DesignCon 2024

200+ Gbps Ethernet Forward Error Correction (FEC) Analysis

Cathy Ye Liu, Broadcom Inc.

cathy.liu@broadcom.com

Abstract:

Recently, the IEEE 802.3 has established the 200 Gbps, 400 Gbps, 800 Gbps and 1.6 Tbps Ethernet task force 802.3dj which is aiming at 200 Gbps per lane link rate. In order to study what is needed and being adopted for the next speed node, this paper covers new forward error correction (FEC) technology, modeling, and performance analysis to aim at 200+ Gbps Ethernet systems and how different FEC options affect signal integrity.

Authors Biography

Cathy Ye Liu, distinguished Engineer, currently heads up Broadcom SerDes architecture and modeling group. Since 2002, she has been working on high-speed transceiver solutions. Previously she has developed read channel and mobile digital TV receiver solutions. Her technical interests are signal processing, FEC, and modeling in high-speed optical and electrical transceiver solutions. She has published many journal articles and conference papers and holds 20+ US patents. Cathy has demonstrated her leadership roles in industry standard bodies and forums. Cathy currently serves as the vice president of the board director of Optical Internetworking Forum (OIF), a member of the board of advisors for the department of Electrical & Computer Engineering (ECE) of University of California at Davis, a member of Signal Integrity Journal editorial advisory board, and the co-chair of the DesignCon technical track of high speed signal processing, equalization and coding. She received DesignCon 2021 engineer of the year award. She received her B.S. degree in Electronic Engineering from Tsinghua University, China and received her M.S. and Ph.D. degrees in Electrical Engineering from University of Hawaii.

1. Introduction

With the fast growth of 5G/6G networks and AI/ML applications, serial link data rates continue to increase due to high speed communication and large bandwidth demands. Recently the IEEE 802.3 has established a 200 Gbps, 400Gbps, 800 Gbps and 1.6 Tbps Ethernet task force 802.3dj [4]. The new task force is aiming at 200 Gbps per lane link rate, doubled from the 100 Gbps per lane rate developed as part of IEEE 802.3bs [1] and 802.3ck [2].

A decade ago the semiconductor industry successfully updated signaling formats from NRZ to PAM4 during the transition from 25 Gbps to 50 Gbps link rates. To offset the signal-to-noise ratio (SNR) penalty caused by higher modulation levels, forward error correction (FEC) has become an essential part of the solution for PAM4 systems. This paper follows and builds from two previous papers: "What is FEC and how do I use it?" [5] and "100+ Gbps Ethernet Forward Error Correction (FEC) Analysis)" [6], and provides updates for the next generation Ethernet rate, 200+ Gbps per lane.

In order to study what is needed and what has been adopted for the next Ethernet speed node of 200 Gbps per lane, this paper investigates different FEC schemes such as end-end, concatenated, and segmented FECs, and how these different FEC schemes affect signal integrity (SI) and performance in different end applications.

With the latest decision made within the IEEE 802.3dj task force for 800GBASE-R and 1.6TBASE-R, in which the Hamming (128,120) inner code has been adopted as a part of the FEC solution for 200 G/s per lane IM-DD optics [7], concatenated FEC modeling and performance analysis have become key to multiple-part link system analysis including the physical medium dependent (PMD) optical channel and two or more Attachment Unit Interface (AUI) electrical channels. Since inner code miss-correction is highly undesirable for overall concatenated FEC performance, its cause, impact, and mitigation schemes need to be studied. Furthermore, soft decision decoding is recommended to achieve better coding gain. In addition, in order to meet low latency requirement especially for AI/ML applications, bypass options to skip certain FEC stages such as the inner code or interleaver, are implementation options to consider.

After analysis on concatenated FEC for chip-module interfaces, the paper will discuss other FEC options for long reach cable backplane and direct attached cable (DAC) interfaces, in which the bandwidth is highly limited and the concatenated FEC might not be suitable due to its overhead.

The paper will conclude by extending the discussion on FEC options to next generation serial link data rates of 400+ Gbps.

2. Refreshed channel error model and FEC performance analysis

In [6], random and burst channel errors and FEC performance analysis models were introduced for 100 Gbps Ethernet rate. In this paper, the same or similar models are used to study the next generation FEC performance. In this section we will add some updates based on the latest developments of 802.3df [3] and 802.3dj [4] task forces.

2.1 FEC architecture and performance for 8×100G PHYs

Recently the 802.3df task force adopted a new Ethernet physical coding sublayer (PCS) and physical medium attachment (PMA) for 800GE with 8×100G PHYs as shown in Figure 1. It is based on two 400GE [1] PCS FEC flows (flow-0 and flow-1) in parallel. Together there are 32 flow lanes, each running at 25 Gbps. Specific flow lanes map to a given PMA output lane such that the 4:1 bit multiplexing is conceptually the same as 400GE.

800GE with 8×100G PHYs has two flows, each containing two Reed Solomon (RS) (544, 514, 15) codewords. To utilize more coding gain, the 802.3df task force defined that each physical lane accesses all four FEC codewords equally, which results in 4-way codeword interleaving instead of 2-way codeword interleaving, as in 802.3bs [1]. From [6] we understand that codeword interleaving can provide better coding gain especially for burst channel errors.

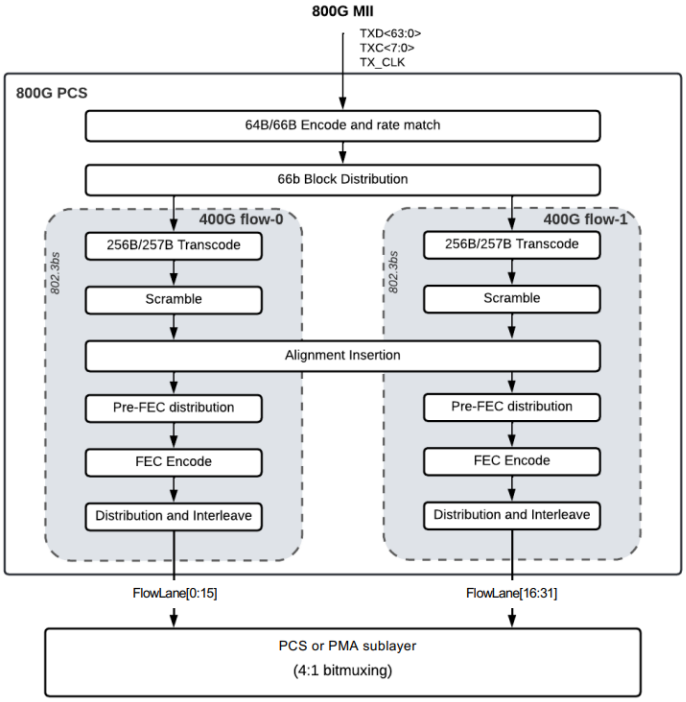


Figure 1. 800GE with 8×100G PHYs PCS transmitter flow

To study the FEC performance of 800GE with 8×100G PHYs with decision feedback equalization (DFE) and different PCS/PMA coding schemes, the Monte Carlo model introduced in [6, section 2.4] is used for analysis in this paper. Real data patterns and DFE tap coefficients are applied in the model. Gaussian noise is added at the decision slicer for a certain detector error ratio or equivalent PAM4 symbol error ratio (SER_{pam4}) (say 1e-4). To shorten the simulation time, an initial error is inserted and the resulting error event through DFE error propagation or other correlated error sources are captured in each trail. Within a reasonable simulation time, a large number of error events are collected and fed into the PMA and PCS layers with specified bit multiplexing (or symbol multiplexing) and codeword interleaving to obtain RS (544, 514, 15)

(as known as KP-FEC) symbol error statistics, {p(1), p(2), p(3), …, p(t)}, which is the probability of a burst error causing 1, 2, 3, … RS symbol errors given an initial error. At the end, post-FEC performance, codeword error rate (CER), can be calculated with inputs {p(1), p(2), p(3), …, p(t)} based on Equation 9 in [6].

The BER objective for 802.3df is 1e-13 and equivalent frame loss ratio (FLR) target is 6.2e-11. The relationship between FLR and CER is determined as following:

- FLR=1.125*CER for single codeword interleaving,
- FLR=2.125*CER for 2 codeword interleaving, and
- FLR=4.125*CER for 4 codeword interleaving.

Figure 2 shows how a group of bits encode to RS symbols and later to PAM4 symbols in the 200GBASE-R and 400GBASE-R transmitter flows with the coding scheme of 4:1 bit multiplexing and 2-way codeword interleaving. Each block represents one bit, and for each block the top number represents the corresponding RS symbol index number that it is encoded to, and the bottom number represents the corresponding bit index number within that RS symbol. Each PAM4 symbol consists of two bits, the lsb (least significant bit) and msb (most significant bit). Unlike symbol multiplexing, lsb and msb bits of adjacent PAM4 symbols could belong to different RS symbols with 4:1 bit multiplexing. The block colors (blue and red) represent which RS codeword it belongs to with 2-way codeword interleaving. The purpose of codeword interleaving is to break a long burst error into two separate codewords and thus to improve the coding gain.

