Investigation on Technical Feasibility of Stronger RS FEC for 400GbE

Xinyuan Wang, Wenbin Yang,Tongtong Wang

HUAWEI TECHNOLOGIES CO., LTD.

IEEE 802.3bs 400 GbE Task Force

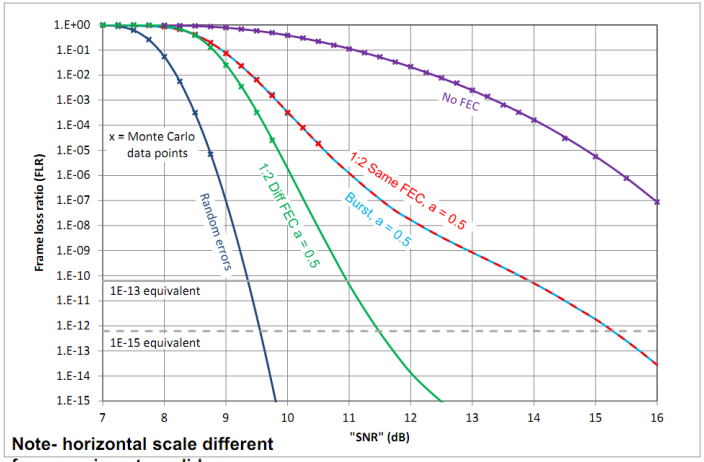
HUAWEI

Introduction and background

- In this presentation, we investigate technical feasibility of stronger RS FEC schemes and compare several possible candidate.

- BCH FEC schemes have quite different FEC performance for random and burst errors, with really poor burst performance.

BCH(2858,2570) all curves



Note- horizontal scale different from previous two slides

anslow_3bs_02_1114

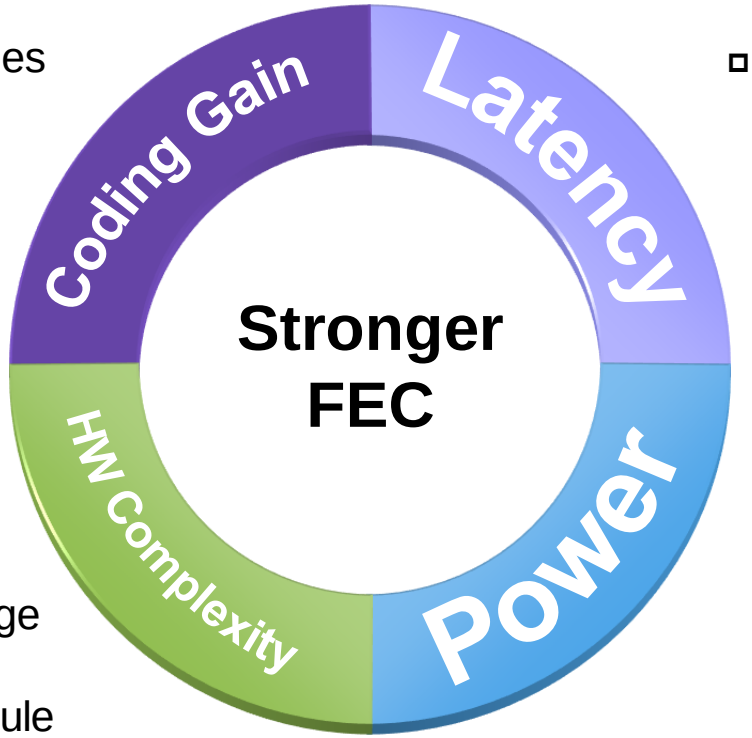
- BCH FEC schemes require much higher logic resource implementation, which probably consume more power.

Type	Codeword	Area (6LUT)	Relative Area
RS KR4	(528,514,7)	10654	1
RS KP4	(544,514,15)	26554	2.5
BCH[1]	(2858,2570,24)	106806	10
BCH[2]	(9193,8192,71)	425000	40

langhammer_3bs_01_1114

400GbE Stronger FEC tradeoffs

- Overhead Vs SerDes rate & technology feasibility.



