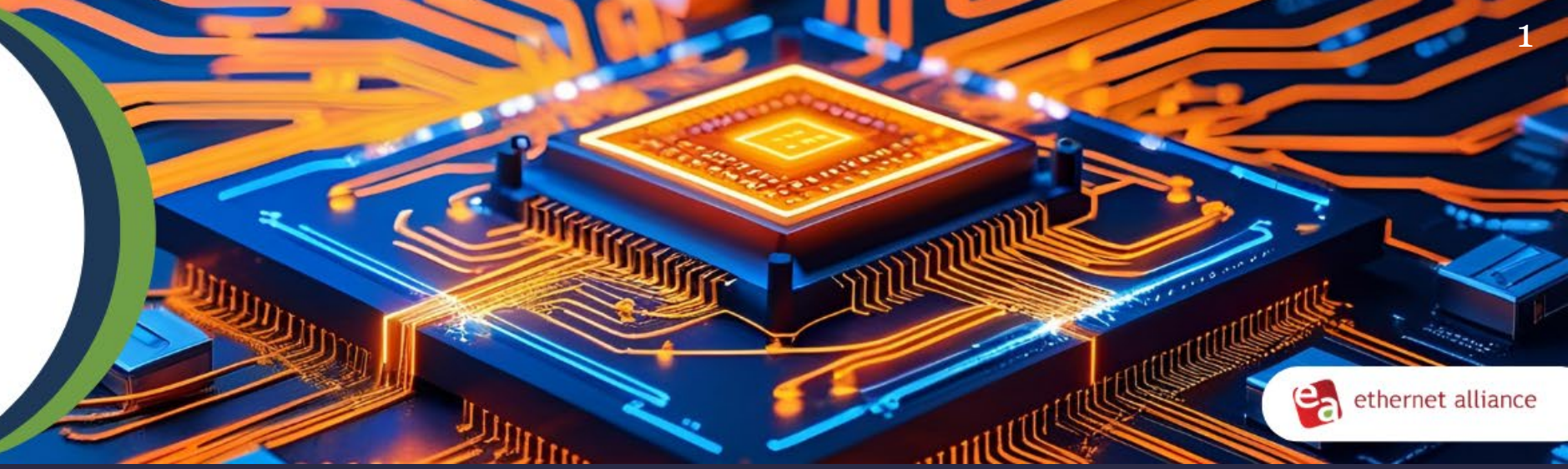
TEF 2025
Ethernet for AI

December 2-3, 2025
Hyatt Centric Mountain View, CA, USA



ethernet alliance

400 Gb/s Signaling for AI Networks From A System Perspective

December 2-3, 2025

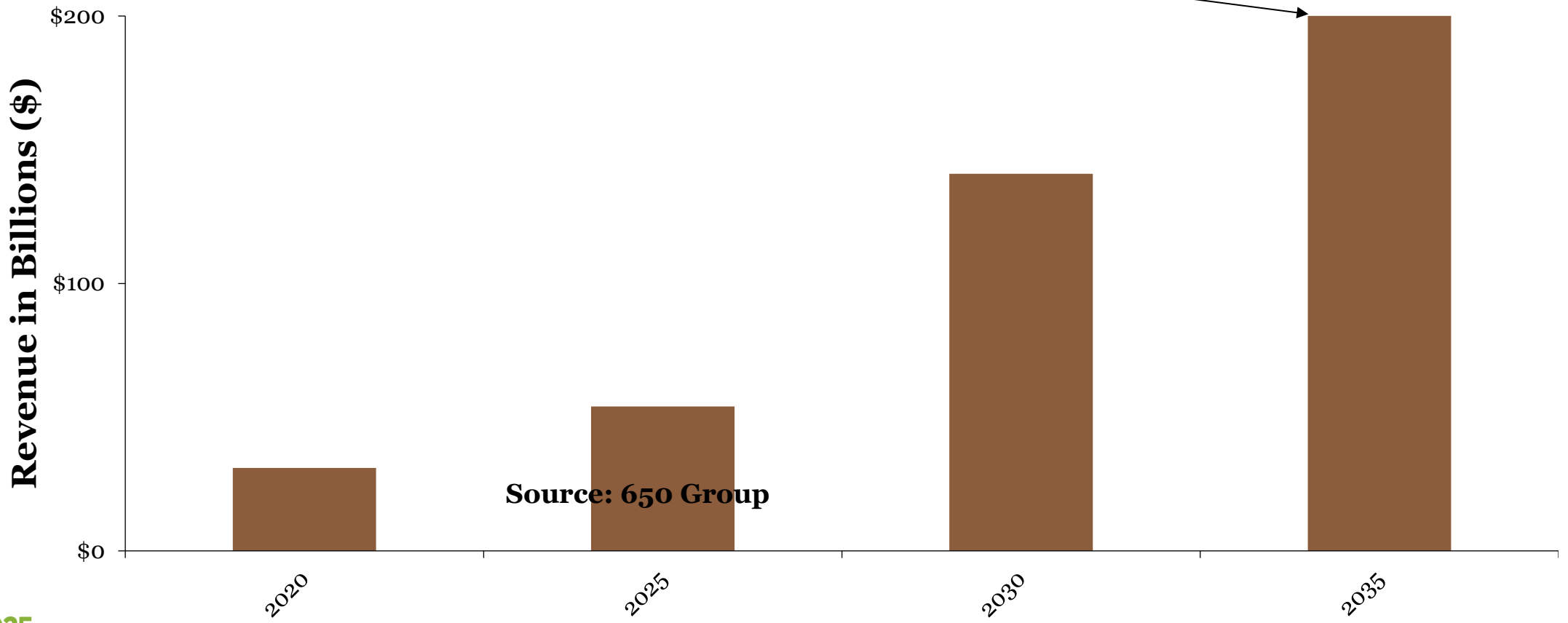
This presentation has been developed within the Ethernet Alliance, and is intended to educate and promote the exchange of information. Opinions expressed during this presentation are the views of the presenters, and should not be considered the views or positions of the Ethernet Alliance

Setting the Stage for Networking in an AI World

Alan Weckel, Founder and Technology Analyst – 650 Group

The Evolution of the Ethernet Switch Market

By 2035, led by AI, The Ethernet Switch Market will Exceed $200B



Includes Total Ethernet Switching Market Campus and DC (Scaleup, Scaleout, Frontend, Scale Across)
Does not include NICs, InfiniBand, NVLink, PCIe

AI and HPC Networking Transition

2024 (x112) -> 2026 (x224) Traditional Cloud Server	
Bandwidth	Technology
100-800G	Ethernet

2024 (x112) -> 2026 (x224) AI Cloud Server (Nvidia)	
Bandwidth	Technology
400-800G	Ethernet
400G-1.6T	InfiniBand / Ethernet
900G-1.8T	NVLink

2024 (x112) -> 2026 (x224) AI Cloud Server	
Bandwidth	Technology
400-800G	Ethernet
400G-1.6T	Ethernet
TBD	• UALink (AMD Infinity Fabric Based) • PCIe/CXL • Ethernet

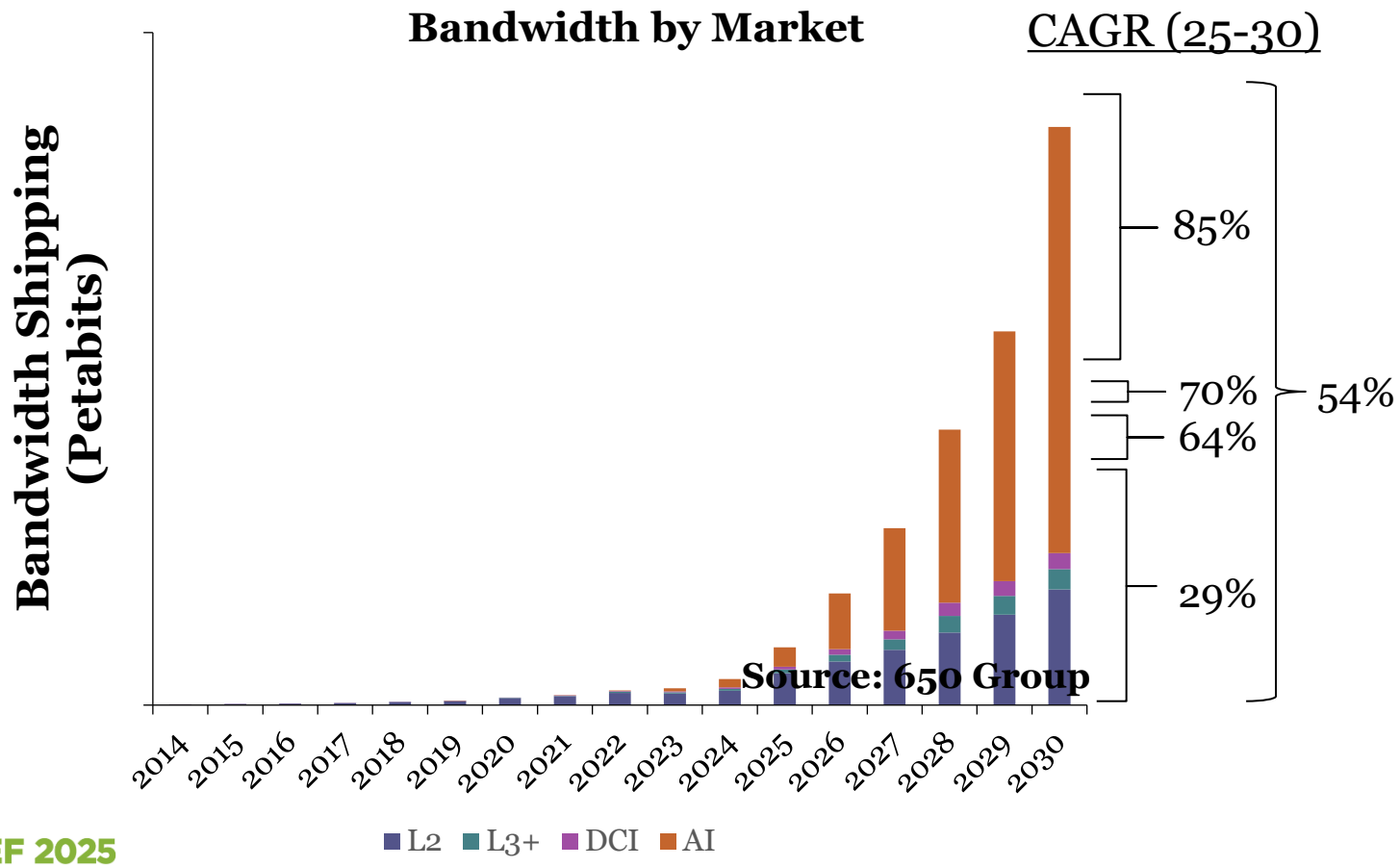
Frontend (1X)

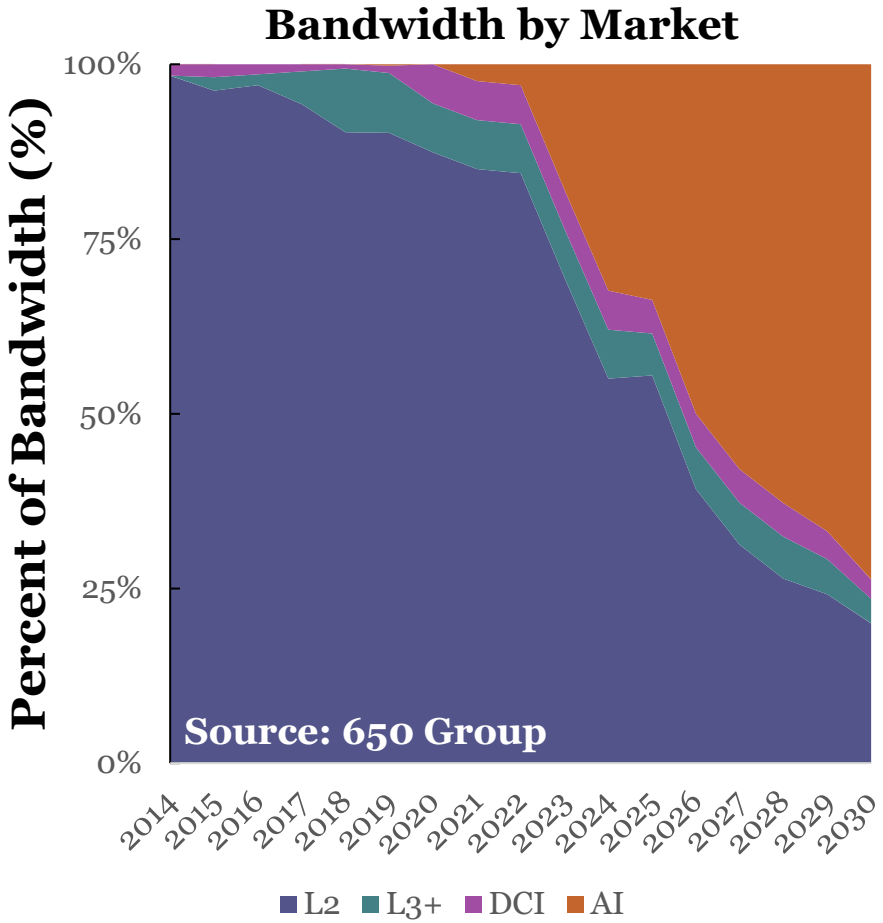
Backend (Scaleout) (10X)

Backend (Scaleup) (100X)

Source: 650 Group

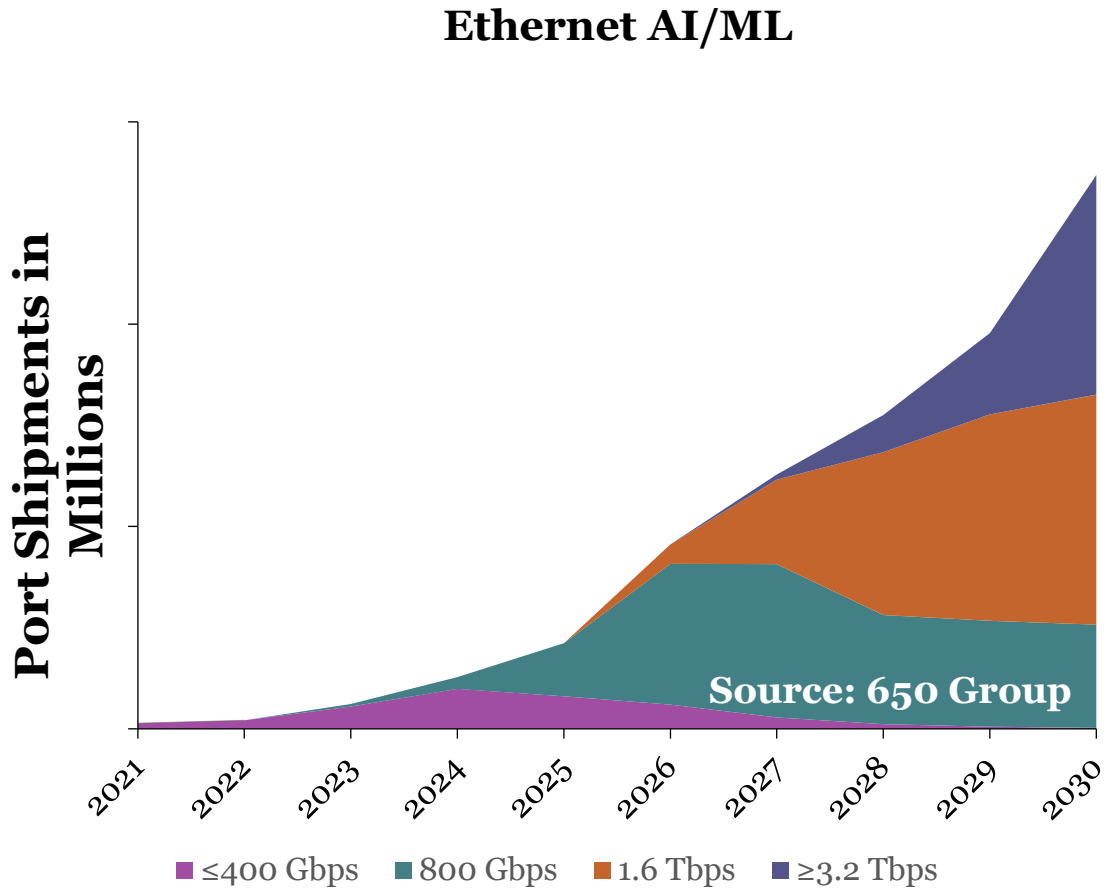
Data Center Networking Bandwidth Ethernet (Back-End) and InfiniBand

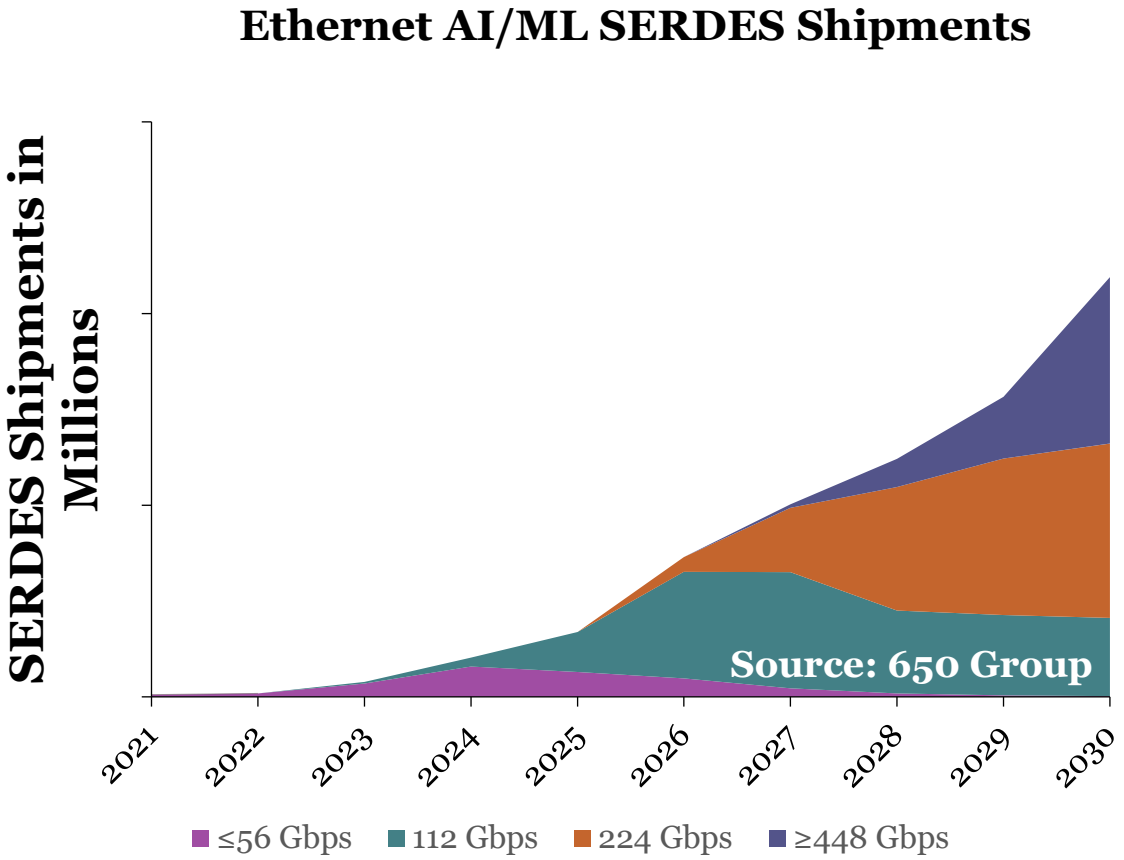




TEF 2025 Ethernet for AI

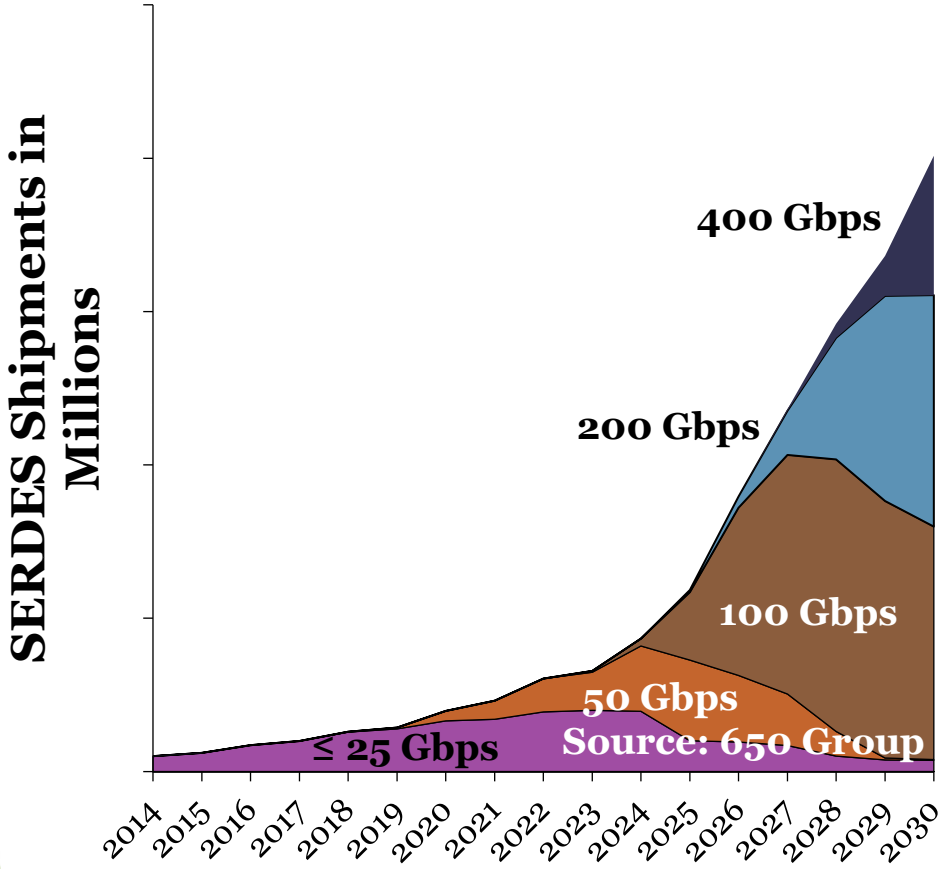
Ethernet Switch – Data Center: AI/ML Port Speeds and SERDES Shipments

