

Ethernet in the Field 50Gb/s Lane Rate Webinar

March 21, 2023

Disclaimer: The views expressed in this panel presentation are those of the presenters and not of the EthernetAlliance.



ethernet alliance

www.ethernetalliance.org

3-Part Series

- Today** Webinar 1: Ethernet in the Field (50Gb/s Lane Rate)
- Webinar 2: Ethernet in Design (100Gb/s Lane Rate)
- Webinar 3: Ethernet in the Future (200Gb/s Lane Rate)



ethernet alliance

www.ethernetalliance.org

Agenda

Welcome Current State of Ethernet	Sam Johnson, Engineering Manager, Intel Corp EA Higher Speed Networking Subcommittee Co-chair
Cabling Nomenclature PAM4 Overview Ecosystem Deployments	Ryan Harris, Sales and Market Manager, High-Speed Cable Assemblies, Siemon Company
50Gb/s Auto Negotiation and Link Training	Craig Foster, Product Line Manager, Storage and Networking, Teledyne LeCroy
Link Establishment Interop Challenges Interop Plugfest Value	Sam Johnson



ethernet alliance

www.ethernetalliance.org

About Ethernet Alliance

Sam Johnson, HSN Subcommittee Co-Chair, Intel



ethernet alliance

www.ethernetalliance.org

The Ethernet Alliance

Global Community of End Users, System Vendors, Component Suppliers & Academia

Our Mission

- **To promote** industry awareness, acceptance and advancement of technology and products based on, or dependent upon, both **existing and emerging IEEE 802 Ethernet standards** and their management.
- **To accelerate industry adoption** and remove barriers to market entry by providing a cohesive, market-responsive, industry voice.
- Provide resources to establish and **demonstrate multi-vendor interoperability.**



Ethernet Alliance Strategy

Expanding the Ethernet Ecosystem, Supporting Ethernet Development

Facilitate interoperability testing & assurance

- Industry Plug fests supporting member and technology initiatives
- PoE Certification Program

Global outreach and collaborative interaction with other industry organizations

- Worldwide Membership
- Multiple SIGs, applications and MSAs
- Industry consensus building

Thought Leadership

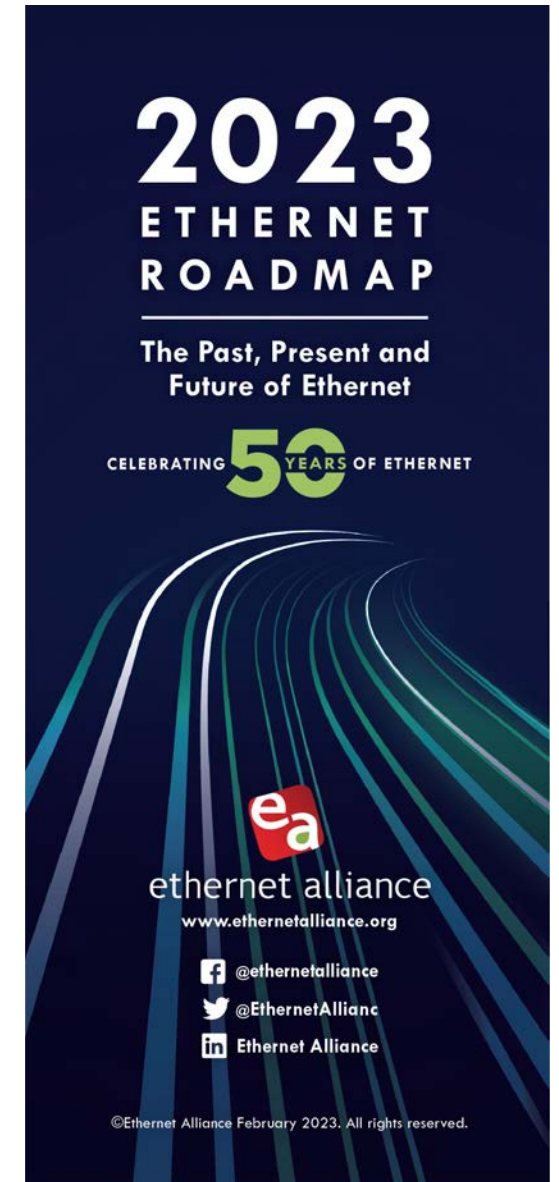
- EA-hosted Technology Exploration Forums (TEFs)
- Technology and standards incubation