- Latency in sensitive applications, such as Finance, DC,……. Especially for short reach solutions,100/500m.

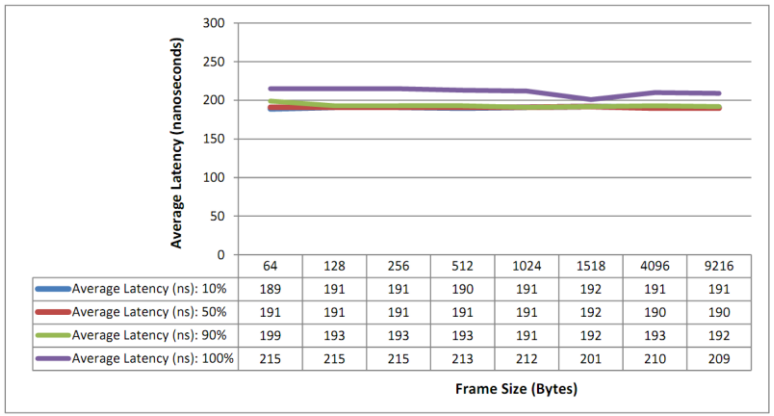
- Difficult to be integrated in host ASIC or FPGA if large resource required.
- QSFP+ & CFP module with silicon chip embedded?

- Impact on small form factor module objective.

What is the latency of Ethernet in history?

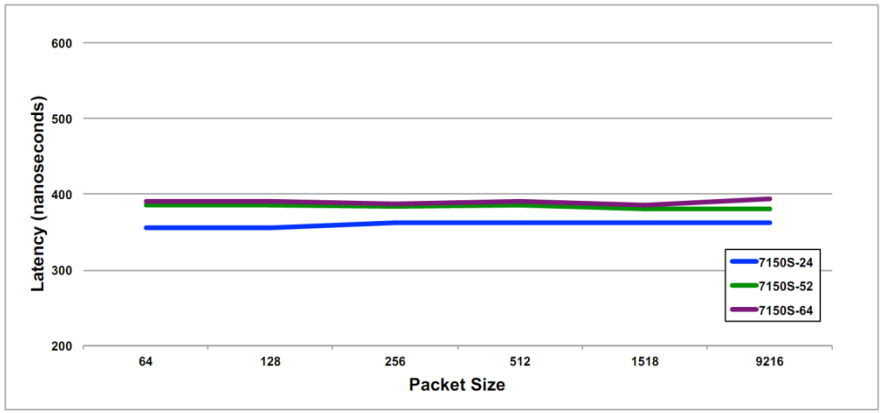
- Cisco Nexus 3548 1/10/40GE switch:

Figure 4: Layer 2 Unicast RFC 2544 Port Pair Latency in Warp Mode



	64	128	256	512	1024	1518	4096	9216
Average Latency (ns): 10%	189	191	191	190	191	192	191	191
Average Latency (ns): 50%	191	191	191	191	191	192	190	190
Average Latency (ns): 90%	199	193	193	193	191	192	193	192
Average Latency (ns): 100%	215	215	215	213	212	201	210	209

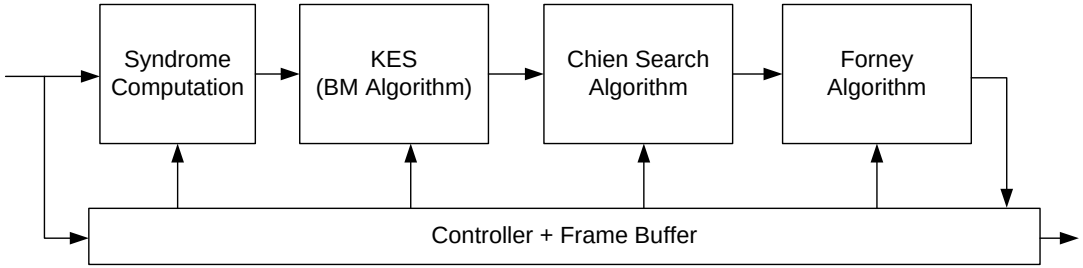
- Arista 7150s 1/10/40GE switch:



- In low latency ethernet switch, 250/350ns in 10GE is tested in several product. Using cut-through to replace store-forward in switch chip is mature technology.
- High frequency trading (HFT) in financial applications, high performance computing(HPC) in DC are especially sensitive to latency.
- Latency in DC is incurred by upper layer protocol (TCP windows, flow control, etc)and much cost on server implementation, especially memory.
- For FEC latency in 400GbE, <500ns is acceptable? <250ns is preferred.

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

- Use 100Gbps KR4 FEC@325MHz for FPGA as baseline in this slides
- Area estimation refer to langhammer_3bs_01_1114 with modification for low latency and larger area
- CG/NCG is based on BERpost=1E-13
- Latency estimation based on t and parallelism(p) on each sub blocks in the following diagram



$t_{syndrome}$ = *n/p1, p1=33 for easy implementation in this slides*

t_{KES} = *2t (if* $t_{KES} > t_{syndrome}$ *, duplicate KES in this slides or decrease p1)*

$t_{chien} + t_{forney}$ = *n/p2+1, p2=33 for easy implementation in this slides, p2≥p1*

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

Summary of Stronger RS FEC Option

RS FEC(n,k,t,m)	CG	NCG	Overhead	Latency(cycle)	Latency(ns)	Area Comparison
Group 1 : Similar RS FEC as KR4 FEC						
RS(528,514,7,10)	5.39	5.28	0%	47	~145ns	1x
RS(544,514,15,10)	6.64	6.39	3.03%	65	~200ns	~3.3x
RS(560,514,23,10)	7.3	6.93	6.06%	81	~249ns	~8.7x
RS(576,514,31,10)	7.76	7.26	9.09%	99	~304ns	~18.6x
Group 2 : Large Block RS FEC						
RS(1056,1028,14,11)	6.07	5.95	0%	93	~286ns	~2.2x
RS(1088,1028,30,11)	7.12	6.88	3.03%	127	~390ns	~10.2x
RS(1120,1028,46,11)	7.7	7.33	6.06%	161	~495ns	~30.2x
RS(1152,1028,62,11)	8.11	7.61	9.09%	195	~599ns	~68.5x
Group 3 : RS(255,239) Like RS FEC						
RS(255,239,8,8)	6.12	5.83	6.7%	33	~102ns	~1.4x
RS(510,478,16,9)	6.85	6.57	6.70%	65	~200ns	~3.7x
RS(1020,956,32,10)	7.34	7.06	6.7%	127	~390ns	~11.4x
Group 4 : 256/257b coding friendly RS FEC*						
RS(800,771,14,10)	6.29	6.13	1.01%	79	~243ns	~3x
RS(816,771,22,10)	6.95	6.71	3.03%	95	~292ns	~6.1x
RS(840,771,34,10)	7.58	7.22	6.06%	121	~372ns	~17.4x
RS(864,771,46,10)	8.02	7.53	9.09%	147	~452ns	~38.9x

*: Need some dummy bits to support FEC lane distribution

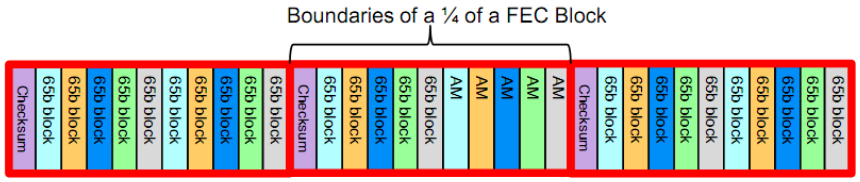
How to Integrate Stronger RS FEC in logic layer