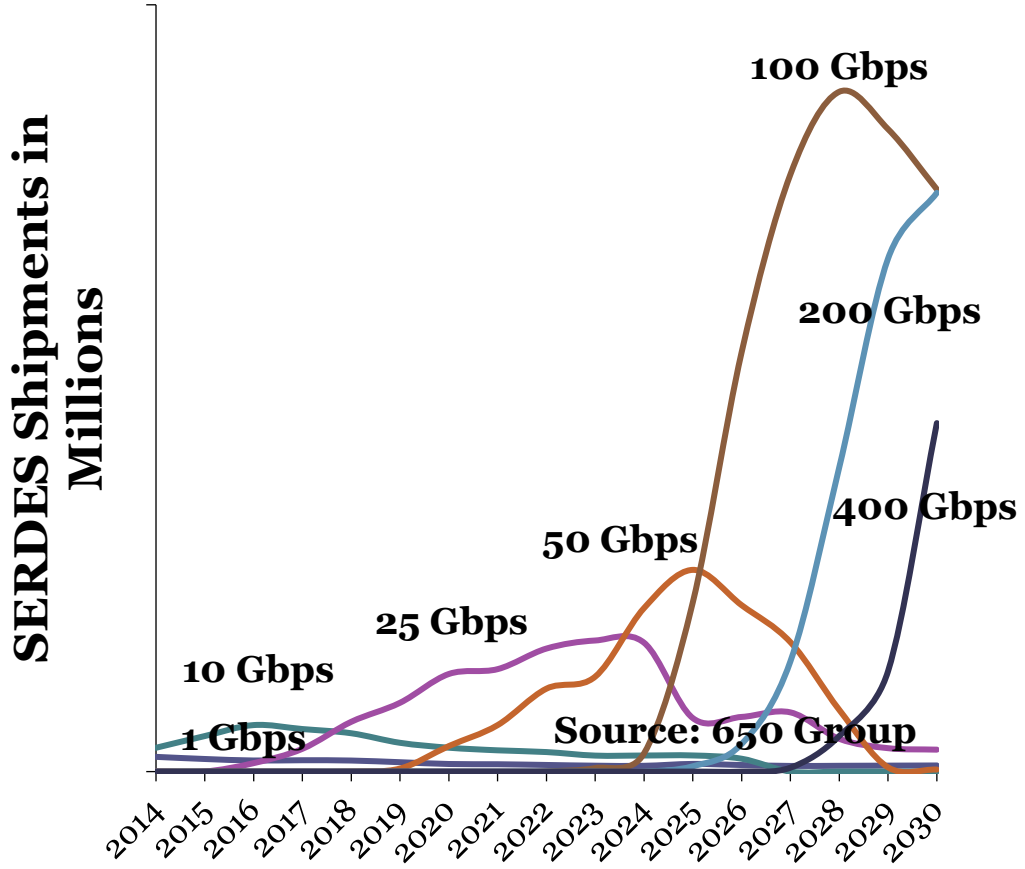




TEF 2025 Ethernet for AI

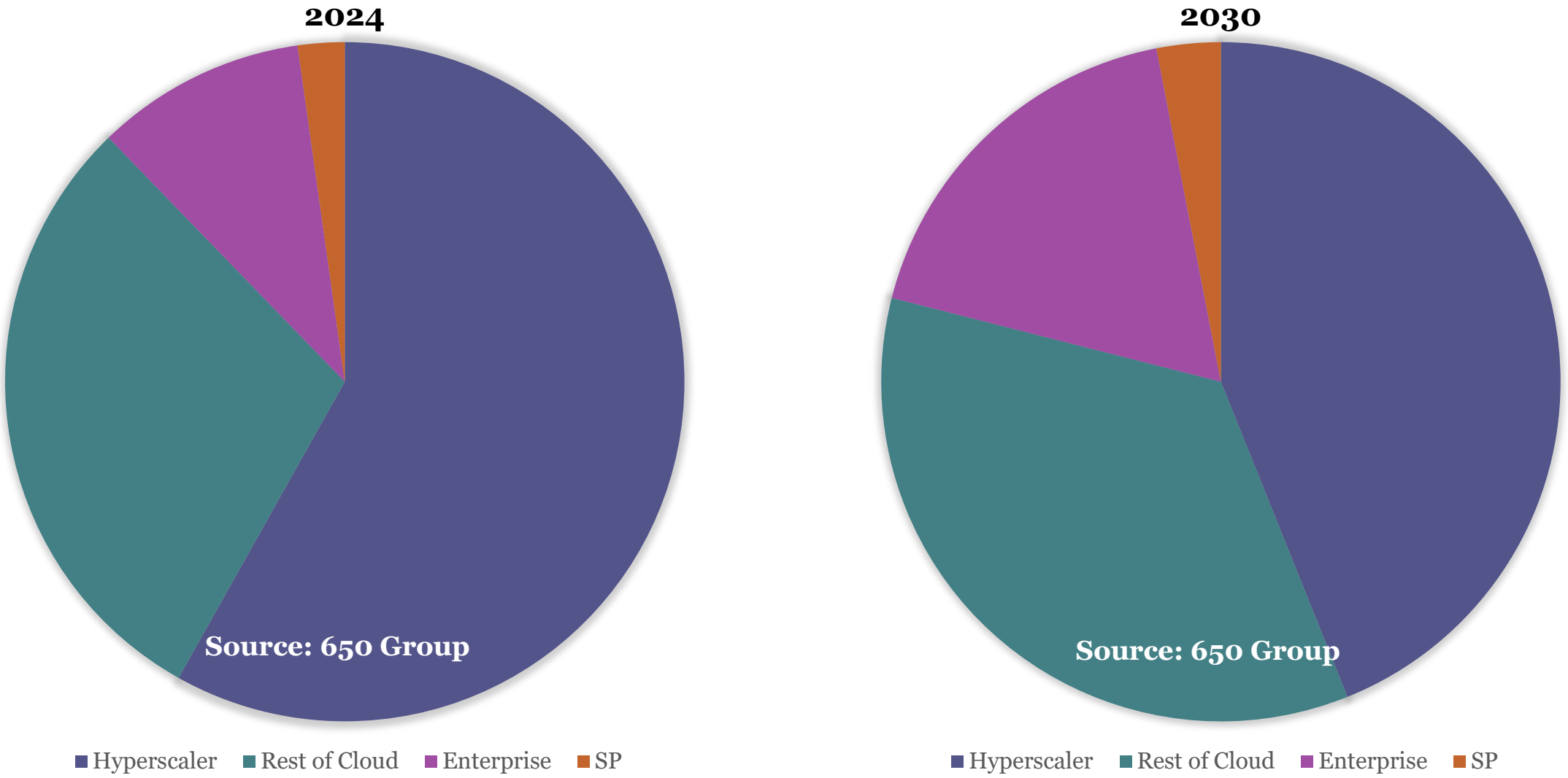
Data Center Switching
Total SERDES Shipments





TEF 2025 Ethernet for AI

Networking for AI Verticals



2025 to 2030 Data Center Switch Stats

2025	Volume/Size
Yearly Switch Port Volume	150-160 M
DC Switch Installed Base	400+ M
Switch Size	51.2T
Power per Rack	100 kW

2030	Volume/Size
Yearly Switch Port Volume	300+ M
DC Switch Installed Base	1+ B
Switch Size	204.8T and higher
Power per Rack	1 MW

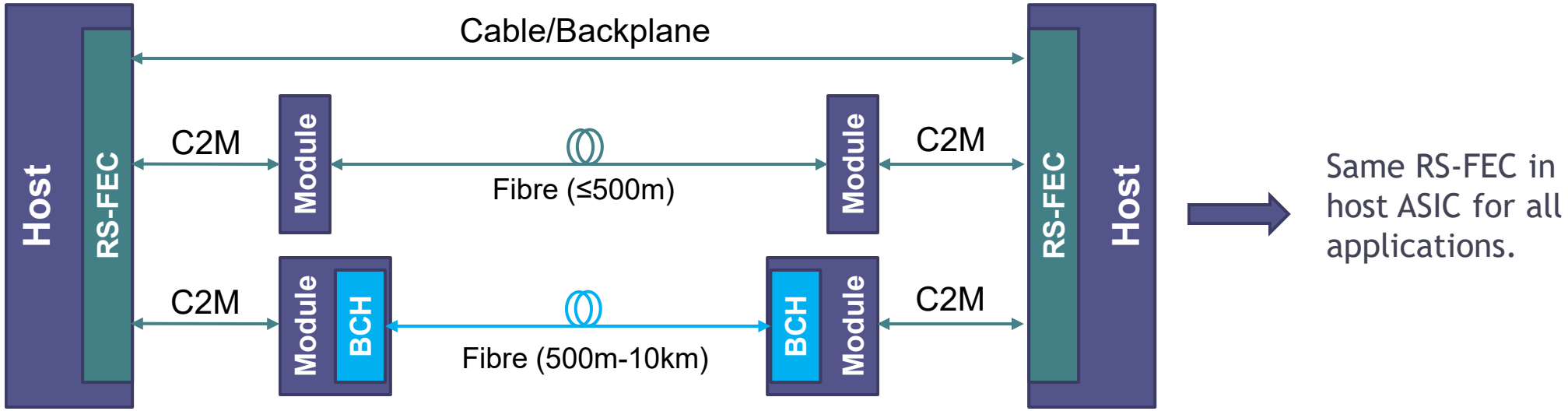
FEC for 448G:
Can Today's FEC Survive Tomorrow's Modulation?

Xiang He, Distinguished Engineer, Huawei Technologies

Co-authors: Congshi Zou, Yuefeng Wu, Shuangxing Dai, Hao Ren, Xuebo Wang, Tong Mu

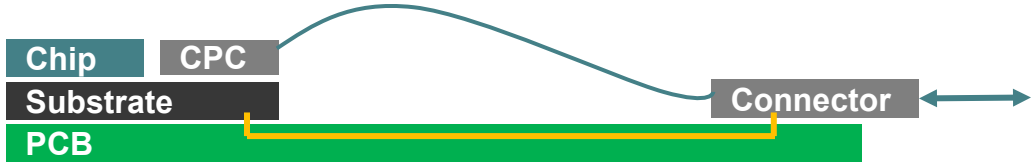
Background

- The end-to-end Reed-Solomon FEC (RS-FEC) has been used for Ethernet for over a decade.
  - The extremely low miscorrection probability of the chosen FEC provided robust and reliable link performance.
- A distributed concatenated FEC scheme (RS+BCH) was introduced for 200G/lane optical links over 500m.
  - Inner FEC using soft-decoding techniques provided lower input BER to the RS-FEC.
  - Host ASIC keeps the RS-FEC
  - Reliability is guaranteed by the outer FEC, despite of higher miscorrection probability of Inner FEC.

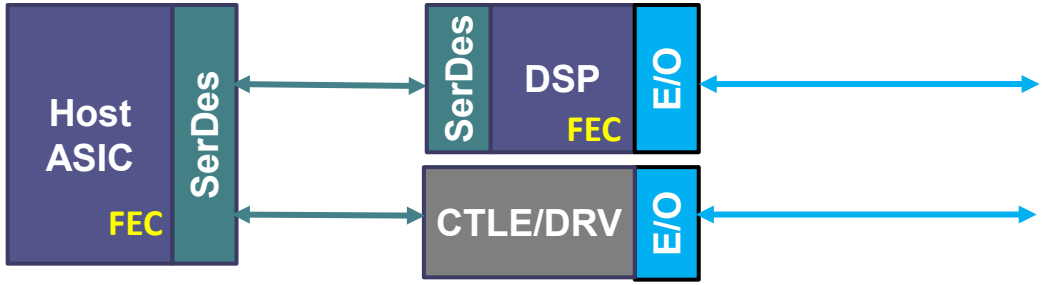


Applications for 448G Electrical Interfaces (and location of FEC)

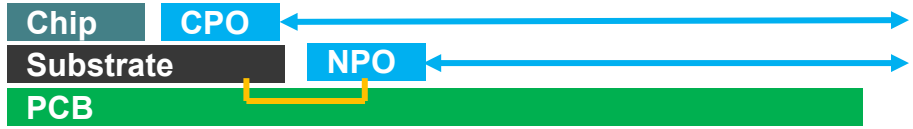
Cable (Incl. CPC/NPC)



FPP Optics



CPO NPO



Passive Channel Improving over the Past Year

- At TEF-2024, almost all channels can only support bandwidth around 80~85GHz.
- Passive component bandwidth has been improved over the past year.
  - OIF and IEEE both had simulated and measured data **>100GHz**.
  - CPC channels **>100GHz** were demoed by all major cable vendors in 2025 OCP Global Summit.
- Co-packaged copper cables are becoming feasible to enable 448G transmission.
- Advanced 2D connectors are promising candidates to replace “gold finger”-type connectors for front panel pluggables.
- OIF high-density connector (HDC) project is considering new form factors and architectures to support 448G interconnects.

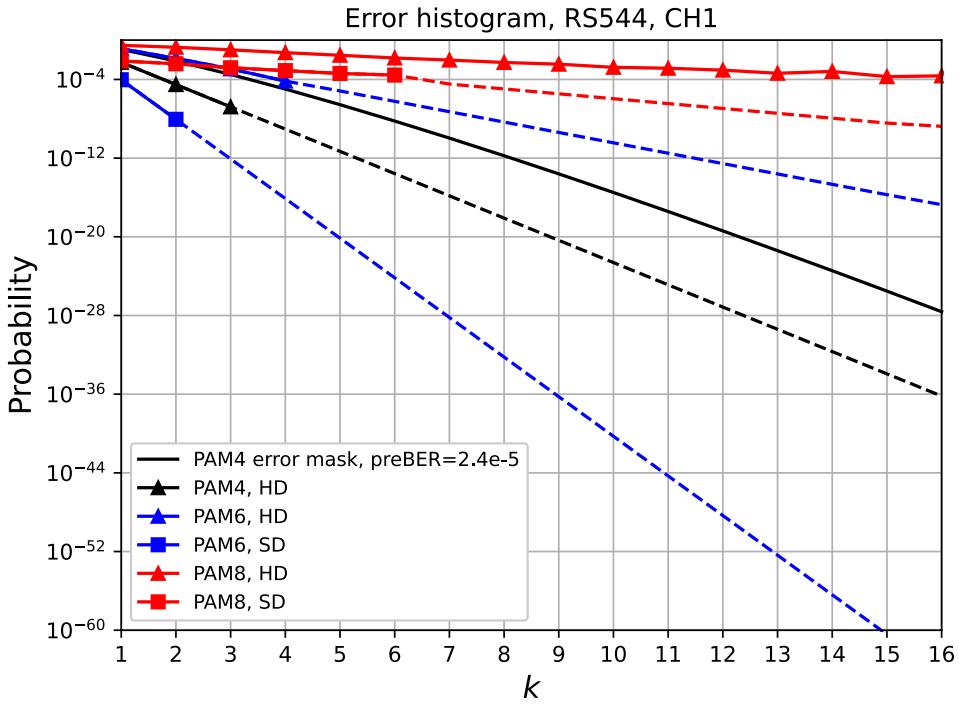
Reusing RS(544,514) for 448G – the "Logical" Choice

- RS(544,514) FEC is everywhere from 400GE to 1.6TE, regardless of the new 448G/lane FEC.
- Breakout is a key (mandatory) requirement for switches.
  - A higher speed Ethernet port can be configured to multiple lower rate PHYs.
  - A switch supporting 3.2TE@448G/lane, will still need to cover 1.6TE and lower rate PCS w/ RS(544,514).
- Reusing RS(544,514) for 448G/lane keeps PCS unchanged and leverages proven silicon.
- **Best case:** completely reuse the existing RS(544,514) FEC as in P802.3dj (for host ASIC).

…but is its error-correction margin still sufficient at higher modulations?

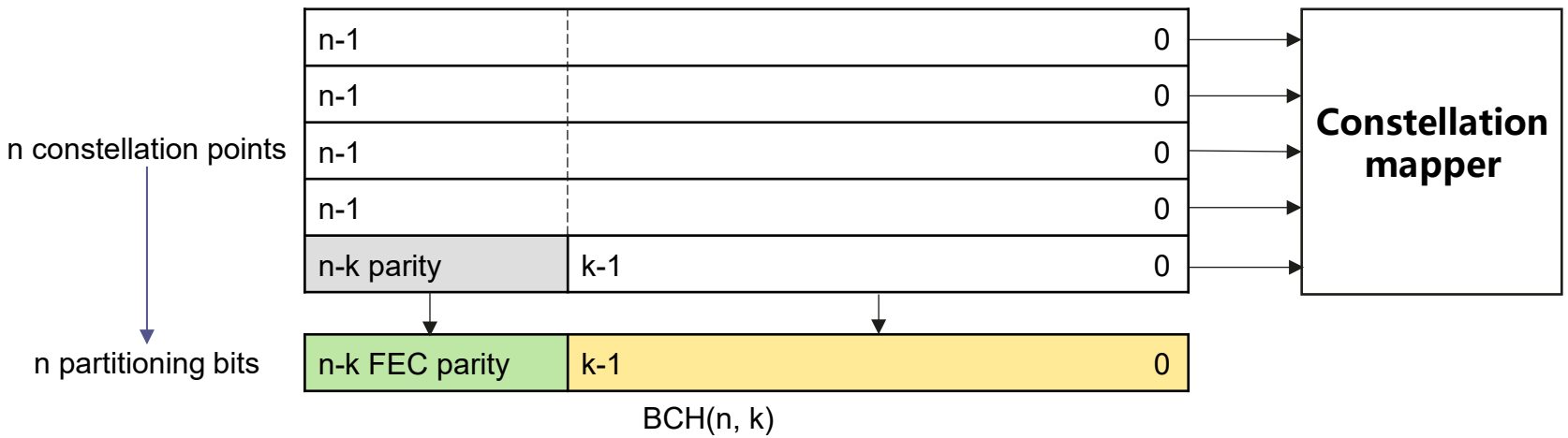
PAM6 or PAM8 Using RS(544,514)

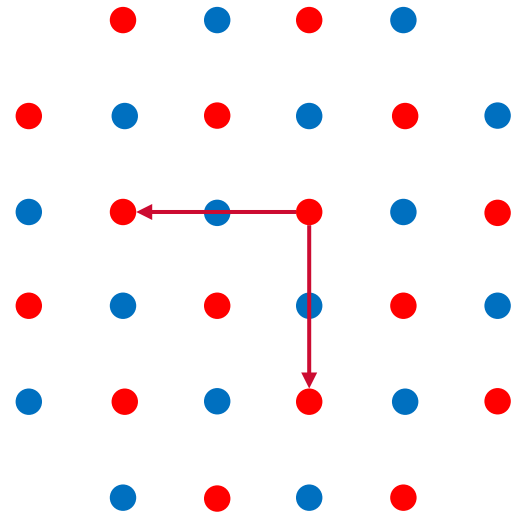
- Burst error measured by bits will be longer for higher modulation, assuming burst error length in terms of PAM-n symbols remains the same.
  - 20-PAM4 symbols: 40bits, affecting 4 RS-FEC symbols.
  - 20-PAM6 symbols: 50bits, affecting 5 RS-FEC symbols.
  - 20-PAM8 symbols: 60bits, affecting 6 RS-FEC symbols.
- 4-way interleaving of RS-FEC codewords adopted by IEEE P802.3dj may be sufficient, but requires more analysis.
  - Our preliminary analysis using block error ratio metric shows risks for PAM6 or PAM8 using RS(544,514) FEC.
- **2nd best case:** Partial reuse – either as the outer-FEC in a concatenated scheme, or a longer RS-FEC that may reuse its enc/dec logic.