Promotion of Ethernet

- Media and industry analysts outreach
- Education
- Marketing (trade shows & panel presentations, white papers, blogs & social media)

2023 Ethernet Roadmap

- **50th Anniversary of Ethernet edition** launched at **OFC 2023**
- **Digital version and graphics** available via the Alliance website
- Also available as part of the “**Ethernet Alliance in a box**” event resources (for members on-demand)



Current State of Ethernet

Sam Johnson, HSN Subcommittee Co-Chair, Intel



ethernet alliance

www.ethernetalliance.org

2023 ETHERNET ROADMAP

The Past, Present and
Future of Ethernet

CELEBRATING 50 YEARS OF ETHERNET



ethernet alliance

www.ethernetalliance.org

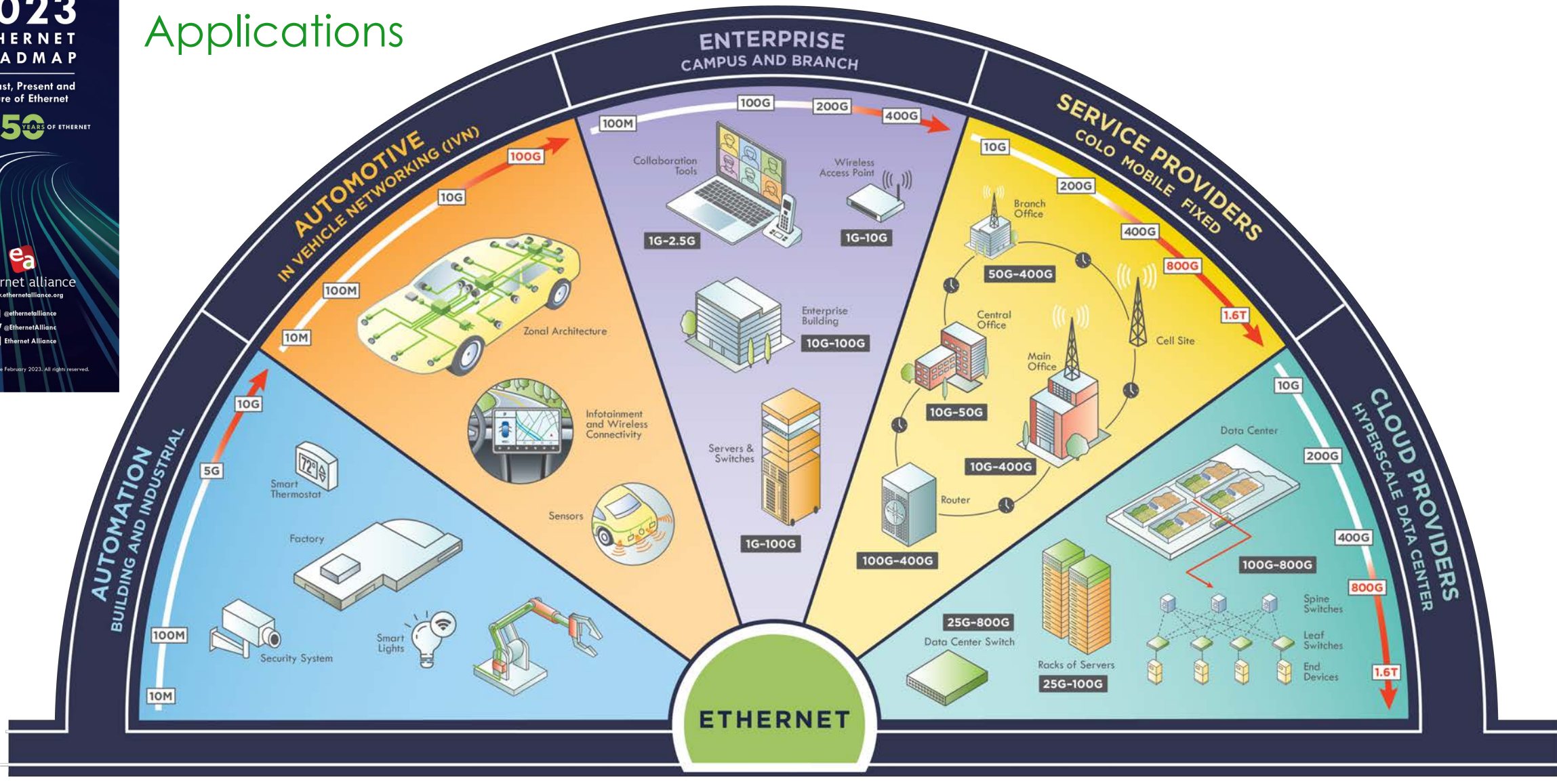
@ethernetalliance

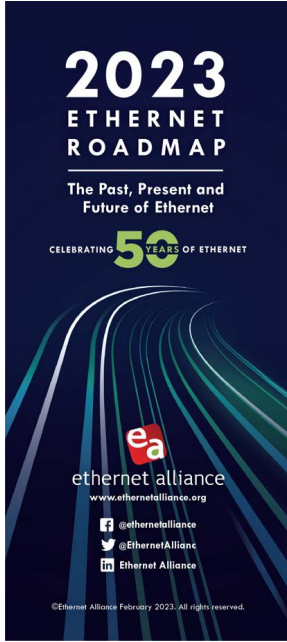
@EthernetAlliance

Ethernet Alliance

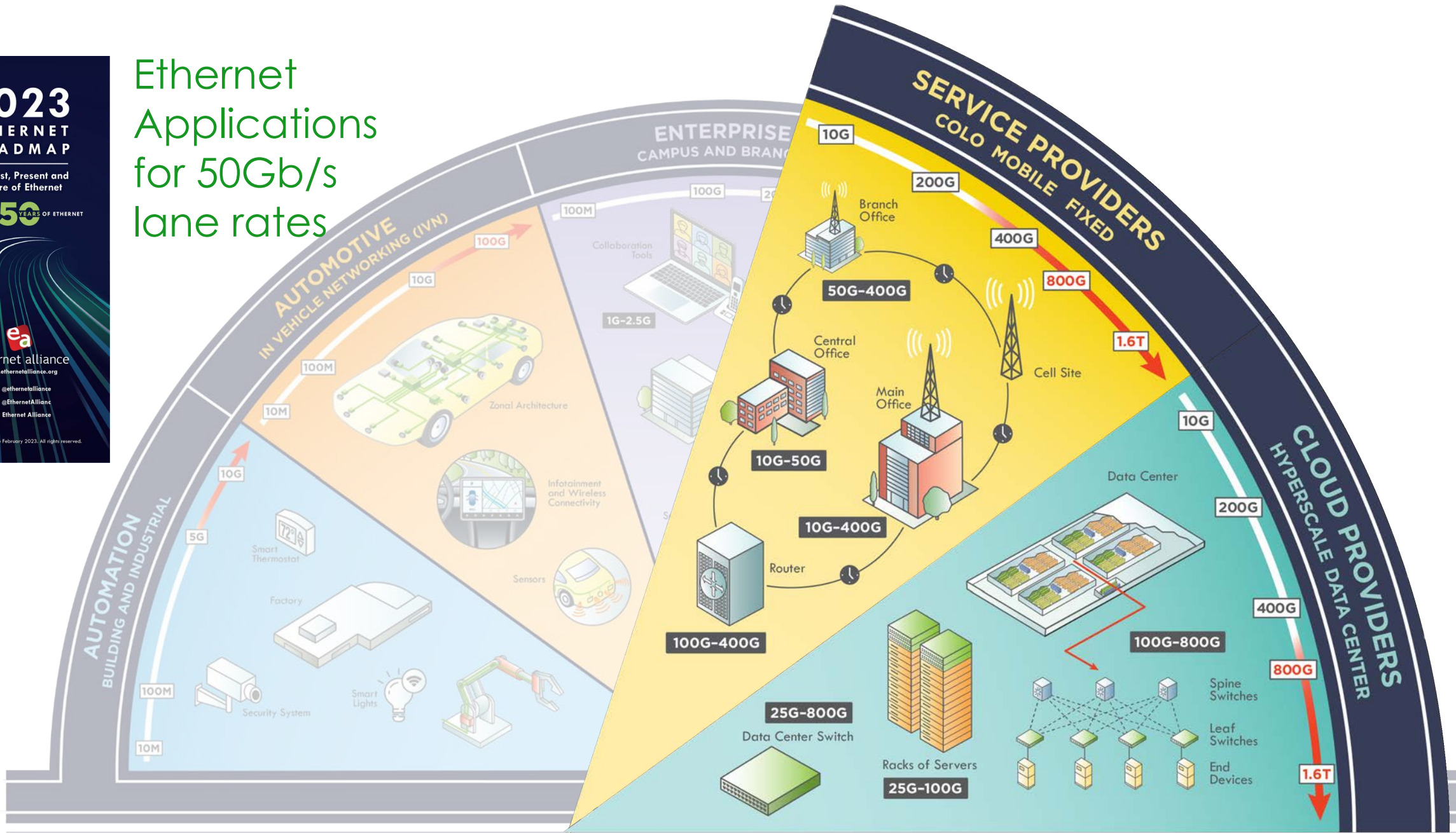
©Ethernet Alliance February 2023. All rights reserved.

