ACCEPTED FROM OPEN CALL

Next Generation Backplane and Copper Cable Challenges

Roy D. Cideciyan, IBM
Mark Gustlin, Xilinx
Mike Peng Li, Altera
John Wang and Zhongfeng Wang, Broadcom

ABSTRACT

This article provides an overview of some of the work that is ongoing in the IEEE P802.3bj 100 Gb/s Backplane and Copper Cable Task Force. The task force is standardizing Ethernet at 100 Gb/s over a 4-lane backplane channel as well as across a 4-lane copper cable. We first describe the background of the market drivers for this technology, and then give an overview of the various technologies that are used to solve the challenging problems of running across the various mediums at a data rate of 25 Gb/s. Also discussed are the details of the forward error correction, transcoding and physical layer coding that are employed to achieve robust links.

INTRODUCTION

The market drivers for a 100 Gb/s backplane standard include an increase in front-panel I/O densities enabled by smaller optical form factors (CFP2 , SFP+, QSFP+, etc.), relentless increases in server-bandwidth capabilities and the related increase in server connection speeds. Second-generation blade servers are adopting 40GBASE-KR4, which was standardized in 2010, whereas third-generation blade servers will need another increase in speed using the same 4-lane backplane connections to keep up with higher server I/O bandwidths. The current 100GBASE-CR10 standard requires 20 twin-axial cables, 10 × 10 Gb/s in each direction. A 4 × 25 Gb/s interface reduces the number of twin-axial cables to 8. This will allow higher density front panels and also reduce the cost of an interface.

The 100 Gb/s Backplane and Copper Cable Study Group spent 10 months studying the feasibility of a 4-lane 100 Gb/s copper backplane and cable interface. What emerged from the subsequent task force were the following objectives:

- Define a 4-lane 100 Gb/s PHY for operation over links consistent with copper twin-axial cables with lengths up to at least 5m
- Define a 4-lane PHY for operation over a printed circuit board backplane with a total channel insertion loss of <= 35 dB at 12.9 GHz (this is consistent with 1 m of high-quality PCB and two connectors)
- Define a 4-lane PHY for operation over a printed circuit board (PCB) backplane with a total channel insertion loss of <= 33 dB at 7.0 GHz (this is consistent with 1 m of medium-quality PCB material and two connectors)

The study group decided to set two 100 Gb/s backplane objectives, one for channels that have an insertion loss of 35dB at 12.9 GHz, which is assumed to use non-return-to-zero (NRZ) signaling, and the other for 33dB at 7.0 GHz, which is assumed to use pulse amplitude modulation-4 (PAM-4). Why two PHYs? Simply put, there are two unique markets and design spaces [1]. The NRZ PHY is specified for new backplane designs in high-end networking and high-performance computing. These designs specify high-end and low-loss backplane materials that are able to support NRZ signaling for up to 1 m at a signaling rate of 25 Gb/s per lane. The PAM-4 PHY is designed to accommodate legacy channels or channels made from lower-performance materials. The two 100 Gb/s Ethernet backplane PHYs allow Ethernet to serve a broad market.

The backplane channels at these signaling rates are challenging environments to run at low bit-error ratios (BERs) so forward error correction (FEC) is also employed to improve the BER and to allow the PHYs to run across very lossy channels. The channels that are defined for 100 Gb/s backplane and copper cable across four lanes require a stronger FEC code with higher coding gain than defined within the original 802.3ba standard. In addition there is a strong desire for very low latency with an unofficial goal of 100 ns of added latency.

After investigating the FEC options and exploring triple tradeoffs between latency, coding gain and complexity, two FEC codes were chosen, one for the NRZ PHY, the other for the PAM-4 PHY. Reed–Solomon (RS) codes were chosen because they have both good random-error and burst-error correction capability. To

0163-6804/13/$25.00 © 2013 IEEE

reduce the latency to within the target goal of 100 ns, it was decided to stripe the FEC code across the four lanes of a given PHY. On the receive side, prior to FEC decoding, the four lanes are aligned by reusing the alignment markers which are part of a 100 Gb/s data stream.

Also a goal was to minimize the over-speed at which a given lane is run to add FEC to the protocol. With that in mind, the decision was made to transcode the data from 64b/66b-encoded data to 256b/257b-encoded data. This aggressive transcoding allows the NRZ PHY to add an RS FEC that has a net coding gain of ~5 dB, and to run without over-speed (the signaling rate still is 25.78125 Gb/s). The PAM-4 PHY runs with the same transcoding and has a net coding gain of ~5.4 dB when running over a very challenging channel and requiring ~5.45 percent over-clocking.

The NRZ and the PAM-4 PHY have most of their processing in common. Both take 20 PCS lanes, transcode the data from 64b/66b to 256b/257b, FEC encode the data, and then stripe the FEC-encoded data to four physical lanes 10 bits at a time. The biggest difference in the PHY architectures is that the NRZ PHY uses an RS(528, 514, $t = 7$, $m = 10$) code with 10-bit symbols ($m = 10$) that has an error-correction capability of $t = 7$ (seven 10-bit symbols can be corrected), whereas the PAM-4 PHY uses an RS(544, 514, $t = 15$, $m = 10$) code with error correction capability of $t = 15$.

