Proposal for 400GbE FEC Architecture

Xinyuan Wang, Tongtong Wang, Wenbin Yang

HUAWEI TECHNOLOGIES CO., LTD.

IEEE 802.3bs 400 GbE Task Force

Introduction and Background

- This presentation investigates the FEC architecture for 1X400Gbps VS 4X100Gbps implementation based on RS FEC
- How to stripe ingress data flow to FEC instance is one of key item to be investigated for moving 400GbE standard forward
- RS FEC seems like a good fit for this project: less complex to implement and better gain in the face of burst errors when compared to a BCH code. KR4/KP4 FEC as example to investigate as mature technology

Big Ticket Items - FEC

- FEC reference presentations
  - wang_x_3bs_01_0115.pdf
- Actions:
  - PMD selection
  - BERin required by PMD
  - Try to eliminate unacceptable FEC options e.g. in wang_x_3bs_01
  - 4x100G or 1x400G FEC striping
  - Impact of overspeed on PMD error rates

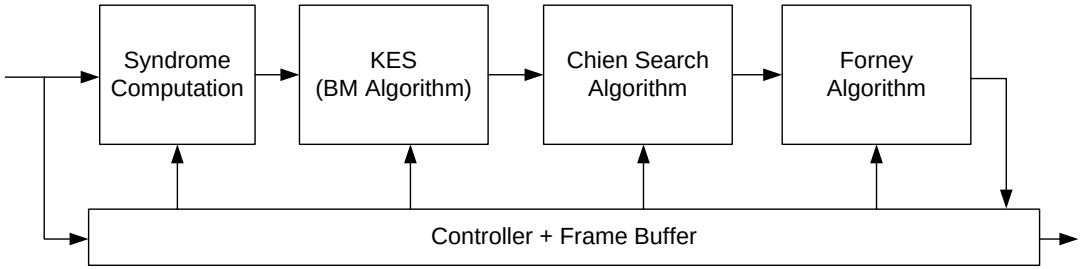
big_ticket_items_3bs_01_0115

RS FEC(n,k,t,m)	CG	NCG	BERin	Overhead	SerDes Rate	Block Time	Latency	Area Ratio	Hardware complexity
RS(528,514,7,10)	5.39	5.28	3.92E-05	0%	25.78125	51.2ns	~87ns	1X	802.3bj
RS(544,514,15,10)	6.64	6.39	3.09E-04	3.03%	26.5625	51.2ns	~112ns	2.9X	802.3bj
RS(560,514,23,10)	7.3	6.93	7.60E-04	6.06%	27.34375	51.2ns	~208ns	14.5X	Implementation compatible with 802.3bj; costs more logic resource
RS(576,514,31,10)	7.76	7.26	1.30E-03	9.09%	28.125	51.2ns	~258ns	33.4X	Implementation compatible with 802.3bj;costs significant logic resource
RS(1088,1028,30,11)	7.12	6.88	6.06E-04	3.03%	26.5625	102.4ns	~315ns	16.7X	costs more logic resource and requires to change AM spacing of 16384; Rule 1 not satisfied
RS(1020,956,32,10)	7.34	7.06	7.95E-04	6.7%	27.5	93.1ns	~304ns	27.2X	cost too more logic resource and require to change AM spacing of 16384; Rule 1,2,5 not satisfied
RS(840,771,34,10)	7.58	7.22	1.10E-03	6.06%	27.34375	76.8ns	~306ns	30.6X	cost too more logic resource and require to change AM spacing of 16384; Rule 1 not satisfied

wang_x_3bs_01a_0115

Latency Estimation of RS(n,k,t,m) FEC

- Use 100Gbps KR4 FEC@644MHz for ASIC as baseline in this presentation
- Latency estimation based on (RS FEC correction ability) t and parallelism(p1/p2) on each sub block in the following diagram;
- FEC Decoder performs error detection with error correction, same as in CL91.5.3.3, aka Mode A in 802.3bj