Ethernet Applications





Ethernet Applications for 50Gb/s lane rates



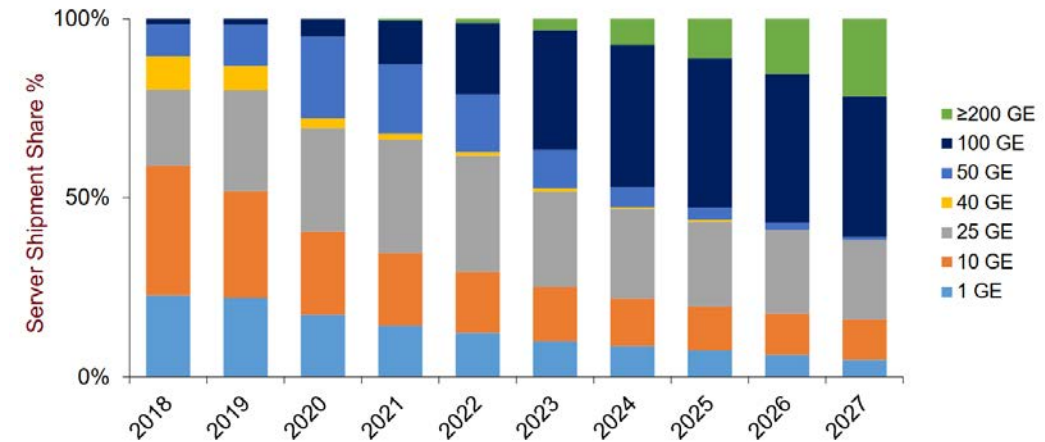
Ethernet Market Forecast

The global Ethernet adapter market size was valued at USD \$4.6B in 2022 and conservatively projected to reach **\$6.3B by 2027** according to the "Ethernet Adapter and Smart NIC 5-Year January 2023 forecast" report from the Dell'Oro Group