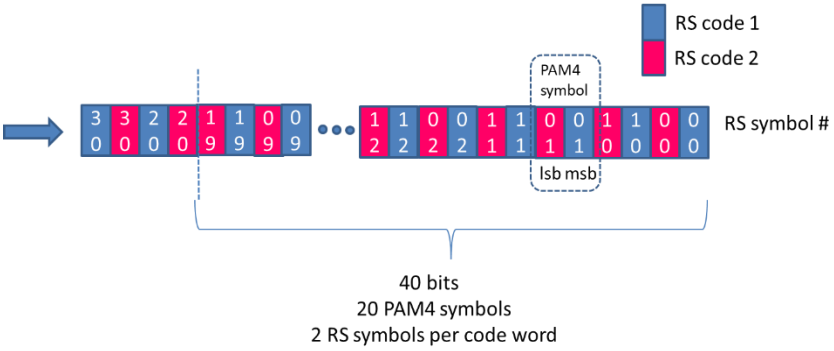


Figure 2. 4:1 bit multiplexing and 2-way codeword interleaving

Figure 3 shows the defined coding scheme in the latest 802.3df 800GBASE-R (8×100G PHYs) specification, 4:1 bit multiplexing and 4-way codeword interleaving. We can see that with 4-way codeword interleaving, adjacent PAM4 errors split to four separate codewords and thus further improve the coding gain.

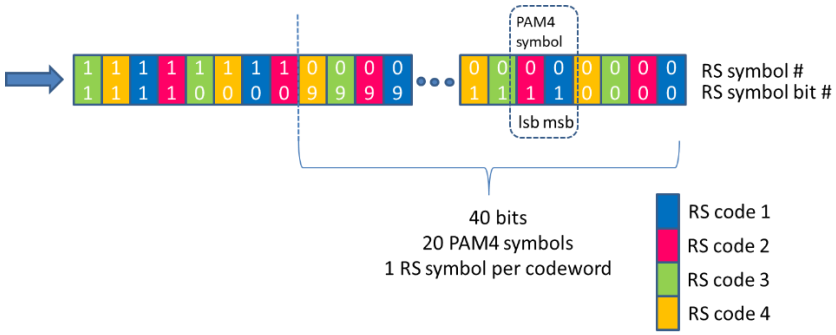


Figure 3. 4:1 bit multiplexing and 4-way codeword interleaving

To compare the coding gains of two coding schemes, 2-way codeword interleaving and 4-way codeword interleaving, post-FEC performance of those two are simulated through the Monte Carlo model over three different types of channel errors:

- Random error a=0
- Medium burst error, a=0.375 such as a 1-tap DFE with tap coefficient equal to 0.5
- Maximum burst error, a=0.75 such as a 1-tap DFE with tap coefficient equal to 1

For this analysis, "a" is the probability of getting an error in the next PAM4 symbol following an initial error. For the maximum error propagation case where a=0.75, 1/(1+D) mod 4 precoding [6] is enabled. For other two cases of a=0 and a=0.375, precoding is disabled.

Figure 4 shows the post-FEC FLR performances vs. slicer SNR values with 2-way and 4-way interleaving schemes, respectively. We can see that 4-way interleaving outperforms 2-way interleaving especially for burst errors with large a values. Such analysis and contributions guided the 802.3df task force to adopt 4:1 bit multiplexing and 4-way codeword interleaving in its PCS/PMA specification.

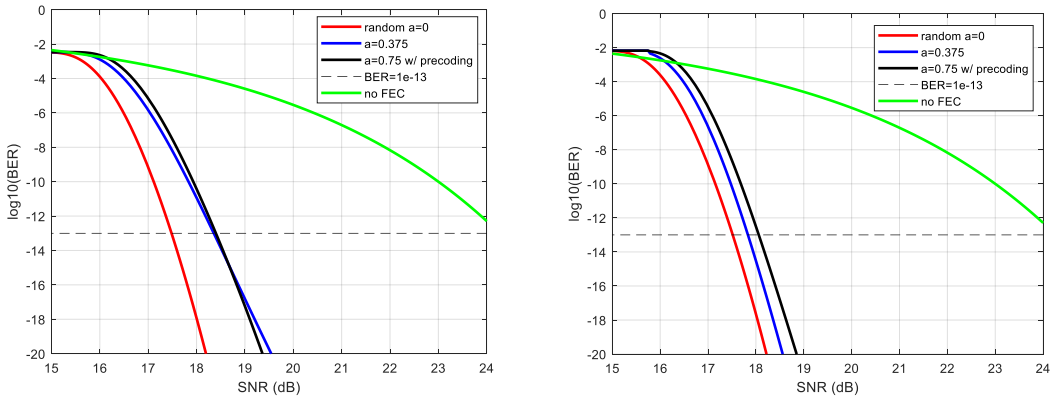


Figure 4. Post-FEC BER performance vs. slicer SNR values for 4:1 bit multiplexing with 2-way (left) and 4-way (right) codeword interleaving coding schemes

3. 200 Gbps FEC schemes and coding algorithms for optical channels

The 802.3dj task force has doubled the data rate from 100 Gbps to 200 Gbps per lane. In this section we will study the potential FEC solutions for chip to optical module interfaces at 200 Gbps data rate, provide performance analysis for different coding schemes and coding algorithms, and discuss their effects to system signal integrity.

A three-part link with two chip-module electrical interfaces (AUI) and one optical link (PMD) as shown in Figure 5 will be focus for analysis in this section.

Figure 5. An example of a three-part link containing electrical (AUI) and optical (PMD) sections

3.1 Three FEC architectures

There are three major FEC architectures in a multi-part link system that have been discussed in the 802.3dj task force and shown in Figure 6:

- Type-1: single FEC spans multiple AUIs and the PMD link, referred to as “end-end FEC”
- Type-2: outer FEC spans multiple AUIs and PMD link (like Type-1) with an additional inner FEC spans PMD link, referred to as “concatenated FEC”
- Type-3: different FECs are dedicated to AUIs and PMD links, referred to as “terminated FEC”

All current 200GBASE-R, 400GBASE-R and 800GBASE-R systems at 100 Gbps per lane use Type-1 FEC, in which electrical AUIs and optical PMD link share a single FEC located at both ends of hosts (such as switch chips). Since the single FEC corrects errors that are contributed by both PMD and AUIs, the BER tradeoff between the AUIs and the PMD is worth investigating. The performance analysis for such end-end FEC systems has been described in [6]. Normally the required BER target on the electrical link is more stringent than the optical link in an end-end FEC architecture since the bulk of the coding gain is allocated to the toughest part of the channel (the optical link), and a relatively smaller coding gain is allocated to the electrical links. In order to meet 1e-13 post-FEC BER target or 6.2e-11 FLR target, the overall input BER to the end-end FEC needs to be 2.8e-4 or lower, with the combination of 2.4e-4 BER target for PMD link and 1e-5 BER target for AUI interfaces.

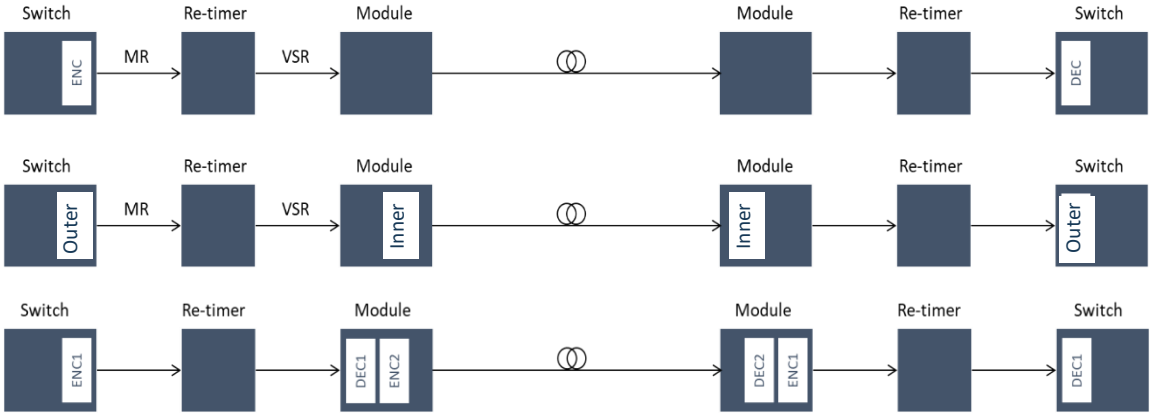


Figure 6. Three types of FEC architectures (top: end-end; middle: concatenated; bottom: terminated)

In order to double the data rate to 200+ Gbps per lane, the PMD link and/or the AUIs may require extra FEC protection (further relaxing pre-FEC BER). Hence, both Type-2 and Type-3 FEC architectures are considered.

For the Type-2 concatenated FEC, an outer FEC spans the whole link from host to host (like Type-1), plus an inner FEC spans only the PMD part. The inner FEC corrects "most" errors contributed by the PMD part while the outer FEC corrects PMD errors not corrected by the inner FEC and errors contributed by the AUIs. The combined effect of inner and outer FECs results in the target BER and FLR for the whole link. This concatenated FEC scheme is new for 802.3.