PAM6 Using Inner FEC w/ Set Partitioning – 2 subsets

- 2D-PAM6 constellation can be split into two subsets such that d_{min} within each subset gets larger[1].
- One partitioning bit can be applied to indicate the subset information and detect probable constellation errors.
- FEC is used to protect the partitioning bit[2] and correct the constellation errors.

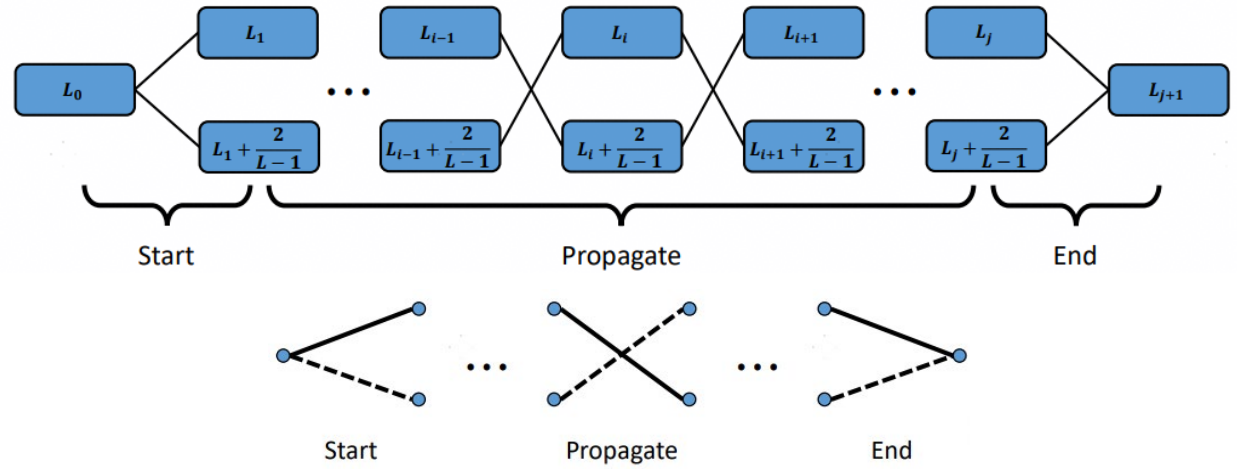




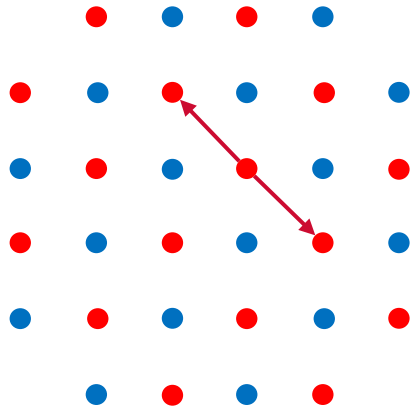
[1] C. Liu, “Performance Analysis at 400+Gbps Over Next-Generation VSR Channels”, Ethernet Alliance Technology Exploration Forum 2024
[2] A. Healey, C. Liu, “Modulations, encoding, and error correction for 448 Gb/s per lane electrical links”, OIF 448Gbps Signaling for AI Workshop

PAM6 MLSE Using Inner FEC w/ 2 subsets – disadvantages

- Error events of an L-PAM $1 + \alpha D$ MLSE are dominated by a zig-zag pattern in the form of alternating between adjacent levels[1]:



$$L_i \in \begin{cases} -1, -1 + \frac{2}{L-1}, \cdots, +1 - \frac{2}{L-1}, +1 & i = 0, j+1 \\ -1, -1 + \frac{2}{L-1}, \cdots, +1 - \frac{2}{L-1} & i = 1, \ldots, j \end{cases}$$

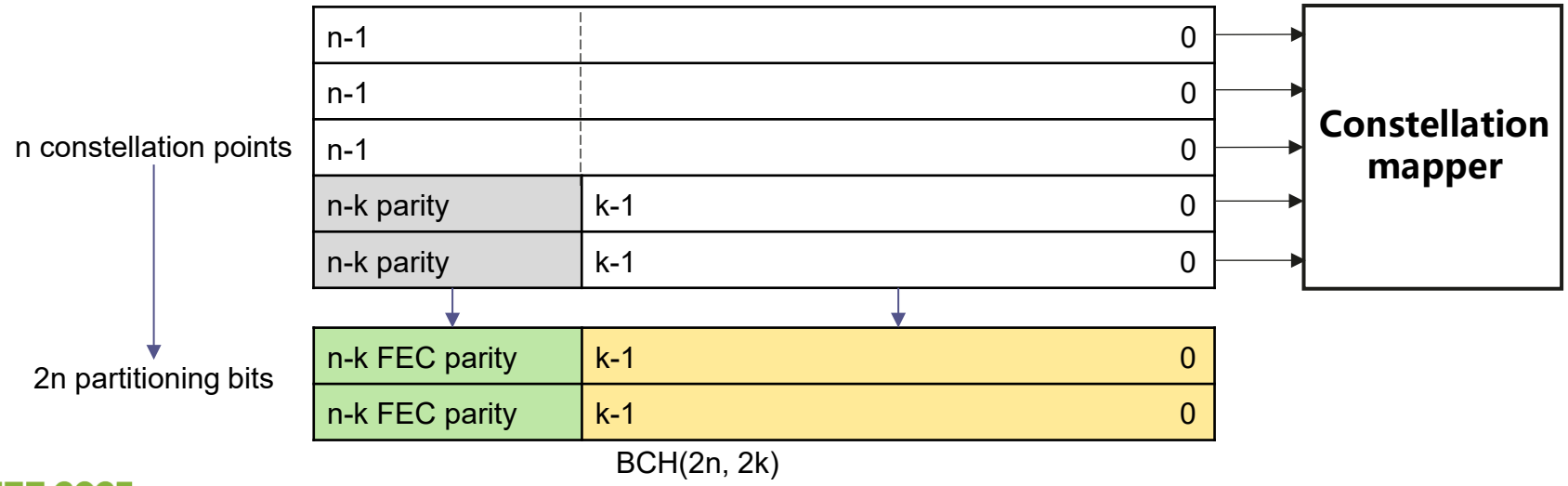


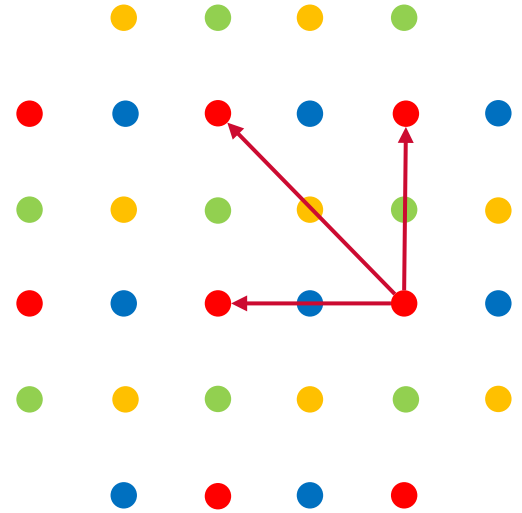
- The error propagation of MLSE corresponds to diagonal constellation errors.
- The diagonal error will be in the same subset with the correct constellation points, thus will have the same partitioning bit.
- → FEC cannot detect or correct the propagation of the diagonal errors.

[1] H. Shakiba, “Analysis of Noise Coloring Effect on MLSE COM Using Error Events”, IEEE 802.3dj task force May 2023

PAM6 Using Inner FEC w/ Set Partitioning – 4 subsets

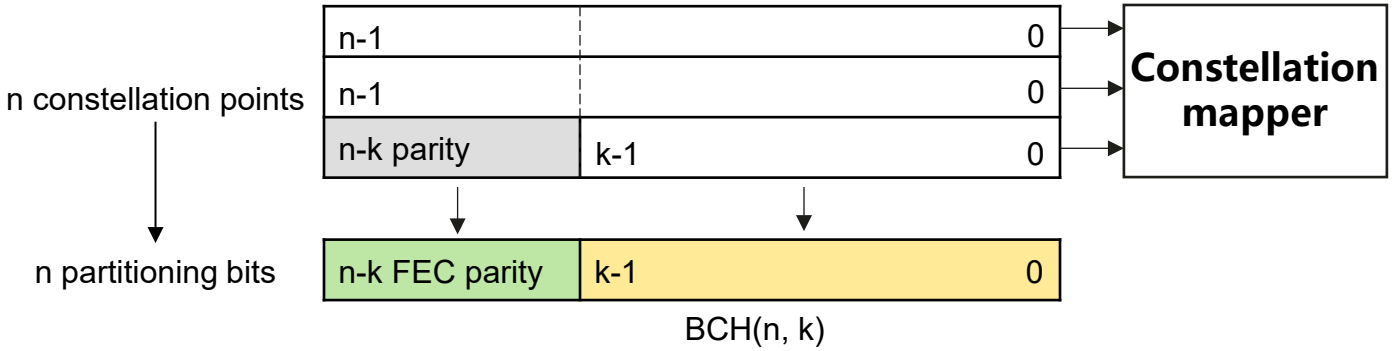
- 2D-PAM6 constellation can be divided into 4 subsets.
- d_{min} of each subset in horizontal, vertical and diagonal directions are all doubled.
- Two partitioning bits required to denote the subset information.
- FEC is used to protect the partitioning bits.
- 2(n-k) parity bits are added such that the corresponding 2n partitioning bits of n constellation points are in a codeword.





PAM8 Using Inner FEC w/ set Partitioning – 2 subsets

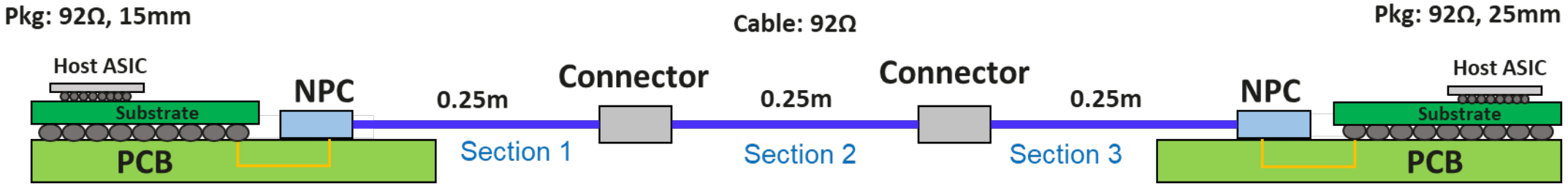
- Gray mapping is considered for PAM8
- Constellation points are divided into 2 subsets[1]
- For a constellation point, XOR of its associated 3 bits gives the partitioning bit
- FEC is adopted to protect the partitioning bits



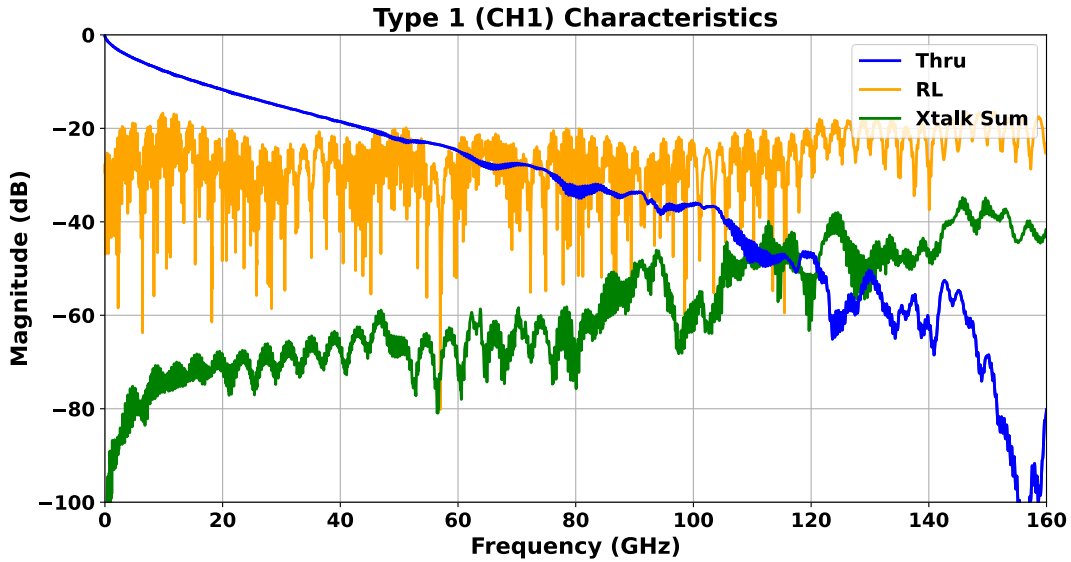
$Subset_0$: $xor(b_2,b_1,b_0) = 0$
$Subset_1$: $xor(b_2,b_1,b_0) = 1$

[1] C. Liu, "Performance Analysis at 400+Gbps Over Next-Generation VSR Channels", Ethernet Alliance Technology Exploration Forum 2024

Example Channel 1#: NPC

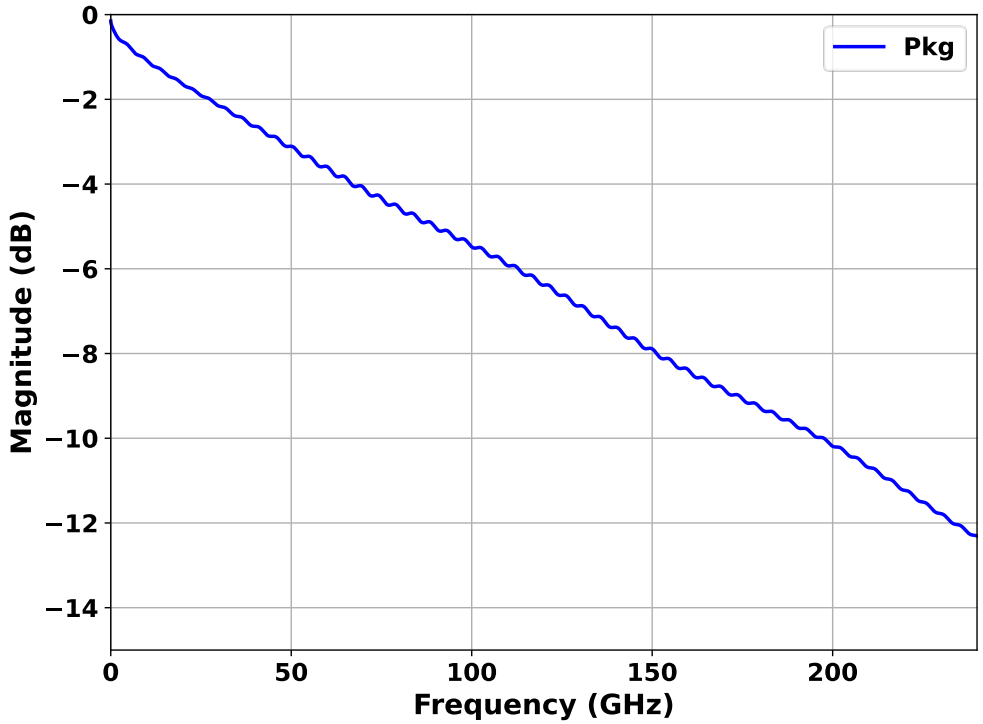


Modulation	Insertion Loss
PAM4	41.3dB
PAM6	33.1dB
PAM8	27.7dB



Package Model Used for Simulation

- Scale from 224G PAM4 based on latest contributions in OIF and IEEE.
- Reference Package Type A with trace length 33mm for both TX and RX.



Modulation	Freq.	Insertion Loss
PAM4- KP4	106.25GHz	5.7dB
PAM6- KP4	80GHz	4.8dB
PAM8- KP4	70.8GHz	4.2dB

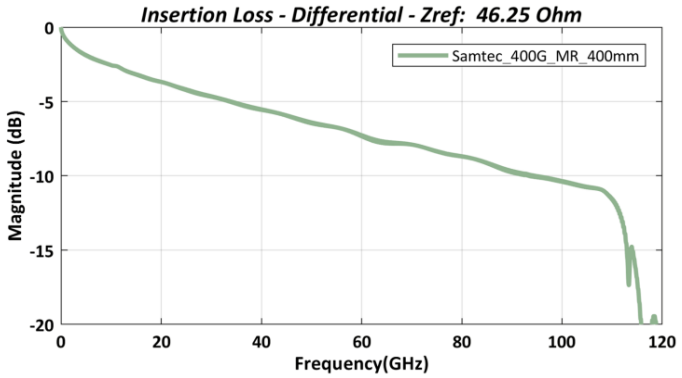
Package model: oif2025.479.00, Q4 2025, Mike Li et al.