- Rule 1: Prefer to keep 16384*66bit*20 AM spacing

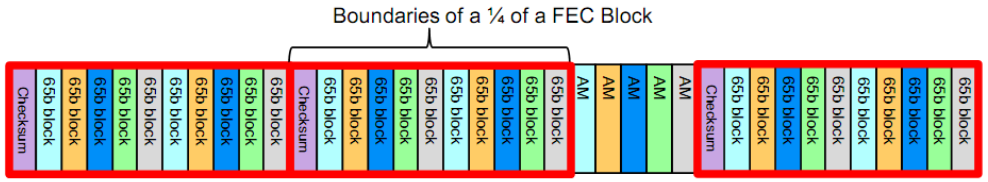
FEC Across Multiple Lanes

At least two implementations are possible:

1. Alignment Markers are included in the FEC blocks



2. Alignment Markers are not included in the FEC blocks



For either case, the Alignment Markers must repetitively be in the same location relative to FEC block starts:

1. (AM Spacing * # PCS Lanes * block size) must be divisible by (FEC block size)
2. ((AM Spacing-1) * # PCS Lanes * block size) must be divisible by (FEC block size)

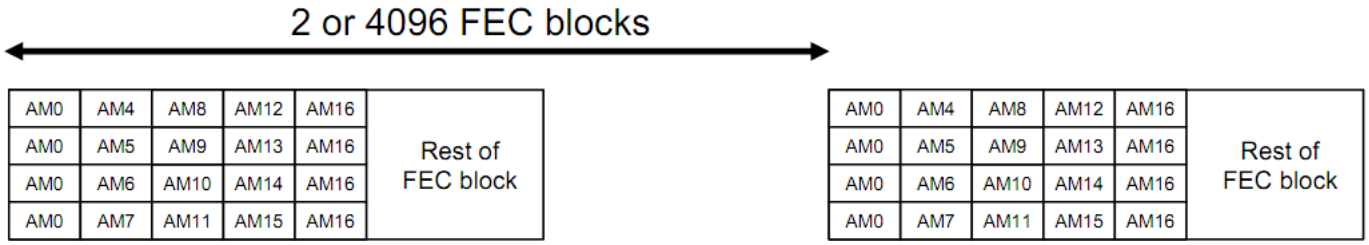
gustlin_02a_0511

Generic Rules for RS(n,k,t,m) FEC in logic layer with *i* FEC lanes (cont'd)

- Rule 2: Alignment marker is uniquely identify for each FEC lanes and friendly Idle delete(64bit) for IPG adjustment. Generally AM length should at least LCM(Least Common Multiple) of "m, i and 64".
- Rule 3: FEC information block: k*m should be divisible by encoder length if no dummy bit added, e.g. 257bit of 256/257 TC/DC, 65bit of 64/65 TC or 513bit of 512/513 TC.
- Rule 4: FEC block: n*m should be divisible by i*m. for example, i=4 in KR4/KP4 FEC.
- Rule 5: Feasible RCM(integer Reference Clock Multiplier) with 156.25MHz. For example, KP4 FEC with 3% over-clocking, RCM=170 for 26.5625Gbps.

RS(576/560/544/528, 514,31/23/15/7,10)

- 4096 FEC blocks in AM period with 0%/3.03%/6.06%/9. 09% over-clocking.