For the Type-3 segmented FEC, dedicated FECs protect different parts of the link such that DEC_1 corrects errors contributed only by one chip-module interface while DEC_2 corrects error contributed only by PMD link. Since each part of the link has its own FEC protection, no BER target tradeoff between the AUIs and the PMD link is required. Both 400GBASE-ZR and 802.3cw use terminated FEC.

Compared to Type-1, both Type-2 and Type-3 might provide better coding gain to PMD link and/or AUIs. However, the Type-3 FEC architecture costs extra latency, power, and complexity due to three FEC segments (three sets of encoders and decoders) to support. While Type-2 FEC provides extra FEC protection for the PMD link with a smaller increment in latency, power, and complexity compared with Type-3 FEC. The intent of Type-2 FEC is to provide a compromise of better performance than Type-1 and lower latency, power, and complexity than Type-3. Therefore, Type-2 concatenated FEC has been adopted as a part of the FEC approach for 200 Gbps per lane IM-DD optics [7]. Details of the proposed concatenated FEC will be discussed later in section 3.3. In next section 3.2, the outer FEC at the host sides will be studied first.

3.2 Host FEC

Transitioning from 100 Gbps per lane to 200 Gbps per lane, industry and the 802.3dj task force preferred to keep a similar PCS structure and RS FEC for maximum reuse and backward

compatibility. Initially, alternative RS codes besides RS (544, 514, 15) were studied for more coding gain. Later, different PMA multiplexing and PCS codeword interleaving schemes were investigated and adopted.

3.2.1 Alternative options to the PCS Reed Solomon (RS) codes

To ease the system design, certain construction rules and assumptions are taken for the PCS RS code selection. For each candidate RS (n, k, t) over GF (2^m) [9]:

- Each RS symbol consists of 8 to 12 bits (e.g., 10-bit symbol for KP-FEC, m=10)
- Assume a 256B/257B block code to avoid additional transcoding
- Message size ($m*k$ bits) corresponds to an integer number of 257-bit blocks
- Codeword spreads evenly across 4, 8 and 16 physical lanes
- Signaling rate is an integer multiple of a 625 MHz reference clock

Table 1 lists a group of candidate RS codes that meet above construction rules and assumptions. Their required slicer SNRs and SER_{pam4} to meet 1e-13 post-FEC BER are calculated for random error case (a = 0). We can see that increasing error correction capacity t (number of RS symbol errors can be corrected per codeword) results in a larger coding gain and relaxed pre-FEC BER target. For example, the RS (576, 514, 31) and RS (4088, 3855, 112) codes can provide more than 1 dB coding gain over the RS (544, 514, 15) code. However, the larger t value implies a lower code rate and/or longer codeword length. Lowering the code rate is undesirable for bandwidth limited copper channels such as backplane or copper cable channels. Meanwhile, the longer codeword length introduces a larger latency and might not be suitable for systems with low latency requirements such as AI and ML applications.

Table 1. Alternative RS codes besides RS (544, 514, 15) and their performance comparisons

RS code (n, k, t)	Bits per symbol	Code rate (k/n)	SNR (dB)	SER_{pam4}
RS (544, 514, 15)	10	0.945	17.45	6.4e-4
RS (560, 514, 23)	10	0.918	16.74	1.6e-3
RS (576, 514, 31)	10	0.892	16.25	2.8e-3
RS (1088, 1028, 30)	12	0.945	17.02	1.1e-3
RS (2176, 2056,60)	12	0.945	16.63	1.8e-3
RS (3264, 3084, 90)	12	0.945	16.46	2.2e-3
RS (4080, 3855, 112)	12	0.945	16.39	2.4e-3

The RS code encoder is a straight forward shift register [9] which contributes negligible encoding latency while the majority of coding latency is from the decoder. As shown in Figure 7, a RS code decoder has 3 stages:

- Syndrome generation: FEC codeword accumulation time at port speed
- Key equation solver: the Berlecamp-Massey algorithm normally has 2t iterations plus a few additional clock cycles
- Chien search and data correction: FEC codeword size and datapath width dependence

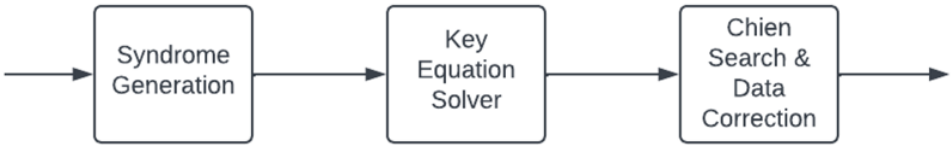


Figure 7. Block diagram of a Reed Solomon code decoder

Overall, the RS code decoding latency is a function of the codeword length *n*, error correction capability *t*, physical lane data rate, PCS FEC decoder cycle time, datapath width, and codeword interleaving depth. Taking assumptions of a 1.56GHz cycle time and 640-bit datapath, Table 2 lists the decoder latency for different FEC codes for selected numbers of physical lanes and codeword interleaving depths. The RS (544, 514, 15) FEC with 4-way codeword interleaving over 4×200G PHYs introduces ~55ns latency.

Table 2. RS code decoder latency

RS FEC (*n*, *k*, *t*)	Link rate (Gbps)	Cycle time (GHZ)	Codeword interleaving	Physical lanes per port	Stage 1 latency (ns)	Stage 2 latency (ns)	Stage 3 latency (ns)	Total latency (ns)
RS (544, 514, 15)	106.25	1.56	1	1	51.20	21.12	8.64	80.96
RS (544, 514, 15)	212.5	1.56	1	1	25.60	21.12	8.64	55.36
RS (544, 514, 15)	212.5	1.56	1	4	6.40	21.12	8.64	36.16
RS (544, 514, 15)	**212.5**	**1.56**	**4**	**4**	**25.6**	**21.12**	**8.64**	**55.36**
RS (576, 514, 31)	225	1.56	1	4	6.40	41.60	8.96	56.96
RS (1088, 1028, 30)	212.5	1.56	1	4	15.36	40.32	16.26	71.94
RS (2176, 2056, 60)	212.5	1.56	1	4	30.72	78.72	29.31	138.75
RS (3264, 3084, 90)	212.5	1.56	1	4	46.08	117.12	42.37	205.57
RS (4080, 3855, 112)	212.5	1.56	1	4	57.60	145.28	52.16	255.04

It is likely that the 802.3dj task force will reuse RS (544, 514, 15) (also known as the KP-FEC) as the PCS FEC code due to its backward compatibility and good tradeoff between performance and latency.

3.2.2 Symbol multiplexing and 4-way codeword interleaving scheme

Analysis in [6] showed 4:1 bit multiplexing has a larger coding gain penalty for burst errors than 2:1 bit multiplexing and symbol multiplexing. We can expect that increasing the bit multiplexing to 8:1 for 200 Gbps per lane would further degrade the performance. Figure 8 shows the FEC performance for the burst error $a = 0.75$ case with different PMA schemes (8:1 bit multiplexing (BM8), 4:1 bit multiplexing (BM4), and symbol multiplexing (SM)) and 4-way codeword interleaving (CI4). We can see that to meet a FLR=6.2e-11 (or BER=1e-13) post-FEC target, the required slicer SNR will be 18.25 dB and 18.05 dB for BM8 and BM4, respectively.

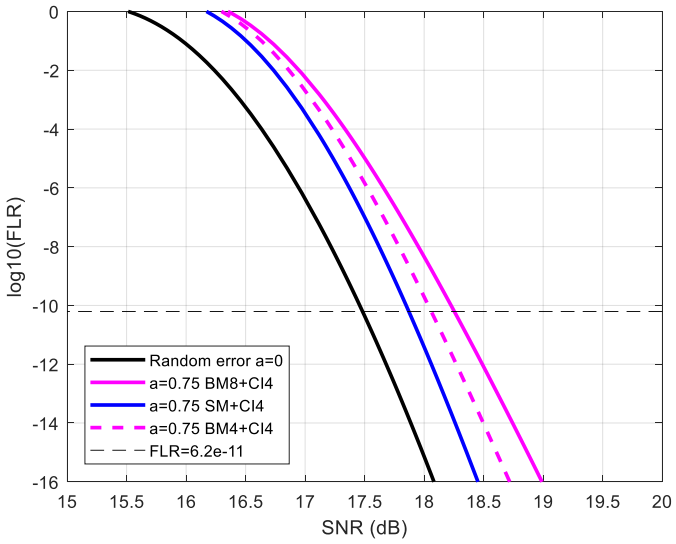


Figure 8. FEC performance for random error a=0 and burst error a=0.75 with 4-way codeword interleaving

To avoid the bit-multiplexing penalty, symbol multiplexing is considered. Figure 9 depicts how a group of bits encode to RS symbols and later to PAM4 symbols for symbol multiplexing and 4-way codeword interleaving, in which a long burst error only impacts one or very few RS symbols per each codeword. It is not surprising to see in Figure 8 that SM+CI4 outperforms BM4 and BM8 for a burst error case a=0.75, and reduces the gap to the random error case of a=0.

