

100G backplane PAM4 PHY encoding (revised)

IEEE P802.3bj

March 2012, Hawaii

Matt Brown – AppliedMicro

Sudeep Bhoja – Broadcom

Contributors and Supporters

- Ran Adee, Intel
- Stephen Bates, PMC-Sierra
- Will Bliss, Broadcom
- David Chalupsky, Intel
- Dariush Dabiri, APM
- Dan Dove, APM
- Howard Frazier, Broadcom
- Ali Ghiasi, Broadcom
- Ziad Hatab, Vitesse
- Dimitrios Giannakopoulos, APM
- Adam Healey, LSI
- Beth Kochuparambil, Cisco
- Kent Lusted, Intel
- Richard Mellitz, Intel
- Venkatesh Nagapudi, APM
- Vasu Parthasarathy, Broadcom
- Jamal Riani, Marvell

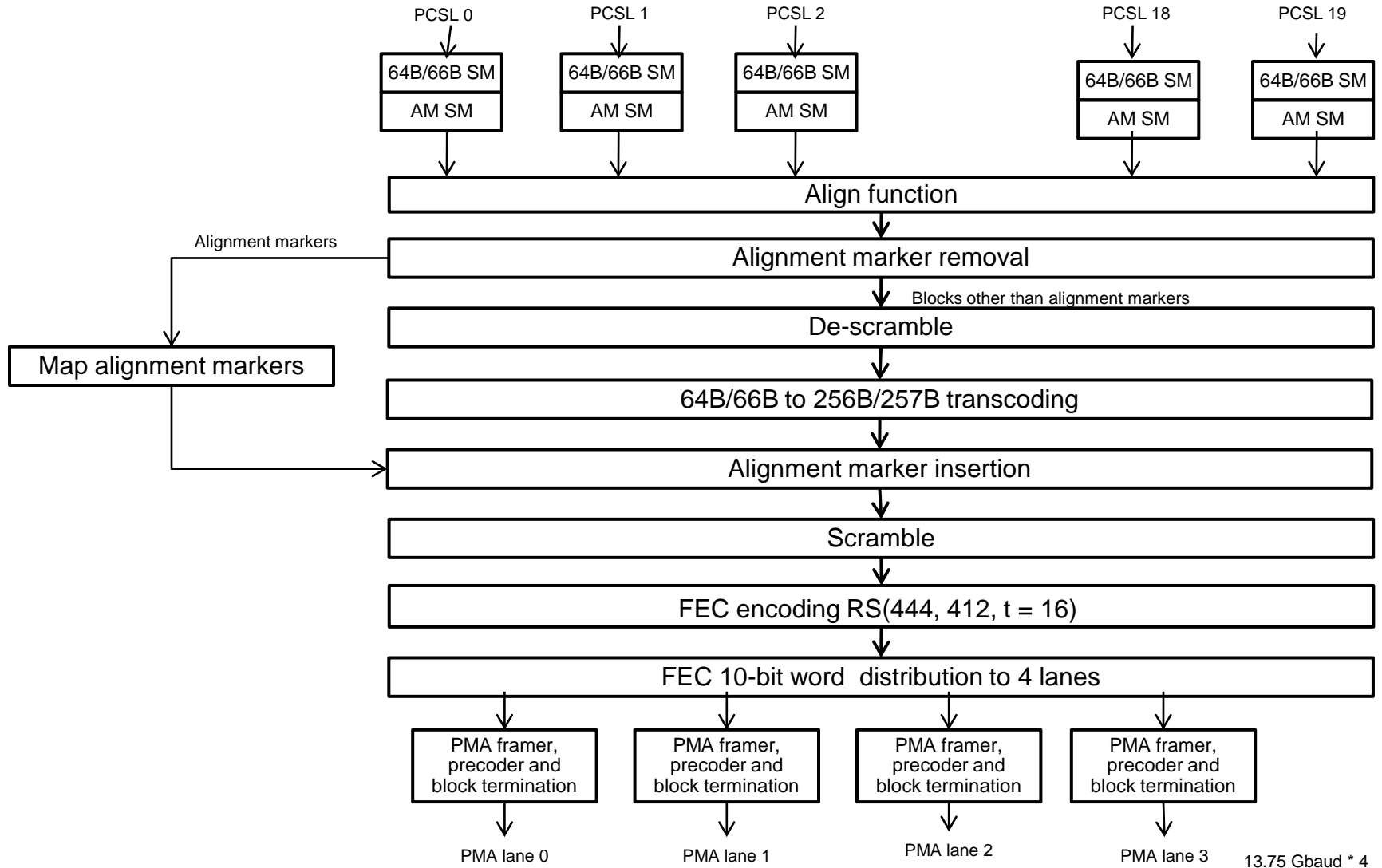
Introduction

- Provide a strawman baseline specification for the PAM4 FEC, PMA, and PMD transmitter encoding.
- Revised from January presentation to incorporate 256B/257B transcoding and alignment marker mapping.