$t_{syndrome}$ = *n/p1, p1=16 for KR4/KP4 FEC implementation in this slides*

t_{KES} = *x2t, (if* $t_{KES} > t_{syndrome}$ *, duplicate KES in this slides)*

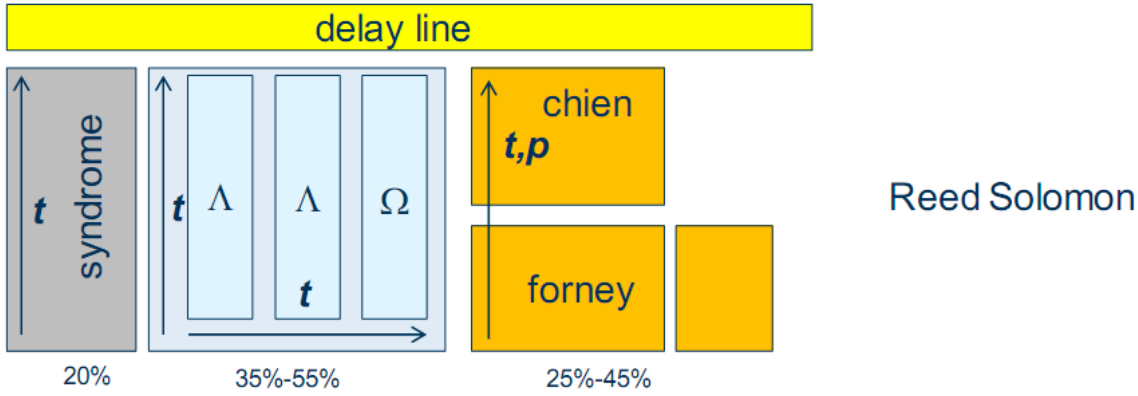
x=1 for t<=15, x=2 for t>15; For longer RS FEC, level of pipelining in the iterative calculation may increase due to longer critical path

$t_{chien} + t_{forney}$ = *n/p2+1, p2=66/68 for KR4/KP4 FEC implementation in this slides, p2≥p1*

FEC Decode Latency = ~($t_{syndrome} + t_{KES} + t_{chien} + t_{forney}$)

Area Estimation of RS(n,k,t,m) FEC

- For area estimation refer to langhammer_3bs_01_1114



- KR4 FEC ASIC area ratio is (modification for low latency target and larger permitted area):

  Syndrome: KES: (Chien+Forney)=20%:40%:40%

- if $t_{KES} > t_{syndrome}$, duplicate KES block to match the throughput of syndrome. This will increase area cost significantly for longer block RS FEC

Comparison of 4X100G & 1X400Gbps for RS(528,514) FEC in 400GbE Logic Layer

RS(528,514,7,10)(100Gbps) 160bit@644MHz(ASIC)	Area	Latency (Cycle)
1. Syndrome(16 parallel)	0.2a	33
2. KES(BM)	0.4a	14
3. Chien(66 parallel)	0.15a	8
4. Forney	0.25a	1
TOTAL	a	56 Cycle(~87ns)

RS(528,514,7,10)(400Gbps) 660bit@624MHz(ASIC)	Area	Latency (Cycle)
1. Syndrome(66 parallel)	0.825a	8
2. KES(BM) (X2 duplication)	0.8a	14
3. Chien(66 parallel)	0.15a	8
4. Forney	0.25a	1
TOTAL	2.025a	31Cycle(~49ns)

- Exact comparison is affected by process node or combinational logic, etc.
- To meet our low latency criteria, size of 1x400Gbps RS FEC@~49ns is around 2X size of 1x100Gbps RS FEC@~87ns
- For real implementation of higher latency & lower parallelism in Chien/Forney in 400Gbps RS FEC, the reasonable area of 1x400Gbps RS(528,514) FEC is ~2.5X size of 1x100Gbps RS(528,514) FEC

Comparison of 4X100G & 1X400Gbps for RS(544,514) FEC in 400GbE Logic Layer

- Based on Low Latency 100Gbps RS FEC with P2=68

RS(544,514,15,10)(100Gbps 160bit@664MHz(ASIC)	Area	Latency (Cycle)
1. Syndrome(16 parallel)	0.1b	34
2. KES(BM)	0.45b	30(2t)
3. Chien(68 parallel)	0.15b	8
4. Forney	0.3b	1
TOTAL	b	73 Cycle(~110ns)