The remainder of this paper will describe the FEC and PMD sublayers in detail. Table 1 has the PHY types. When this paper refers to all three PHYs together, the term 802.3bj PHYs will be used.

PHY Name	Medium Type	Signaling
100GBASE-KR4	High Quality PCB	NRZ
100GBASE-KP4	Standard PCB	PAM-4
100GBASE-CR4	Twin-axial cable	NRZ

Table 1. *PHY types.*

TRANSCODING

In IEEE 802.3 Ethernet, 64b/66b coding is performed in the physical coding sublayer (PCS) of the PHY. The 64b/66b code has an overhead of two bits, and the 64-bit payload is randomized by means of a self-synchronizing scrambler characterized by the polynomial $1 + x^{39} + x^{58}$. The incoming 64-bit payloads are associated with either a data block or a control block. Therefore, two header bits, 01 or 10, are appended at the beginning of the payload to indicate the type of the 66-bit block.

Data blocks and control blocks start with the header bits 01 and 10, respectively. The 64-bit content of a data block is eight data bytes denoted D_0, D_1, ..., D_7. The first byte of the 64-bit content of the control block is the block-type field (BTF) indicating the type of a control block. The relationship between the first and the second nibble of BTF can be characterized by a one-to-one mapping. Note that there is only one type of control block indicating the start of a frame because a frame always starts at the first byte of a block, whereas there are eight types of control blocks indicating frame termination depending on the eight possible locations of the last byte of a frame.

Transcoding refers to translating a data format, such as the 64b/66b-coded format, to a new more compact data format that is used for transmission; it is typically used to enable FEC to be added to the data stream without increasing the data rate. Transcoding achieves a relatively low compression ratio and is usually performed prior to FEC coding, which inserts redundancy into transmitted data to correct errors that occur during data detection at the receiver. In [2], 512b/513b transcoding in conjunction with RS coding for 100 Gb/s NRZ backplane while keeping the line rate at (66/64) × 25 Gb/s = 25.78125 Gb/s, i.e., 0 percent over-clocking, was proposed. In 100GBASE-KR4 and 100GBASE-CR4, a similar approach has been adopted to transmit 256b/257b-transcoded and RS-encoded data at 25.78125 Gb/s in each of the four physical lanes. Although the transcoding scheme is common to all 802.3bj PHYs, the PAM-4 transmission rate of 27.1875 Gb/s is 5.45 percent higher than the transmission rate of 100GBASE-KR4 and 100GBASE-CR4 because of the larger overhead required to withstand the higher insertion loss of the 100GBASE-KP4 channels.

Fixed-rate transcoding schemes convert N incoming 66-bit blocks each with 64-bit payload, into a single $(N \times 64 + L)$-bit block, where the compression ratio is $(N \times 66)/(N \times 64 + L)$. Conventional fixed-rate transcoding schemes collect N consecutive 66-bit data and control blocks and check whether there is at least one control block among the incoming N blocks. If all N blocks are data blocks, then an L-bit header is added to the N 64-bit data payloads to form the transcoded block. If there is at least one control block among the N blocks, the 64-bit contents of the blocks are then reshuffled by moving all control blocks to the start of the transcoded block and the data blocks to the end. The block-type field of every control block in the transcoded block is then mapped into a byte that contains a 3-bit field specifying the original position of the control block in the N blocks, a 4-bit field indicating the type of the control block, and a 1-bit flag bit pointing to the type of the next 64-bit block in the transcoded block (e.g., control:0 and data:1). Several $M \times 512b/(M \times 512 + L)b$ transcoding schemes, which rearrange the order of the incoming blocks, have been designed for $1 \le M \le 4$ and $1 \le L \le 3$. Note however that the latency, the implementation complexity and the power dissipation associated with these schemes are relatively high.

The 802.3bj FEC sublayer architecture is required to enable an implementation with less than 100 ns increase in total latency due to FEC processing at Tx and Rx. Latency is a parameter that is of critical importance in link design. Consequently, it was decided that transcoding is performed at the rate of 100 Gb/s as in the case of FEC encoding, i.e., not on a per-physical-lane

There is only one type of control block indicating the start of a frame because a frame always starts at the first byte of a block, whereas there are eight types of control blocks indicating frame termination depending on the eight possible locations of the last byte of a frame.