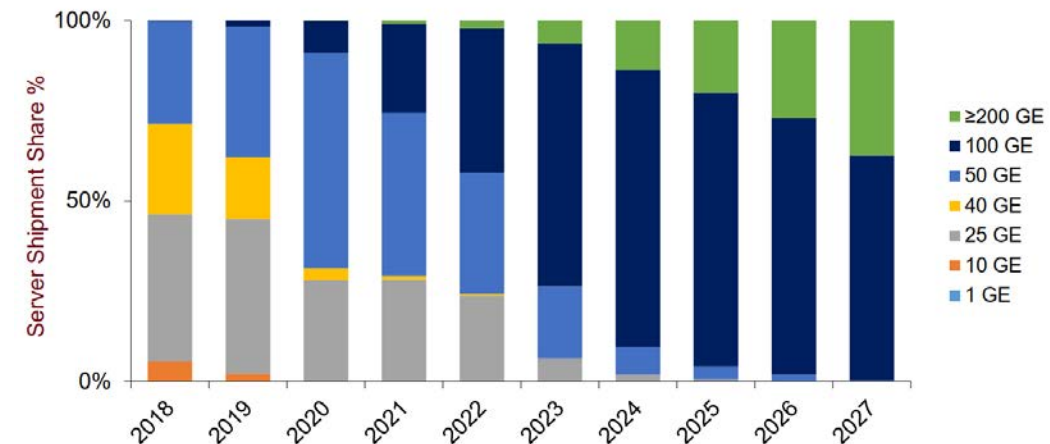
Additional predictions for 2027:

- 200 Gb/s and higher-speed ports will account for 44% of the server network revenue.
- The Smart NIC market will reach \$2B.
- **100GbE will surpass 25GbE** in port shipments as the mainstream server port speed.