RS(544,514,15,10)(400Gbps) 680bit@625MHz(ASIC)	Area	Latency (Cycle)
1. Syndrome(68 parallel)	0.425b	8
2. KES(BM) (X4 duplication)	1.8b	30
3. Chien(68 parallel)	0.15b	8
4. Forney	0.3b	1
TOTAL	2.675b	47Cycle(~75ns)

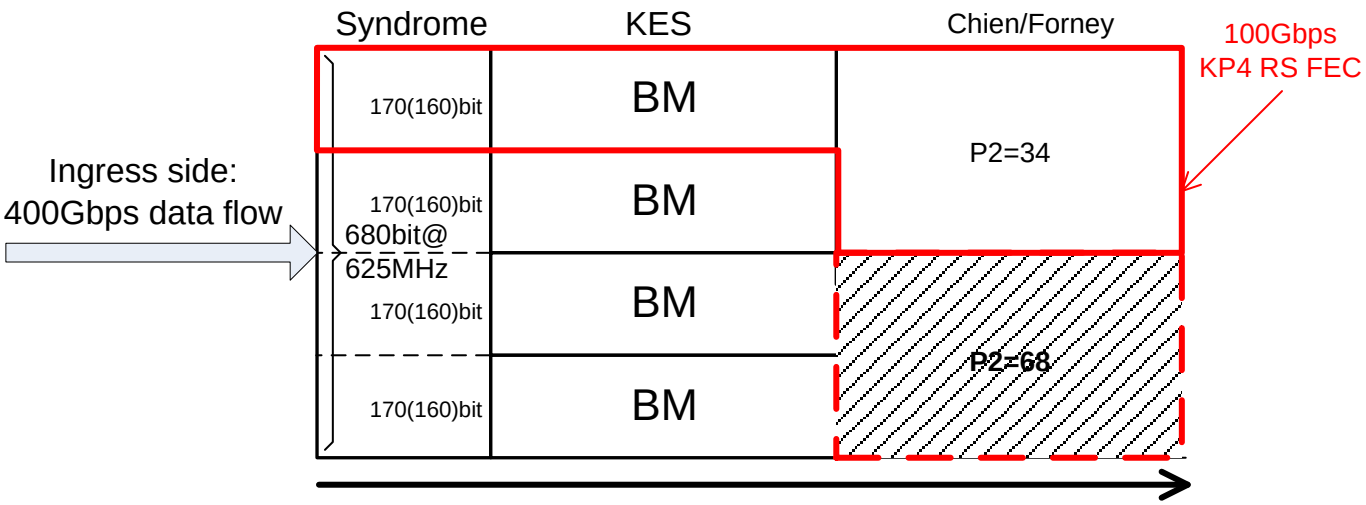
- Based on Smaller Area100Gbps RS FEC with P2=34, which is closer to real implementation

RS(544,514,15,10)(100Gbps 160bit@664MHz(ASIC)	Area	Latency (Cycle)
1. Syndrome(16 parallel)	0.1c	34
2. KES(BM)	0.5c	30(2t)
3. Chien(34 parallel)	0.1c	16
4. Forney(34 parallel)	0.3c	1
TOTAL	c	81 Cycle(~122ns)

RS(544,514,15,10)(400Gbps) 680bit@625MHz(ASIC)	Area	Latency (Cycle)
1. Syndrome(68 parallel)	0.425c	8
2. KES(BM) (X4 duplication)	2c	30
3. Chien(68 parallel)	0.2c	8
4. Forney(68 parallel)	0.6c	1
TOTAL	3.225c	47Cycle(~75ns)

- Generally, the reasonable area of 1x400Gbps RS(544,514) FEC is ~3X/3.5X size of 1x100Gbps RS(544,514) FEC

How to implement 1X400Gbps RS(544,514) FEC?