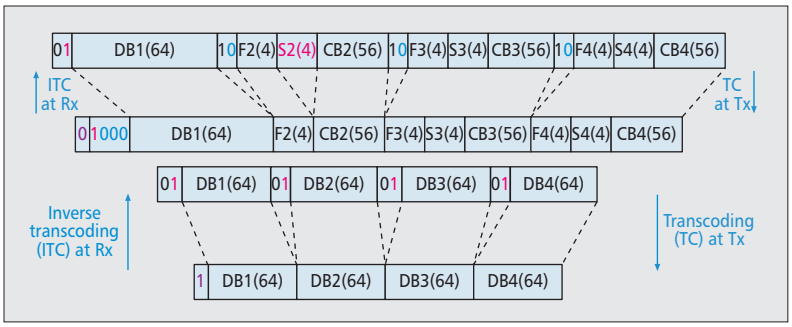

Figure 1. *256b/257b transcoding.*

basis. Therefore, the latency associated with $(N \times 66)/(N \times 64 + L)$ transcoding is about $(N/8) \times 10$ ns, i.e., about 10 ns for 512b/513b and 512b/514b transcoding ($N = 8$) corresponding to 10 percent of the maximum FEC latency budget. As both latency and compression ratio increase with N and a low latency and a high compression ratio are desirable, N should be selected only after careful consideration of the trade-off between low latency and high compression ratio. Another important consideration for the selection of N is the implementation complexity involved in the processing of alignment markers (AMs) in the FEC sublayer. 66-bit AMs are periodically inserted at the start of an RS code word once every 4096 RS code words to provide the means for FEC block and transcoded block synchronization at the receiver because an AM period contains 4096 RS code words and an RS code word contains 20 transcoded blocks. As there are 20 PCS lanes containing one AM every 2^{14} 66-bit blocks each, the implementation complexity of AM processing will be low if 20 is divisible by N. Therefore, it was decided that the payload granularity associated with transcoding shall be 256b, corresponding to $N = 4$, which provides an optimum trade-off between low latency and high compression ratio in conjunction with maximum FEC capability while enabling a total FEC and transcoding latency of less than 100 ns.

The 256b/257b transcoding scheme is described in [3]. Transcoding is performed without reshuffling the order of the four incoming 66-bit blocks. 256b/257b reshuffling-free transcoding reduces the latency, power dissipation and implementation complexity of the FEC sublayer when compared to 512b/514b conventional transcoding [3]. If all four incoming blocks are data blocks, 256b/257b transcoding is performed by stripping off the 2-bit headers of all four 66-bit data blocks and appending a header bit of 1 to the four 64-bit data payloads, as shown in Fig. 1. As in conventional transcoding, there is no reshuffling of the order of the data blocks after transcoding. If there is at least one control block among the four incoming blocks, 256b/257b transcoding is performed by first stripping off the 2-bit headers of all four incoming 66-bit blocks and appending a header bit of 0 to the four payloads of the four blocks. In the second step, the second 4-bit nibble in the BTF of the first control block in a transcoded block is deleted, whereas the first 4-bit nibble the BTF of the first control block, indicating the type of the first control block, is kept without change. In the final step, a 4-bit header $x_1\, x_2\, x_3\, x_4$ following the overall header bit 0 is inserted, where $x_i = 0$ indicates that the i-th block is a control block CBi and $x_i = 1$ indicates that the i-th block is a data block DBi. Figure 1 illustrates the situation when the data block DB1 arrives first and is followed by three control blocks, CB2, CB3, and CB4.

Forward Error Correction

Various FEC codes have been used in modern digital communication systems. Among them, turbo codes and low-density parity-check (LDPC) codes are the two most promising error-correction codes. Both can approach the Shannon limit with iterative soft-decision decoding. For 100 Gb/s backplane and copper-cable applications, both latency and power consumption are critical. Thus codes with soft-decision decoding are not very well suited. Product codes, such as CI-BCH codes and SP-BCH codes [4], can achieve outstanding decoding performance with hard-decision decoding. However, the block size has to be very large when the code rate is high. For example, the net coding gain of either product code in [4] is larger than 9.3 dB at a target BER of 10^{-15}, whereas the code block size is close to 1 million bits. Therefore it seems we are left with simple block codes, including RS, BCH, and Fire codes, for this application.

A Fire code was used for 10GBASE-KR PHYs. However, it is not well suited for 802.3bj PHYs because of its limited error-correcting capability (its gain is ~ 2 dB, and a much higher coding gain is needed for these PHYs). For a block code with the same redundancy ratio and similar block size, BCH codes normally outperform RS codes in terms of net coding gain for an additive white Gaussian noise (AWGN) channel. However, error propagation is a major concern in choosing an FEC code for the 10GBASE-KR PHY. At the receiver side, with a 1-tap decision feedback equalizer (DFE), the probability of error propagation can be up to 0.5 with NRZ modulation. It can reach 0.75 with PAM-4 modulation. As the 802.3bj PHYs will have similar error propagation issues, the FEC code best suited for these channels is an RS code.

To choose an optimal block code, specifically in our case a kind of RS code, we have taken into account various tradeoffs such as clocking rate, coding gain, hardware complexity, and FEC latency. In general, the higher the clocking rate (i.e., the more redundancy to be added), the more coding gain can be achieved, up to a certain bound on the clocking rate. The larger the block size, the higher the coding gain, but also the higher the processing latency. More parallelism can reduce processing latency, but will increase hardware complexity. For a detailed analysis of these tradeoffs, the interested reader is referred to [5].