Server Speed Migration, Total Market

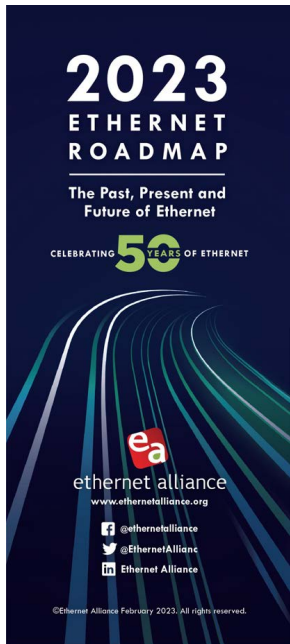
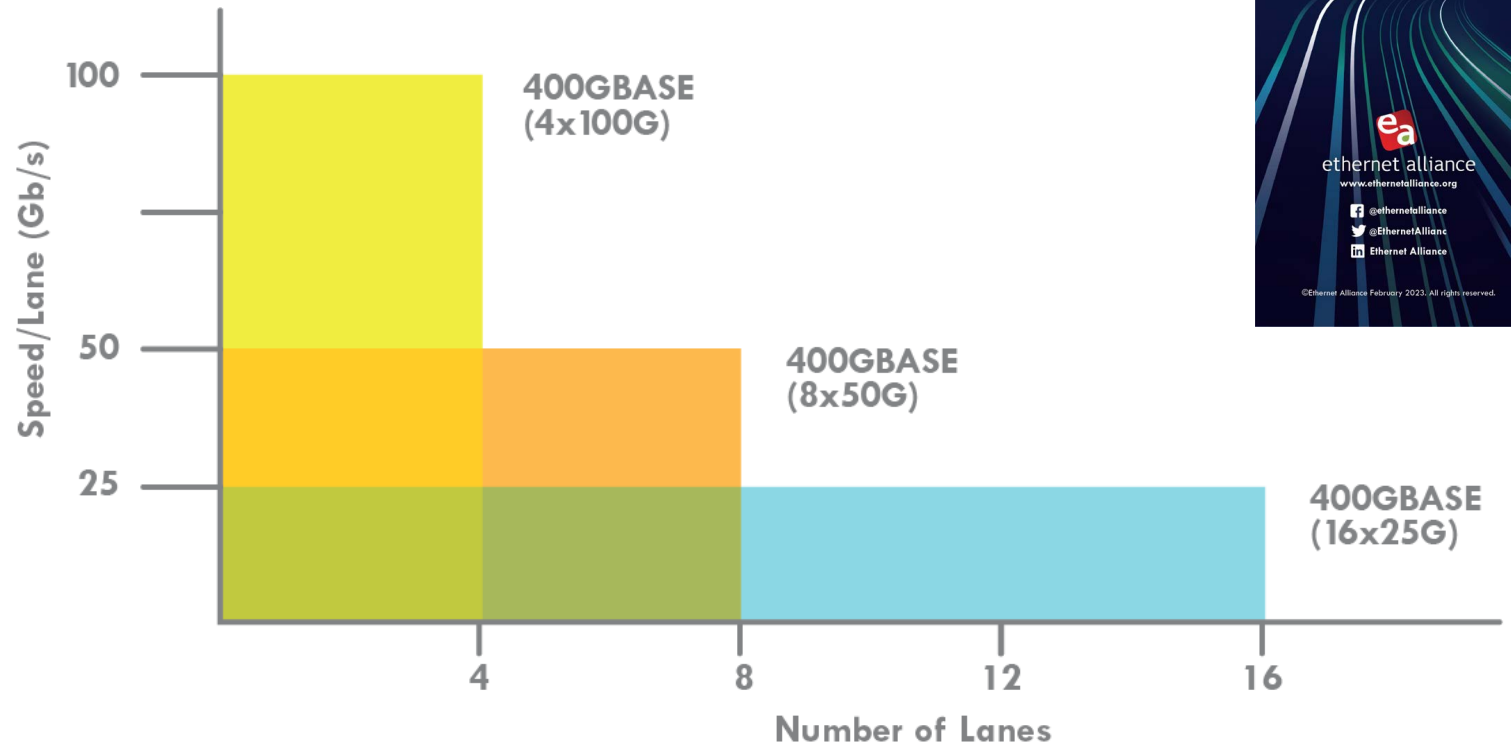
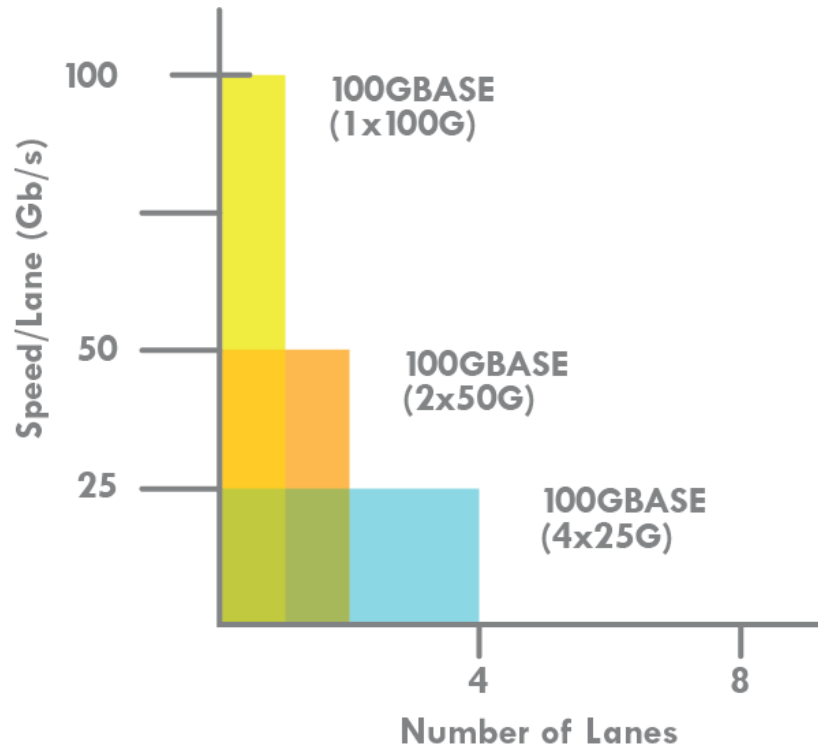