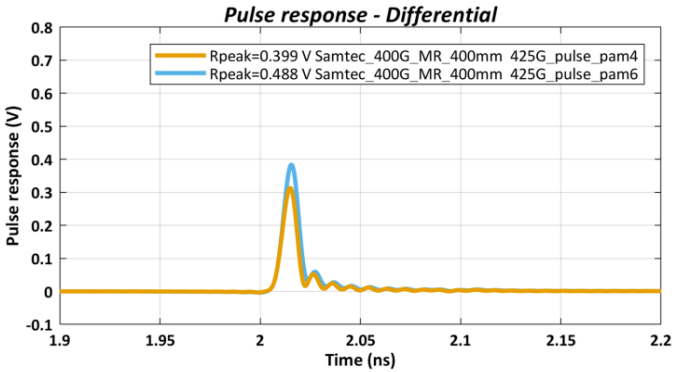
Simulation Results for Channel #1

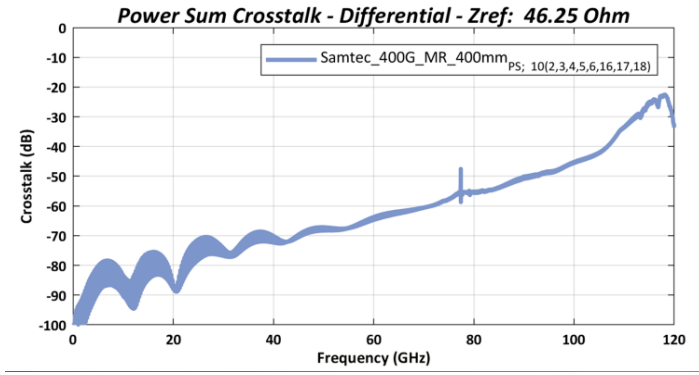
Modulation		PAM4	PAM6			PAM8		
FEC		KP4	KP4	KP4+ SP4(180,170)	KP4+ BCH(128,120)	KP4	KP4+ SP2(126,120)	KP4+ BCH(128,120)
Data rate [Gbps]		425	425	450	454	425	446	454
Insertion Loss [dB]		41.3	33.1	33.7	33.6	27.7	28.7	29.5
ICN@bump [mV]		0.85	0.37	0.42	0.43	0.23	0.24	0.25
Alpha		0.98	0.46	0.51	0.53	0.27	0.35	0.36
BER (MLSE HD)		2.97E-7	1.47E-6	3.62E-6	5.48E-6	1.97E-5	3.37E-5	1.39E-5
BER (after Inner FEC HD)		-	-	7.36E-7	2.37E-7	-	6.83E-7	1.73E-7
BER (after Inner FEC SD)		-	-	<1E-8	<1E-8	-	<1E-8	<1E-8
KP4 error histogram*	Bin 1	1.62E-3	5.54E-3	2.07E-3	7.25E-4	9.05E-2	2.33E-3	6.70E-4
	Bin 2	-	-	-	8.06E-5	6.44E-3	4.46E-4	6.70E-5
	Bin 3	-	-	-	-	5.41E-4	-	-
	Bin 4	-	-	-	-	4.91E-5	-	-

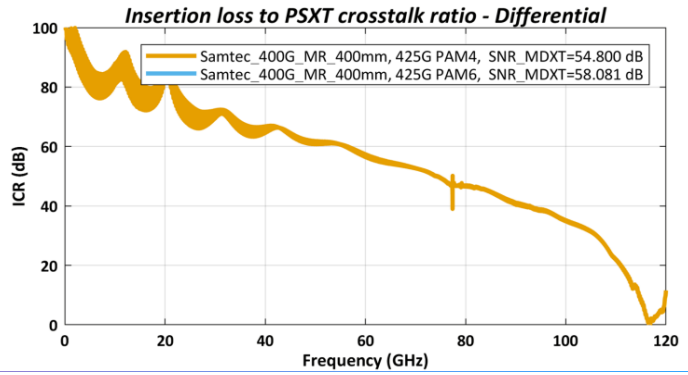
* HD decoded results were used for histogram

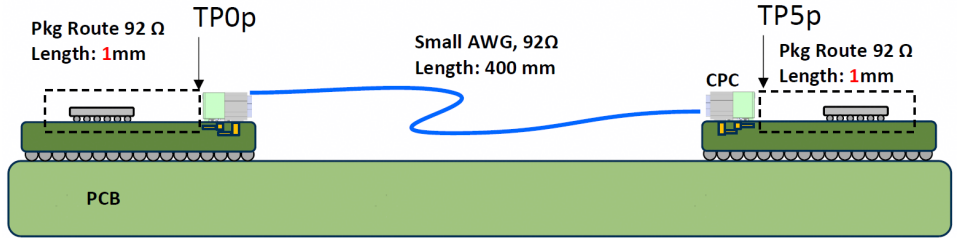
Example Channel #2 - CPC











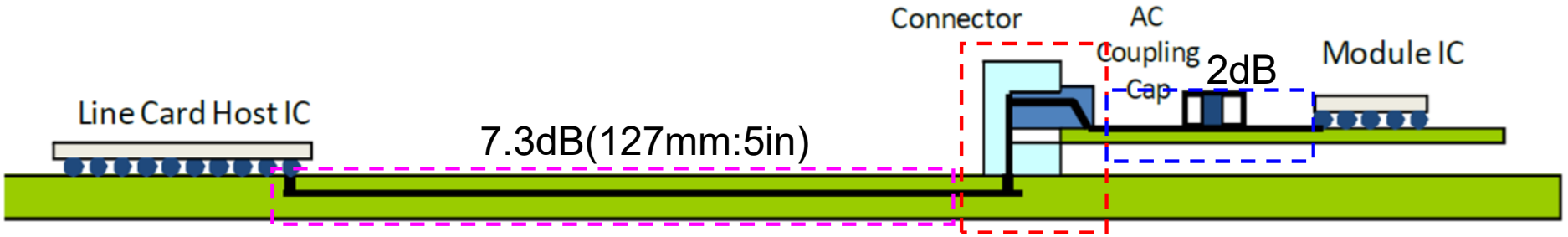
- 400mm CPC channel with smooth roll-off till ~110GHz.
  - Package model not included
- Added our package model on both ends.

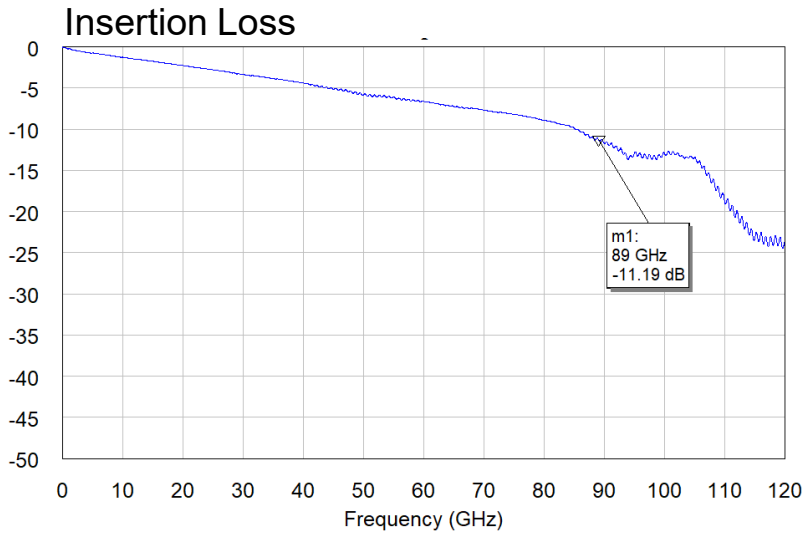
Channel source: "2025114_Samtec_CPC_Channel_Model", Tom Palkert, SNIA/SFF-TA-1043 Copper for AI.

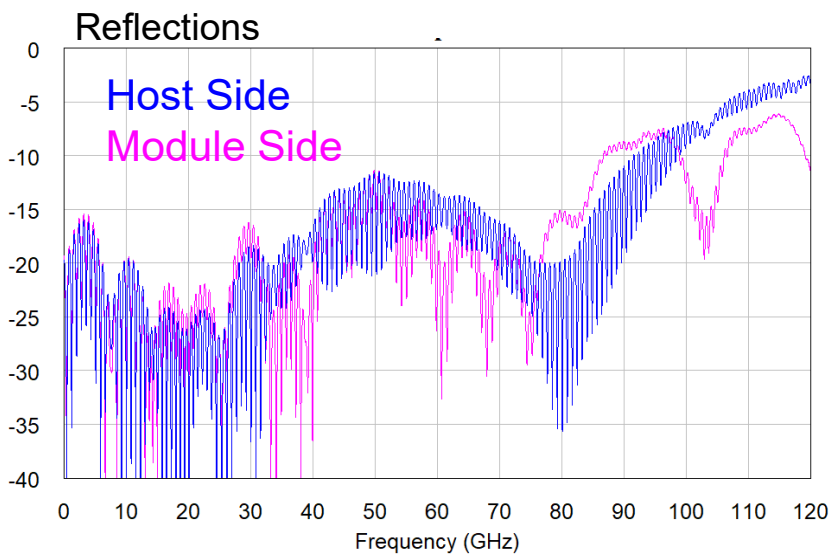
Simulation Results for Channel #2

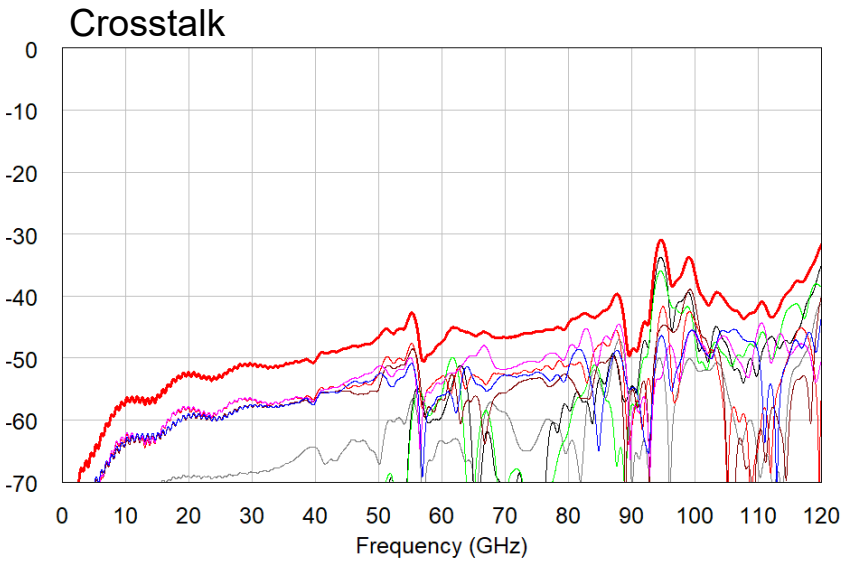
Modulation	PAM4	PAM6			PAM8		
FEC	KP4	KP4	KP4+ SP4(180,170)	KP4+ BCH(128,120)	KP4	KP4+ SP2(126,120)	KP4+ BCH(128,120)
Data rate [Gbps]	425	425	450	454	425	446	454
Insertion Loss [dB]	8.7	7.3	8.0	8.2	6.5	6.7	6.8
ICN@bump [mV]	1.18	0.30	0.39	0.41	0.10	0.12	0.13
Alpha	0.46	0.12	0.14	0.15	0.10	0.12	0.12
BER (w/ MLSE HD)	<1E-8	<1E-8	<1E-8	<1E-8	8.79E-6	1.84E-6	2.09E-6
BER (after inner FEC HD)	-	-	<1E-8	<1E-8	-	2.73E-8	2.46E-8
BER (after inner FEC SD)	-	-	<1E-8	<1E-8	-	<1E-8	<1E-8

Example Channel #3 – VLC PCB VSR channel







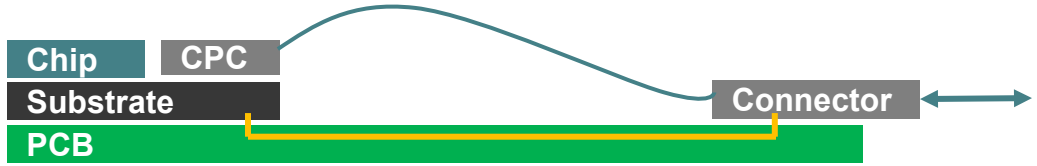


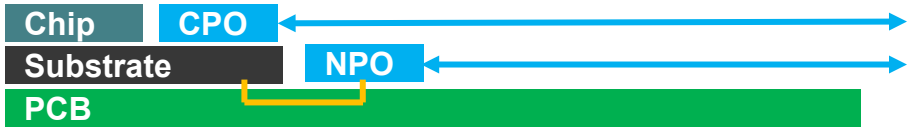
Channel source: OIF2025.178.11, Q4 2025, Toshiyasu Ito, Yamaichi Electronics

Simulation Results for Channel #3

Modulation	PAM4	PAM6			PAM8		
FEC	KP4	KP4	KP4+ SP4(180,170)	KP4+ BCH(128,120)	KP4	KP4+ SP2(126,120)	KP4+ BCH(128,120)
Data rate [Gbps]	425	425	450	454	425	446	454
Insertion Loss [dB]	13.7	9.4	10.3	10.5	7.5	7.9	8.0
ICN@bump [mV]	0.86	0.58	0.61	0.61	0.5	0.5	0.51
Alpha	0.59	0.14	0.13	0.13	0.12	0.13	0.14
BER (MLSE HD)	<1E-8	1.1E-8	2.99E-8	5.73E-8	3.91E-6	1.25E-6	1.77E-6
BER (after inner FEC HD)	-	-	<1E-8	<1E-8	-	2.72E-8	<1E-8
BER (after inner FEC SD)	-	-	<1E-8	<1E-8	-	<1E-8	<1E-8

Application Complications for Different Modulations





PAM4

- **Higher loss** may limit the reach of copper cables. Recent development in the industry provides more confidence.
- **Module connectors** needs improvement to support PAM4 bandwidth. HDC is a viable solution.
- Seems OK technically, but **time-to-market** may be to far away.

PAM6/PAM8

- Additional Inner FEC increases **power consumption in host ASIC**, which can be challenging for highly-integrated switch chips.
- Optical is going with PAM4, adding PAM6/8-to-PAM4 **gear box** with Inner-FEC termination for each C2M section will increase end-to-end **power**.
- **Cannot support CPO/NPO** due to different modulations between E/O.

Summary

- Achieving sufficient channel bandwidth is the primary enabler for a robust 448G/lane electrical interface – as always.
- Maintaining the existing RS(544,514) FEC is highly desirable to preserve architectural continuity and reduce ASIC complexity for next-gen switch designs.
- **PAM4, PAM6, and PAM8 are all feasible** with sufficient bandwidth, but higher-order modulations demand significantly higher SNR, stronger equalization, more FEC gain, etc., resulting in increased power.
- PAM4 remains a feasible option, offering design simplicity, supporting higher levels of integration needed for AI super-pods and large-scale systems.

Channel-Aware Modulation and FEC Selection for 400G+ Ethernet

Tony Chan Carusone
CTO, Alphawave Semi

December 2, 2025

TEF 2025
Ethernet for AI

400G Technology Progress

- Electrical interconnect bandwidth has seen steady progress
  - Channel loss at 106GHz has dropped from 60+dB to 40-45dB
  - Notch frequencies have increased from 90GHz to 110+GHz
- A wide diversity of optical modulation technologies are being looked at for 400G links
  - Mature demonstrations of TFLN and InP modulation bandwidth exceeding 100GHz
  - Research demonstrating SiPho modulation bandwidth of 80+GHz
- This evolution has implications for modulation and FEC choice

Electrical 400G COM Parameters

- COM 4.8 is used in our analysis (Open-source tool): IEEE 802.3 Channel Operating Margin (COM) Open Source Project Ad Hoc Public Area
  - N_qb is used to model ADC quantization noise
- Here are the nominal values of the parameters we set for our analysis:
  a) Percentage reduction in C_d/C_b/Ls parameters compared to 200G parameters
  b) Assuming a 4th order Butterworth filter
  c) DC gain values are [-10:2:0] and up to 10dB of boost is achieved by moving zero to a location below this frequency
- 3dB COM implementation penalty is not included in the bit-error-rate
  - 10^{-7} is roughly equivalent to 3dB COM margin for BER=2.4e-4

Parameter	Nominal Value
Front-end Improvement[a]	40%
TX SNR	33dB
TX RLM	0.95
RX Bandwidth[b]	100 GHz
CTLE P1/Z[c]	75GHz
CTLE P2	140GHz
No of FFE Pre-cursors	20
No of FFE Post-cursors	50
ADC ENOB	7bit
RX Noise Density	4e-9 V^2/GHz
Random Jitter	70fs
Dual-Dirac Jitter	150fs

Simulation Setup – COM Parameters

Table 93A-1 parameters			
Parameter	**Setting**	**Units**	**Information**
f_b	See Prev. Slide	GBd	
f_min	0.05	GHz	
Delta_f	0.01	GHz	
C_d	See Prev. Slide	nF	[TX RX]
L_s		nH	[TX RX]
C_b		nF	[TX RX]
R_0	50	Ohm	
R_d	[50 50]	Ohm	[TX RX]
PKG_NAME	**PKG_MODEL PKG_MODEL**		TX RX
z_p select	1		[test cases to run]
L	4		
M	32		
filter and Eq			
f_r	See Prev. Slide	*fb	
c(0)	0.55		min
c(-1)	[-0.4:0.05:0]		[min:step:max]
c(-2)	[0:0.05:0.1]		[min:step:max]
c(-3)	0		[min:step:max]
c(1)	0		[min:step:max]
N_b	1	UI	
b_max(1)	0.75		As/dffe1
b_max(2..N_b)	0.3		As/dfe2..N_b
b_min(1)	0		As/dffe1
b_min(2..N_b)	-0.15		As/dfe2..N_b
g_DC	[-10:2:0]	dB	[min:step:max]
f_z	75	GHz	
f_p1	75	GHz	
f_p2	140	GHz	
g_DC_HP	0		[min:step:max]
f_HP_PZ	1.32815	GHz	
Bessel_Thomson	0	logical	Bessel filter
Raised_Cosine	0	logical	RaisedCosine filter
Butterworth	1	logical	Butterworth filter
RC_Start	6.70E+10	Hz	start freq for RCos
RC_end	1.23E+11	Hz	end freq for RCos

.START	PKG_MODEL	

Table 93A–3 parameters			
Parameter	**Setting**	**Units**	Information
package_tl_gamma0_a1_a2	[5e-4 0.00065 0.000293]		
package_tl_tau	0	ns/mm	
package_Z_c	[87.5 87.5; 95 95; 100 100; 100 100]	Ohm	
R_d	[50 50]	Ohm	
z_p (TX)	[0; 0; 0; 0]	mm	[test cases]
z_p (NEXT)	[0; 0; 0; 0]	mm	[test cases]
z_p (FEXT)	[0; 0; 0; 0]	mm	[test cases]
z_p (RX)	[0; 0; 0; 0]	mm	[test cases]
C_p	[0 0]	nF	[TX RX]
A_v	0.413	V	vp/vf=
A_fe	0.413	V	vp/vf=
A_ne	0.45	V	

.END		

I/O control		
DIAGNOSTICS	0	logical
DISPLAY_WINDOW	0	logical
CSV_REPORT	0	logical
RESULT_DIR	\results\400g\400G_{date}\	
SAVE_FIGURES	0	logical
Port Order	**[1 3 2 4]**	
RUNTAG	400G_sanity_	
COM_CONTRIBUTION	0	logical
TDR and ERL options		logical
TDR	0	logical
ERL	0	logical
ERL_ONLY	0	ns
TR_TDR	0.01	
N	4000	logical
TDR_Butterworth	1	
beta_x	0	
rho_x	0.618	
TDR_W_TXPKG	0	UI
N_bx	20	
fixture delay time	[0 0]	
Tukey_Window	1	
Noise, jitter		UI
sigma_RJ	See Prev. Slide	UI
A_DD		V^2/GHz
eta_0	4.00E-09	dB
SNR_TX	33	
R_LM	0.95	
11-2022 BenArtsi pkg	oif2022.065.02	

Table 93A–3 parameters			
Parameter	**Setting**	**Units**	Information
package_tl_gamma0_a1_a2	[5e-4 0.00065 0.0003]		
package_tl_tau	0.006141	ns/mm	
package_Z_c	[92 92 ; 70 70; 80 80; 100 100]	Ohm	
z_p (TX)	[12 30 45; 1 1 1 ; 1 1 1 ; 0.5 0.5 0.5]	mm	[test cases]
z_p (NEXT)	[12 30 45; 1 1 1 ; 1 1 1 ; 0.5 0.5 0.5]	mm	[test cases]
z_p (FEXT)	[12 30 45; 1 1 1 ; 1 1 1 ; 0.5 0.5 0.5]	mm	[test cases]
z_p (RX)	[12 30 45; 1 1 1 ; 1 1 1 ; 0.5 0.5 0.5]	mm	[test cases]
C_p	[0.4e-4 0.4e-4]	nF	[TX RX]
A_v	0.413	V	vp/vf=
A_fe	0.413	V	vp/vf=
A_ne	0.45	V	
Operational			
ERL Pass threshold	10	dB	
COM Pass threshold	3	db	
DER_0	2.40E-04		
T_r	2.35E-03	ns	
FORCE_TR	1	logical	
PMD_type	C2C		
EW	1		
MLSE	1		
ts_anchor	1		
sample_adjustment	[-8 8]		
Local Search	2		
DER_CDR	1.00E-02		Maximimum DER_DFE that MLSE will be evaluated
Q	0.00E+00	ns	MLSE Implementation penalty
Filter: RX FFE			
ffe_pre_tap_len	20	UI	
ffe_post_tap_len	50	UI	
ffe_pre_tap1_max	1		
ffe_post_tap1_max	1		
ffe_tapn_max	1		
FFE_OPT_METHOD	MMSE		FV-LMS or MMSE
num_ui_RXFF_noise	2048		
T_O	0	mUI	Needed for C2M VEC calculations
Floating Tap Control			0 1 2 or 3 groups
N_bg	0		taps per group
N_bf	4		UI span for floating taps
N_f	80		max DFE value for floating taps
bmaxg	0.2		rss tail tap limit
B_float_RSS_MAX	0.1		start of tail taps limit
N_tail_start	25	UI	