The overall latency associated with FEC processing is a major aspect when using FEC in the 802.3bj PHYs. The lower the latency, the better it is from the application's point of view. However, a small latency not only limits the block size

of the FEC code, which in turn will limit the performance of the code, but may also impact the decoder complexity. In [5], a total latency of 100 ns was suggested as the upper bound, and this goal was informally agreed upon by the standard body as the target latency (the most demanding applications want to minimize the latency, and 100 ns was seen as a compromise that can meet most applications' needs and enable a reasonable FEC implementation). The target net coding gain for the 802.3bj PHY's FEC was set at 5dB, and the overall hardware was expected to be less than 300K ASIC equivalent gates (or, an area smaller than 0.1 mm^2 in 28-nm CMOS).

Several RS codes were proposed for the 802.3bj PHYs, many that require over-clocking and some that don't (see e.g. [5, 6]). Among the various candidate FEC codes, RS(528, 514, t = 7, m = 10) based on 512b/514b transcoding presented in [7] attracted much interest. This code achieves the largest coding gain of all candidate codes that do not require over-clocking and that meet the latency constraint [6]. One FEC block consists of 10 transcoded blocks. Later in the project, the transcoding block size was halved to 256b/257b transcoding [8]. This change facilitates the processing of AM blocks and also helps reduce latency in the transcoding stage. The overall processing latency associated with this code was estimated to be less than 95 ns with an example implementation in 40-nm CMOS ASIC.

The effective coding gain under burst-error condition was estimated to be close to 5 dB. In brief, this code meets all target requirements discussed above. In addition, it has some nice properties. First of all, the payload size is exactly eighty 66-bit blocks or exactly 20 transcoded blocks. Second, the total number of coded bits per block is divisible by 40 (5280/40 = 132). This ensures that each physical lane will get the same number of coded RS symbols per FEC block. Third, an RS code defined over the Galois field (GF)(2^{10}) is power efficient because its trinomial field generator polynomial leads to low-complexity multiplication in the Galois field. Most significantly, when using the RS(528, 514, t = 7) code, no over-clocking for the line data rate is necessary when combined with 256b/257b transcoding.

An in-depth discussion of data striping was presented in [8]. We first take four 66-bit blocks from four PCS lanes, and use 256b/257b transcoding to generate 257 compressed bits. Then we use a gearbox[1] to convert these 257 bits to a multiple of 40 bits, e.g., 160 bits or 240 bits. After RS encoding, we distribute the encoded data 10 bits per physical lane (PL) as shown in Fig. 2, where S_i (i = 0, 1, 2, …) denotes the source symbol sequence, and P_i (i = 0, 1, 2, …) denotes the parity symbol sequence. In this way, we ensure that a single burst of up to 11 bit (or $n \times 10 + 1$ in general) will not cause more than two (or $n + 1$ in general) RS symbol errors, which will limit the performance degradation of the RS code in burst channels.

To increase the commonality of the two PHYs used for two different modulation schemes, the Standard body adopted the transcoding scheme of 100GBASE-KR4/CR4 also for 100GBASE-KP4. Considering the significantly higher SNR loss PAM-4 modulation incurs compared with NRZ modulation, the FEC code targeting the 100GBASE-KP4 system has to be much stronger than the FEC code for 100GBASE-KR4/CR4. The task force chose RS(544, 514, t = 15, m = 10) [9] based on several practical considerations. The overall processing latency associated with this code can be close to 100 ns at the cost of additional complexity. Its payload size is exactly the same as that in the 100GBASE-KR4 case. Also, a bus width of 160 bit can be chosen for the codec of this code, which can maximize the hardware sharing between two different (RS(t = 7) and RS(t = 15)) decoders. Note that 5440/160 = 34, i.e., we need exactly 34 cycles to receive an FEC block, which will simplify the decoder design, specifically in parallel syndrome computation and parallel Chien [10] search. In addition, the over-clocking rate for introducing such a strong FEC is only 3 percent. The effective coding gain of this FEC code under burst-error conditions is 5.4 dB. The data striping for this code is also shown in Fig. 2.