Server Speed Migration, Top 4 US Cloud

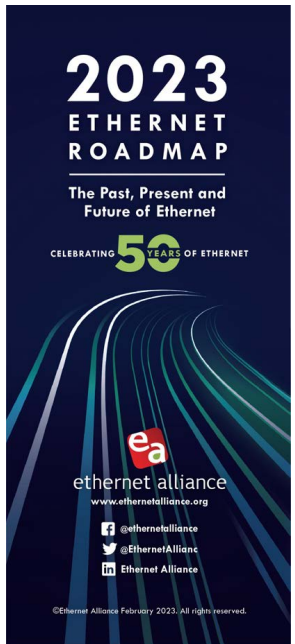
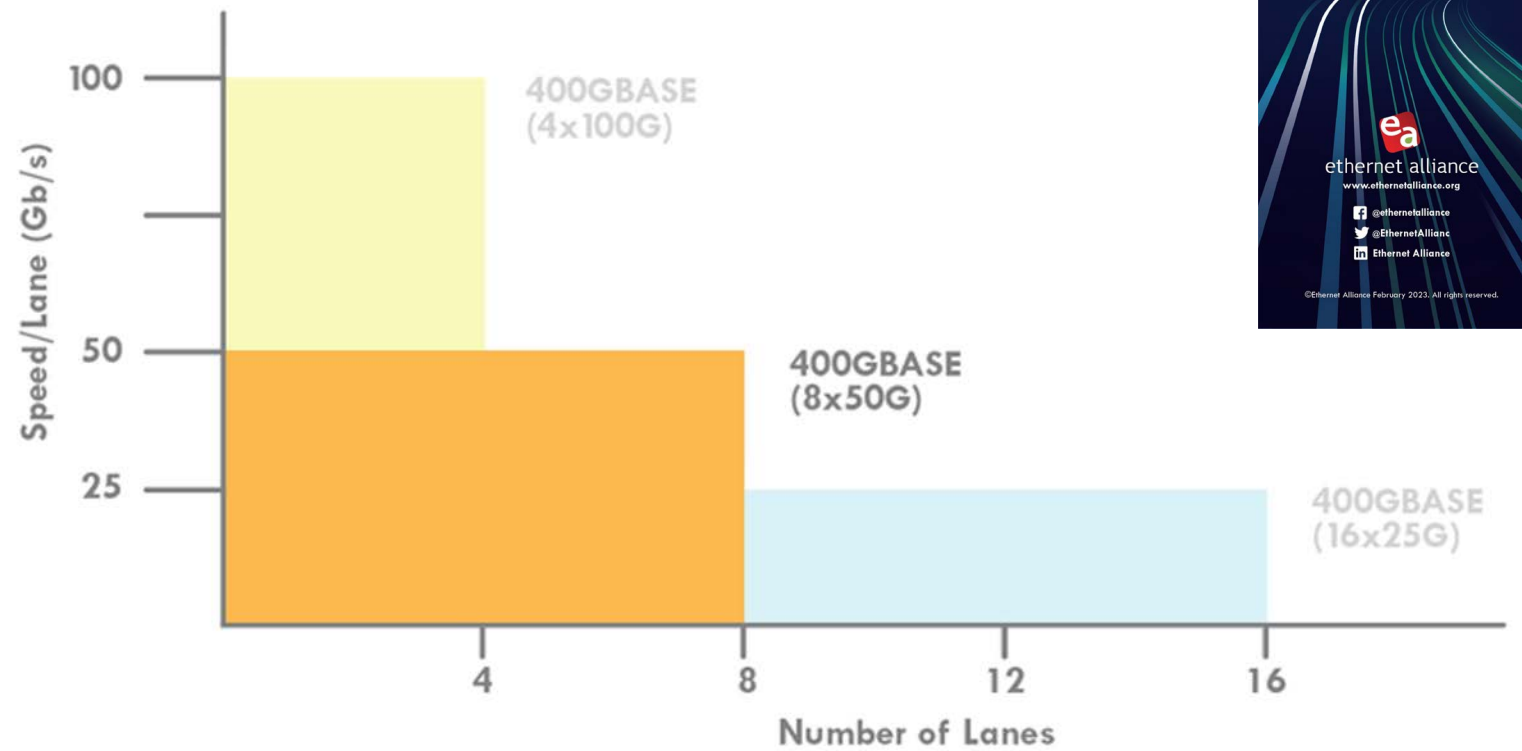
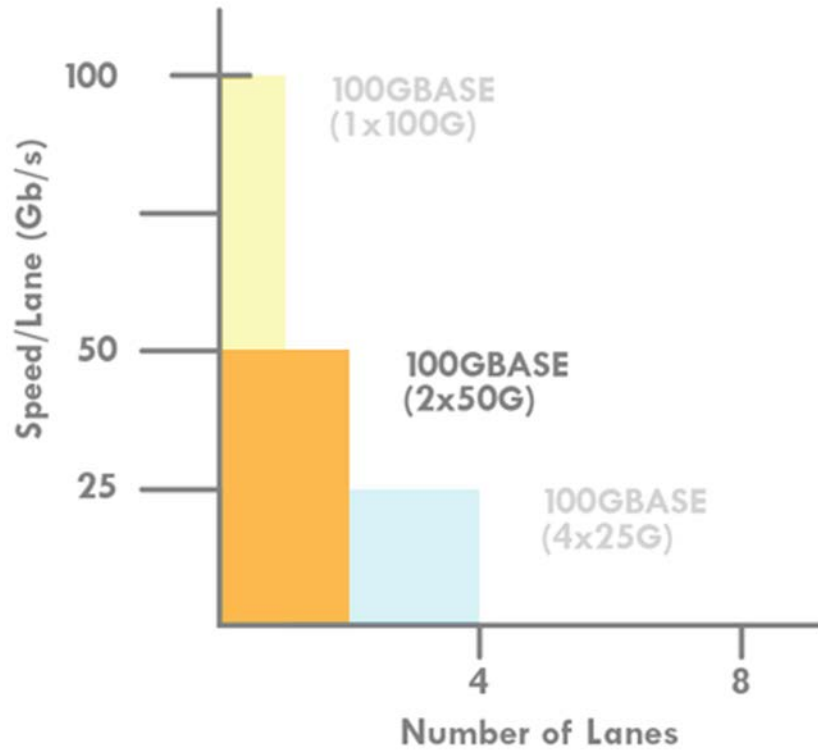