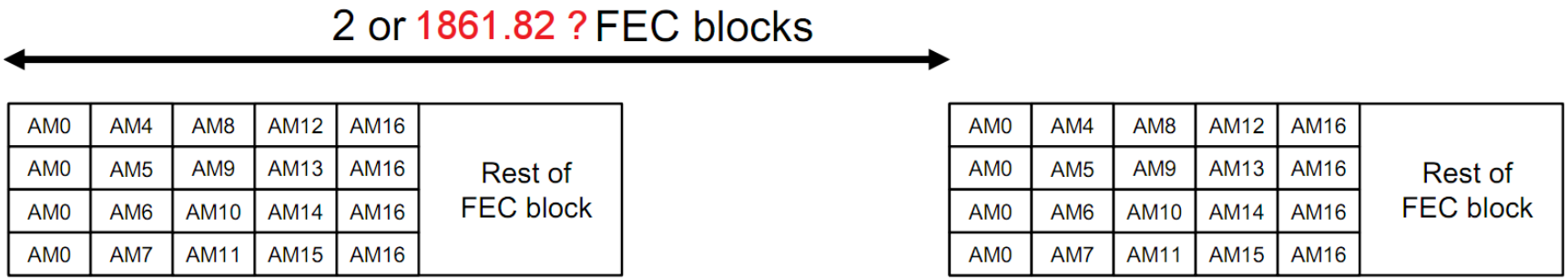


gustlin_400_02a_1113

- AM=320bit;
- FEC Information Block=5140bit=257*20 with 256/257 TC/DC;
- FEC Block=(576/566/544/528)*10=(144/140/136/132)*4*10;
- RCM=180/175/170/165@156.25MHz.

RS(1152/1120/1088/1056,1028,62/46/30/14,11)

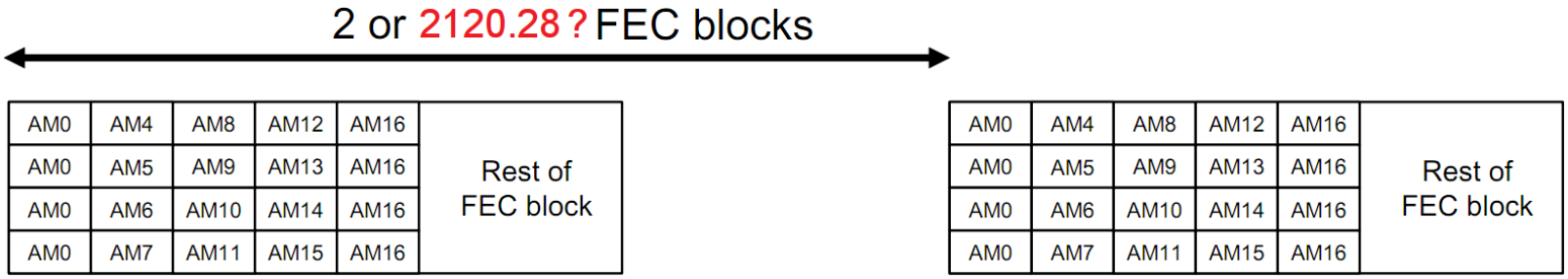
- Not an integer number of FEC blocks in AM spacing!
  - Change AM distance? Or
  - Overlap 1st FEC Block with part of AM area? Not a good option for coupling AM with FEC blocks.



- AM=319bit with 1 dummy bit ;
- FEC Information Block=1028*11bit=257*44 with 256/257 TC/DC;
- FEC Block=(1152/1120/1088/1056)*11=(288/280/272/264)*4*11;
- RCM=180/175/170/165@156.25MHz.

RS(1020,956,32,10)