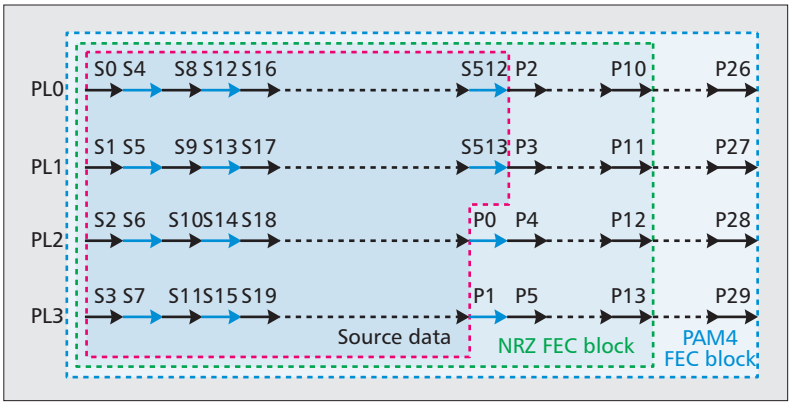

Figure 2. *FEC lane striping.*

NRZ PHY

One of the project objectives is to design a backplane PHY for NRZ signaling that supports channels with an insertion loss of 35 dB at 12.9 GHz. This section focuses on the 4×25G backplane market requirements and associated NRZ (or PAM2) transceiver signaling capability (also called 100GBASE-KR4). Note that the same PHY specification is used for driving the copper cable, called 100GBASE-CR4. The insertion loss (IL) for the backplane is material dependent. For improved FR-4, a typical IL is 1.04 dB/in at 12.9 GHz (Nyquist of 25.75 Gb/s). For Megtron-6 material, a typical IL is 0.68 dB/in. For comparison, we list a "better material" whose loss is 0.95 dB/in, i.e., slightly better than the improved FR-4. Assuming that the maximum IL allowed at Nyquist is 35 dB, we can calculate the backplane chip-to-chip distances, and the results are listed in Table 2.

Table 2 indicates that for a link IL budget of 35 dB, our "better material" can support a distance of up to 34 inches and Megtron-6 a distance of up to 48 inches. For a ~30 dB link budget, Megtron-6 can support distances of up to ~40 inches. Note that when also manufacturing process variations, including surface rough-

[1] *A gearbox is a shift function to switch between one bit width to another*

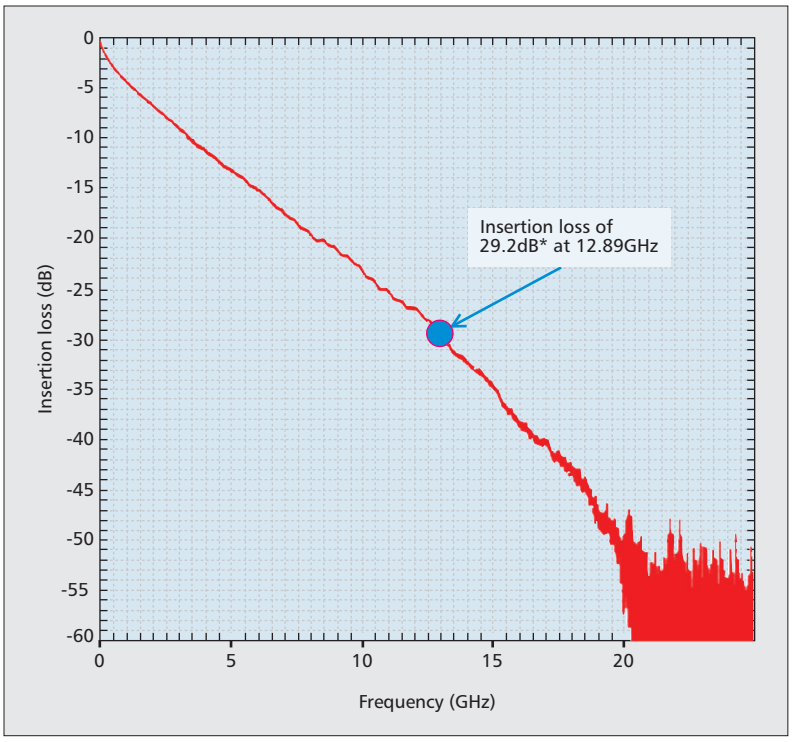

Figure 3. *IL for a 1-m 25 Gb/s backplane.*

ness, are taken into account, the loss per inch could of course be better or worse than what we have discussed.

The capability of the NRZ transceiver is evaluated with practical backplanes via collaborative simulations. This process provides the necessary calibration and validation of products from various transceiver providers to ensure accuracy and confidence.

A major task of the task force was to perform feasibility simulation studies for a 1-m 25 Gb/s backplane for server application provided by a server backplane manufacturer [11]. This backplane has an IL of 29.2 dB at 12.89 GHz and is shown in Fig. 3. Other channel characteristics, such as return loss (RL), insertion loss deviation (ILD), and insertion loss to crosstalk ratio (ICR) can be found in [11]. Note the frequency of interest is at the Nyquist frequency that corresponds to the average power level of the signal and is at half of the baud rate for NRZ signaling.

The simulation for this 1-m backplane was conducted by four 25G NRZ backplane transceiver providers [11], and consistent results were obtained. Simulations used a 3–4 tap (1-2 pre-tap, 1–2 post tap) transmitter feed-forward equalizer (FFE), a 12–20 dB gain continuous time linear equalizer (CTLE) or equivalent, and 5–12 taps of DFE. The simulation results suggest that the NRZ transceiver can support 29.2 dB/1 m server backplane, with > 29 mV eye-height (EH), and 24.7 percent eye-width (EW) at 10^{-12} BER without the need of FEC.

Further simulation studies for backplanes with loss > 30 dB have been conducted, and the results can be found in [12]. Note that in those simulations FEC was included and that FEC is currently mandatory within the draft standard.