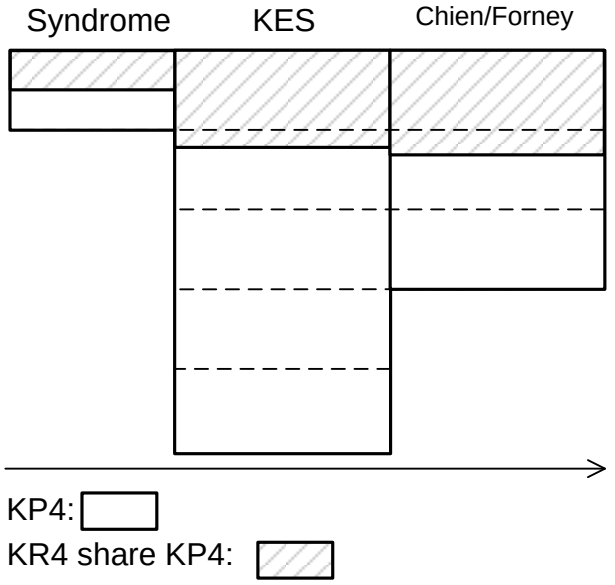
- Area of Syndrome implementation depends on data rate and needs similar logic resource for 1X400Gbps and 4X100Gbps@similar clock rate
- Duplication KES(BM) block based on t(KES) vs t(Syndrome, or Block time)
- Area of Chien/Forney is related with data throughput. For lower P2(~34 vs 68), 1X400bps will approach 4X100Gbps implementation
- Even for future higher speed Ethernet, much lower block time or higher performance FEC with large t will lead to parallel implementation in most function block in FEC architecture

Estimate of Area Ratio for KR4 vs KP4 FEC

- Area ratio for KP4 FEC vs KR4 FEC, 2.9X:1X in "wang_x_3bs_01a_0115"

RS FEC(n,k,t,m)	CG	NCG*	BERin	Overhead	SerDes Rate	Block Time	Latency**	Area Ratio
Group 1 : Similar RS FEC as KR4 FEC								
RS(528,514,7,10)	5.39	5.28	3.92E-05	0%	25.78125	51.2ns	~87ns	1X
RS(544,514,15,10)	6.64	6.39	3.09E-04	3.03%	26.5625	51.2ns	~112ns	2.9X

- Is KP4 FEC a superset of KR4 FEC?
  - Almost yes, ~5% additional logic resource for KP4 FEC to support KR4 FEC
  - Assume area of KP4 FEC will roughly cover KR4 FEC

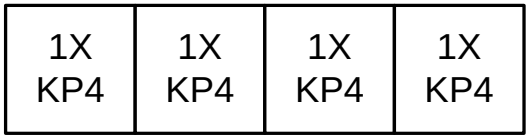


Area Estimate of 1x400 & 4x100GbE Compatible FEC

- 4x100Gbps *VS* 1x400Gbps+4x100Gbps

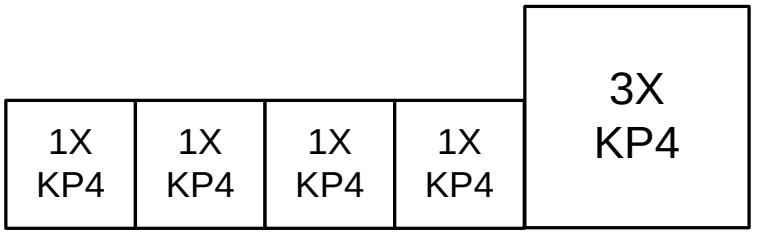
Scenario 1: KP4 FEC only in both 100GbE and 400GbE

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



- FEC architecture Option 1:
- 4X100Gbps KP4 FEC

  4X=4b=4X2.9a=11.6a



- FEC architecture Option 2:
- (4x100Gbps +1X400Gbps) KP4 FEC:

  7X=7b=7X2.9a=20.3a

- Using 4x100Gbps FEC(Option 1) for 400GbE is more area efficient in compatible FEC design
- If scale up from more realistic 100Gbps FEC*, the area for Option 2 is enlarged to 21.75a

**second approach on slide 6*

Area Estimate of 1x400 & 4x100GbE Compatible FEC

- 4x100Gbps *VS* 1x400Gbps+4x100Gbps

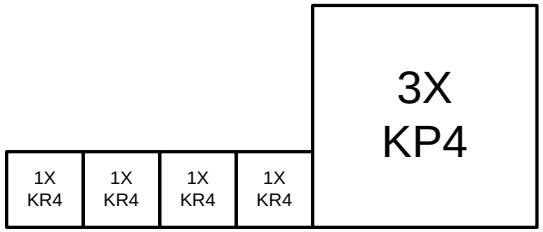
Scenario 2: KR4 FEC in 100GbE and KP4 FEC in 400GbE