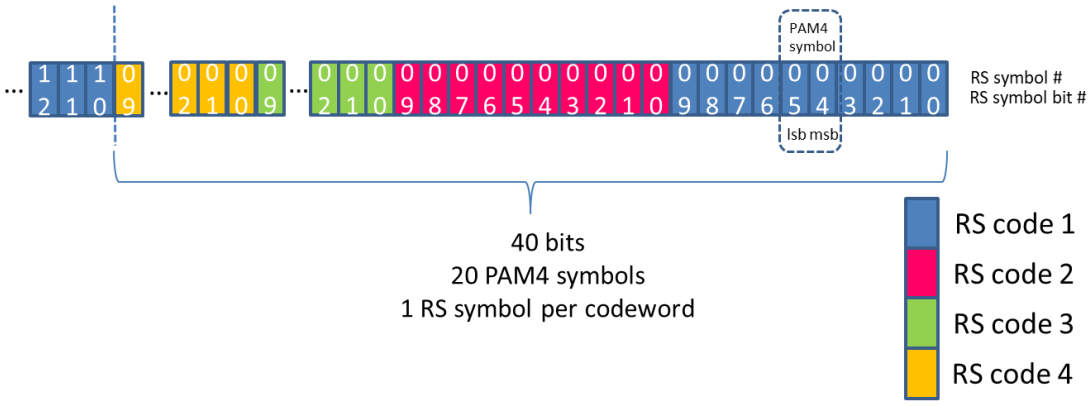


Figure 9. Symbol multiplexing and 4-way codeword interleaving

With the help of symbol multiplexing and 4-way codeword interleaving, 200 Gbps link allows a higher DFE tap weight. Figure 10 shows the simulation results of a 1-tap DFE with tap coefficient h_1 swept from 0 to 0.5. We can see that SM allows larger DFE weights than BM8, for example, h_1 is up to 0.5 for SM vs. 0.3 or lower for BM8.

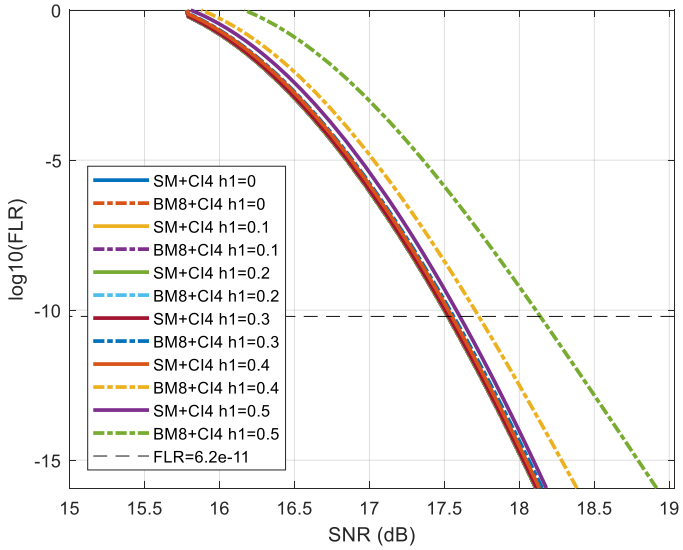


Figure 10. FEC performance for SM and BM8 with 4-way codeword interleaving over different DFE h_1 coefficients

It is likely that the 802.3dj task force will adopt RS (544, 514, 15) with symbol multiplexing and 4-way codeword interleaving as its host PCS and PMA coding scheme.

3.3 Concatenated FEC: Inner code and decoding algorithms for the optical PMD

In this section, we will focus on the Type-2 concatenated FEC that has been adopted by 802.3dj as a part of the FEC approach for 200 G/s per lane IM-DD optics.

3.3.1 Hamming code and its encoder

For a concatenated FEC, the inner code should be short and with small overhead such that the overall concatenated FEC has reasonable code rate and relatively low latency. Hence, Hamming codes or BCH codes [9] were considered.

Recently, the 802.3dj task force adopted the Hamming (n=128, k=120) code as its inner code for the concatenated FEC. This inner code is based on the Hamming code (127, 120) by adding one extended parity check bit.

With the extended parity check bit, the minimum distance d_{min} = 4. This improves the error detection capability to 3 bits per codeword, while the error correction capability is still 1 bit per codeword.

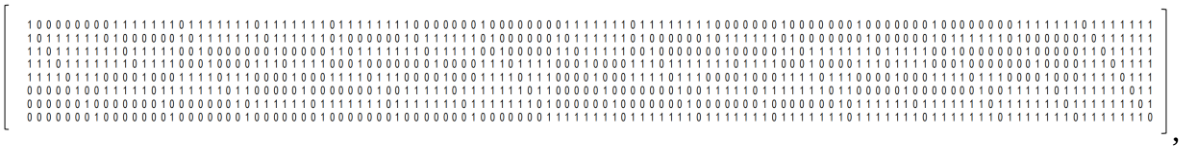
The encoding process of a linear block code or a Hamming code can be defined as a matrix operation: $\boldsymbol{c} = \boldsymbol{u} \cdot G_{k,n}$, where $\boldsymbol{c}$ is the codeword sequence, $\boldsymbol{u}$ is the information part of the codeword and G is the generator matrix that uniquely defines the linear block code. A desirable property for a linear block code to possess is the systematic structure of the codewords, in which a codeword is divided into two parts: the message part, and the parity check part. The message part consists of k unaltered information (or message) digits, and the parity check part consists of

n-k redundant digits, which are linear sums of the information digits. A linear systematic (*n*, *k*) code is completely specified by the *G* matrix in the following form:

$$G_{k,n} = [P_{k,n-k} \; I_{k,k}]. \qquad \text{Eq. 1}$$

Let $I_{k,k}$ denote the *k*-by-*k* identity matrix and $P_{k,n-k}$ denote the parity matrix. As defined in [11],

$P^{T}_{120,8}$ =



,

where P^{T} is the transpose of the matrix *P*.

3.3.2 Hamming code decoder

The decoding procedure for a Hamming code consists of three steps:

1. Compute the syndrome sequence ***s*** of the received vector ***r*** to detect errors
2. Identify the location of the error
3. Correct the error

In the first step, the syndrome sequence ***s*** is calculated as

$$\boldsymbol{s} = H_{n-k,n} \cdot \boldsymbol{r} \qquad \text{Eq. 2}$$

where ***r*** is received noisy sequence and $H_{n-k,n}$ is the parity check matrix formed as

$$H_{n-k,n} = [I_{n-k,n-k} \; P^{T}_{k,n-k}]. \qquad \text{Eq. 3}$$

If the syndrome sequence ***s*** is all-zero, a correct transmission is assumed. Otherwise, errors are detected. The Hamming code (128, 120) with $d_{min} = 4$ is able to detect up to three bits in error per codeword.

The second step of the decoding is generally the hardest part, to identify the error location. For a Hamming code with $d_{min} = 4$, it can only correct one bit error per codeword. Since there are eight bits in the syndrome sequence $\boldsymbol{s} = (s_0, s_1, s_2, s_3, s_4, s_5, s_6, s_7)$ and s_0 is the extended parity bit, there are 2^7 possible patterns of $\boldsymbol{s}' = (s_1, s_2, s_3, s_4, s_5, s_6, s_7)$. For a hard decision decoding decoder, each of these ***s***' corresponds to a unique error location in ***r***. After the error location is identified, the error can be corrected easily.

If more than one bit error per codeword does occur, such a decoding procedure could detect errors but miss-correct them in wrong locations, referred to as miss-correction.

3.3.3 Miss-correction of inner code

For concatenated codes, miss-corrections of the inner decoding could significantly degraded performance. In particular, let $\boldsymbol{r} = \boldsymbol{c} + \boldsymbol{n}$, where $\boldsymbol{c}, \boldsymbol{n} \in \{0, 1\}$ denote an inner codeword and a random or a burst error vector. A single error correction Hamming code yields the correct codeword $\boldsymbol{c} \in C$ if and only if $d_H(\boldsymbol{r}, \boldsymbol{c}) = w_H(\boldsymbol{n}) <= t$, where d_H and w_H denote the Hamming distance and weight [9]. Conversely, if $w_H(n) > t$, the decoding either fails or there exists another codeword $\boldsymbol{c}' \in C$ such that $d_H(\boldsymbol{r}, \boldsymbol{c}') <= t$. In the latter case, we say that a miss-correction occurs, meaning that decoding is technically successful but the decoded codeword $\boldsymbol{c}'$ is not the correct one. Miss-corrections are highly undesirable because they introduce additional errors (on top of channel errors) into the outer code decoding and can make miss-corrections more difficult to resolve for the outer code decoder.

In practice, decoding is implemented by first calculating the syndrome $\boldsymbol{s}$. There are 2^{n-k} possible syndromes. Each syndrome is associated with an estimated error vector $\boldsymbol{n}'$, where $w_H(\boldsymbol{n}) <= t$, or a decoding failure. For the first case, the decoded output is computed as $\boldsymbol{r} + \boldsymbol{n}'$ and $\boldsymbol{n}' = \boldsymbol{n}$. The second case corresponds to an undetected error or miss-correction where $\boldsymbol{n}' \neq \boldsymbol{n}$. In particular, for binary t-error-correcting block codes, the probability of miss-correction is simply the ratio of the number of decodable syndromes and the total number of syndromes:

$$\frac{\sum_{i=0}^{t}\binom{n}{i}}{2^{n-k}}. \qquad \text{Eq. 4}$$

This could be a large value when t is small (say t =1) and/or n-k is small. For the Hamming code (128, 120), the probability of miss-correction is as high as 0.5039 so definitely not negligible.