Simulations have shown that for backplanes with an IL > 30 dB, FEC is needed to maintain a 15 mV EH and 15 percent EW margin at 10-12 BER. Also, with the FEC, the transceiver can support backplanes with an IL of 35 dB with > 35 mV EH and 35 percent EW margin as well as backplanes with an IL of 40 dB with > 15 mV EH and 15 percent EW margin, both at 10^{-12} BER. All simulations have both random and DFE-induced burst errors. The minimum FEC for achieving the minimum EH and EW margins for a 40-dB IL backplane is RS (352, 342, t = 5, m = 12). The strongest FEC achieving the minimum EH and EW margins for a 45 dB IL backplane is RS (248, 228, t = 10, m = 9), where t is number of correctable symbols and m the symbol size in bits for the RS FEC.

For the transmitter, the differential peak-to-peak voltage max, the output waveform steady-state voltage minimum and the maximum are affected by the presets of a 3-tap FFE and can be obtained via pulse- and equalizer-based linear-fitting. Common-mode AC and DC voltages, differential and common mode return losses (RLs), far-end output noise, minimum rise/fall times, and various jitters (DCD, RJ, and TJ

	Measured Megtron6	Improved FR4	"Better Material"	Improved FR4	"Better Material"	Measured Megtron6
Loss Budget (dB)	29.2	30	30	35	35	35
Loss of 2 Connectors (dB)	1.3	1.3	1.3	1.3	1.3	1.3
Loss from 6Vias (dB)	1.2	1.2	1.2	1.2	1.2	1.2
Loss Budget for Traces (dB)	26.7	27.5	27.5	32.5	32.5	32.5
Material Loss (dB/inch)	0.68	1.04	0.95	1.04	0.95	0.68
Trace Distance (in)	39.26	26.44	28.95	31.25	34.21	47.79
Trace Distance (m)	1.00	0.67	0.74	0.79	0.87	1.21

Table 2. *Maximum backplane distance.*

Modulation	PAM-2	PAM-3	PAM-4	PAM-5	PAM-6	PAM-7
Baud rate (GHz)	25.8	15.2	12.9	11.5	10.3	9.4
Salz SNR (dB)	15.8	23.6	29.2	30.3	31.1	31.9
Required SNR (dB)	15.5	19.8	22.5	24.5	26.2	27.5
SNR Margin (dB)	0.3	3.8	6.7	5.8	5.9	4.5

Table 3. *SNR margin for different choices of PAM-N.*

One of the objectives of the IEEE 802.3bj Standard is to define a four-lane 100 Gb/s PHY that can operate over backplanes that can currently transfer data at 10 Gb/s over a single lane using the IEEE 802.3ap standard (10GBASE-KR).

(with DDJ excluded)) will be specified at the package pin/ball.

The channel specification method we used specifies the insertion-loss deviation (ILD) derived from channel IL model vs. IL measurement fitting, integrated crosstalk noise (ICN), and various channel RLs (differential, common-mode, and differential-to-common mode). More recently, a time-domain channel specification method was proposed and adopted by the task force that uses a single figure of merit (FOM), called channel operation margin (COM, see [13]), to qualify a channel. The time-domain COM method assumes a reference transmitter and receiver, and comprises all channel impairments, such as IL, crosstalk, and RL, as well as the trade-offs between them, thus providing a better accuracy and channel margin.

The receiver is specified with various RLs (differential, common-mode, and differential-to-common mode) and an interference tolerance. The latter covers various worst-case transmitter signaling and jitter combinations plus worst-case channel conditions (IL, ILD, RL, ICN, noise, and combinations thereof) and ensures that the receiver can recover the data at the target BERs.

PAM-4 PHY

One of the objectives of the IEEE 802.3bj Standard is to define a four-lane 100 Gb/s PHY that can operate over backplanes that can currently transfer data at 10 Gb/s over a single lane using the IEEE 802.3ap standard (10GBASE-KR). This objective implies that the data rate per lane must increase by a factor of 25 Gb/s/10 Gb/s = 2.5. To achieve this goal, the baud rate, the number of bits per symbol, or both must be increased. The requirement that 100 Gb/s PHYs must work over existing backplanes prevents us from increasing the baud rate significantly because the behavior of these backplanes is not well-defined above 5 GHz. Measurements show that existing backplanes can have 65 dB insertion loss at 12.5 GHz, which renders 25 GHz NRZ signaling impossible [14]. In the following, we will describe how a combination of higher-order modulation, FEC, and a small increase in baud rate enables 100 Gb/s data transfers across these challenging channels.

Historically, backplane communication has relied on NRZ modulation. Higher-order modulation (e.g., PAM-M) provides an interesting trade-off. As the number of modulation levels increase, the voltage difference between the signal levels decreases so that a larger signal-to-noise ratio (SNR) is necessary to keep the error rate constant; however, this effect is balanced by a reduction of the signaling rate, which reduces the IL of the channel. To determine the best line-code to operate over the worst-case 10GBASE-KR channel, we compute the Salz SNR [15] for PAM-2, PAM-3, … assuming an extrapolated 10GBASE-KR ICR limit line. The Salz SNR, which is the SNR of an ideal infinite-span decision DFE, is a popular metric in communication theory, because it is a useful upper bound of the DFE receiver performance. Table 3 shows the Salz SNR margin for several choices of PAM-M for the extrapolated 10GBASE-KR channel assuming that a single error-event error-correction code is used [16]. PAM-4 provides an efficient operating point with the largest SNR margin.