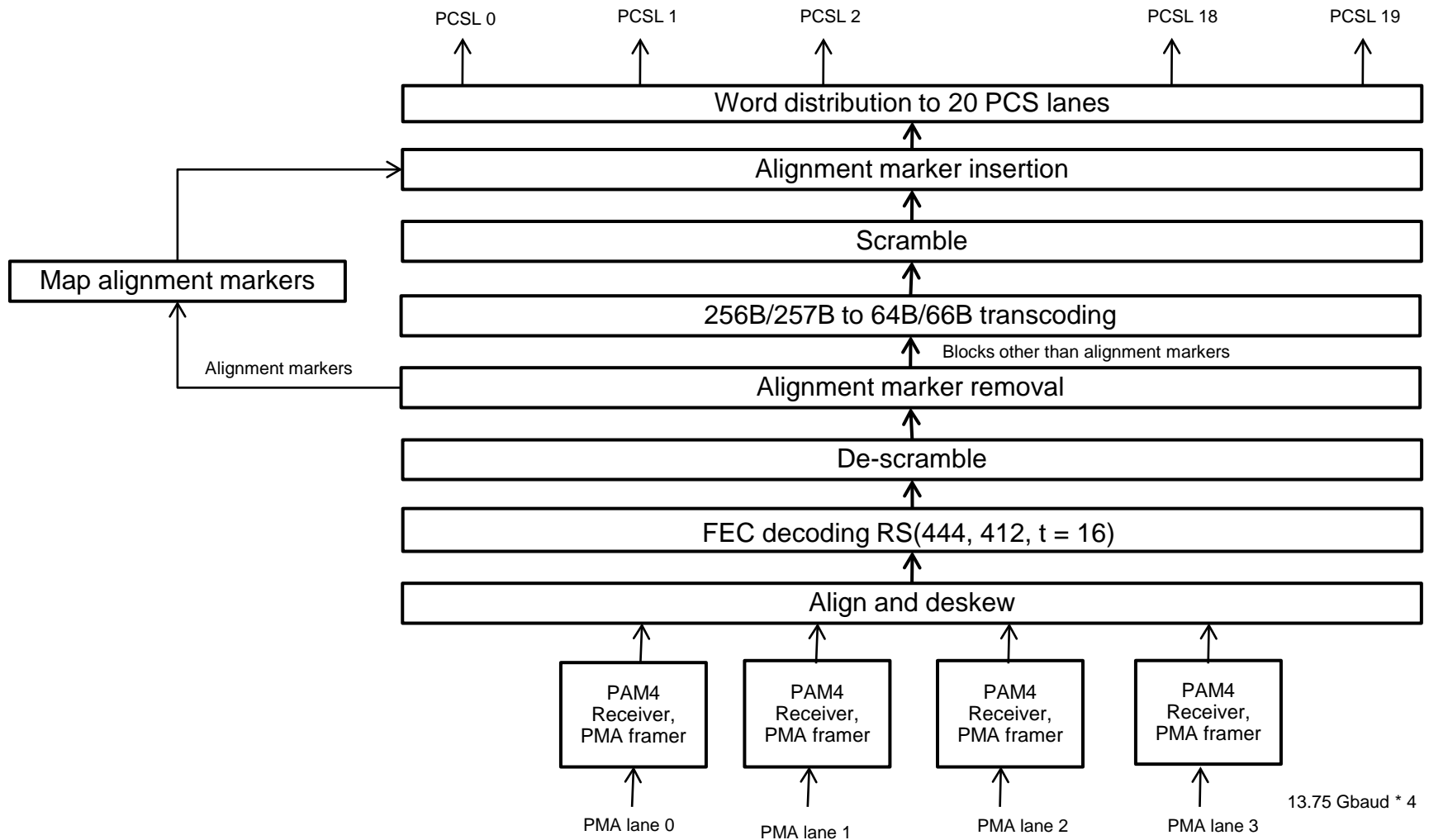
Transmitter process

- Transcoding: 256B/257B (was 512B/514B)
 - Aligns with NRZ (gustlin_01_0312)
- FEC: RS(444,412,T=16,M=10)
- PAM4 Symbols: Gray mapping,
 - $\{+1,+1/3,-1/3,-1\}$ map to $\{10,11,01,00\}$
- Precoding: $1/(1+D)$ MOD 4
- PAM4 block termination: 1 PAM4 termination symbol per 32 PAM4 symbols
 - 63 data bits per 32 PAM4 symbols
- PAM4 symbol rate: $88 * 156.25 \text{ MHz} = 13.75 \text{ Gbaud}$
- Tx pre-emphasis: 3 taps, one pre, one post
 - same structure as for 10GBASE-KR
- PAM4 test methodology and parameters addressed in bliss_01a_0911.

Tx encoding flow



RX decoding flow



PCS Lane Processing

- Synchronize to 64B/66B blocks on each PCS lane per 802.3ba 82.2.11.
- Synchronize to PCS alignment markers (64B/66B blocks) on each PCS lane per 802.3ba 82.2.12.
- Align (or deskew) and re-order PCS lanes based on alignment markers per 802.3ba 82.2.12.
- Descramble 64B/66B blocks per 82.2.15.
 - Required for transcoding.
- Same as for NRZ PHY.

Transcoding

- 256B/257B transcoding per cideciyan_01_0312.
- Map 64B/66B blocks to 256B/257B per gustlin_01_0312.
 - Alignment markers will not be transcoded, but instead will be re-mapped.
- Same as for NRZ.
- MTTFPA > 3.9E15 years
 - Post-FEC BER $\leq 1\text{E-}12$, RS(444,412,16,10) FEC
 - Analysis on slide 36.
 - FYI Lifetime of universe $\sim 13\text{E}9$ years.

Scrambling

- Use self-synchronizing scrambler
 - Same scrambler as for PCS in 802.3ba 82.2.5.
 - All data bits including the 256B/257B header bits and alignment markers are scrambled.
- Same as for NRZ except...
 - Alignment markers are scrambled as well.
 - Need AM mapping to PAM4 to be balanced, randomized, and clock rich.
 - May be able to re-map AM's so that scrambling is not required.
 - Analysis required.
 - Ideally, re-mapping would be common to NRZ and PAM4.

FEC

- RS(444,412,T=16,M=10) code format
 - single, efficient, dual-purpose (NRZ/PAM4) FEC core is possible if FEC generator math specified similarly for both
- FEC frame content
 - correctable payload = $412 \times 10 = 4120$ bits
 - parity = $32 \times 10 = 320$ bits
 - data = 64x 64B/66B blocks transcoded to 16x 256B/257B blocks
 - total data = 4112 bits
 - 8 dummy bits (4120-4112) per FEC frame required
 - 8 zeros added (assumed) for parity calculation
 - Payload words 408-411 will contain 8 data bits and 2 dummy bits.
 - one 8-bit word will end up on each of the 4 PMA lanes
 - dummy bits not transmitted
- FEC encoding is mandatory; negotiation is not required.