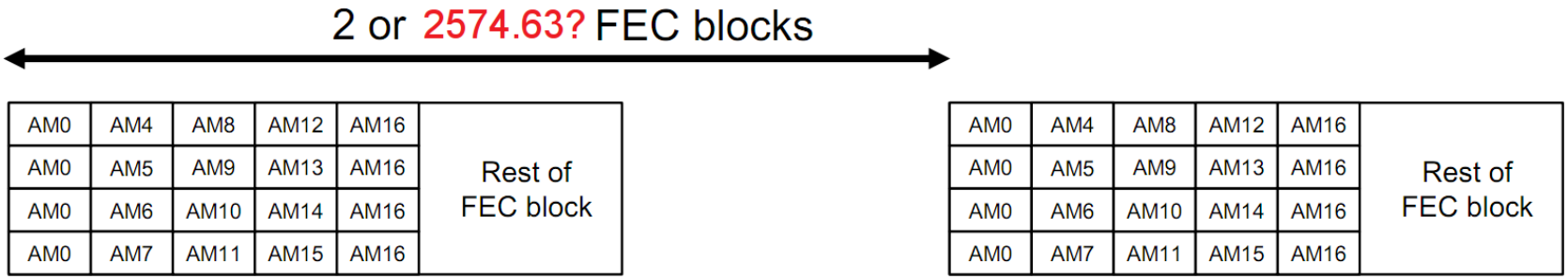
- Not an integer number of FEC blocks in AM period!



- AM=320bit;
- FEC Information Block=9560bit, not an integer number of 65,66,257,513bit; Change to 9570bit for adapting to 66bit block.
- FEC Block=(1020)*10=255*4*10;
- RCM, Not an integer number @156.25MHz.

RS(840,771,34,10)

- Extend FEC block to 840*m for easy implementation with 10bit dummy bit;
- Not an integer number of FEC blocks in AM period!



- AM=320bit;
- FEC Information Block=771*10bit=257*30 with 256/257 TC/DC;
- FEC Block=(840)*11=(210)*4*11;
- RCM=175@156.25MHz. Same over-clock as RS(560,514,23,10).

Compare of Possible Stronger RS FEC

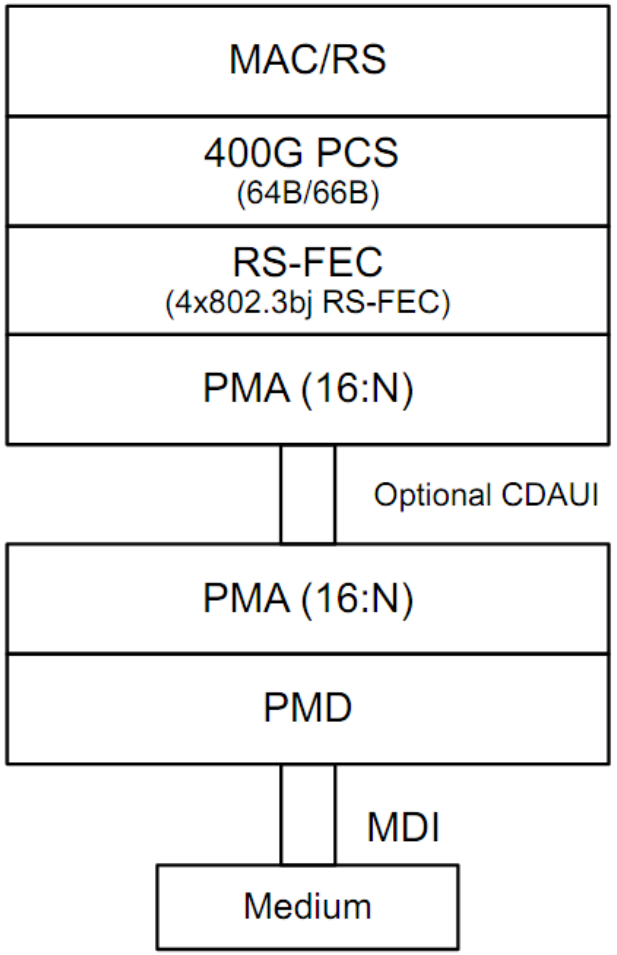
RS FEC(n,k,t,m)	CG	NCG	Overhead	Latency(ns)	Area	Hardware complexity
RS(528,514,7,10)	5.39	5.28	0%	~145ns	1x	802.3bj
RS(544,514,15,10)	6.64	6.39	3%	~200ns	~3.3x	802.3bj
RS(560,514,23,10)	7.3	6.93	6.06%	~249ns	~8.7x	Implementation compatible with 802.3bj
RS(576,514,31,10)	7.76	7.26	9.09%	~304ns	~18.6x	Implementation compatible with 802.3bj
RS(1120,1028,46,11)	7.7	7.33	6.06%	~495ns	~30.2x	cost too more logic resource and require to change AM spacing of 16384
RS(840,771,34,10)	7.58	7.22	6.06%	~372ns	~17.4x	cost more logic resource and require to change AM spacing of 16384
RS(864,771,46,10)	8.02	7.53	9.09%	~452ns	~38.9x	cost too more logic resource and require to change AM spacing of 16384