In order to model the inner code miss-correction and to its impact to overall concatenated FEC performance, real inner code encoders and decoders are added to the Monte Carlo simulation with outer RS (544, 514, 15) code including symbol multiplexing and 4-way codeword interleaving.

Figure 11 shows concatenated FEC performance with hard decision (HD) decoding for different channel error profiles and DFE coefficients of h_1=0 and h_1=0.5. We can see that for a random error case (h_1=0), the inner code can improve slicer BER 1-2 orders of magnitude and improve post-FEC BER 5-10 orders of magnitude over the KP-FEC only, this is equivalent to 1.5dB of coding gain to meet a 1e-13 post-FEC BER. However, for the correlated error case (h_1=0.5), the performance of the concatenated FEC with HD inner code decoding is degraded significantly due to the miss-corrections. In this case, the coding gain reduces to only 0.3dB.

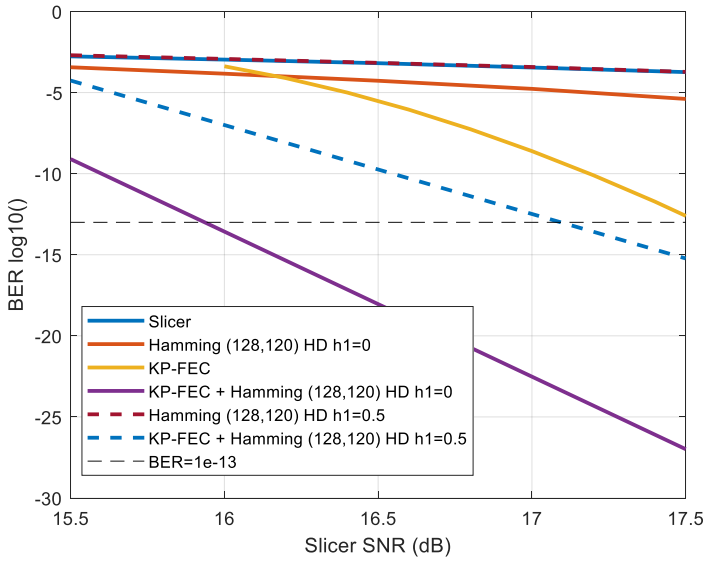


Figure 11. Concatenated FEC performance with HD inner code decoding

In order to mitigate the inner code miss-correction impact and improve the concatenated FEC performance, soft decision (SD) decoding algorithms and further interleaving schemes can be considered.

3.3.4 Soft decision decoding (SD)

All the decoding algorithms discussed so far are based on hard decision outputs of the receiver, that is, the input to the decoder for each bit is quantized in two levels, denoted as 0 and 1. Then the decoder processing this binary received sequence is referred to as hard decision (HD) decoding. If the outputs of the receiver are unquantized or quantized into more than two levels, a sequence of soft decision inputs can be taken to the decoder to process soft decision (SD) decoding.

Because the decoder uses the additional information to recover the transmitted codeword, SD decoding provides better FEC performance than HD decoding. In general, soft decision maximum likelihood decoding (MLD) of a code has about 3dB of coding gain over HD decoding [9]. However, MLD can be much harder to implement than HD decoding and requires more computational complexity and decoding latency.

To achieve a better trade-off between performance and decoding complexity, some practical suboptimal SD decoding algorithms can be applied. Chase introduced three algorithms in [10], namely, algorithm-1, algorithm-2, and algorithm-3, with different complexity levels. This paper uses Chase algoritm-2 as the SD decoding algorithm for the inner Hamming decoder.

Let $\boldsymbol{r} = (r_0, r_1, \ldots, r_{63})$ be a soft decision received sequence at the output of the receiver slicer. Each receiver symbol r_i with $0 \leq i \leq 63$ is determined independently to z_i, $z_i \in \{0, 1, 2, 3\}$ for

PAM4 signaling. Then, the magnitude of slicer error, $|r_i - z_i|$, can be used as a reliability measure of the HD bit z_i. The larger $|r_i - z_i|$ is, the less reliable the hard decision z_i is. Based on the reliability measure of the received symbols, a group of least reliable positions (LRPS) can be identified. The precision of the slicer errors is implementation determined. 5-bit fixed-point values are used in this paper for the slicer error, $|r_i - z_i|$.

Let $\boldsymbol{z} = (z_0, z_1, \ldots, z_{63})$ be the hard decision received sequence obtained from $\boldsymbol{r}$. Then, errors are more likely to occur in the LRPs. The errors in the LRPs can be reduced or removed by modifying the hard decision received sequence in these positons.

Let E be a set of error patterns with errors defined only in the LRPs. For each error pattern $\boldsymbol{e}$ in E, $e_i \in \{-1, 1\}$, we modify $\boldsymbol{z}$ by adding $\boldsymbol{e}$ to $\boldsymbol{z}$. Consequently, there are error patterns in E that reduce the number of errors at the LRPs of $\boldsymbol{z}$, and very likely, the modified test pattern $\boldsymbol{z}+\boldsymbol{e}$ contains either no errors or a number of errors within the error-correcting capability t of the code.

In Chase algorithm-2 the number of LRPs locations to consider is $\lfloor d_{min}/2 \rfloor$. In our case, $d_{\min} = 4$ for the Hamming (128, 120) code. Hence, there are 2^2 possible test patterns, including the all-zero pattern. For each of these four test patterns, a hard decision decoding is applied to generate a list of at most four candidate codewords. For each generated candidate codeword, a SD decoding metric can be calculated. The candidate codeword with the best metric is selected as the decoded solution. This SD decoding procedure is shown in Figure 12 and in following steps:

1. Form the hard decision received sequence $\boldsymbol{z}$ from $\boldsymbol{r}$ and assign a reliability value to each symbol of $\boldsymbol{z}$.
2. Generate the error patterns in E one at a time, possible in likelihood order. For each error pattern $\boldsymbol{e}$, form the test patterns $\boldsymbol{z}+\boldsymbol{e}$.
3. Decode each test pattern into a codeword using HD decoder.
4. Compute the SD decoding metric for each generated candidate codeword.
5. Select the candidate codeword with the best metric as the decoded solution.

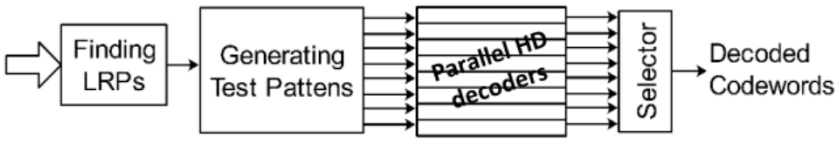


Figure 12. Block diagram of a Chase-like soft-decision (SD) decoder

There are different ways to compute SD decoding metric for each generated candidate codeword. As an example, we can add $1-|r_i - z_i|$ of selected LRPs in each test pattern and corresponding HD corrected position. The smaller the summation value, the more likely the generated codeword candidate is.

Figure 13 shows concatenated FEC performance with HD and SD inner code decodings for random error case $h_1=0$. We can see that SD decoding outperforms HD decoding. SD decoding can improve the slicer BER 2-3 orders of magnitude and provide more than 2 dB coding gain to meet 1e-13 post-FEC BER compared with KP-FEC only. In the rest part of the paper, only Chase algorithm-2 SD inner code decoding is used in the simulation results.

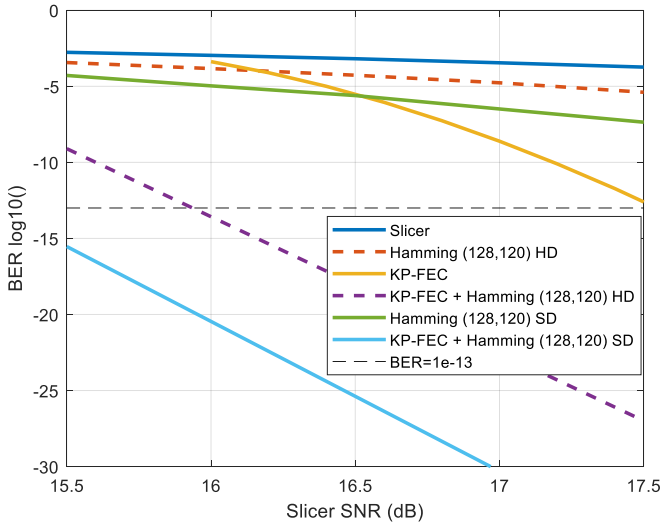


Figure 13. Concatenated FEC performance with HD and SD inner code decodings for h_1=0

In the previous section 3.2, we learned that an end-end FEC architecture with symbol multiplexing and 4-way codeword interleaving can tolerate certain DFE error propagation, say h_1=0.5. Here, we'll study the impact of DFE error propagation with a concatenated FEC architecture. Instead of the random error case as shown in Figure 13, we set simulation with 1-tap DFE with coefficient h_1 sweeping from 0 to 0.5. From Figure 14 we see that the proposed concatenated code with SD decoding has a performance degradation of 1.5dB when DFE coefficient h_1 increases to 0.5, which is much more severe than an end-end FEC. The main reason of such sensitivity to the DFE error propagation is that the inner Hamming code (128, 120) is relatively weak for burst error correcting capability and thus resulting in miss-corrections.