13.75GBaud Precoding/FEC Summary

RS(444, 412, t = 16)	Delta (dB)	Coding Gain (dB) <u>BER = 1E-15</u>
Random Error		7.12
DFE Burst Error Penalty	-0.88	6.24
Extended KR channel 6.7% over clocking loss	-1.0	5.24 (<100ns total latency)

- ~6.7% over clocking (88*156.25 MHz)
- 5.24 dB Coding gain for Extended KR channel
- Overhead includes FEC parity & PAM4 block termination

Comparison of RS FEC candidate codes

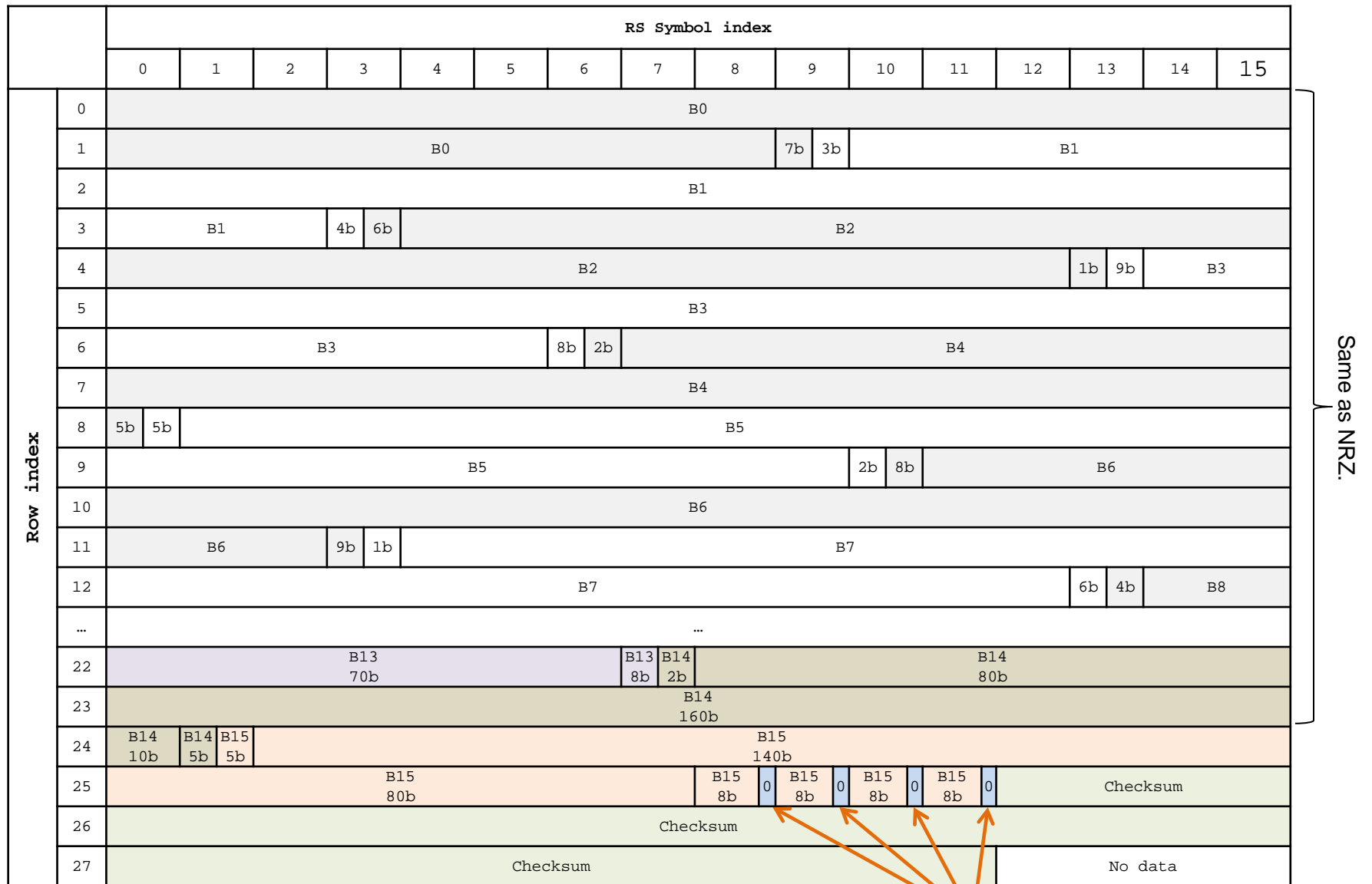
FEC codes GF(2 ¹⁰)	Total Coding Gain (dB)	Burst Coding Gain (dB)	Latency (ns)
RS(444, 412, t = 16)	5.24	6.24	82 - 123
RS(550, 520, t = 15)	5.1	5.9	102 - 154
RS(546, 520, t = 13)	4.9	5.6	102 - 154
RS(544, 520, t = 12)	5.0	5.6	102 - 154
RS(540, 520, t = 10)	4.9	5.2	102 - 154

- Codes in bhoja_01_0911 and cideciyan_01_1111 (found using computer search)
- RS(444, 412, t = 16) has best coding gain within 100ns target latency
 - Example implementation of 460K gates in 40nm CMOS has 99.9ns latency