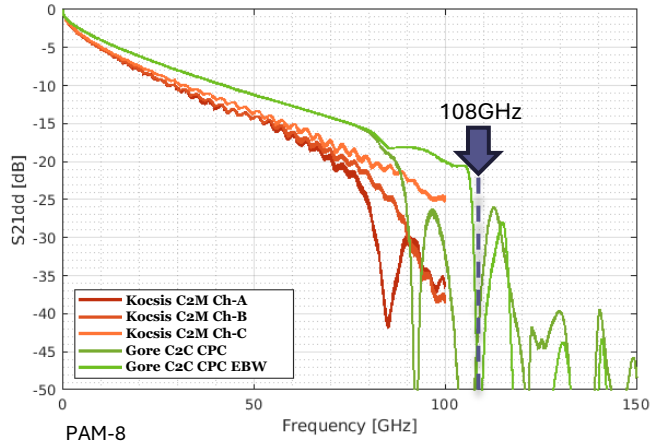
SAVE_CONFIG2MAT	0	
Receiver testing		
RX_CALIBRATION	0	logical
Sigma BBN step	5.00E-03	V
ICN parameters		
f_v	0.139	Fb
f_f	0.139	Fb
f_n	0.139	Fb
f_2	123.250	GHz
A_ft	0.450	V
A_nt	0.450	V

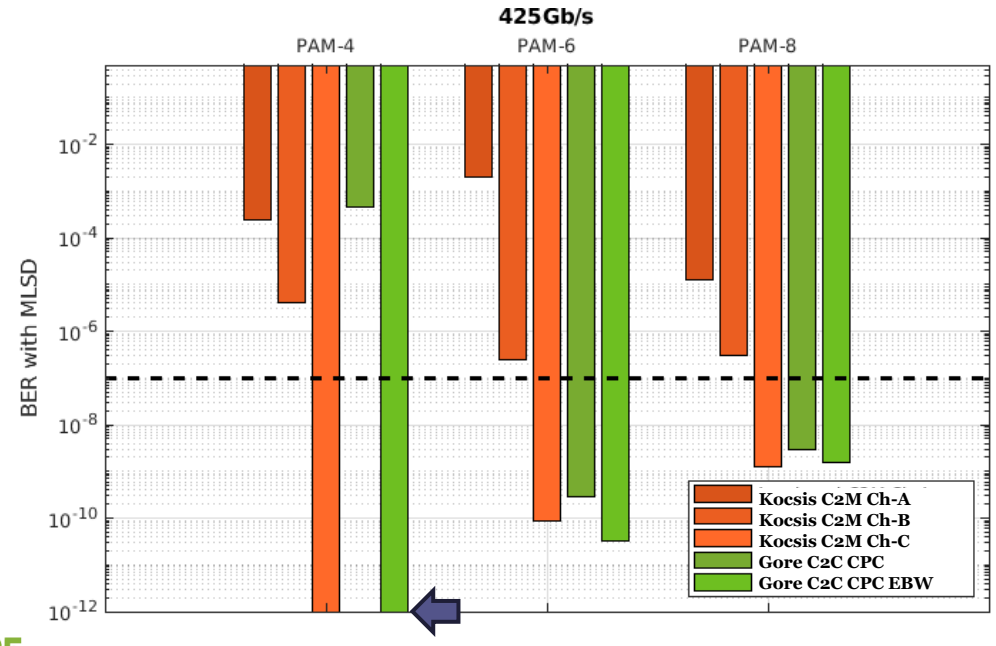
Parameter	**Setting**	
board_tl_gamma0_a1_a2	*[0 6.44084e-4 3.6036e-05]*	1.5 db/in @ 56G
board_tl_tau	*5.790E-03*	ns/mm
board_Z_c	100	Ohm
z_bp (TX)	*32*	mm
z_bp (NEXT)	*32*	mm
z_bp (FEXT)	*32*	mm
z_bp (RX)	*32*	mm
C_0	[0.2e-4 0]	nF
C_1	[0.2e-4 0]	nF
Include PCB	**0**	**logical**
Seletions (rectangle, gaussian,dual_rayleigh,triangle		
Histogram_Window_Weig	gaussian	selection
Qr	*0.02*	UI

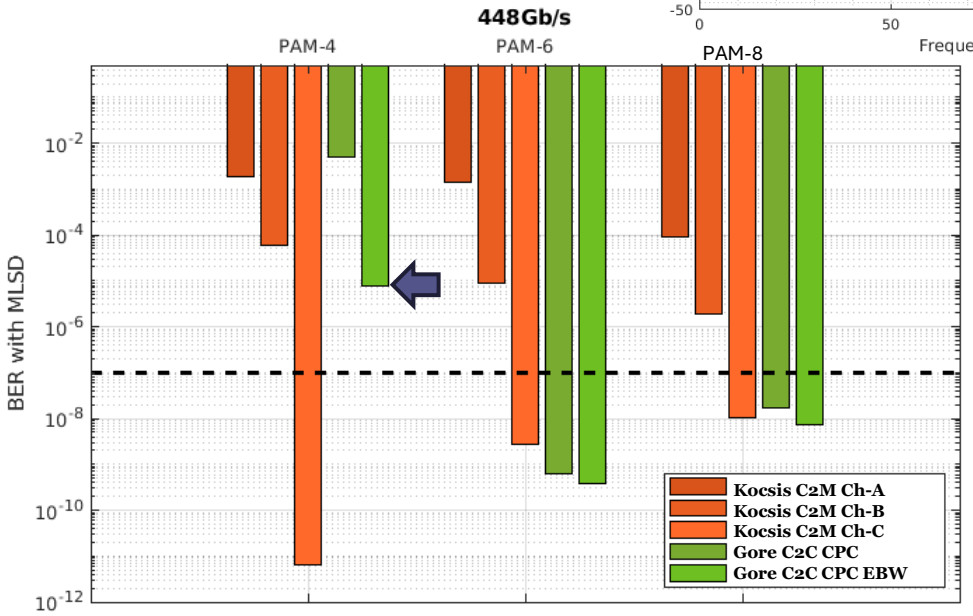
Pre-FEC BER Results

- CPC-based C2C (0.5m) and C2M (0.4m) channels
- ➢With 7% additional overhead, a pronounced pre-FEC BER increase is observed for some channels

IEEE E4AI Channel Data:
gore_e4ai_01a_250529
kocsis_e4ai_01_250327



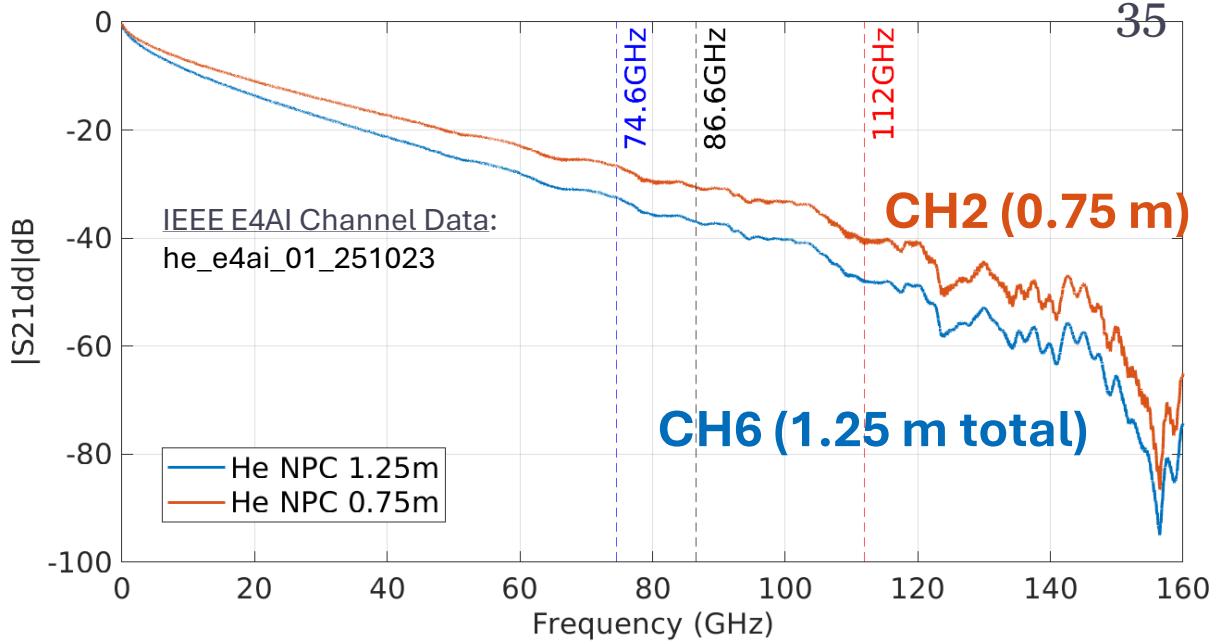


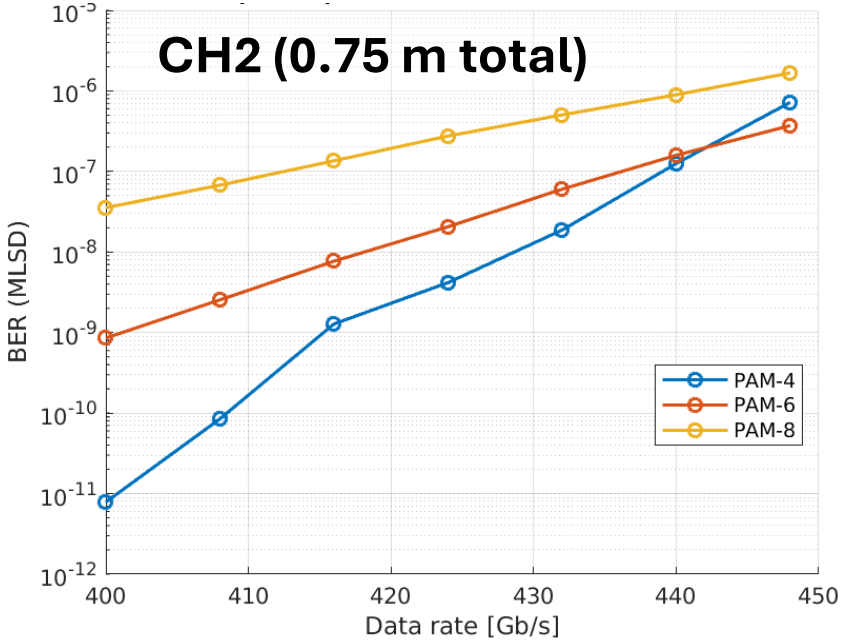


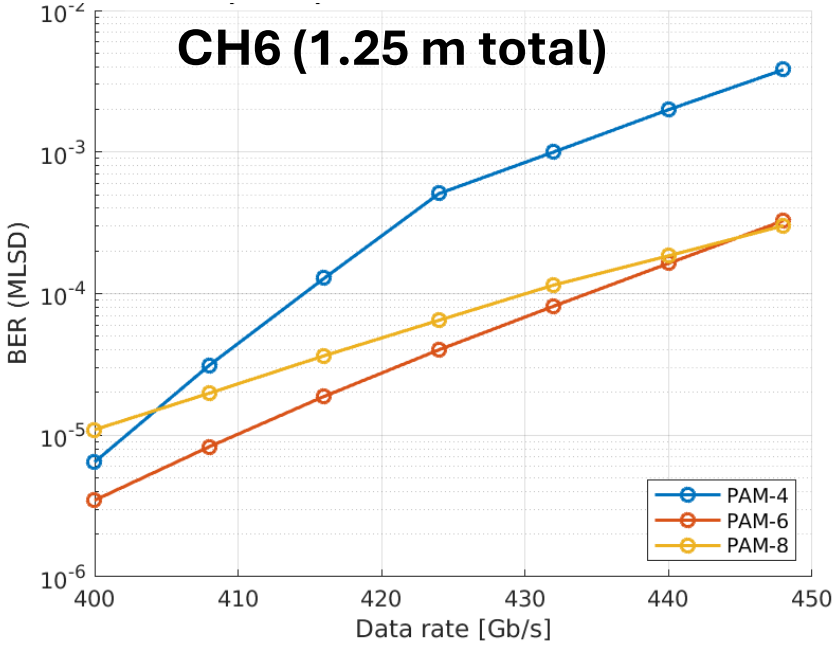
TEF 2025
Ethernet for AI

Pre-FEC BER Results

- Channels composed of NPC and 1 or 2 connectors
- ➢ Less pronounced, though still significant, pre-FEC BER increase is observed for channels with smoother frequency response