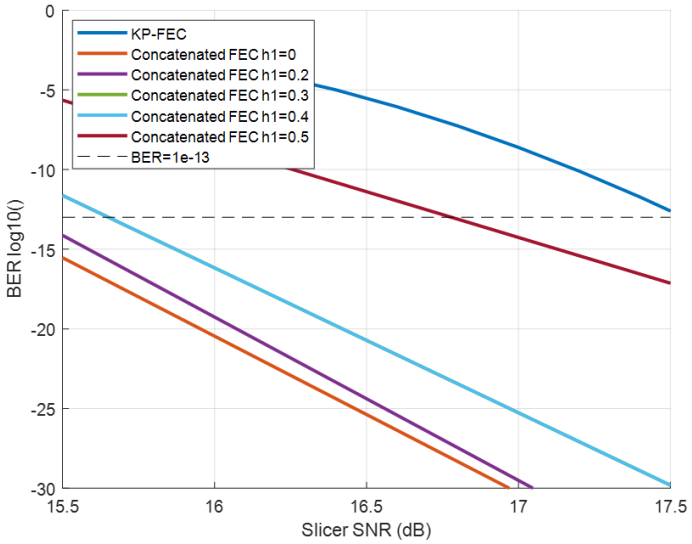


Figure 14. Concatenated FEC performance with SD inner code decoding for different h_1

In order to improve the concatenated FEC performance especially over burst channel errors, further interleaving schemes within the PMD inner code sublayer are proposed.

3.3.5 Inner code interleaving schemes

First, we can consider a block interleaver between the PMD channel and inner Hamming (128, 120) code. Figure 15 shows receiving and decoding of L interleaved inner codewords, $L \geq 1$. It simply arranges L Hamming inner codes into L rows of a rectangular block and then transmitting/receiving the block column by column. Even though the minimum distance of the interleaved block is still $d_{min} = 4$ as an individual Hamming (128, 120) code, this channel block interleaving can break a long burst PAM4 error into L different codewords.

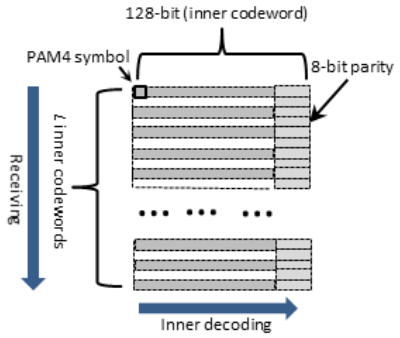


Figure 15. Hamming channel block interleaver

By doing this, we expect that the concatenated code can tolerate longer burst errors or more DFE error propagation. Figure 16 shows the concatenated FEC performance with channel block interleaving of L=1, 2 ... and up to 8 for h_1=0.5. We can see that with L>4 the channel interleaving improves concatenated coding gain about 1.5 dB over KP-FEC only.

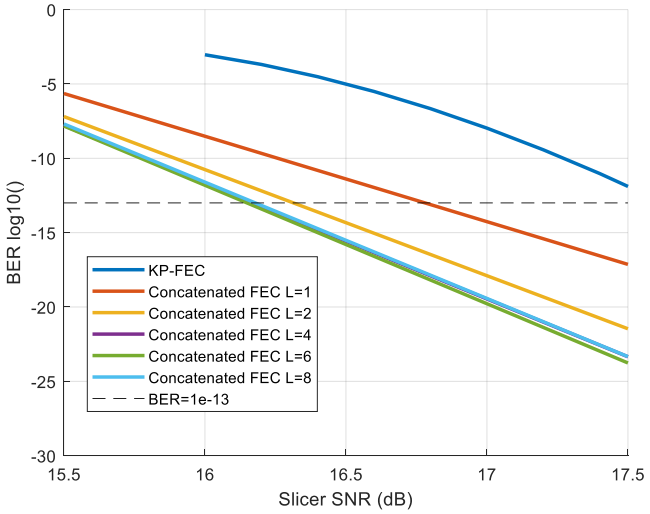


Figure 16 Concatenated FEC performance with channel block interleaving for h_1=0.5

To break the long burst errors and improve concatenated code performance, there are multiple interleaving schemes proposed in [7] for PMD inner code sublayer:

- Hamming interleaver L=8 to break the long burst errors into different inner codewords
- Circular shift block to maximize the distance in bauds between transmitted PAM4 symbols from two different RS symbols in the same RS (544, 514, 15) outer code
- Convolutional interleaver to guarantee that the 12x10 bit payload of the Hamming encoder comes from 12 distinct RS codewords

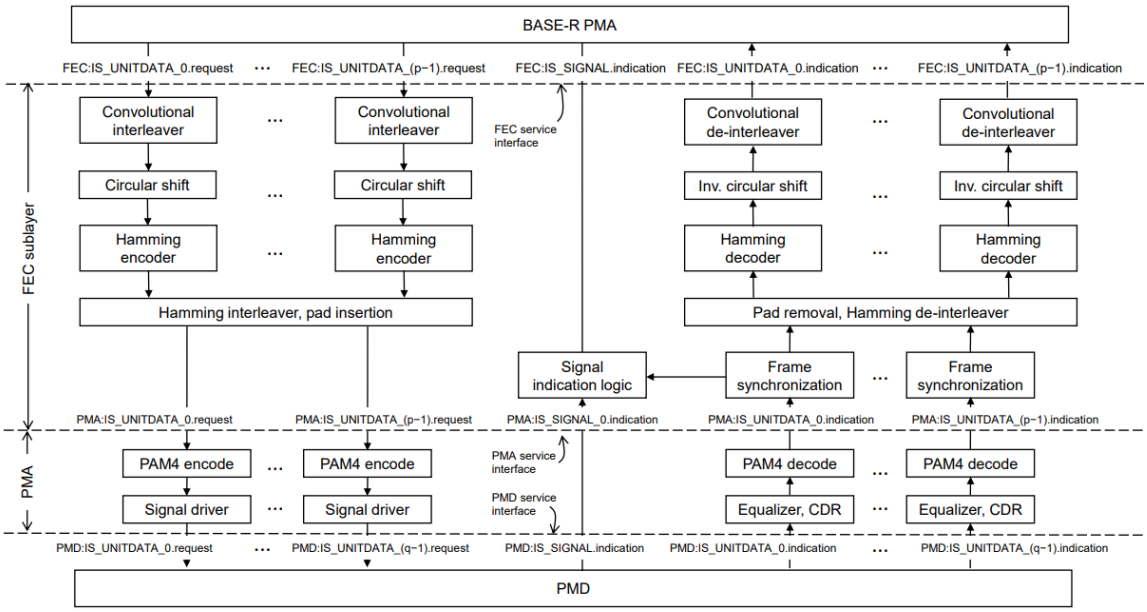


Figure 17. Block diagram of inner code sublayer

The proposal [7] describes the above three interleaving functions in detail, with the flow as shown in Figure 17. It is a good reference material before the 802.3dj force task publishes the final specifications. The proposal claims that the inner Hamming code (128, 120) with the above three interleaving functions could relax the PMD optical BER target from 2.4e-4 to 4.8e-3, more than one order of magnitude.

However, the major cost of adding the inner code is the latency, especially with the convolutional interleaver. The latency of the inner code itself including encoder and decoding is about 10ns, while the convolutional interleaver for 800GBASE-R with 4-way RS codeword interleaving case increases the latency to 56ns. For 800GBASE-R/400GBASE-R with 2-way RS codeword interleaving case the latency is further increased to 140ns.

3.3.6 Low latency PMD PHY

To cut the latency for shorter distances of PMD optical channels or high performance optical modules, two different FEC modes have been discussed and proposed to the 802.3dj task force:

- Mode_FECo: The optical link is run with the RS (544, 514, 15) FEC protection alone, the same as end-end FEC.

- Mode_FECi: The optical link is runs with the RS (544, 514, 15) FEC protection operating as an outer code, supplemented by a Hamming (128,120) code protection operating as an inner code.

Of course, the PMD BER target would be different for these two FEC modes provided that the error statistics are sufficiently random:

- Mode_FECo: The BER of the PMD link shall be less than 2.4e-4 when processed with an 800GBASE-R/1.6TBASE-R PCS.
- Mode_FECi: The BER of the PMD link shall be less than 4.8e-3 when processed with an 800GBASE-R/1.6TBASE-R PCS and an inner code sublayer.

Basically, we need two separate PHY specifications. One is associated with Mode_FECi for optical channels longer than 2km, and the other is associated with Mode_FECo for either short reach (say less than 500m) or co-packaged optic (CPO) and linear pluggable optic (LPO) types of interfaces.