Mapping 256B/257B blocks to FEC frame

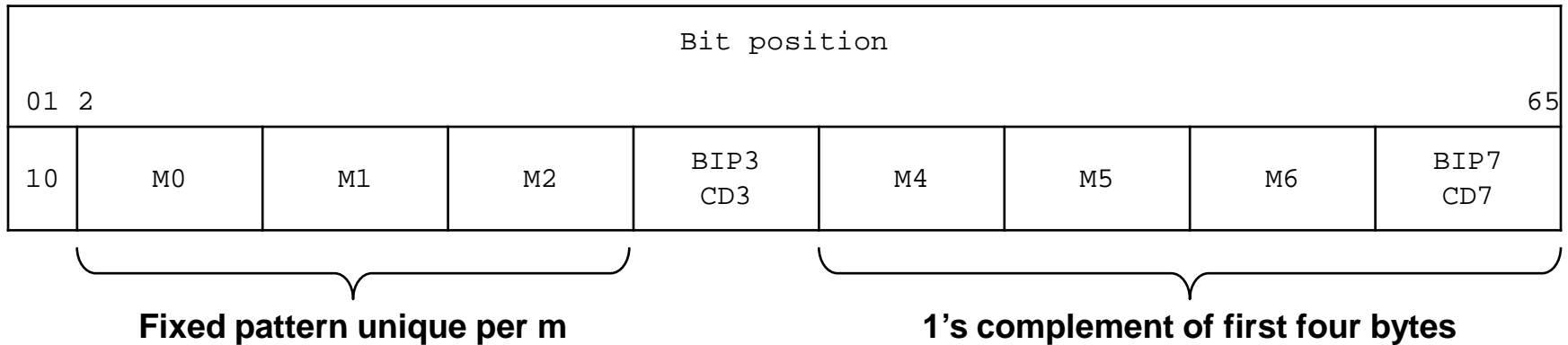
- 256B/257B blocks are concatenated and organized into a series of 10-bit FEC words.
 - Except for last four FEC words which are 8 data bits with 2 pad bits each (see FEC slide).

FEC frame structure



Alignment markers

66-bit alignment marker m, 64-bit payload denoted as AM



FEC frame structure with AMs

		RS Symbol index																
		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
Row index	0	A0 ₀	A1 ₀	A2 ₀	A3 ₀	A0 ₁	A1 ₁	A2 ₁	A3 ₁	A0 ₂	A1 ₂	A2 ₂	A3 ₂	A0 ₃	A1 ₃	A2 ₃	A3 ₃	
	1	A0 ₄	A1 ₄	A2 ₄	A3 ₄	A0 ₅	A1 ₅	A2 ₅	A3 ₅	A0 ₆ A4 ₆	A1 ₆ A5 ₆	A2 ₆ A6 ₆	A3 ₆ A7 ₆	A4 ₇	A5 ₇	A6 ₇	A7 ₇	
	2	A4 ₈	A5 ₈	A6 ₈	A7 ₈	A4 ₉	A5 ₉	A6 ₉	A7 ₉	A4 ₁₀	A5 ₁₀	A6 ₁₀	A7 ₁₀	A4 ₁₁	A5 ₁₁	A6 ₁₁	A7 ₁₁	
	3	A4 ₁₂ A8 ₁₂	A5 ₁₂ A9 ₁₂	A6 ₁₂ A10 ₁₂	A7 ₁₂ A11 ₁₂	A8 ₁₃	A9 ₁₃	A10 ₁₃	A11 ₁₃	A8 ₁₄	A9 ₁₄	A10 ₁₄	A11 ₁₄	A8 ₁₅	A9 ₁₅	A10 ₁₅	A11 ₁₅	
	4	A8 ₁₆	A9 ₁₆	A10 ₁₆	A11 ₁₆	A8 ₁₇	A9 ₁₇	A10 ₁₇	A11 ₁₇	A8 ₁₈	A9 ₁₈	A10 ₁₈	A11 ₁₈	A8 ₁₉ A12 ₁₉	A9 ₁₉ A13 ₁₉	A10 ₁₉ A14 ₁₉	A11 ₁₉ A15 ₁₉	
	5	A12 ₂₀	A13 ₂₀	A14 ₂₀	A15 ₂₀	A12 ₂₁	A13 ₂₁	A14 ₂₁	A15 ₂₁	A12 ₂₂	A13 ₂₂	A14 ₂₂	A15 ₂₂	A12 ₂₃	A13 ₂₃	A14 ₂₃	A15 ₂₃	
	6	A12 ₂₄	A13 ₂₄	A14 ₂₄	A15 ₂₄	A12 ₂₅ A16 ₂₅	A13 ₂₅ A17 ₂₅	A14 ₂₅ A18 ₂₅	A15 ₂₅ A19 ₂₅	A16 ₂₆	A17 ₂₆	A18 ₂₆	A19 ₂₆	A16 ₂₇	A17 ₂₇	A18 ₂₇	A18 ₂₇	
	7	A16 ₂₈	A17 ₂₈	A18 ₂₈	A19 ₂₈	A16 ₂₉	A17 ₂₉	A18 ₂₉	A19 ₂₉	A16 ₃₀	A17 ₃₀	A18 ₃₀	A19 ₃₀	A16 ₃₁	A17 ₃₁	A18 ₃₁	A19 ₃₁	
	8	P 5b	B5 5b	B5 150b														
	9	B5 100b										B5 2b	B6 8b	B6 50b				
...	...																	
24	B14 10b	B14 5b	B15 5b	B15 140b														
25	B15 80b								B15 8b	0	B15 8b	0	B15 8b	0	B15 8b	0	Checksum 40b	
26	Checksum 160b																	
27	Checksum 120b																	

Same as NRZ.

