Highly-Resilient 50G Gen-Z PHY and Channel Design

Karl Bois, Chris Brueggen, Elene Chobanyan, Michael Krause, Pete Maroni, James Regan
{karl.bois, chris.brueggen, elene.chobanyan, mkrause, pete.maroni, jregan}@hpe.com

Previously in *Highly-Resilient 25G Gen-Z PHY* paper a case for high resilience fabric performance while avoiding Forward Error Correction (FEC) was presented. This paper is discussing the performance that can be achieved at 50GT/s by limiting channel characteristics such as insertion/reflections losses and crosstalk. It will be demonstrated that platforms can achieve raw Bit Error Ratio (BER) <1E-9. With a low latency (<2ns for a four-lane link) Gen-Z Phit FEC, the effective BER is reduced to 1E-15.

Channel Topologies and Performance

The *50G Gen-Z Physical Layer Specification* does not normatively specify channel frequency masks. Instead, to ensure interoperability, Gen-Z specifies an industry-adopted method for validating channel compliance – Channel Operating Margin (COM). To provide designers with a starting point, Gen-Z defines informative insertion loss masks shown in dashed black lines in *Figure 1*. *Gen-Z-E-PAM4-50G-Fabric* defines a Gen-Z physical layer capable of a line-rate at 53.125 GT/s per lane, inclusive of the overhead with Phit FEC, using 4 level Pulse Amplitude Modulation (PAM4) signaling over Fabric media (~20dB channel). *Gen-Z-E-PAM4-50G-Local* is designed for Local media (~10dB channel).

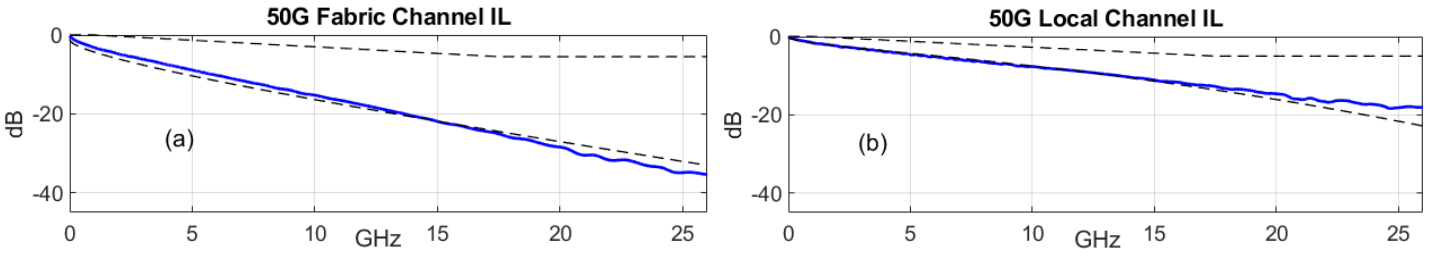

Figure 1: Channel insertion loss (IL): (a) Gen-Z-E-PAM4-50G-Fabric, (b) Gen-Z-E-PAM4-50G-Local

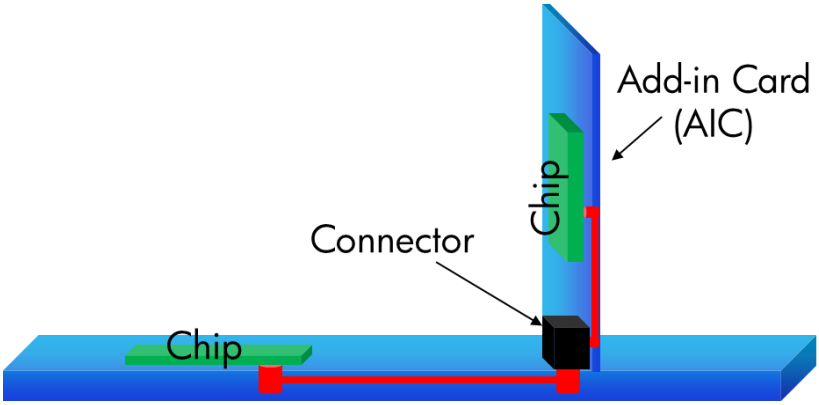

Figure 2: Channel Topology

For each electrical specification (i.e. *50G Local* and *50G Fabric*) the channel shown in *Figure 2* was designed to the respective maximum loss budgets up to 13.28125 GHz. Both channels were optimized within the actual product implementation feasibility to minimize the insertion loss to crosstalk ratio (ICR) shown in *Figure 3* and insertion loss deviation (ILD) shown in *Figure 4*. ICR and ILD are measures of the channel noise and reflectivity, respectively, and are computed as specified in *IEEE Standard for Ethernet 802.3™-2015*.

