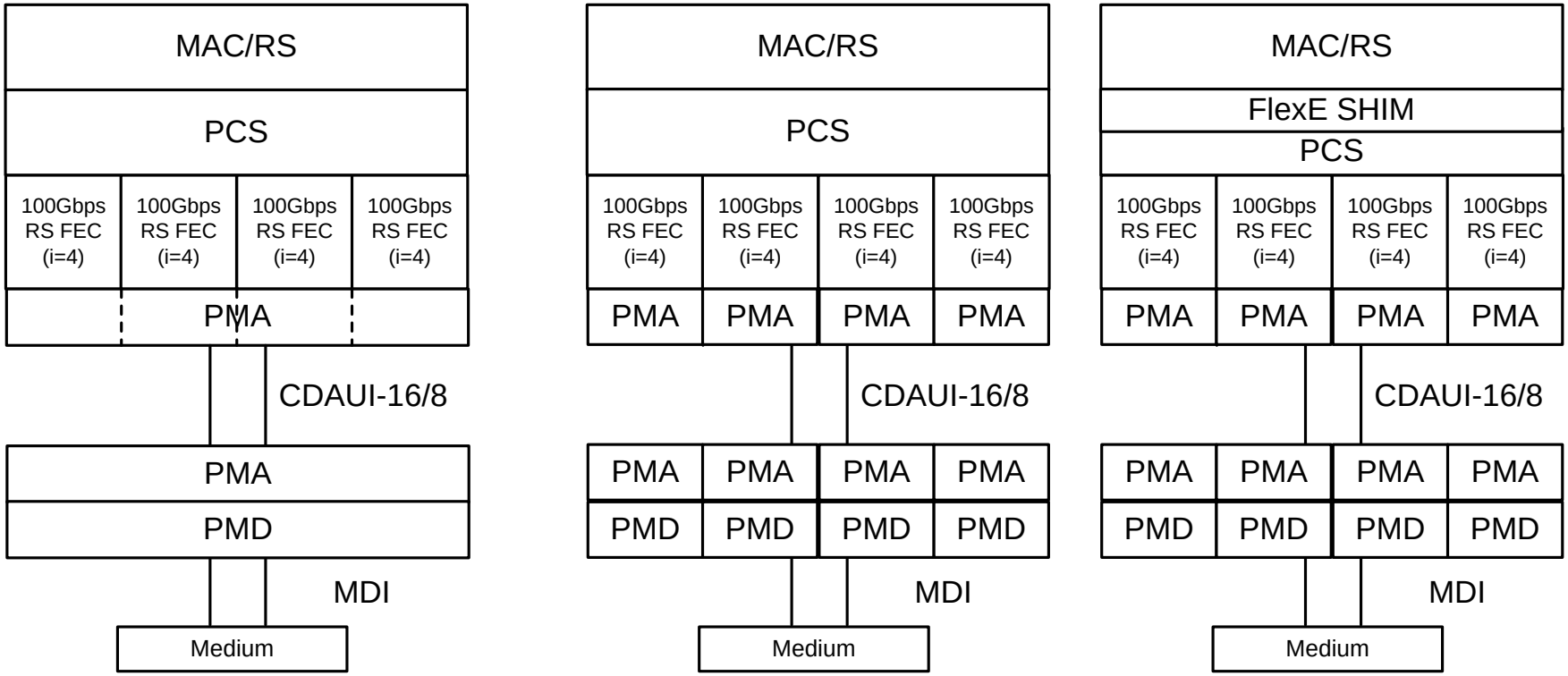


Copy and paste from oif2015.127.01

- For 400Gbps FlexE, it is 4X100G FEC architecture in logic layer

ONE FEC Architecture in Ethernet

- 400GbE with 4 pipeline 4X100G FEC
- Breakout into 4X100GbE
- 400Gbps FlexE with 4 PHY bonding



- 4X100Gbps FEC proposal is an excellent option to unify FEC architecture in Ethernet, even in ITU B100G is one of potential candidate yet.

Summary

- The FEC architecture proposal with 4X100Gbps FEC in parallel is a more simple solution, it will not only lower total area cost in 400GbE & 4X100GbE compatible design and also enable breakout feature, reuse IP cores and unified line card design and lead to broader market potential

Thank you