Implementation of a PAM-4 digital receiver with a conventional FFE and DFE is straightforward using a look-ahead DFE structure. Because each symbol can have four different values, the complexity of a look-ahead DFE grows as 4^N, where N is the number of DFE taps. This exponential growth in DFE complexity suggests that a digital receiver should implement a large number of FFE taps and as few DFE taps as possible.

DFE receivers cause error bursts when one or more of the DFE taps are large because an incorrect decision is likely to cause subsequent decisions to be wrong. These error bursts can significantly reduce the coding gain provided by FEC. Block interleaving can break up the error bursts, but the additional latency caused by interleaving is unacceptable. Pre-coding [17] reduces the effect of DFE error propagation for a 1-tap DFE by breaking up DFE error bursts so that each error burst turns into two single errors after decoding. Pre-coding is less effective in breaking error bursts generated by multi-tap DFEs, but as mentioned earlier, hardware complexity strongly favors the use of one-tap DFEs.

The 802.3bj standard specifies a RS FEC code that provides more than 7 dB of raw coding gain for the 100GBASE-KP4 PHY. As increasing the baud rate reduces the Salz SNR and therefore the net coding gain, the standard uses 256b/257b transcoding to minimize the baud rate needed to transmit the FEC parity bits. When one accounts for the effects of DFE error propagation, pre-coding, increased baud rate, and channel insertion loss, FEC provides a net coding gain of approx. 5 dB. Taking the coding gain into account, PAM-4 systems should have a margin of approx. 6 dB for non-idealities and imple-

Transcoding is used to reduce the over-speed that the system must run with FEC, and FEC is applied across multiple lanes to reduce the latency to acceptable levels for the demanding high-performance applications targeted.

mentation penalty on the worst-case 10G KR channels. The 802.3bj standard also specifies the insertion of periodic block termination bits [18] to simplify the optional implementation of more complex equalizers, such as maximum likelihood sequence estimation (MLSE), which could provide 1.5 dB of additional margin.

Conclusions

This article provided an overview of the work that is ongoing in the IEEE P802.3bj 100 Gb/s Backplane and Copper Cable Task Force. The task force faced with numerous demands from the industry, such as 100 Gb/s across four lanes of legacy backplanes, new high-performance backplanes, and copper cable. With two PHYs, one supporting NRZ modulation and the other supporting PAM-4 modulation, the 802.3bj PHYs are able to handle the needs of many applications. Transcoding is used to reduce the over-speed that the system must run with FEC, and FEC is applied across multiple lanes to reduce the latency to acceptable levels for the demanding high-performance applications targeted.

The authors would like to acknowledge the editorial suggestions provided by C. Bolliger.

References

Study group presentations can be found at sg = http://www.ieee802.org/3/100GCU/public
Task force presentations can be found at tf = http://www.ieee802.org/3/bj/public

[1] H. Frazier, "Two Markets, Two Channels, Two PHYs," tf/jan12/frazier_01a_0112.pdf, Jan. 2012.
[2] R. Cideciyan, "512b/513b Transcoding and FEC for 100 Gb/s Backplane and Copper Links," tf/sep11/cideciyan_01a_0911.pdf, Sept. 2011.
[3] R. Cideciyan, "256b/257b Transcoding for 100 Gb/s Backplane and Copper Cable," tf/mar12/cideciyan_01a_0312.pdf, Mar. 2012.
[4] Z. Wang *et al.*, "Communication Device Employing binary Product Coding with Selective Additional Cyclic Redundancy Check (CRC) Therein," US Patent 2012/0220474, Sep. 2010.
[5] Z. Wang and C. J. Chen, "Feasibility of 100G KR FEC," sg/may11/wang_01_0511.pdf, May 2011.
[6] R. Cideciyan and J. Ewen, "Transcoding/FEC Options and Trade-offs for 100 Gb/s Backplane and Copper Cable," tf/nov11/cideciyan_01a_1111.pdf, Nov. 2011.
[7] Z. Wang, H. Jiang and C.J. Chen, "Further Studies of FEC Codes for 100G KR," tf/nov11/wang_01a_1111.pdf, Nov. 2011.
[8] M. Gustlin *et al.*, "Backplane NRZ FEC Baseline Proposal," tf/mar12/gustlin_01_0312.pdf, Mar. 2012.
[9] M. Brown and Z. Wang, "100G Backplane PAM4 PHY FEC/PMA Encoding Enhancements," tf/may12/brown_01a_0512.pdf, May 2012.
[10] R. T. Chien, "Cyclic Decoding Procedures for the BCH Codes," *IEEE Trans. Inform. Theory*, Oct. 1964.
[11] P. Patel *et al.*, "PAM 2 on a 1 Meter Backplane Channel," tf/sep11/patel_01b_0911.pdf, Sept. 2011.
[12] T. Beukema, M. Meghelli and J. Ewen, "Line Signaling Performance Comparison on Emerson Channel," tf/jan12/meghelli_01_0112.pdf, Jan. 2012.
[13] R. Mellitz *et al.*, "Time Domain Channel Specification: Proposal for Backplane Channel Characteristic Sections," tf/jul12/mellitz_01_0712.pdf, July 2012.
[14] H. Frazier and V. Parthasarathy, "Study of 100 Gb/s on 40GBASE KR4 Channel," sg/jan11/parthasarathy_01_0111.pdf, Jan. 2011.
[15] J. Salz, "Optimum Mean Square Decision Feedback Equalization," *Bell Sys. Tech. J.*, Oct. 1973.
[16] W. Bliss, "25 Gb/s Communication over 10BASE KR Channels," sg/jan11/bliss_01_0111.pdf, Jan. 2011.
[17] P. Kabal and S. Pasupathy, "Partial Response Signaling," *IEEE Trans. Commun.*, Sept. 1975.
[18] D. Dabiri, "Enabling Improved DSP Based Receivers for 100G Backplane," tf/sep11/dabiri_01_0911.pdf, Sept. 2011.