3.4 Multi-part link BER allocations

A three-part link including two AUI (electrical) interfaces and one PMD (optical) link with two FEC modes is shown in Figure 18. As we know that for all 200GBASE-R, 400GBASE-R and 800GBASE-R with 100 Gbps per lane the overall input BER to the end-end FEC needs to be 2.8e-4 or lower, with the combination of 2.4e-4 BER target for PMD link and 1e-5 BER target for AUI interfaces, now for 200 Gbps per lane, how do we allocate PMD and AUI BER targets?

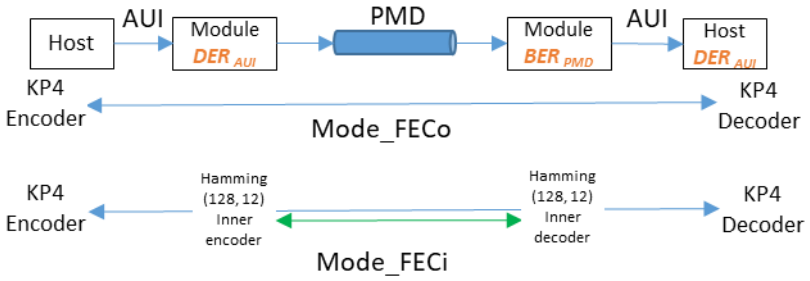


Figure 18. Three-part link including two AUIs and PMD with two FEC modes

For the final stage KP4 FEC performance,

$$CER = \sum_{i>15}^{544} SER_{RS}(i) \qquad \text{Eq. 5}$$

where $SER_{RS}(i)$ is the probability of having i RS symbols in error per RS codeword. To achieve post-FEC BER target of 1e-13 or FLR target of 6.2e-11, the CER needs to be less than FLR/4.125=1.5e-11 with 4-way codeword interleaving.

In the three-part link system shown in Figure 18, each part introduces RS symbol errors independently. Hence, the final $SER_{RS}(i)$ is the convolution of all three parts,

$$SER_{AUI}(i) = \sum_{j=0}^{i} SER_{AUI_1}(j) \cdot SER_{AUI_2}(i-j), \quad \text{Eq. 6}$$

$$SER_{AUI+PMD}(i) = \sum_{j=0}^{i} SER_{AUI}(j) \cdot SER_{PMD}(i-j). \quad \text{Eq. 7}$$

The $SER_{RS}(i)$ of each part can be obtained through Monte Carlo simulations, in which different channel error profiles (random or burst) with symbol multiplexing and 4-way codeword interleaving are applied.

Figure 19 shows the simulation results of the three-part link with different AUI decision error ratio (DER), error free and 2.67e-5 that have been potentially considered by the 802.3dj task force till July 2023 [15], note that the AUI DER is defined by channel operating margin (COM) and $DER=SER_{PAM4}/1.5$ for PAM4 signaling. To achieve an FLR target of 6.2e-11 (horizontal dash line) and provided that the PMD link error statistics are sufficiently random, the PMD BER target is:

- 2.8e-4 if there is no AUI part or AUI parts are error free
- 2.4e-4 for AUI DER=2.67e-5 with random error case (a=0)
- 2.2e-4 for AUI DER=2.67e-5 with burst error case (a=0.75)

It is noted that the PMD BER target is set for Mode_FECo or after inner decoding for Mode_FECi. Based on analysis like this, the 802.3dj task force will decide the final AUI DER and PMD BER allocations and targets.

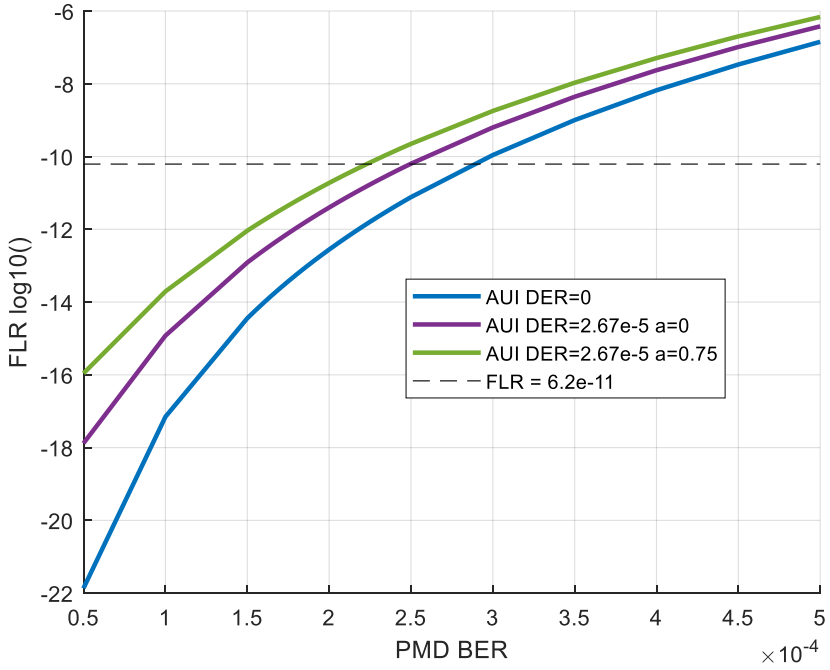


Figure 19. Three-part link FEC performance vs. PMD BERs with different AUI error profiles and DER targets

3.5 How FEC options affect the system signal integrity?

From the discussions and analysis we conclude that the different FEC options provide different coding gains with different levels of costs including coding overhead, coding latency and complexity. Those factors will affect system signal integrity through different aspects.

First of all, the FEC can relax the BER target. The larger coding gain the FEC provides, the lower BER target can be relaxed to. For example, [7] claims that the proposed concatenated code can relax the PHY's BER target from 2.4e-4 to 4.8e-3 for 200 Gbps. Table 3 lists the required slicer SNR for different BER targets for a PAM4 signaling format over AWGN channel. In other words, relaxing the BER target is equivalent to tolerating more noise, xtalk and jitter. Without increasing signal energy or equalization power, the proposed concatenated FEC can help us to tolerate 2.8 dB more noise, jitter, crosstalk or combined.

Table 3. Required slicer SNR for different BER targets for PAM4 over AWGN channel

BER targets	1e-13	1e-6	1e-5	1e-4	2.4e-4	4.8e-3
SNR (dB)	24.3	20.4	19.5	18.2	17.7	14.9

Secondly, strong interleaving schemes at the host PCS and/or the PMD inner code sublayers can effectively mitigate DFE error propagation, low frequency jitter, baseline wander, and other sources of correlated errors.

Of course, such high coding gain and long burst error tolerance of certain FEC options have an inevitable cost. The proposed concatenated FEC provides up to 2.8 dB coding gain to the PMD link but increases the link rate more than 6.7%, and adds more than 50ns latency. The increased link rate could introduce higher channel loss and xtalk significantly for a bandwidth limited channel over long reach copper cable. The increased latency could make the FEC option unsuitable for AI and ML applications in which low latency requirement is critical.

Last but not least, interoperability between different PHYs with different FEC options and FEC modes is highly desirable. The 802.3dj task force is developing the specifications to ensure such interoperability exists.

4. FEC options for long reach copper channels

In section 3, we discussed a few alternative FEC schemes to further improve coding gain, such as more powerful RS codes like RS (576, 514, 31) or longer RS codes.

RS codes with lower code rates than RS (544, 514, 15), or high overhead concatenated FECs are not suitable for long reach copper channels like cable backplanes or direct attached cables (DAC) due to their bandwidth limitation. However, the longer RS codes with the same overhead as RS (544, 514, 15), such as the four longer RS codes with $n > 1000$ in Table 2 can be considered to get better coding gain without further bandwidth expansion, of course with the cost of larger latency. Figure 20 depicts the trade-off between latency and coding gain for different RS codes for random error case ($a = 0$). We see that RS (1088, 1028, 30) code can provide 0.5 dB coding

gain over RS (544,514, 15) code, with 18ns more latency. RS (4088, 3855, 112) code can provide even larger coding gain, about 1.5dB, but with 200ns more latency.