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

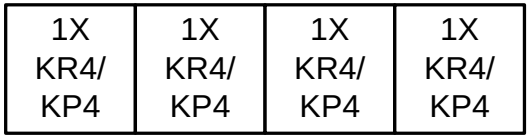
**second approach on slide 6*

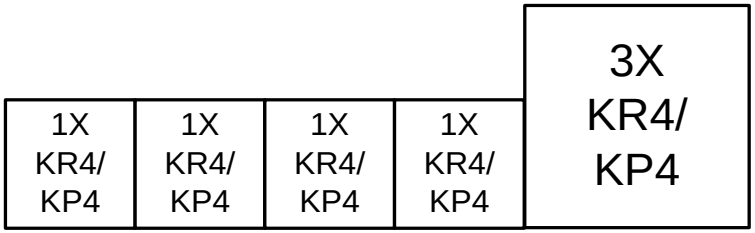
Area Estimate of 1x400 & 4x100GbE Compatible FEC

- 4x100Gbps *VS* 1x400Gbps+4x100Gbps

Scenario 3: KR4/KP4 FEC in both 100GbE and 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+1X400Gbps KR4/KP4 FEC:

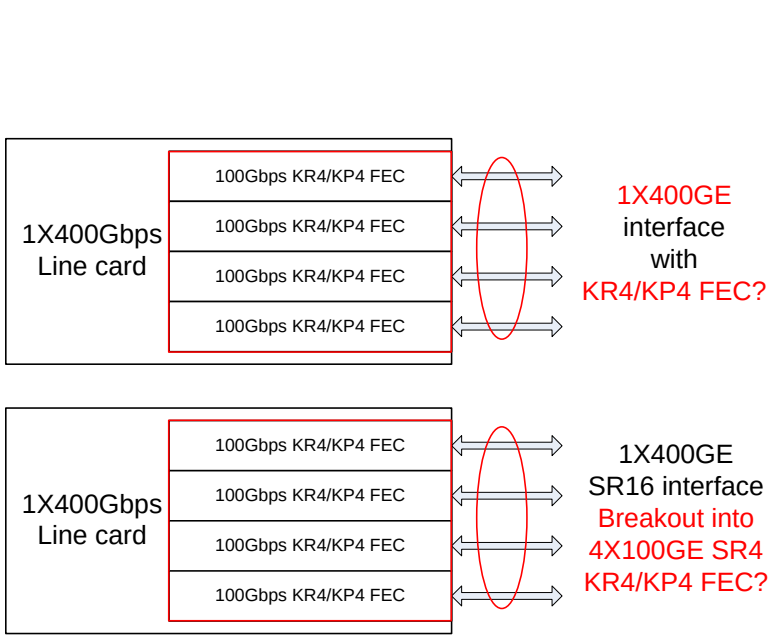
  7X=7X2.9a=20.3a

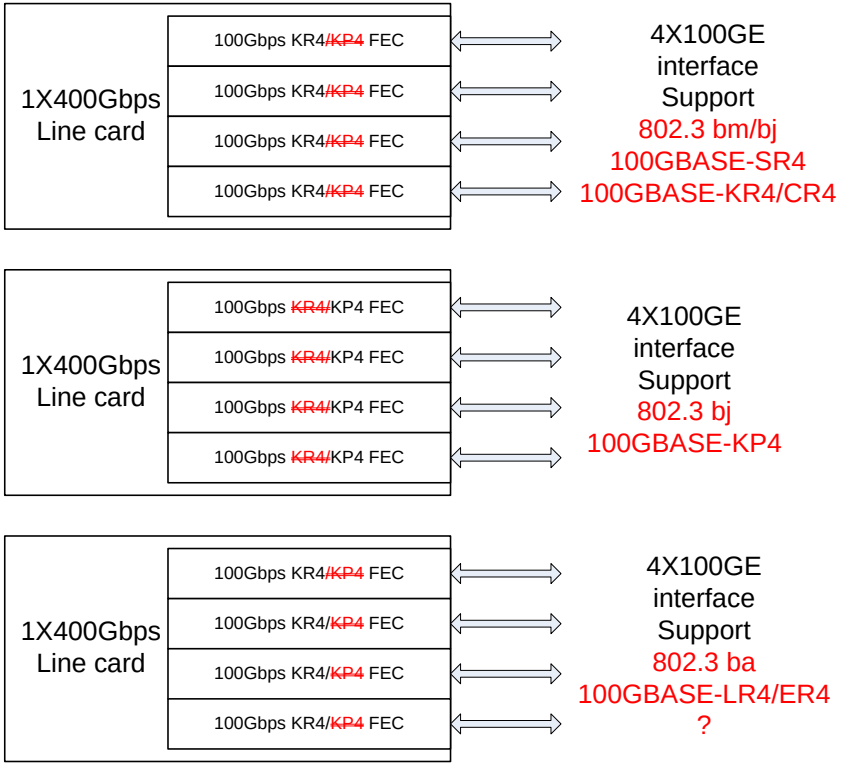
- Using 4x100Gbps FEC(Option 1) for 400GbE is more area efficient in compatible FEC design.
- If scale up from more realistic 100G FEC* , the area for Option 2 is enlarged to 21.75a