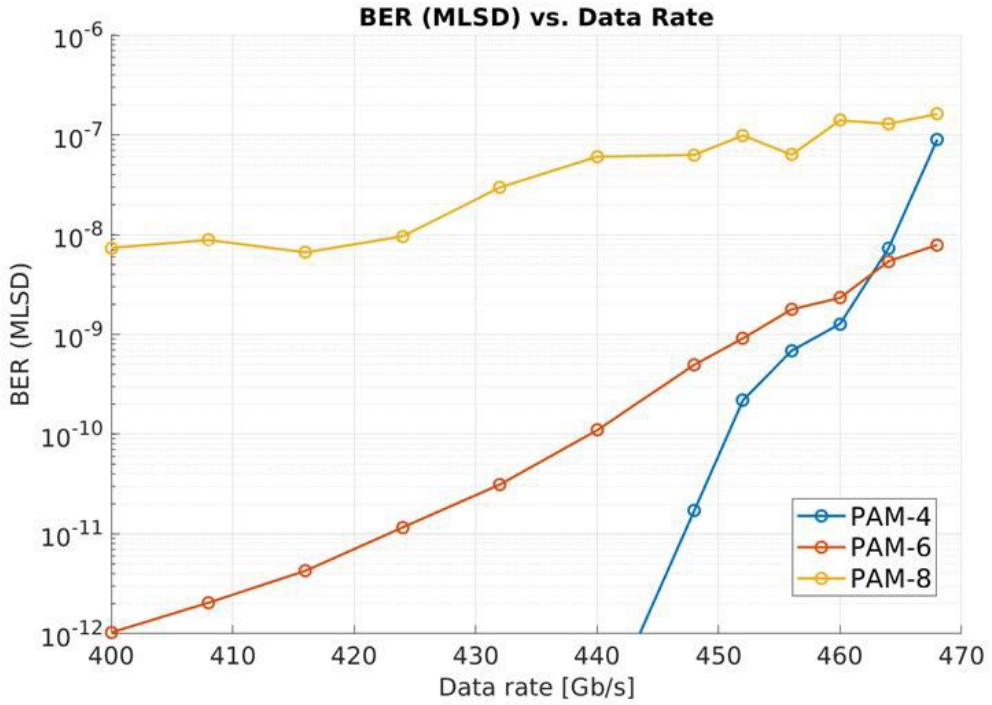


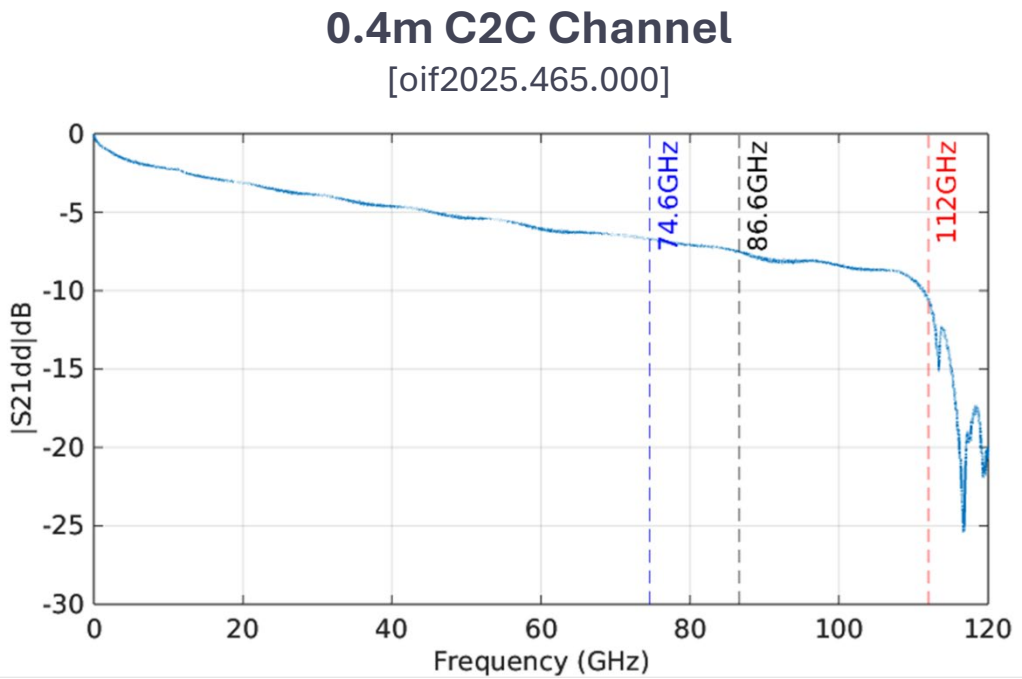


Pre-FEC BER Results

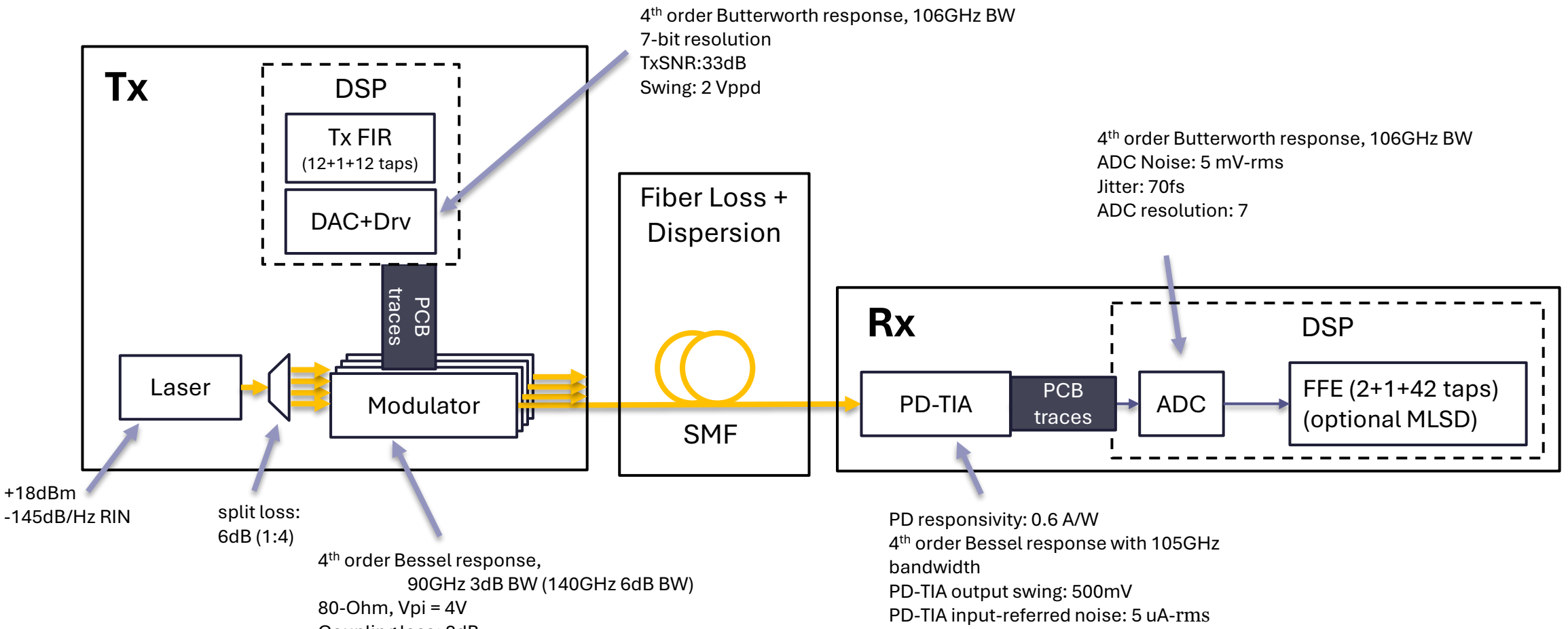
- Optimized CPC-based 0.4m C2C channel

➢A 7% overhead costs 1 to 1.5dB COM for PAM-4

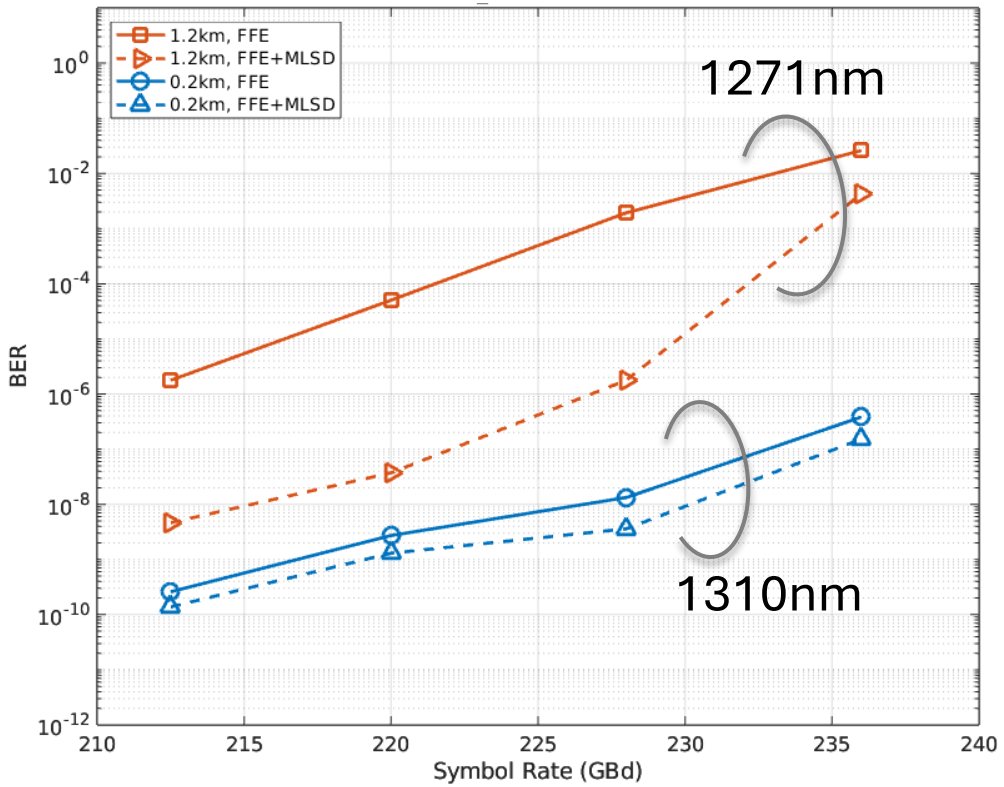




4-PAM Optical Link Model



Pre-FEC BER Results

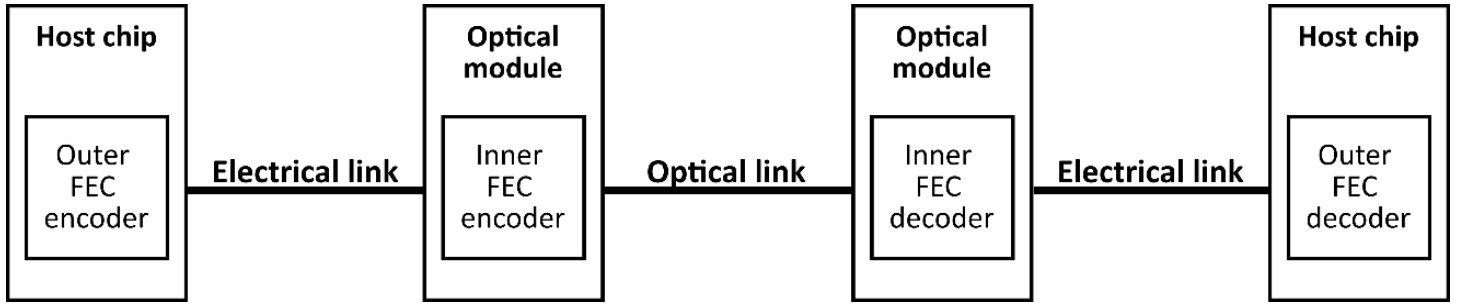


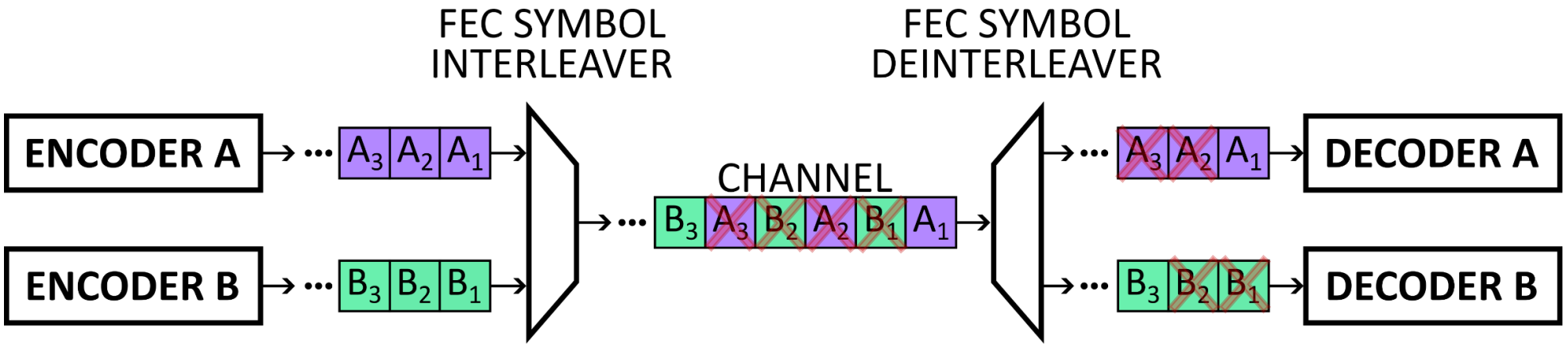
- BER increases up to 5 orders of magnitude with 7% additional overhead

Concatenated FEC

- Introduction of concatenated FEC at 200G
  - Additional coding gain was afforded only to the optical links
- Correlated errors may be introduced by the channel and inner FEC decoder

Example: 200G FECi

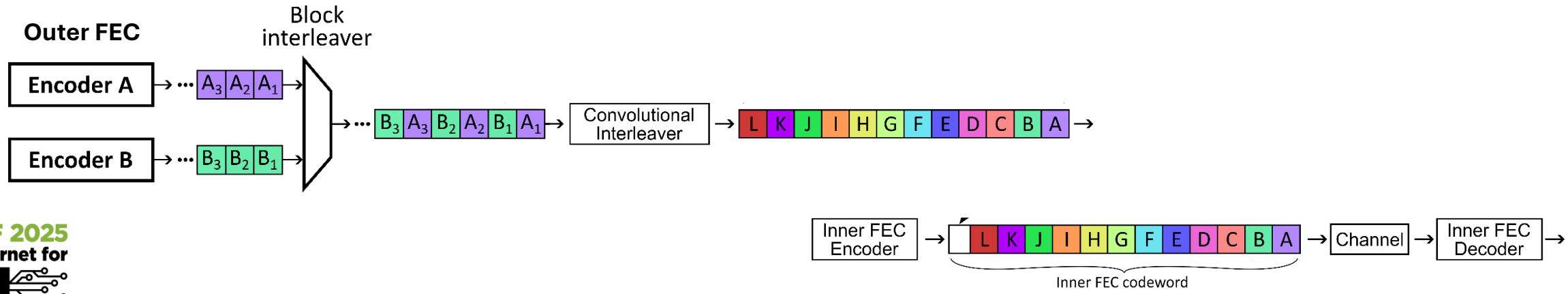




Example: 2-way block interleaving

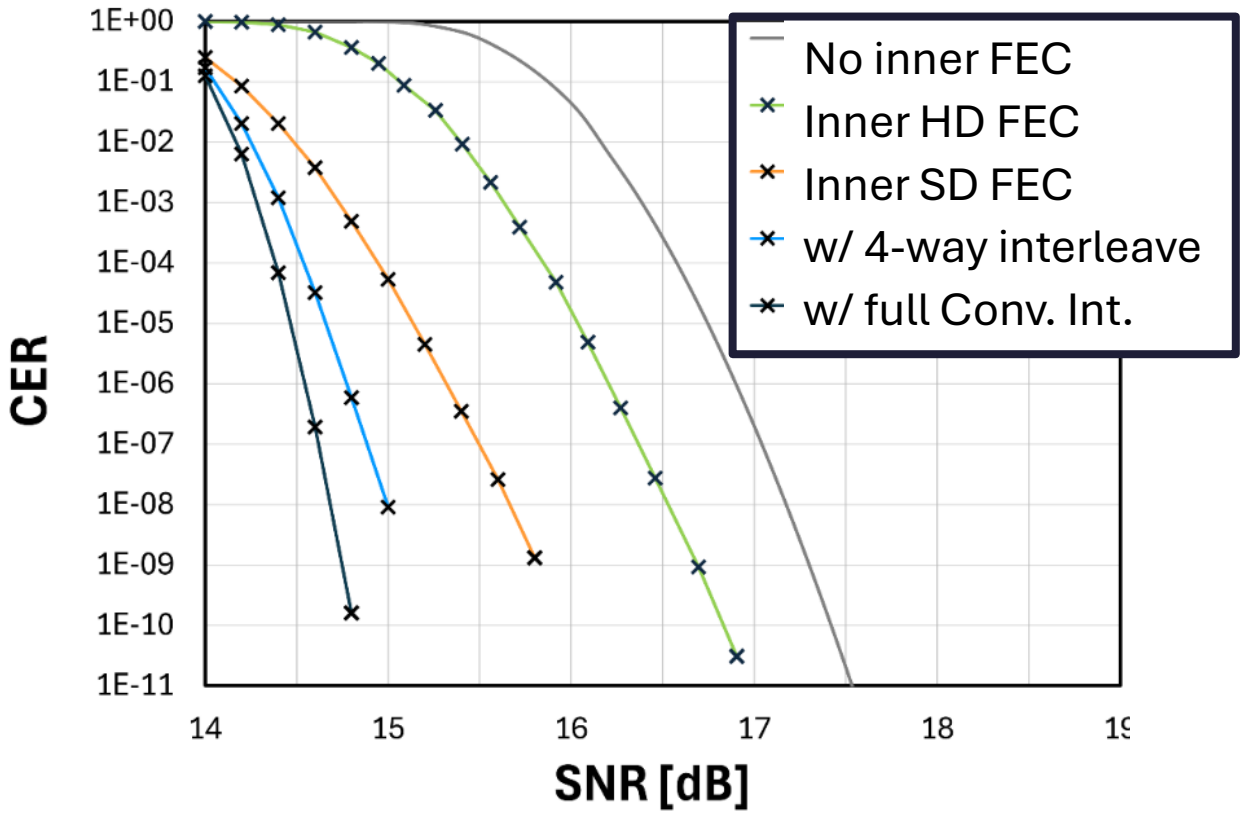
Interleaving

- Interleavers spread correlated errors across many outer codewords
- Interleavers add power, area, and latency
- Optional hardware bypasses allow for tradeoffs in performance vs. power & latency
- Example below: combined block & convolutional interleaving



AWGN Channel Example

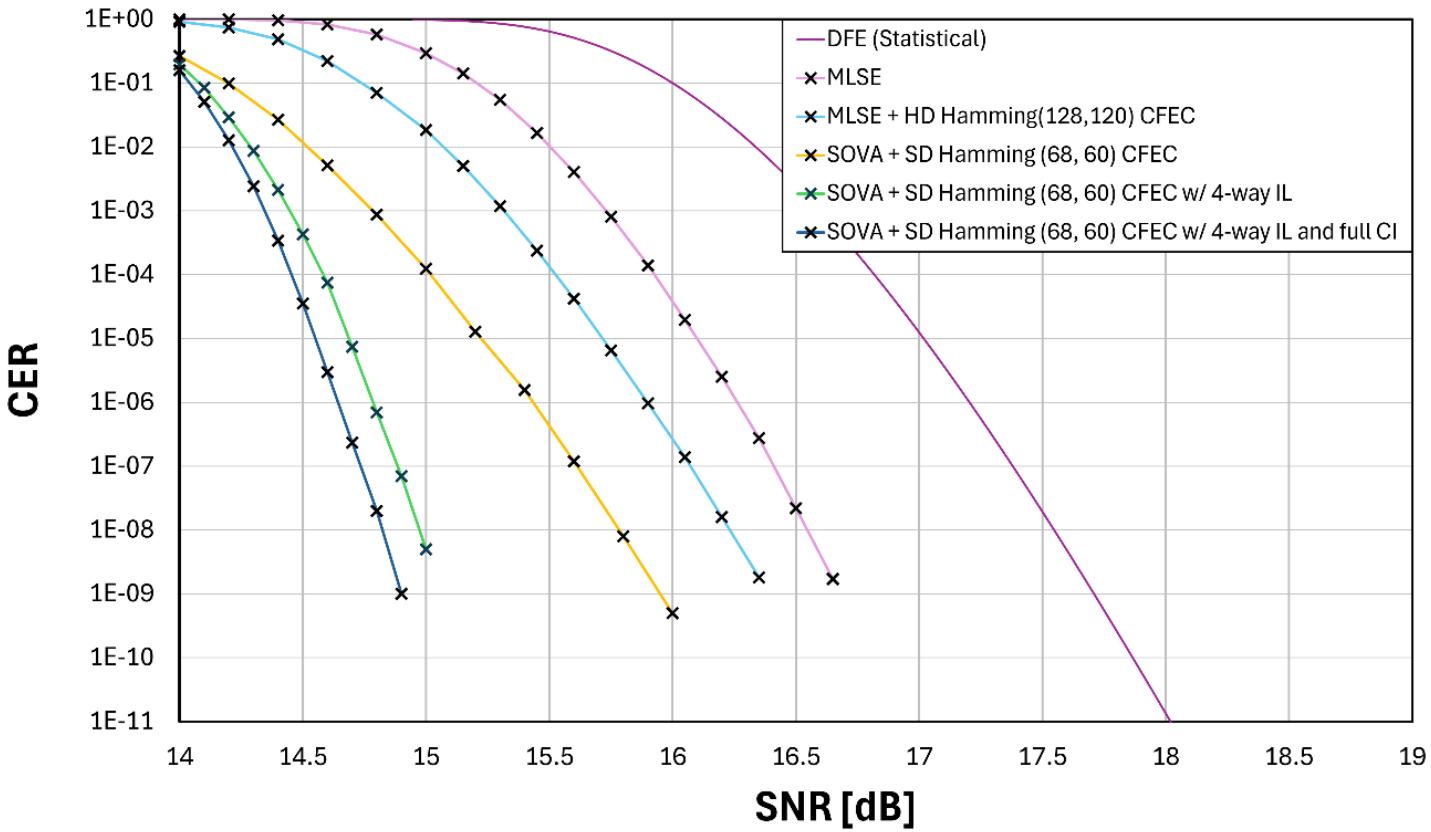
- Approximately 2.5dB SNR performance scaling can be obtained in the FEC alone for an AWGN channel depending on:
  - Inner FEC enablement & decoding
  - Interleaver complexity



[Barrie et al, DesignCon 2025]

$(1 + 0.5z^{-1})$ partial response channel w/ MLSD

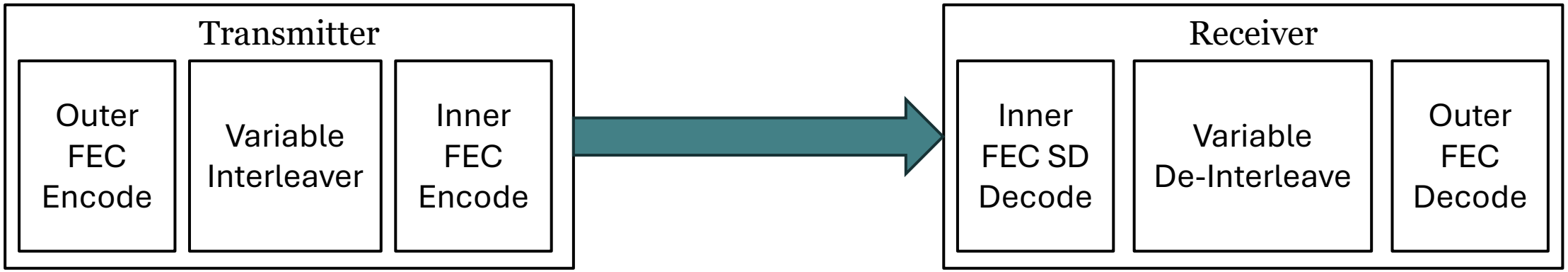
- Similar performance scaling is observed for a partial response channel
- Note, performance with the interleaver is practically the same as the AWGN channel



[Barrie et al, DesignCon 2025]

FEC Responsive to Channel Conditions

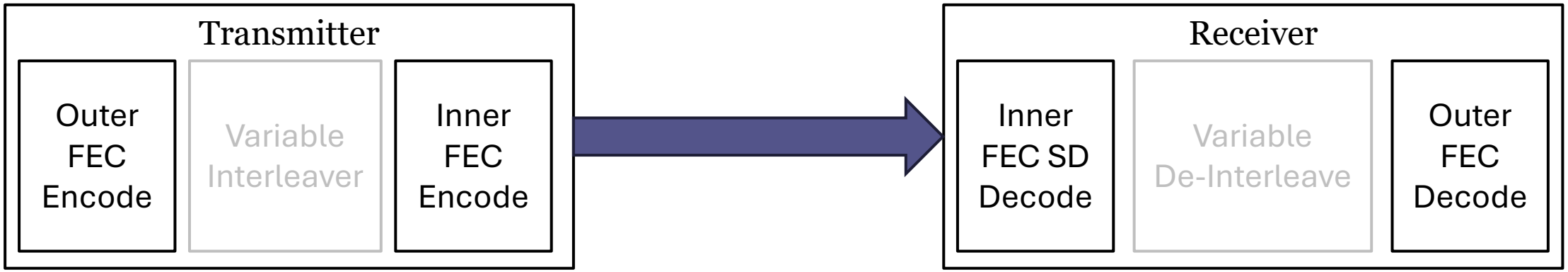
- Challenging links can benefit from soft decoding of an inner FEC plus a sufficient interleaver to maximize coding gain



Full Coding Gain

FEC Responsive to Channel Conditions

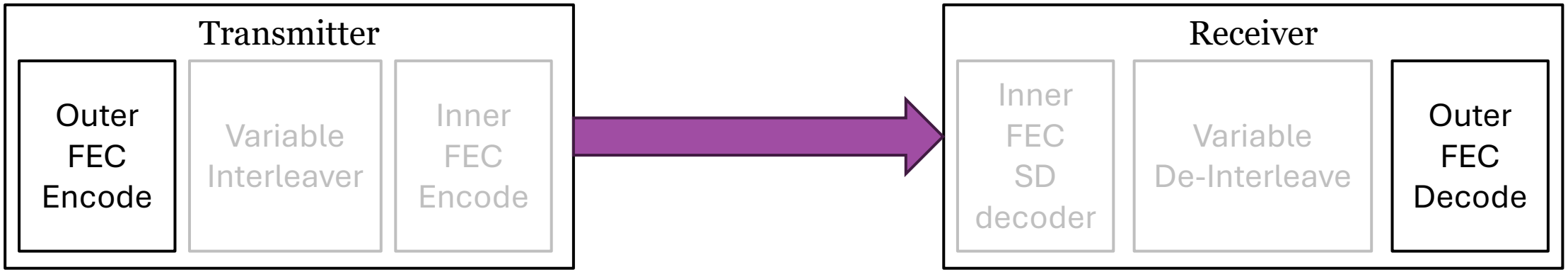
- Improved channel responses can benefit from lower latency bypassing the interleaver & de-interleaver