Time domain simulations to obtain COM values at BER ranging from 1E-12 to 1E-4 were performed with two far-end and three near-end aggressors (crosstalk contributors). The rest of the COM tool computation settings for 50G Local channel were leveraged from *CEI-56G-MR-PAM4* clause in *IA # OIF-CEI-04.0* and for 50G Fabric channel from *CEI-56G-LR-PAM4* clause in *IA # OIF-CEI-04.0* with the required modifications of signaling rate (f_b) = 26.5625 Gsym/s.

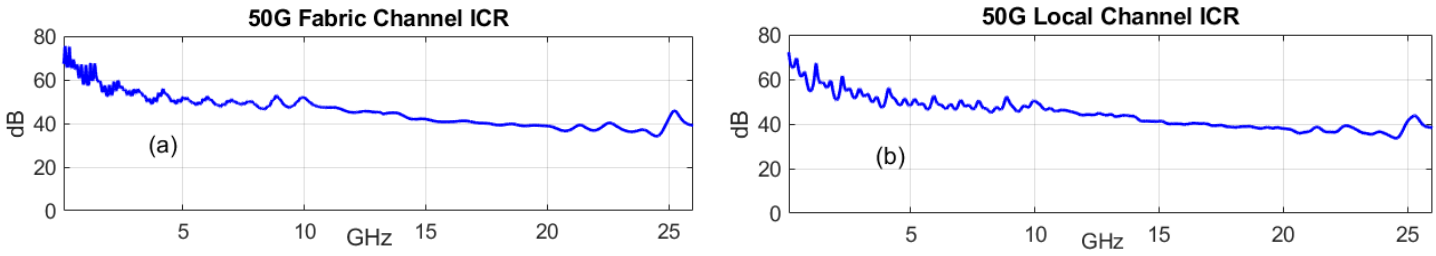

Figure 3: Insertion loss to cross-talk ratio (ICR): (a) Gen-Z-E-PAM4-50G-Fabric, (b) Gen-Z-E-PAM4-50G-Local

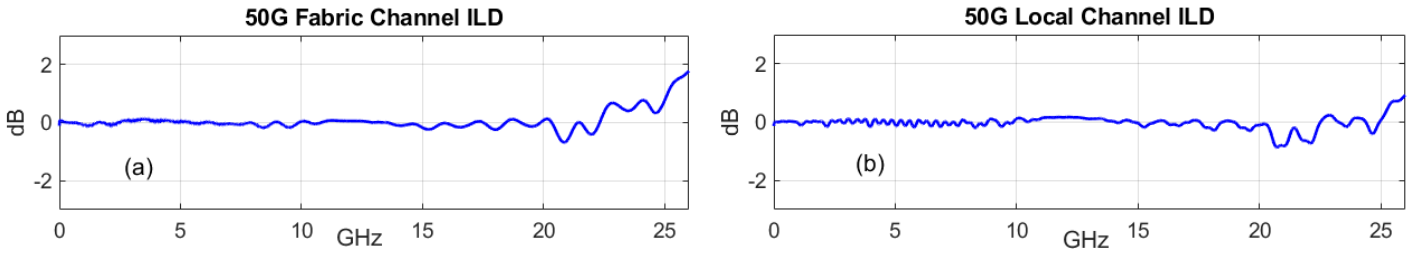

Figure 4: Insertion loss deviation (ILD): (a) Gen-Z-E-PAM4-50G-Fabric, (b) Gen-Z-E-PAM4-50G-Local

Compliant channels can be built to this specification when COM ≥ 3 dB (failing channels with COM ≤ 3 dB are highlighted in bold red in Table 1).

Table 1. Channel Operating Margin (COM)

BER	50G Fabric	50G Local
1e-4	7.512	7.492
1e-5	6.324	6.318
1e-6	5.390	5.408
1e-7	4.633	4.656
1e-8	3.990	4.031
1e-9	3.430	3.484
1e-10	**2.939**	3.005
1e-11	**2.497**	**2.579**
1e-12	**2.098**	**2.190**

Further, these results demonstrate that if decision feedback equalization (DFE) based designs (12-tap DFE with 0.7 constraint for 1st tap and 0.2 for subsequent taps) are used, then Gen-Z 50G PAM4 for both Local and Fabric channels can achieve a raw BER of 1E-9. The combination of an optimized channel design, Gen-Z's Phit FEC, strong packet-based CRC, and end-to-end and link-local packet retry mechanisms yield highly reliable channels with negligible solution-visible performance impacts. Details of Gen-Z's Phit FEC and incurred latency are discussed in the *Phit FEC* section.