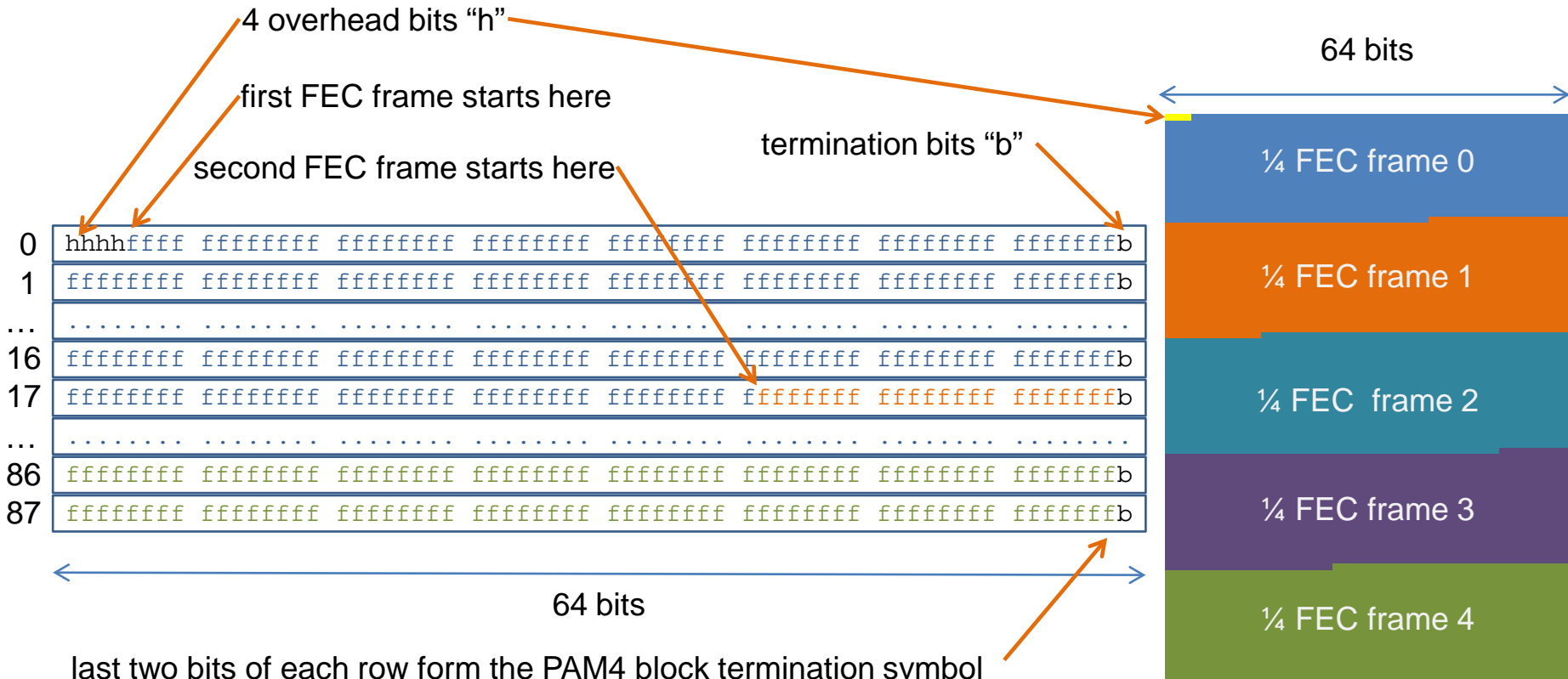
Mapping FEC to PMA lanes

- Cycle through FEC 10-bit words through each of the 4 PMA lanes.
 - The FEC frame contains 444 10-bit words
 - For each FEC frame, 111 10-bit words are destined for each of the four PMA lanes.
 - FEC words $(i+j*4)$ go to lane i
 - i is $\{0,1,2,3\}$, where i represents the lane #
 - j is $\{0,1,2,\dots,110\}$, j indexes the FEC words destined for each lane
 - Note that for FEC words 408 to 411, only the 8 data bits are transferred to each lane.

PMA Frame

- PMA frame generated for each PMA lane.
- PMA frame is composed of...
 - 5 quarter FEC frames, $5 \cdot (4440 - 8) / 4 = 5540$ bits
 - 4 overhead bits
 - essential to give a resultant PAM4 symbol rate of $88 * 156.25$ MHz
 - various possible applications discussed on subsequent slide
 - 88 PAM4 block termination bits
 - 1 termination bit per 63 data bits
 - 5632 bits total

PMA frame structure (one per lane)



last two bits of each row form the PAM4 block termination symbol

Each pair of bits, map to one PAM4 symbol.

For the PAM4 block termination symbol, we want "b" and the preceding bit "f" to indicate +1 or -1 so ...

For gray mapping, b = 0, always!

if the preceding bit is 1, then 10 maps to +1

if the preceding bit is 0, then 00 maps to -1

Legend:

"f" = bits from 5 FEC frames

"h" = overhead bits

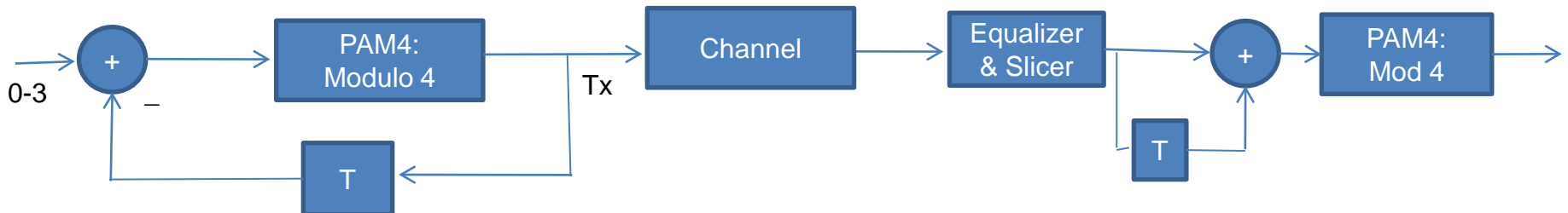
"b" = block termination bits

PMA Frame Overhead Bits

- Each PMA per-lane frame has 4 overhead bits.
- Must be randomized or at least “friendly”.
- Various applications ...
 - PMA frame alignment (see previous slide)
 - lane identification
 - control channel for remote transmitter control
 - vendor specific use