Low Latency

FEC Responsive to Channel Conditions

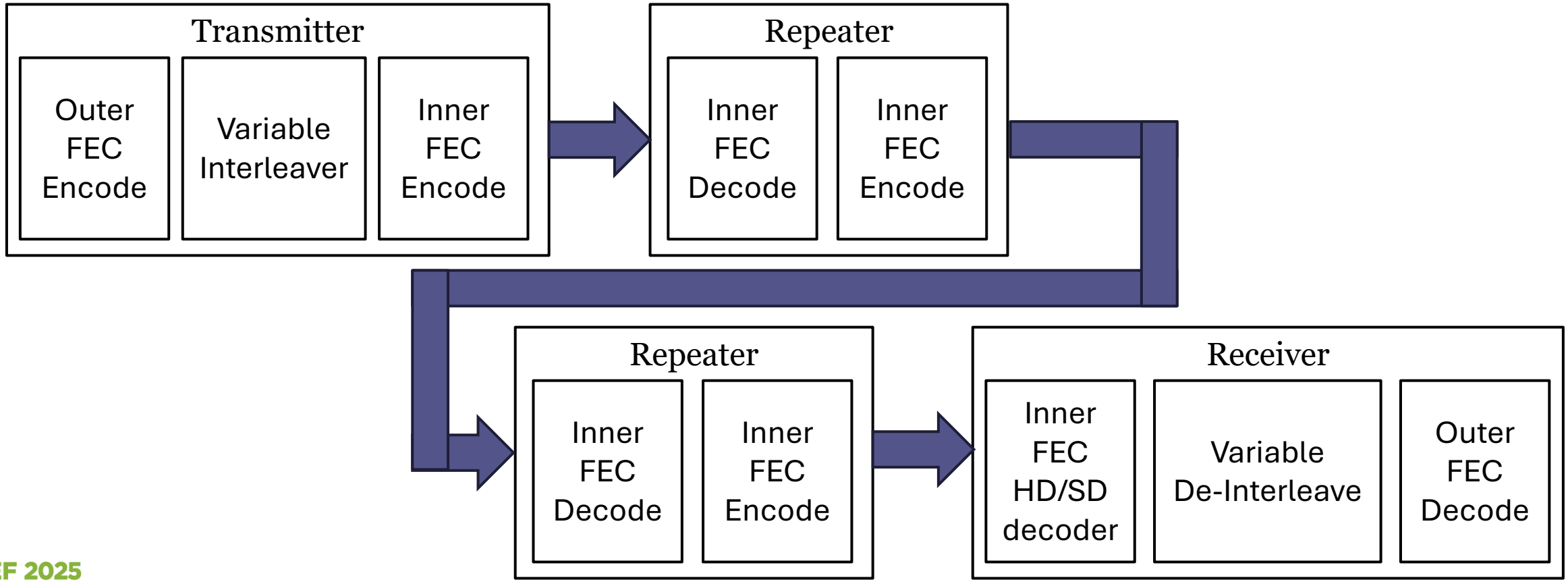
- Highly optimized interconnect may completely bypass the inner FEC to minimize latency and power consumption



Minimum Latency & Power

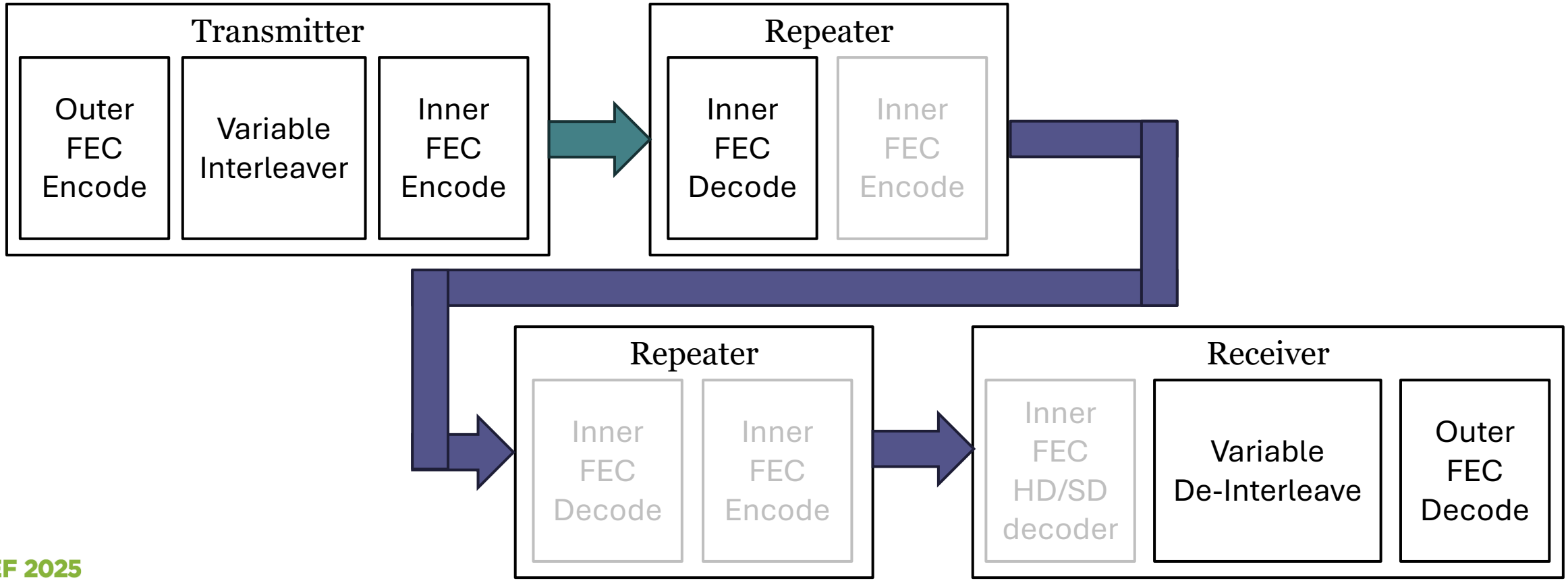
Repeater Links

FEC termination can be introduced to allow for tailored solutions



Repeater Links

For example, maximize coding gain on only the challenging PHY



Conclusion

- 400G component and interconnect technologies continue to evolve
- Most powerful FECs (e.g. staircase, OFEC) introduce added power & latency that is likely to be unacceptable
- KP4 RS outer FEC provides a strong basis and backward compatibility
- Concatenating an inner FEC (e.g. 200G FECi) provides extensibility
- Can be made to support different modulations and/or inner FECs for different physical layers
- Allows for flexibly trading coding gain for power and/or latency reductions
  - Hard / Soft inner decoding
  - Variable length (or bypassed) interleaving

Benefits and limitations of inner error-correcting codes for 400 Gb/s per lane electrical links

Adam Healey
Fellow, Physical Layer Products Division
Broadcom Inc.

TEF 2025
Ethernet for AI

Electrical interface evolution

Data rate per lane, Gb/s	10	25	50	100	200	400
Modulation	PAM-2	PAM-2	PAM-4	PAM-4	PAM-4	?
Nominal cable reach, m	7	5	3	2	1~2[1]	?
Technology added	DFE	RS FEC	Stronger RS FEC	Floating-tap DFE	MLSD	Inner FEC?
Year[2]	2007	2014	2018	2022	2026 (est.)	?

- New tools pulled from the toolbox with each generation
- Used to address challenges presented by doubling the data rate
- What if the next set of tools includes an inner error correcting code?

TEF 2025 Ethernet for AI

[1] Nominal cable reach is a function of host class
[2] Year of publication of the IEEE 802.3 Ethernet standard.

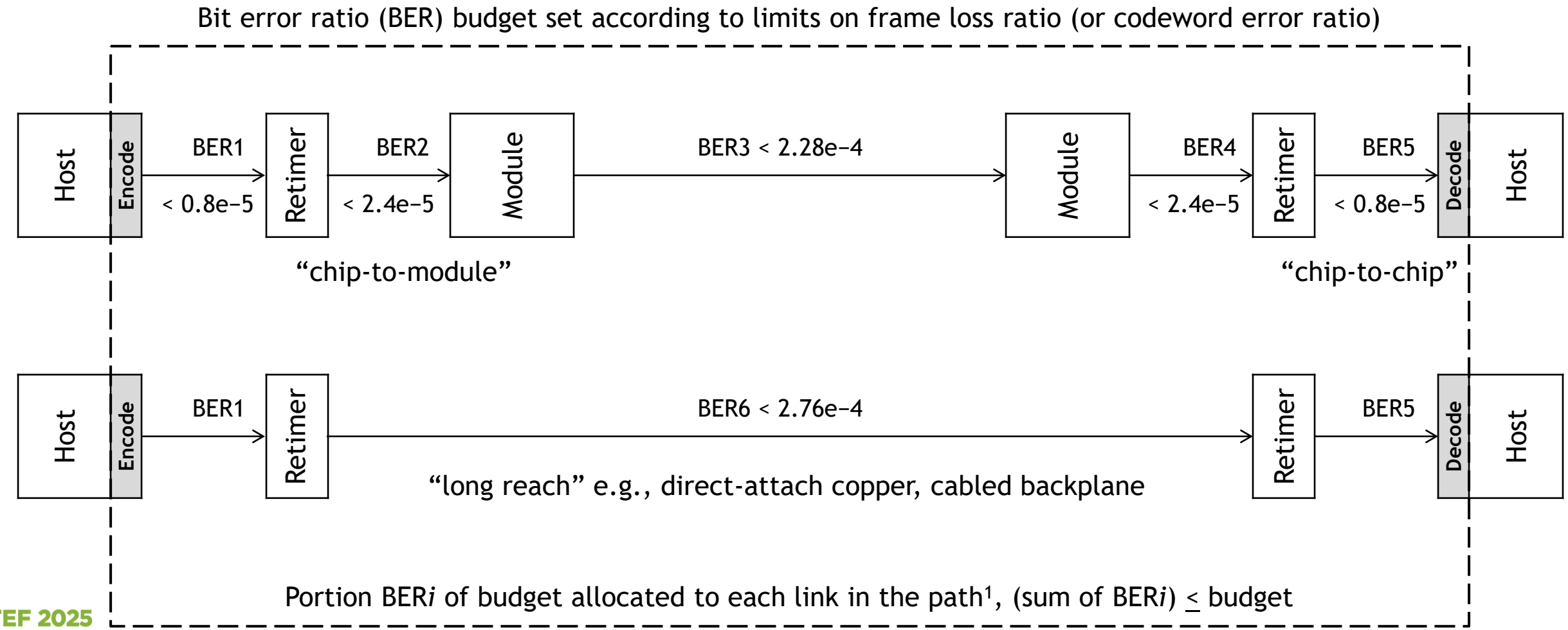
DFE = decision feedback equalizer
FEC = forward error correction
MLSD = maximum-likelihood sequence detector
RS = Reed-Solomon

Inner error correcting code overview

- Not a new concept, but relatively new to Ethernet
- Introduced for challenging 200G/lane IM-DD[1] optical links
- Improves the SNR[2] margin of a link
- Improvements may enable extension of link distance
- Can enable 400G/lane links to fit within the established infrastructure
- Can be by-passed on higher-performing links to reduce latency/power
- Design is closely tied to modulation

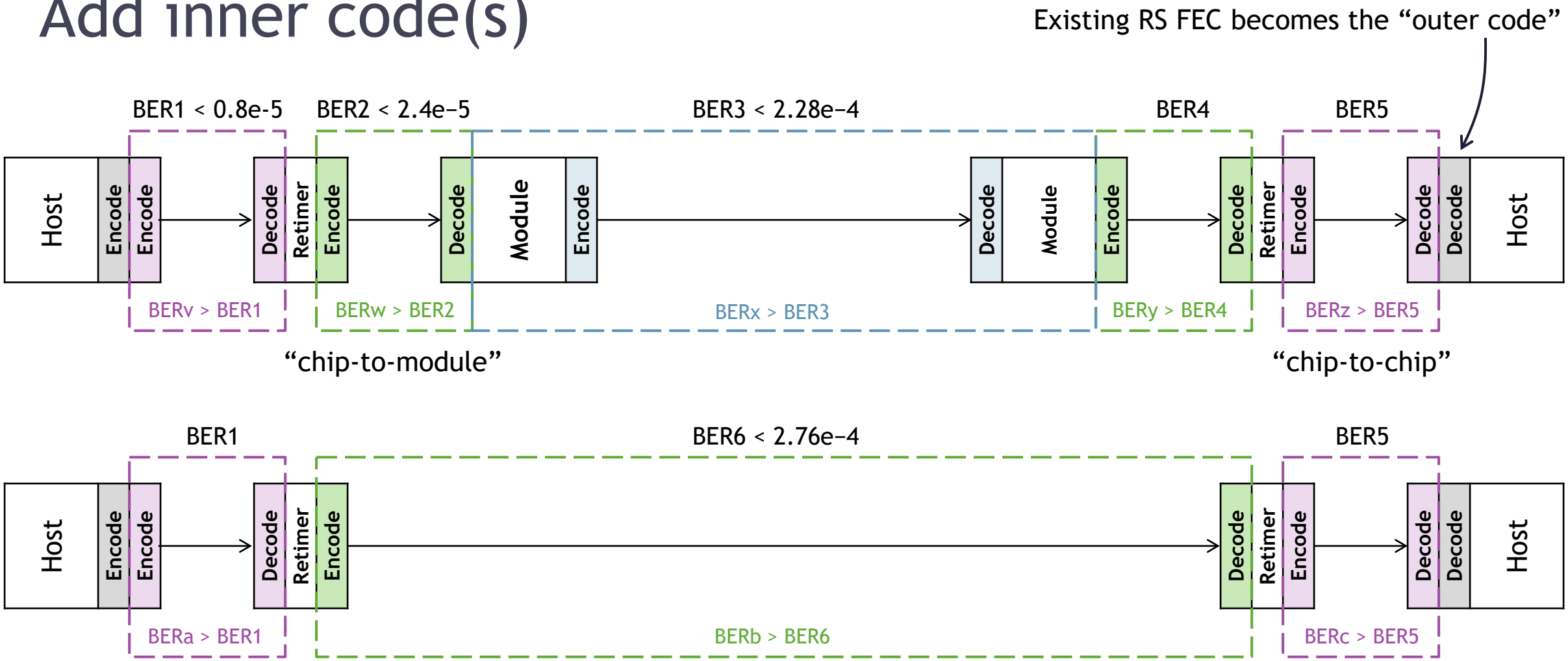
[1] Intensity modulation, direct detection
[2] Signal-to-noise ratio

Established Reed-Solomon encoding infrastructure

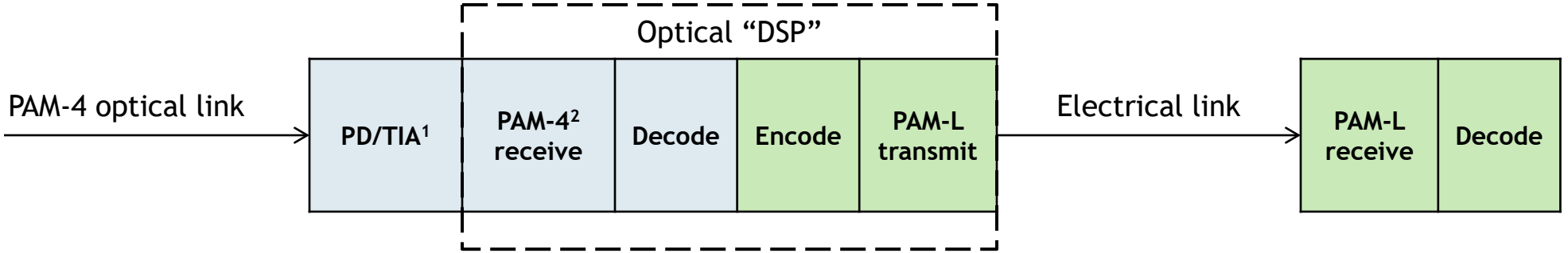


[1] Link BER may need to be less than BER*i* if errors occur in a way that impairs the performance of the decoder

Add inner code(s)



Modulation considerations

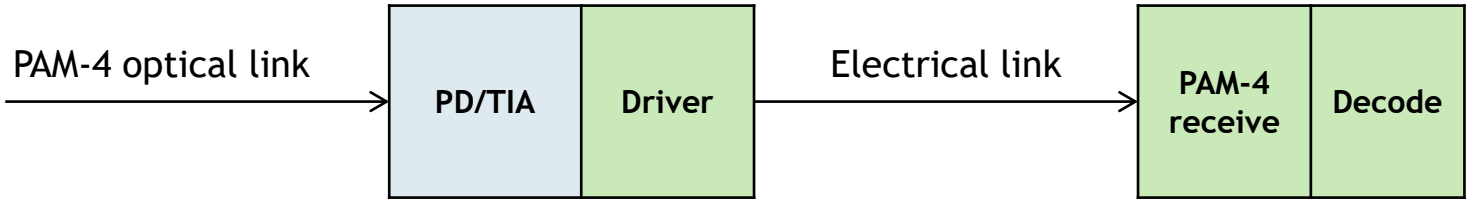


- While 400G/lane IM-DD optical links are expected to use PAM-4, electrical link encoding/modulation can be freely chosen to address the challenges presented by the electrical channel

Additional complexity?	Retimers/gearboxes include sophisticated signaling processing engines and are becoming increasingly data-aware. Incremental increase in complexity can be justified by higher performance.
Need to decode and re-encode data?	Soft-decision decoding needs to be done near the receiver, so the inner code will likely be decoded anyway.
Need to reconcile two different signaling rates?	Solved problem at 200G/lane. Consider a 113.475 GBd optical link with inner code served by a 106.25 GBd chip-to-module link.

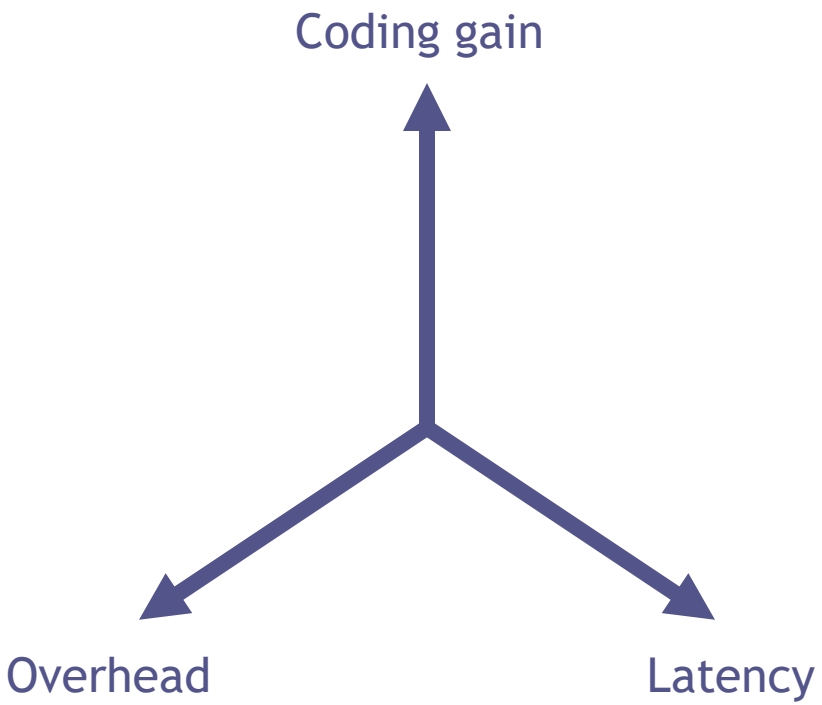
[1] Photodetector and transimpedance amplifier
[2] 4-level pulse amplitude modulation

One notable exception…



- A linear optical receiver would require the electrical link to support the same modulation and encoding as the optical link
- If the optical link requires an inner code, then the electrical link operates at the signaling rate required by that inner code
- Trade-offs are similar to other applications that employ an inner code

Triple trade-off for error-correcting codes



- Design space is a trade-off between coding gain (performance improvement), overhead, and latency
- Improvement in one area typically comes at the expense of other area(s)

Overhead

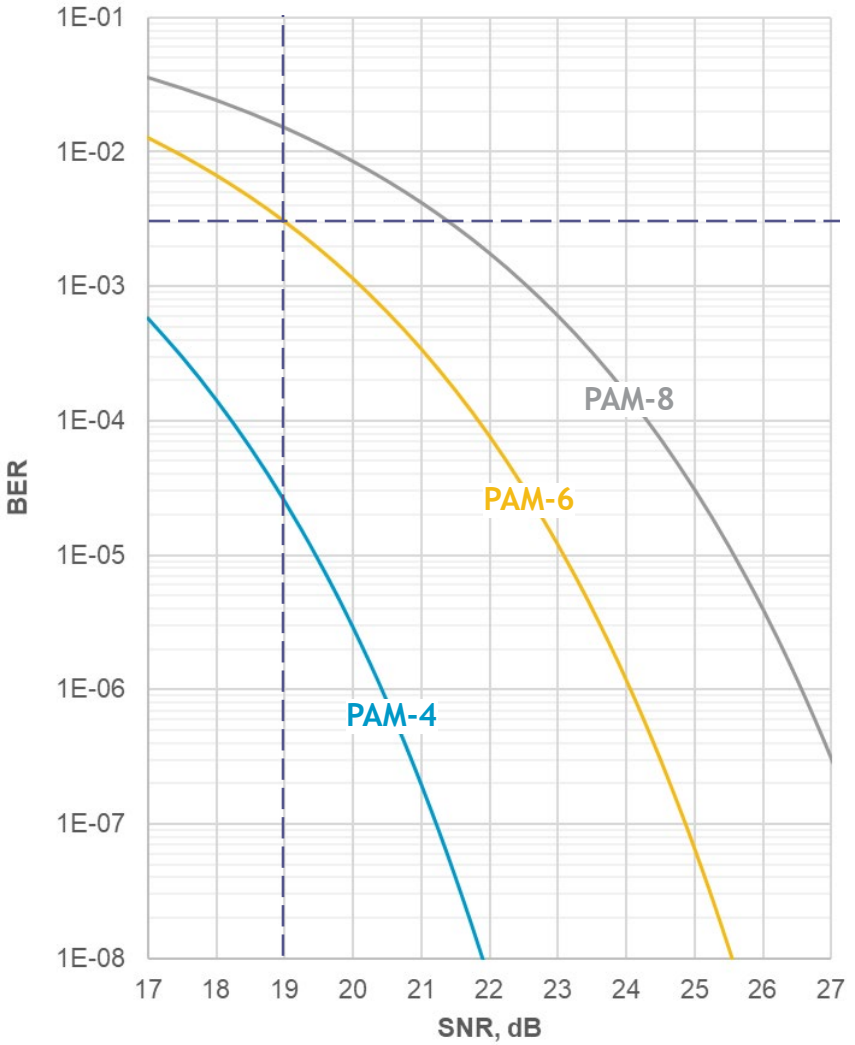
- Error ratio is proportional to the minimum distance between coded signals
- Inner code adds redundancy to increase the minimum distance
- Redundancy can be added by increasing the signaling rate and/or the size of the signal constellation
- The added redundancy is “overhead” that may have adverse effects on link performance
- The SNR penalty due to overhead is often assessed as $10\log_{10}(r)$ where r is the code rate[1]
- This tends to be an optimistic assessment for bandwidth-limited electrical channels