**second approach on slide 6*

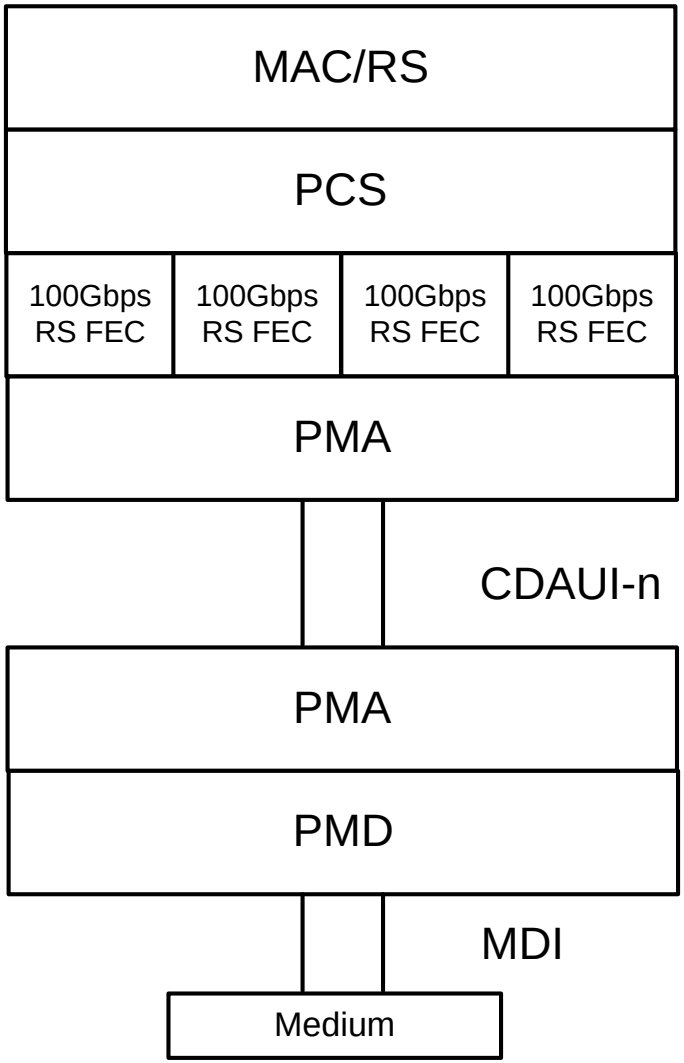
From System/ASIC Perspective: 400GbE with 4X100Gbps FEC Architecture

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





Proposal for 400GbE Logic Layer with RS FEC



- 4X100Gbps RS FEC in the PCS to provide a single FEC in the system
- RS(528,514)/RS(544,514) is most reasonable candidate

Summary

- The FEC architecture proposal with 4X100Gbps parallel will lower total area cost in 400GbE & 4X100GbE, in addition to enable breakout, IP core reuse and unified line card and lead to broad market potential
- RS(528,514), RS(544,514) FEC can share most of logic implementation. Even RS(560,514) and RS(576,514) FEC, if higher coding gain needed, are still in the same FEC family with similar functional blocks.

Thank you