Compare of 4X100G/1X400Gbps RS FEC in 400GbE logic layer

RS(528,514,7,10)(100Gbps） 330bit@325MHz(FPGA)	LUT6	DELAY (Clk)
1. Syndrome(33 parallel)	6125	16
2. BM	2832	14
3. Chien(33 parallel)	3924	16
5. Forney	2181	1
TOTAL	15062	47(145ns)
RS(528,514,7,10)(400Gbps） 4*330bit@325MHz(FPGA)	**LUT6**	**DELAY (Clk)**
1. Syndrome(132 parallel)	6125*4	4
2. BM(duplicate x4)	2832*4	14
3. Chien(132 parallel)	3924*4	4
4. Forney	2181	1
TOTAL	53705	19(60ns)

- The logic resource of 1X400Gbps RS FEC is ~3.5X of 100Gbps RS FEC.
- 4X100Gbps RS FEC in parallelism is preferred in 400GbE for architecture robust.

Proposal of 400GbE Logic Layer with RS FEC



Or, other reasonable stronger RS FEC as in this slides

gustlin_400_02a_1113

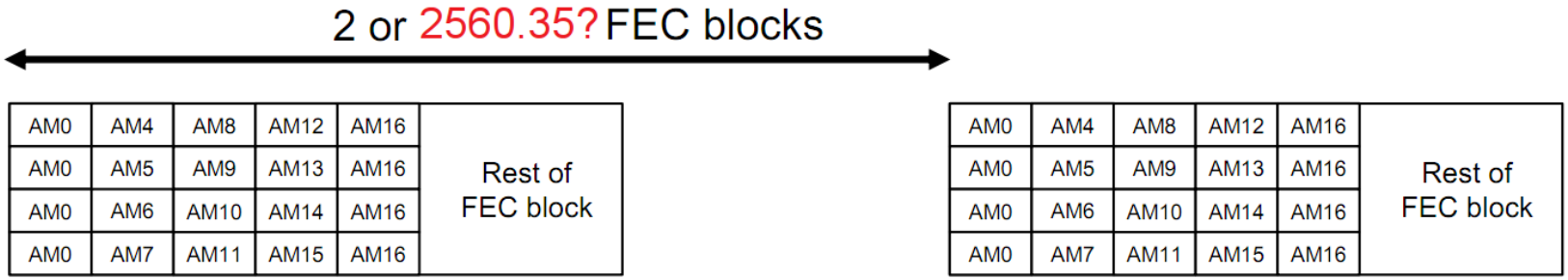
Summary

- RS FEC is more technical feasible in burst performance, hardware complexity, latency and friendly accommodate into 400GbE logic layer.
- Prefer to keep 256/257 transcoding to lower over-clocking or achieve more coding gain.
- The proposal of End-to-End FEC strategy with 4X100Gbps RS FEC parallelisming will make 400GbE system simple and cost efficient.

Thank you

RS(816,766,25,10)

- Not an integer number of FEC blocks in AM period!



- AM=320bit;
- FEC Information Block=7660bit, not an integer number of 65,66,257,513; Change to 7710bit for adapting to 257bit block.
- FEC Block=(816)*10=204*4*10;
- RCM, Not an integer number @156.25MHz.