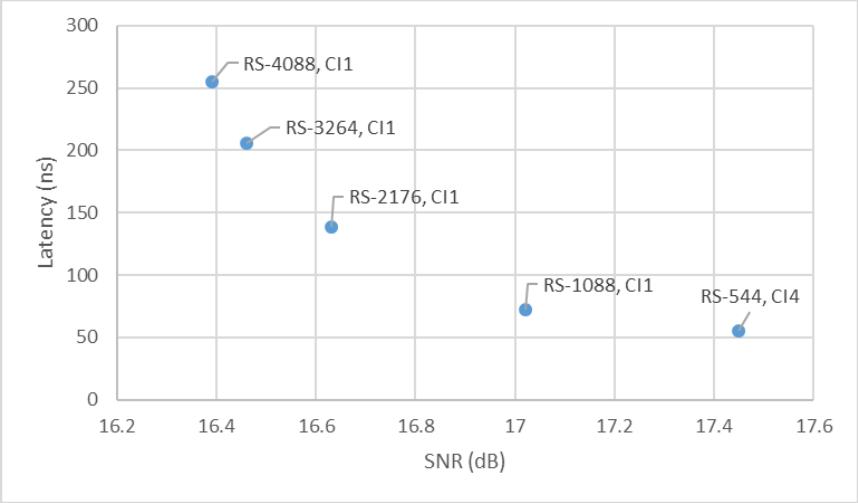


Figure 20. Latency vs. coding gain with different RS codes

For low latency applications, SD decoding of the RS (544, 514, 15) code could be alternatively considered as well. For instance, the erasure decoding algorithm described in [9] can be applied to RS codes. For a *t*-error correcting RS code, the erasure decoding can correct all combinations of *v* symbol errors and *e* symbol erasures provided that the inequality

$$v+e/2 \leq t \qquad \text{Eq. 8}$$

holds. For RS (544, 514, 15) code with t =15, erasure decoding can erase up to 30 symbols per codeword which can improve the coding gain over HD decoding. Because the erased positions are known, erasure decoding only needs to find the values of the erased *e* symbols. Hence, there is small increment of decoding complexity over HD decoding. The erasure positions can be associated with LRPs similar to Chase algorithms for inner code decoding. Such erasure decoding algorithms require relatively lower latency than longer RS codes with relatively small incremental implementation complexities due to the soft information generation, routing and processing.

5. Advanced FEC options for 400 Gbps per lane

In order to increase the link rate from 200 Gbps to 400 Gbps, the bandwidth requirement will be continuously doubled if we stay with the current PAM4 modulation scheme. Based on the feasibility study of the existing analog-front-end (AFE) bandwidth and ADC sampling rate, PAM8 or higher modulation level is preferred.

Table 4 lists key parameters for different PAM schemes, such as the number of bits per symbol, signaling rate, unit interval, fundamental frequency, and SNR penalty. We can see that

increasing the modulation levels reduces the bandwidth requirement and the sampling rate, but with the cost of higher SNR penalty and sensitivity to noise and jitter.

Table 4. Key parameters for different PAM schemes for 400+ Gbps data rates

Data rate, Gbps	**212.5**	**425**					
Number of PAM levels	**4**	**4**	**5**	**6**	**7**	**8**	**16**
Bits per symbol *	2	2	2.25	2.5	2.75	3	4
Signaling rate, Gbaud	106.25	212.50	188.89	170.00	154.55	141.67	106.25
Unit interval, ps	9.41	4.71	5.29	5.88	6.47	7.06	9.41
Fundamental frequency, GHz	53.13	106.25	94.44	85.00	77.27	70.83	53.13
Required SNR at slicer, dB **	18.23	18.23	20.21	21.81	23.15	24.30	30.23
SNR penalty, dB	0	0	1.99	3.59	4.92	6.08	12.01

* Assumes an efficient mapping of bits to PAM symbols.
** For BER = 1e−4

To offset the SNR penalty introduced by higher modulation levels, more advanced FECs can be used. However, traditional FECs alone can hardly compensate for the 6dB SNR penalty from PAM4 to PAM8 modulation expansion. Furthermore, their FEC coding gains are normally achieved at the expense of larger overhead which increases the bandwidth requirement even more.

For bandwidth limited channels, coding schemes combined with modulation techniques could be considered. One such coding scheme is called coded modulation, which achieves its coding gain without bandwidth increasing. Trellis coded modulation (TCM) [12] is one of such techniques. TCM has been adopted in existing telecom DSL/ADSL technology.

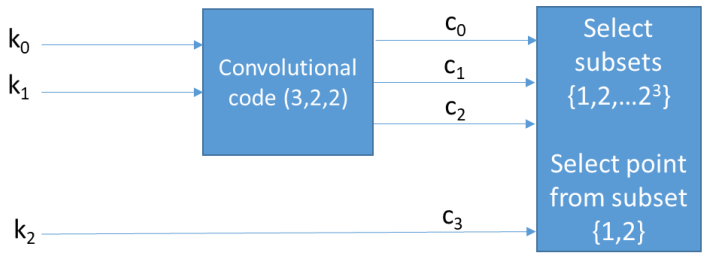


Figure 21. Example of 4-state rate 2/3 trellis-coded PAM8

Figure 21 shows an example of 4-state rate 2/3 trellis-coded PAM8 signals. Here, three message bits are split into two parts. The first 2 bits (k_0 and k_1) are convolutional encoded into 3 bits (c_0, c_1, and c_2) and the other bit (k_2) is uncoded to c_3. At the last stage, the four bits (c_0, c_1, c_2 and c_3) are mapped to a PAM16 symbol. By doing this, a sequence of PAM8 symbols are coded to a sequence of PAM16 symbols. The baud rate remains at 141.67 Gbaud for 425 Gbps link rate.

The decoding of a coded modulation is normally trellis based, such as the Viterbi algorithm. The convolutional code (3, 2, 2) defines the trellis diagram with 4 states and the k_2 bit creates parallel transitions in the trellis diagram. The coding gain of a TCM over uncoded PAM-8 is determined by the Euclidean distance [9] between the TCM coded and uncoded PAM-8.

A set of optimum TCM codes based on the foregoing search procedure are listed in Table 1 in [12]. The codes were bounded by computer search. We can see that the TCM coding gain increases with:

- Larger number of states
- Higher modulation levels expansion
- Lower code rate of convolutional code

As an example, we apply trellis coded modulation on PAM8 signaling, by good set partitioning and good convolutional code selection which could result in a coding gain of about 3.5dB with 4-state, close to 4 dB with 8-state, and up to 6 dB with 128-state. Such coding gain could be useful to compensate for the 6 dB SNR penalty from PAM4 to PAM8 modulation expansion while the channel bandwidth requirement is relaxed from 212.5 Gbaud to 141.67 Gbaud. Of course, the cost of the TCM is the complexity of the receiver which requires Viterbi trellis decoding.

As discussed in [13], other high coding gain and Shannon limit approaching FEC codes, such as low density parity check (LDPC) codes and Turbo product codes (TPC), have already applied to long-distance optical transmission systems [14]. For next-generation 400 Gbps link rate, we could leverage optical system work by deploying higher coding gain FECs if needed, but with the cost of coding overhead, latency and complexity.

6. Conclusions

In this paper, channel error models and FEC performance analysis have been updated. Different Ethernet coding schemes have been studied and simulated for 800GE and 1.6GE systems with 200 Gbps per lane. Concatenated FEC with SD decoding for inner code to protect 200 Gbps optical link is investigated. How different FEC options affect system signal integrity is discussed as well. Advanced FEC schemes for long copper channels and 400 Gbps per lane are further explored.

Reference:

[1] IEEE Std 802.3bs-2017 IEEE Standard for Ethernet: Media Access Control Parameters, Physical Layers and Management Parameters for 200 Gb/s and 400 Gb/s Operation

[2] IEEE Std 802.3ck-2022 IEEE Standard for Ethernet: Physical Layer Specifications and Management Parameters for 100 Gb/s, 200 Gb/s, and 400 Gb/s Electrical Interfaces Based on 100 Gb/s Signaling

[3] IEEE P802.3df 400 Gb/s and 800 Gb/s Ethernet Task Force: https://www.ieee802.org/3/df/index.html

[4] IEEE P802.3dj 200 Gb/s, 400 Gb/s, 800 Gb/s, and 1.6 Tb/s Ethernet Task Force: https://www.ieee802.org/3/dj/index.html

[5] Cathy Liu, “What is FEC, and How Do I Use It?” Signal Integrity Journal, https://www.signalintegrityjournal.com/articles/1284-what-is-fec-and-how-do-i-use-it

[6] Cathy Liu, “100+ Gb/s Ethernet Forward Error Correction (FEC) Analysis,” DesignCon 2019.

[7] FEC baseline proposal for 200 Gbps per Lane IM-DD Optical PMDs, https://www.ieee802.org/3/dj/public/23_03/patra_3dj_01b_2303.pdf

[8] BER objective format for 400GbE, https://www.ieee802.org/3/400GSG/public/adhoc/logic/jun26_13/anslow_01_0613_logic.pdf#page=9

[9] S. Lin and D. Costello, *Error Control Coding,* Prentice Hall, February 2002.

[10] D. Chase, “A Class of Algorithms for Decoding Block Codes with Channel Measurement Information”, IEEE Trans. on information theory, Vol. IT-18, NO. 1, January 1972.

[11] Proposal for a specific (128,120) extended inner Hamming Code, https://www.ieee802.org/3/df/public/22_10/22_1005/bliss_3df_01_220929.pdf

[12] G. Ungerboeck, “Trellis-coded modulation with redundant signal sets Part II: State of the art”, IEEE communications magazine, Vol. 25, No 2, February 1987.

[13] Cathy Liu, “Salz SNR & Shannon Limit Study for the Next Speed Node Beyond 112Gbps (and up to 224Gbps)”, DesignCon 2021.

[14] G. Tzimpragos, and et. al., “A Survey on FEC Codes for 100G and Beyond Optical Networks”, IEEE communications surveys & tutorials, Vol. 18, Issue 1, 2016.

[15] Consensus proposal for AUI error requirements, https://www.ieee802.org/3/dj/public/23_05/ran_3dj_02_2305.pdf