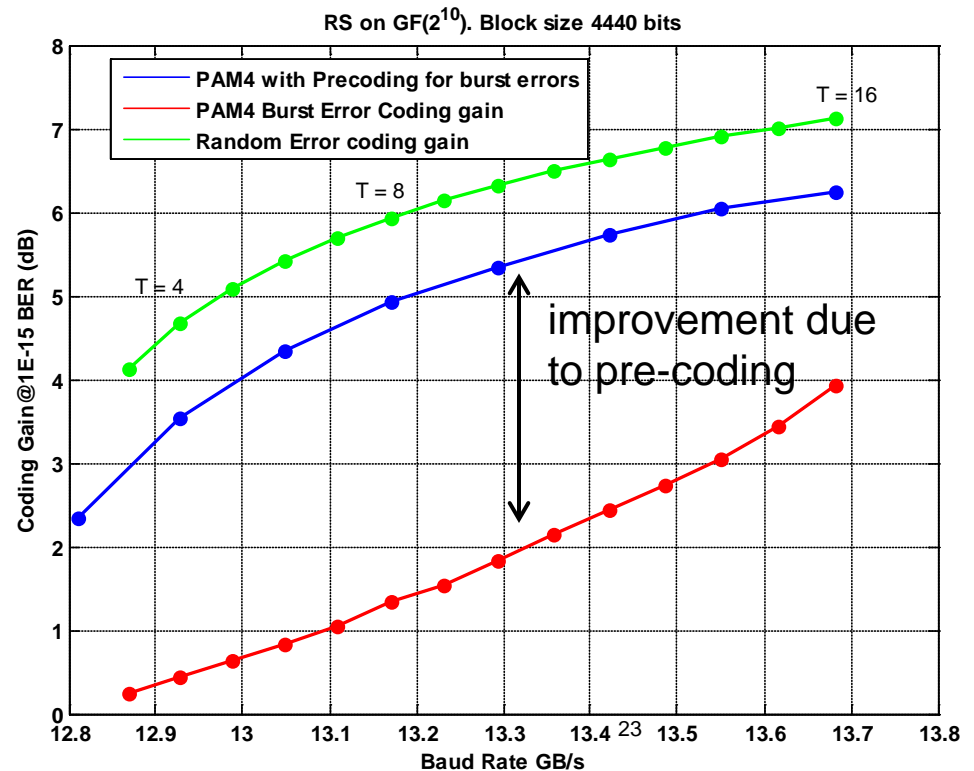
Pre-Coding

- $1/(1+D)$ modulus 4 pre-coding
 - See bliss_01_0311, “Signaling Terminology; PAM-M and Partial Response Precoders”
 - Rx uses a $(1+D)$ mod 4 after slicing
- Simple to implement
- Very low Complexity; similar complexity to duo-binary precoder.
- Pre-coding is mandatory; negotiation is not required.



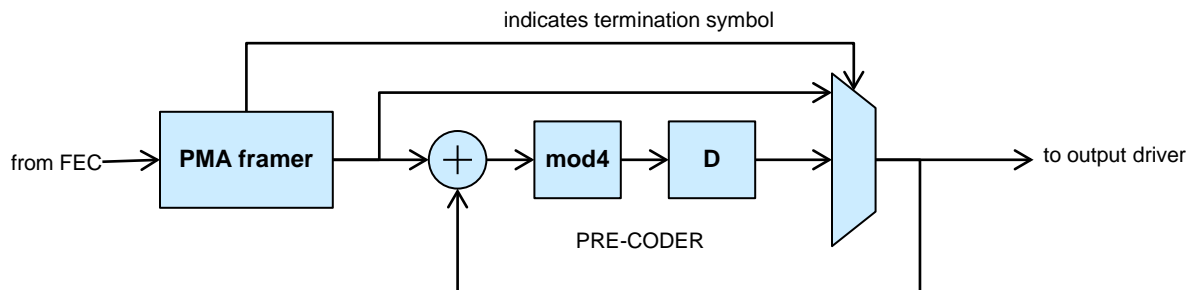
Motivation for pre-coding

- Pre-coding mitigates error propagation in DFE and MLSD receivers.
 - Greatly reduces number of errors per burst.
 - For 1-tap DFE, reduces burst to two errors, one at beginning and one at end
 - For MLSD see dabiri_01_0911 “Enabling Improved DSP Based Receivers for 100G Backplane”
- Graph shows improved coding gain (blue) due to precoding.
- The delta between burst error and random error is $\sim 1.0\text{dB}$ with $1/(1+D) \bmod 4$ precoding



PAM4 Block Termination

- PAM4 block termination symbol every 32 PAM4 symbols
 - For efficiency, each PAM4 termination symbol transmits one data bit.
 - 63 data bits sent every 32 PAM4 symbols
 - Increases baud rate by 64/63.
 - Each PAM4 block termination symbol is mapped to either +1 or -1.
 - At the transmitter, termination added within the precoder.
 - At the detector, termination removed after the detector.
- See dabiri_01_0112.
- PAM4 block termination encoding is mandatory; negotiation is not required.



Functional representation of block termination and pre-coding

Motivation for PAM4 Block Termination

- Block termination by transmitting known PAM4 symbols on a regular cycle enables...
 - efficient and effective MLSD, maximum likelihood sequence detection (dabiri_01_0911)
 - parallel DFE implementations
 - Keshab K. Parhi, Pipelining of parallel multiplexor loop and Decision Feedback Equalizers, ICASSP, 2004

PAM4 encoding

- Gray mapping
 - pre-coder output {10, 11, 01, 00} maps to {+1,+1/3,-1/3,-1}
 - based on 2B1Q coding used in HDSL and ISDN

PMA synchronization

- Lock to PAM4 termination blocks by searching for PAM4 termination symbols
 - PAM4 termination symbols (1 in 32) are always either +1 or -1.
 - Similar to framing on 10 or 01 sequence for 64B/66B, can borrow and modify 64B/66B synchronization state machine.
- Lock to PMA frame
 - Use known content of overhead bits.
 - Once locked to the PAM4 termination blocks, look for 4 bits (2 PAM4 symbols) every 88 rows.
 - Again, similar to 64B/66B synchronization.