[1] Code rate is ratio of uncoded bits to coded bits and is less than or equal to 1

Example of increasing constellation size

- Begin with 170 GBd PAM-6
- Increase the constellation size to PAM-8 with no change to the signaling rate
- Results in 20% overhead which would enable inclusion of a relatively powerful inner code (r = 5/6)
- However, PAM-8 suffers a performance penalty relative to PAM-6
- “Net coding gain” is the gain of the code minus the 2.4 dB modulation penalty

Modulation	BER at SNR = 19 dB	Δ	SNR for BER = 3e−3	Δ
PAM-6	3e−3	—	19	—
PAM-8	1.5e−2	5x	21.4	+2.4

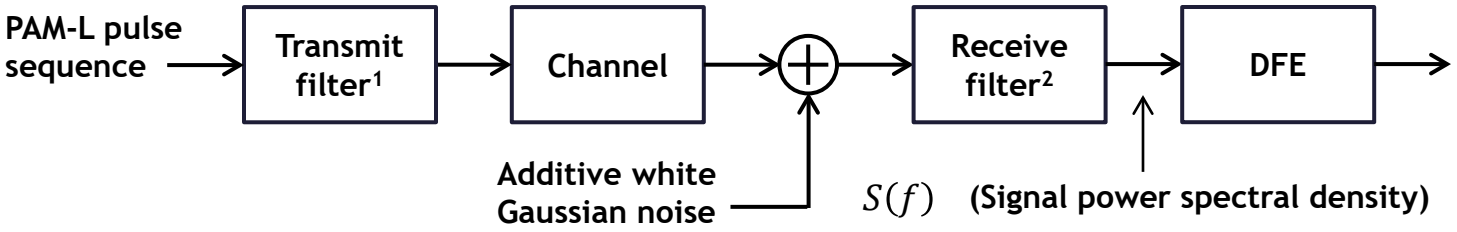


Considerations for the choice of signaling rate

- Prefer integer multiples of the typical 156.25 MHz reference clock
- Consider that the VCO[1] frequency may be a fraction of the signaling rate e.g., 1/2, 1/4
- Consider that the reference frequency may be a multiple of 156.25 MHz e.g., 312.5 MHz, 625 MHz
- So, integer multiples of 2, 4, 8, etc. are even more preferred
- Consider that power dissipation increases with increasing frequency
- Consider that there is channel-dependent performance degradation with increasing frequency

[1] Voltage-controlled oscillator

Consequences of higher signaling rate

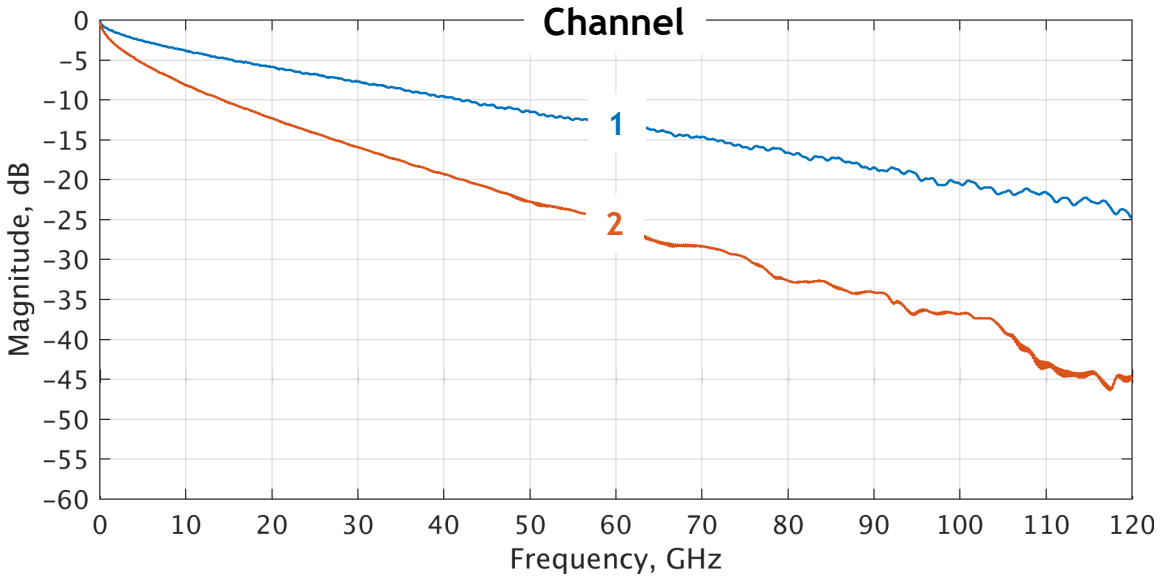


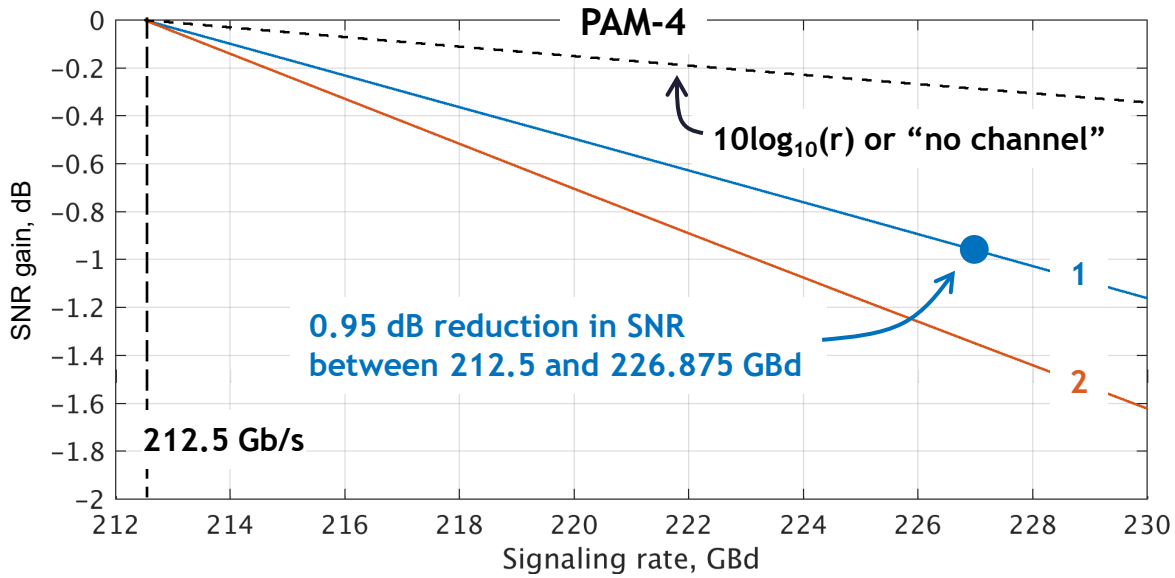
$S(f)$ (Signal power spectral density)

$N(f)$ (Noise power spectral density)

$$SNR = \frac{1}{2\pi}\int_{-\pi}^{\pi} 10\log_{10}\left(\frac{S_F(\theta)}{N_F(\theta)} + 1\right) d\theta$$ (Salz SNR)

$$X_F(\theta) = \frac{1}{T}\sum_m \left| X\left(\frac{\theta + 2\pi m}{2\pi T}\right)\right|^2$$ (Folded power spectral density)

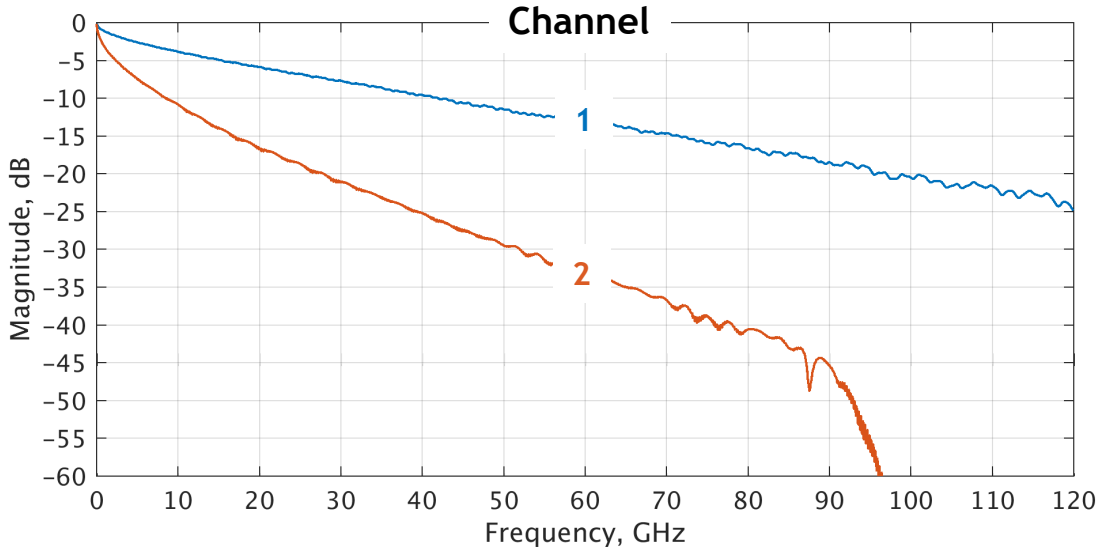


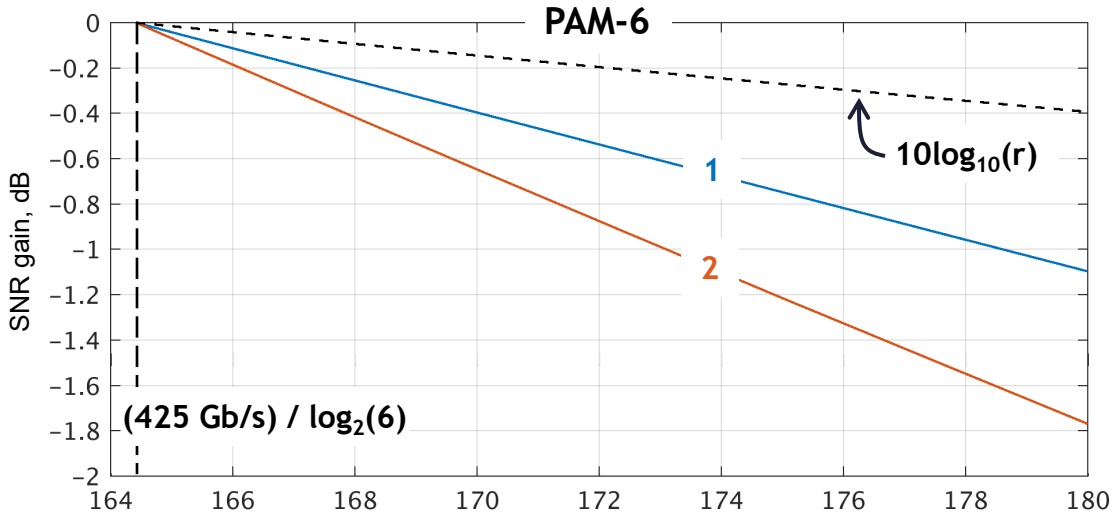


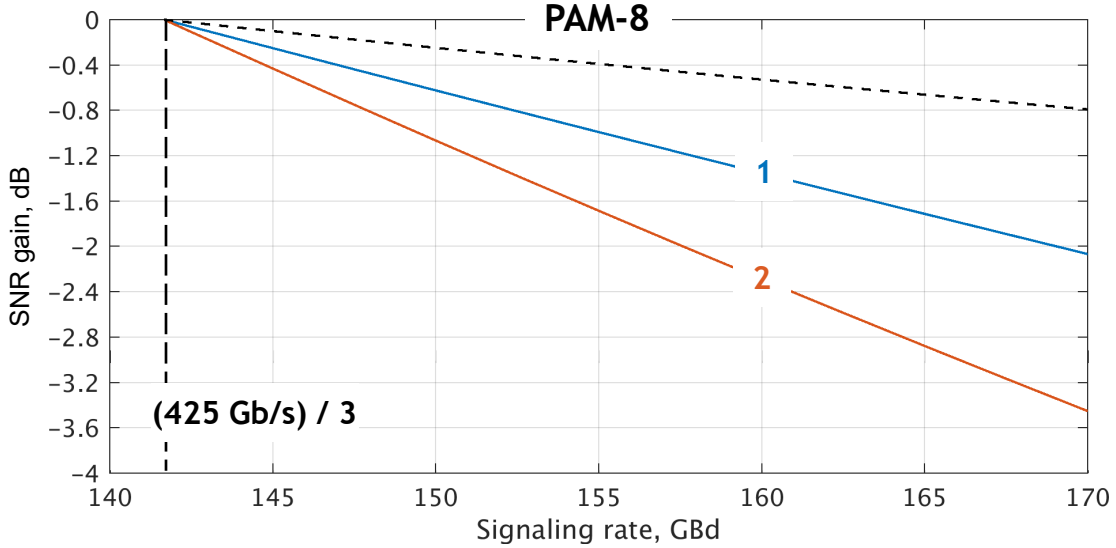
[1] Gaussian filter with 20-80% rise time equal to T/2 where T is the unit interval

[2] 8th order Butterworth filter with −3 dB bandwidth 1/(2T)

Consequences of higher signaling rate, continued







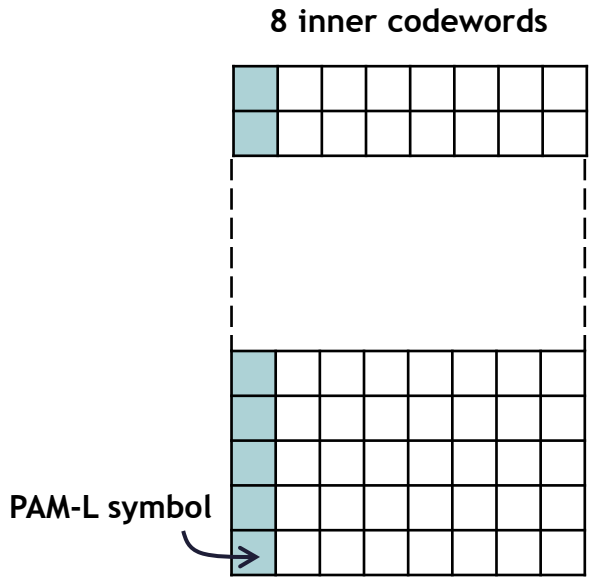
- SNR penalty with increasing signaling rate is considerably higher than $10\log_{10}(r)$
- Actual SNR penalty with increasing signaling rate depends on implementation details but similar trends are expected

Latency

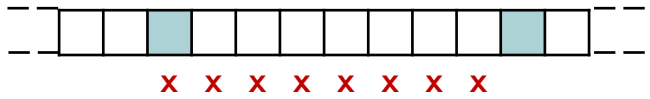
- Inner code encoding and decoding operations add latency to the link
- Interleaving is a significant contributor to the total latency
- It is used to disperse correlated errors into more random error patterns
- Codes tend to perform best with random errors
- Interleaving can be considered for both the inner and outer codes

Inner code interleaving

- Receiver may produce “clumps” of errors that can defeat the inner code
- Bursts of errors from DFE or MLSD, or periods of elevated error rate due to to low-frequency jitter or interference
- Interleave multiple inner codewords to distribute clumps of errors among different codewords
- Relatively low latency cost since inner codewords tends to be shorter



Round-robin multiplexing of PAM-L symbols

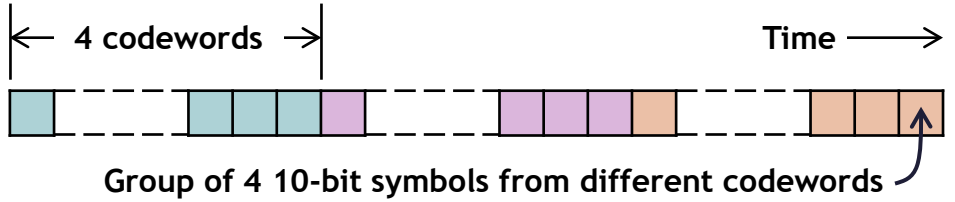


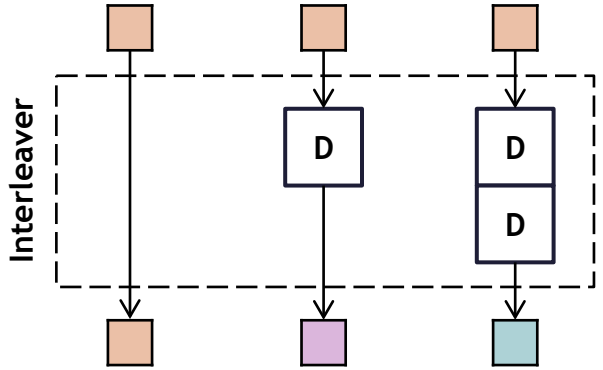
Burst of errors with length ≤ 8 only impacts one symbol from any given codeword

For example, a 64-symbol inner codeword is approximately 282.1 ps at 226.875 GBd.

Outer code interleaving

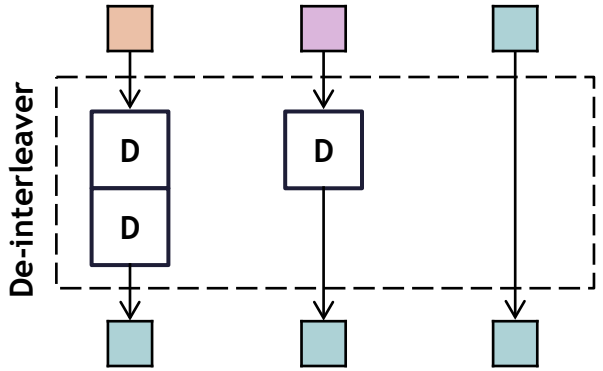
- 200G/lane links feature 4-way Reed-Solomon codeword interleaving using 10-bit symbol multiplexing
- Reed-Solomon interleaving “depth” can be increased to improve resiliency to mis-correction by the inner code
- Lower total latency using convolutional interleavers
- If all inner code instances in a link use the same depth, then the interleave and de-interleave operations need only be done once





Inner code operates on group of 12 symbols from different codewords

Inner code error impact at most 1 symbol from a given codeword



≥ 8 codeword delay

D ≥ 4 codeword delay

Note that 1 codeword = 12.8 ns for 400 Gb/s Ethernet

Summary

- Inner error correcting code may be the next tool pulled from the toolbox for 400G/lane electrical links
- It is a consideration regardless of the choice of modulation
- Lower-overhead codes are preferred for bandwidth-limited electrical links
- “Net coding gain” needs to be the focus
- Drives to soft-decision decoding for better SNR gains with lower overhead
- Interleaving can be used to maximize the performance of the inner code at the expense of latency
- Triple trade-off needs to be carefully considered to find the best balance of performance gain and added latency

QUESTIONS?