Biographies

ROY D. CIDECIYAN received the Dipl.-Ing. degree in electrical engineering from Aachen University of Technology, Germany, in 1981 and the M.S.E.E. and Ph.D. degrees in electrical engineering from the University of Southern California, Los Angeles, in 1982 and 1985, respectively. He joined IBM Research, Zurich, Switzerland, in 1986, where he worked on various projects related to data transmission, magnetic storage on hard disk and tape, and non-volatile memory technologies. In January 2010, he was elected Fellow of the IEEE. Since 2011, he contributed to the 100 Gbit/s Ethernet backplane and copper cable standards.

MARK GUSTLIN (mark.gustlin@xilinx.com) is a principal system architect at Xilinx working with customers on system architecture and design, and working with the Xilinx silicon team on next generation products and IP. Previously he spent 15 years at Cisco Systems architecting and designing high end routers. He is an active participant in the IEEE, one of the inventors of the 40/100GbE PCS, and was one of the editors of the 802.3ba standard. He is currently involved in the IEEE 802.bj and 802.3bm task forces, working on next generation 100GbE interfaces, and he is participating in the 400 Gb/s study group. He holds a B.S.E.E. in Electrical Engineering from San Jose State University.

MIKE PENG LI [F] has been with Altera Corporation since 2007 and currently is an Altera Fellow. He is a corporate expert and adviser, as well as CTO office principal investigator, on high-speed link technology, standard, SERDES and I/O architecture, electrical and optical signaling, silicon photonics, optical FPGA, high-speed simulation/debug/test, jitter, noise, signal and power integrity. He was the Chief Technology Officer (CTO) for Wavecrest Corporation from 2000-2007. He is an affiliated professor at the Department of Electrical Engineering, University of Washington, Seattle. He holds a Ph.D. in physics (1991), an M.S.E (1991), in electrical and computer engineering and an M.S. in physics (1987), from the University of Alabama, Huntsville. He also holds a B.S (1985) in space physics from the University of Science and Technology of China. He was a Post Dr. and then a research scientist at the University of California, Berkeley (1991–1995). He has close to 100 publications on refereed journals and conferences and close to 20 patents.

JOHN WANG is an associate technical director at Broadcom Corporation in San Jose, California. He received a B.S degree in electrical engineering and computer science from Caltech and S.M., E.E. and Ph.D. degrees in electrical engineering from the Massachusetts Institute of Technology. He was a senior fellow at Acuson Corporation (now Siemens Ultrasound) and implemented color and spectral Doppler and digital image compression for the Sequoia ultrasound platform. At Big Bear Networks, he developed 40 Gb/s transceivers and a 10 Gb/s LRM receiver ASIC. Next, he worked on 10GBASE-T at Aquantia. He is currently working on 100 Gb/s electrical and optical PHYs at Broadcom.

ZHONGFENG WANG received both Bachelor and Master degrees from Tsinghua University. He obtained the Ph.D. degree from the Department of Electrical and Computer Engineering at the University of Minnesota, Minneapolis. He is now an Associate Technical Director at Broadcom Corporation, California. Before that, he was an Assistant Professor at Oregon State University. Even earlier he was working for National Semiconductor Incorporation. He is a world-recognized expert on Low-Power High-Speed VLSI Design for Signal Processing Systems. He has published over one hundred technical papers with two best paper awards received in the IEEE Circuits and Systems (CAS) society. He has edited one book "VLSI" and filed over thirty U.S. patent applications and disclosures. In the past ten years, he has served as Associate Editor for IEEE Trans. on CAS-I, CAS-II, and VLSI Systems for many terms. He has been serving in three technical committees in the IEEE CAS society and Signal Processing society. In 2013, he served in the annual best paper award selection committee in the IEEE CAS society. His current research interests are in the area of Low Power/High Speed VLSI Design for High-Speed Networking Systems.