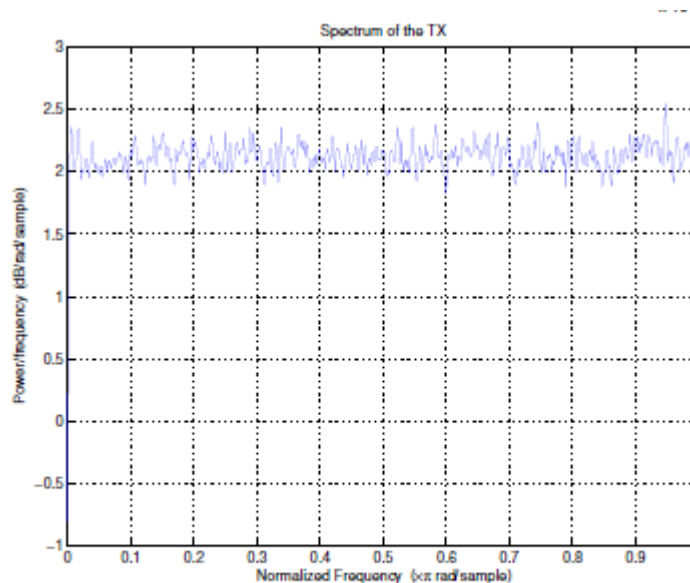
Energy Efficient Ethernet Operation

- Fast synchronization for REFRESH and WAKE.
 - Synchronize on PAM4 termination symbols.
 - Use prescribed sequence to accelerate synchronization.
- For REFRESH, PCS and FEC not required.
 - Replace with scrambled sequence.
 - Similar to EEE/LPI for 10GBASE-T.
- For WAKE, rapid alignment markers not required by the PMA/PMD receiver.
 - Will still be required at the PCS RX at the PCS end point.
- No significant impact to work being done in EEE consensus group.
 - Compatible and complementary with PCS state machine in Gustlin_02_1111.

Thanks!

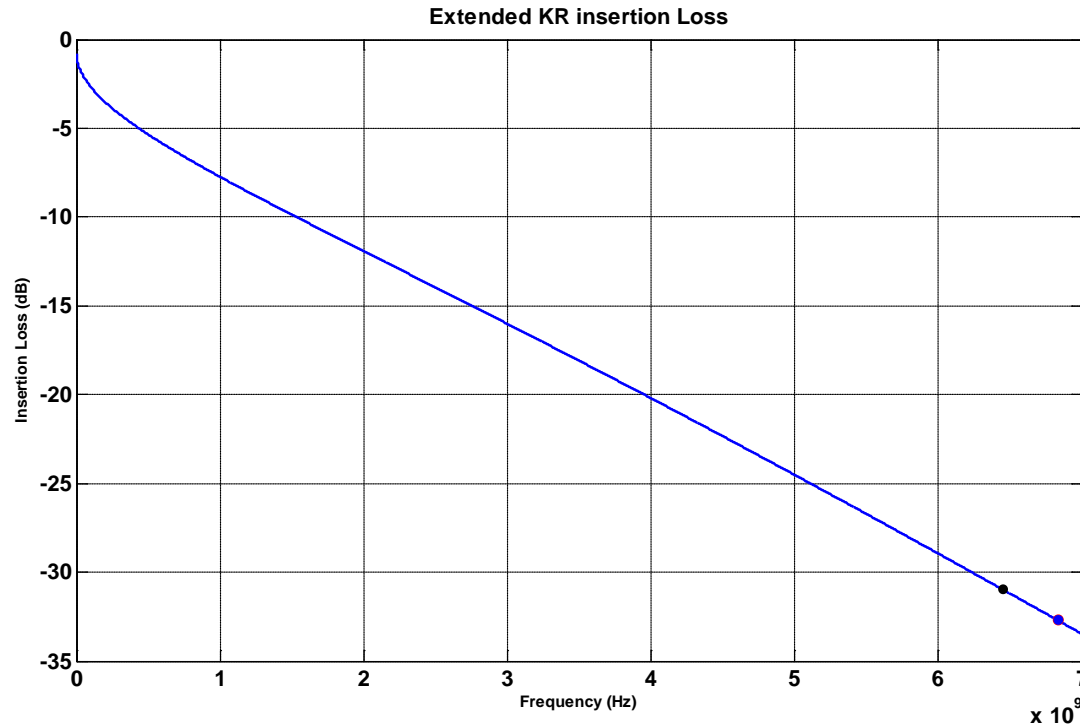
BACKUP SLIDES

Power spectrum with PAM4 block termination symbols



- The simulated spectrum above shows no spectral content due to block termination symbols.
- Pattern is repeating structure (not content) of 32 PAM4 symbols...
 - 31 random PAM4 symbols in $\{-1, -1/3, +1/3, -1\} * 3$
 - 1 random PAM4 symbol in $\{-1, +1\} * 3$