Phit FEC

The relatively high BER associated with PAM4 technology necessitates the use of FEC for performance and reliability. With no FEC:

- Errors detected on packet CRC checks would result in frequent packet retransmissions and link resynchronizations leading to customer-visible performance degradation in the form of reduced bandwidth and increased instantaneous memory access latency. Non-memory semantic fabrics can support recovery events at effective BERs of 1E-12 or higher, but memory applications running on a Gen-Z fabric require a 1E-15 effective BER to minimize fabric congestion and memory access latency.
- There is a higher probability of errors which may not be detected by packet CRC checks, leading to reduced Mean Time to False Packet Acceptance (MTTFPA).

Gen-Z 50G PAM4 local and fabric use a 288-bit **Ph**ysical Dig**it**, or Phit, which contains the FEC codeword. Phits differ from traditional **Fl**ow Control Dig**its**, or Flits, in that Phits do not involve in flow control. The use of Phits allows FEC to be handled completely by the Physical Layer. This allows the Link Layer to support a multitude of Physical Layers with or without FEC.

The Phit FEC is a light-weight (low latency, low decoder area and complexity) BCH encoding and provides good performance and reliability for BER <= 1e-9. It relies on gray coding + pre-coding such that a burst error resulting from DFE error propagation is converted to 2 bit errors (entry + exit). The FEC allows correction of 2 bit errors (i.e. one independent error) per 288-bit Phit. The correction capability of the BCH code is artificially restricted for strong detection of uncorrectable errors (resulting from >1 independent error per Phit). Uncorrectable errors are reported to the Link Layer and combined with the packet CRC check for increased reliability.

Coding Gain:

FEC coding gain can be calculated as follows. Given a particular raw BER, the probability of N errors in a given block size may be estimated by assuming a binomial distribution. At BER=1e-9 with random error distribution, the probability of N independent errors in a 288-bit block can be calculated as:

N=1: $p_1 = \binom{288}{1} (1e\text{-}9)^1 (1\text{-}1e\text{-}9)^{287} = \sim 2.88e\text{-}7$

N=2: $p_2 = \binom{288}{2} (1e\text{-}9)^2 (1\text{-}1e\text{-}9)^{286} = \sim 4.13e\text{-}14$

…

Given a particular link width and transfer rate (e.g. x4 at 53.125 GT/s) we can calculate the total number of 288-bit Phits transferred per unit time. With the estimated probabilities of N errors per Phit, we can calculate how many Phits per unit time will have 1 error, 2 errors, and so on.

With no FEC, any error is uncorrectable (passed along to packet CRC check), and this occurs ~1.27e4 times per minute. With Phit FEC, phits with >1 error are uncorrectable and are generally flagged as containing an error as they proceed to the packet CRC check. This occurs ~1.83e-3 times per minute. The coding gain is thus 1.27e4 / 1.83e-3 = ~6.97e6

"Corrected BER" = 1e-9 / 6.97e6 = ~1.43e-16

Also, the reporting of FEC uncorrectable errors improves overall error detection such that MTTFPA = ~1e19 years.

Performance:

The Phit FEC consists of 256 bits of Gen-Z traffic, 4 Physical Layer Control bits, and 28 FEC redundancy bits resulting in an 11.1% bandwidth overhead. This bandwidth overhead is higher than the 5.5% overhead of the Ethernet RS(544,514) codeword but has significantly lower latency.

The single-lane link latency is 5.42ns (288b / 53.125Gbps), which is much lower than the Ethernet RS(544,514) latency of 102.4ns. The Phit FEC codeword is spread across the entire link, so latency improves with wider link widths. For example, the four-lane link latency is one-fourth the latency at 1.36ns.

The Phit FEC decoder is much smaller and less complex than traditional RS FEC implementations, and error correction can be implemented in a 3-cycle pipeline at 1.2GHz core frequency, or 2.5ns correction latency. Correction latency is hidden by correction bypass for error-free phits. When an error occurs, the correction pipeline fills up and drains in less than 10ns by running the core faster than the link bandwidth, at which point the correction latency is no longer present. This results in a negligible average latency impact since errors only occur every 1.18ms on a x16 link.