PAM4 SNR Loss due to Over clocking

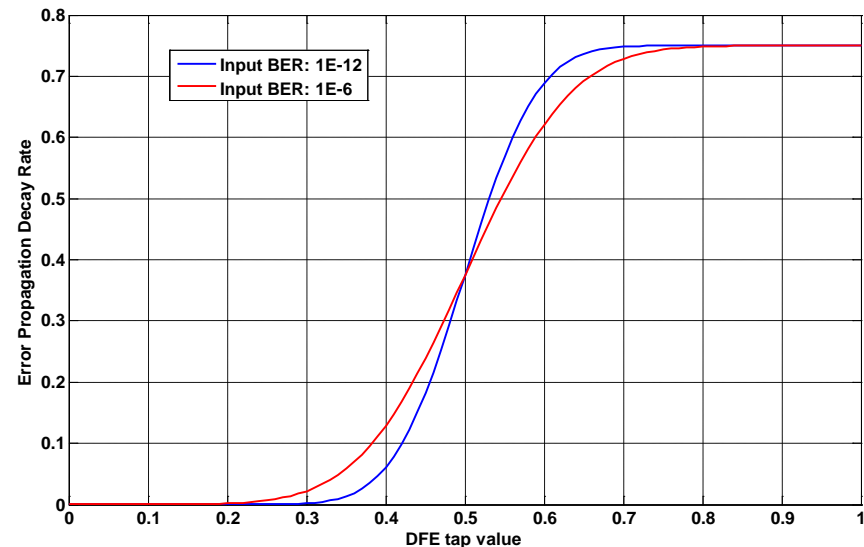


For FEC baud rate of 13.67G, the SNR loss due to over clocking

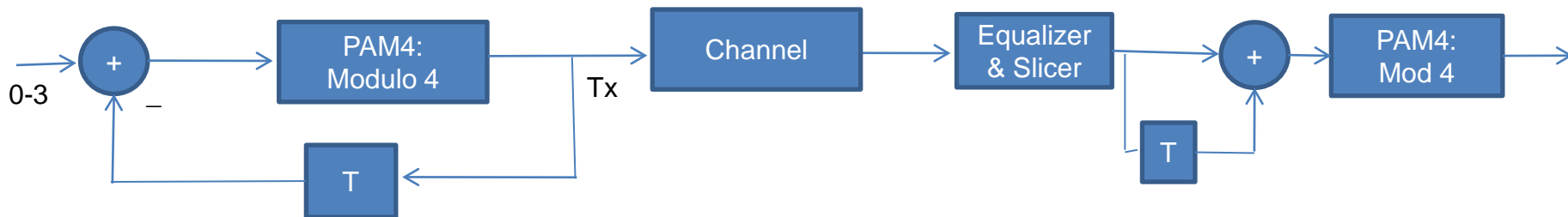
$$\triangleright \text{SNR}_{\text{delta}} = (\text{IL}_{6.84\text{GHz}} - \text{IL}_{6.45\text{GHz}}) / 2 = 0.9\text{dB}$$

Precoding Motivation: PAM4 DFE bursts

- DFE's are well known to multiply errors in the feedback loop
 - A single error will become a burst error
- Consider PAM4 1-tap DFE with tap coeff = 1
 - If previous decision is wrong, then there is 3/4 probability of making a successive error
 - i.e. Probability of K consecutive errors = $(3/4)^k$
- Lower 1st DFE tap between 0.6 to 1 have similar burst length as tap coefficient of 1
 - Tap of 1: 0.75^k
 - Tap of 0.7: 0.72^k
 - Tap of 0.6: 0.62^k
- A single random error may consume multiple Reed Solomon words
 - Burst error coding gain is lower than coding gain for random errors

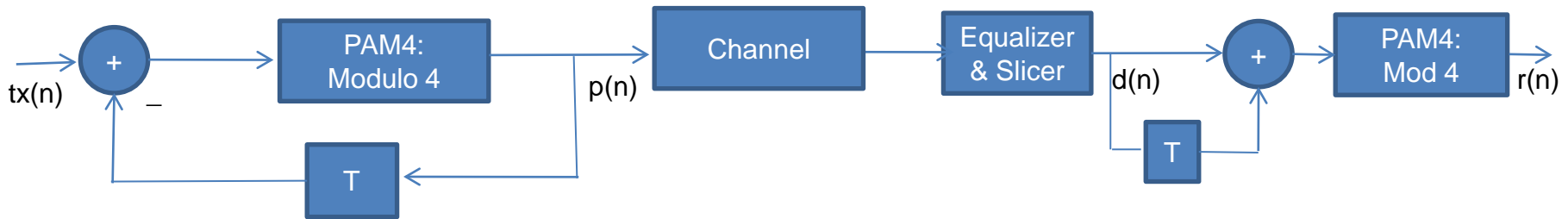


$1/(1+D)$ Precoding for DFE burst errors



- The burst error length of the DFE error events for PAM4 can be reduced by using precoding
- PAM4 Tx precoding uses a $1/(1+D) \bmod 4$
 - See bliss_01_0311, “**Signaling Terminology; PAM-M and Partial Response Precoders**”
 - Rx uses a $(1+D) \bmod 4$ after slicing
- Simple to implement
- Very low Complexity; similar complexity to duo-binary precoder
- Reduces 1 tap DFE burst error runs into 2 errors per error event
 - One error at the entry, one error at the exit

1/(1+D) Precoding worked example



- Precoder Input : $tx(n)$
 - 2 2 2 2 0 3 2 0 1 3 3 0 0 0 0 2 3 0 3
- Precoder Output : $p(n)$
 - 0 2 0 2 2 1 1 3 2 1 2 2 2 2 2 0 3 1 2
- DFE, Slicer Output : $d(n)$
 - 0 1 1 1 3 0 2 2 3 0 3 1 3 1 3 0 3 1 2
- Error Event : $p(n) - d(n)$
 - 0 1 -1 1 -1 1 -1 1 -1 1 -1 1 -1 1 -1 0 0 0 0
- Decoder Output after 1+D at Rx : $r(n)$
 - 2 1 2 2 0 3 2 0 1 3 3 0 0 0 0 3 3 0 3

↖ Entry Error ↖ Exit Error

This example does not include the PAM4 block termination.

Mean Time To False Packet Acceptance (MTTFPA)

- Assume any FEC frame known to be in error is marked.
 - Any 64B/66B blocks within the marked FEC frame are replaced with error blocks.
 - The errored packets are then eventually discarded by the downstream MAC.
 - Only FEC frames without error detected (falsely decoded) may result in falsely accepted packets.
- Probability of a FEC false decode, P_{FFD} (i.e. outputting a false codeword)
 - $P_{FFD} = 1/t!$, where t is the strength of the code
 - The output codeword will generally contain $2t+1$ errors
 - Ethernet CRC32 cannot guarantee detection for $2t+1$ errors
 - A false CRC32 match is random with probability 2^{-32}
- Probability of false packet acceptance, P_{FPA}
 - $P_{FPA} \sim P_{FFD} * BER_{OBJ} * 2^{-32} * N = 1.1E-35 * N$
 - N = average number of packets affected by each FEC frame, somewhere between 0 and 7
 - BER_{OBJ} = FEC BER objective = $1E-12$
- For mandatory PAM4 FEC, RS(444, 412, $t = 16$)
 - $MTTFPA \sim 1/P_{FPA} * 1 / (13.75E9 * 2 * 4) * 1 / (60*60*24*365)$ years = $2.6E16 / N$ years
 - For $N = 7$, $MTTFPA \sim 3.9e15$ years
 - Lifetime of universe is $\sim 13E9$ years.