Source: Dell'Oro Group 2/2/23 Press Release: <https://www.delloro.com/news/smart-nic-market-to-approach-2-billion-by-2027/>

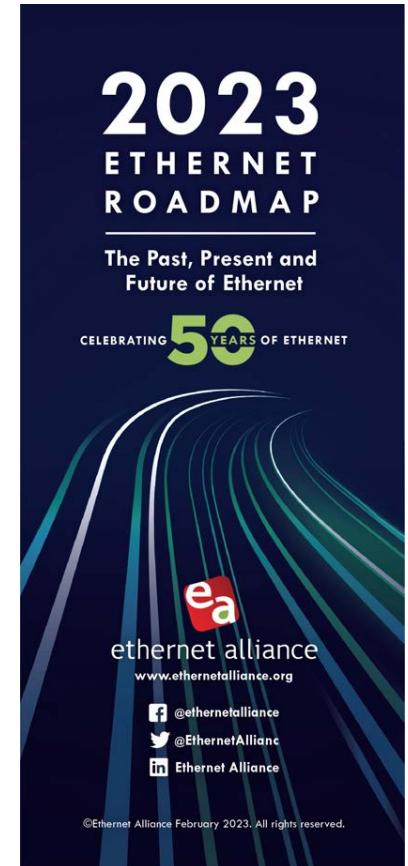
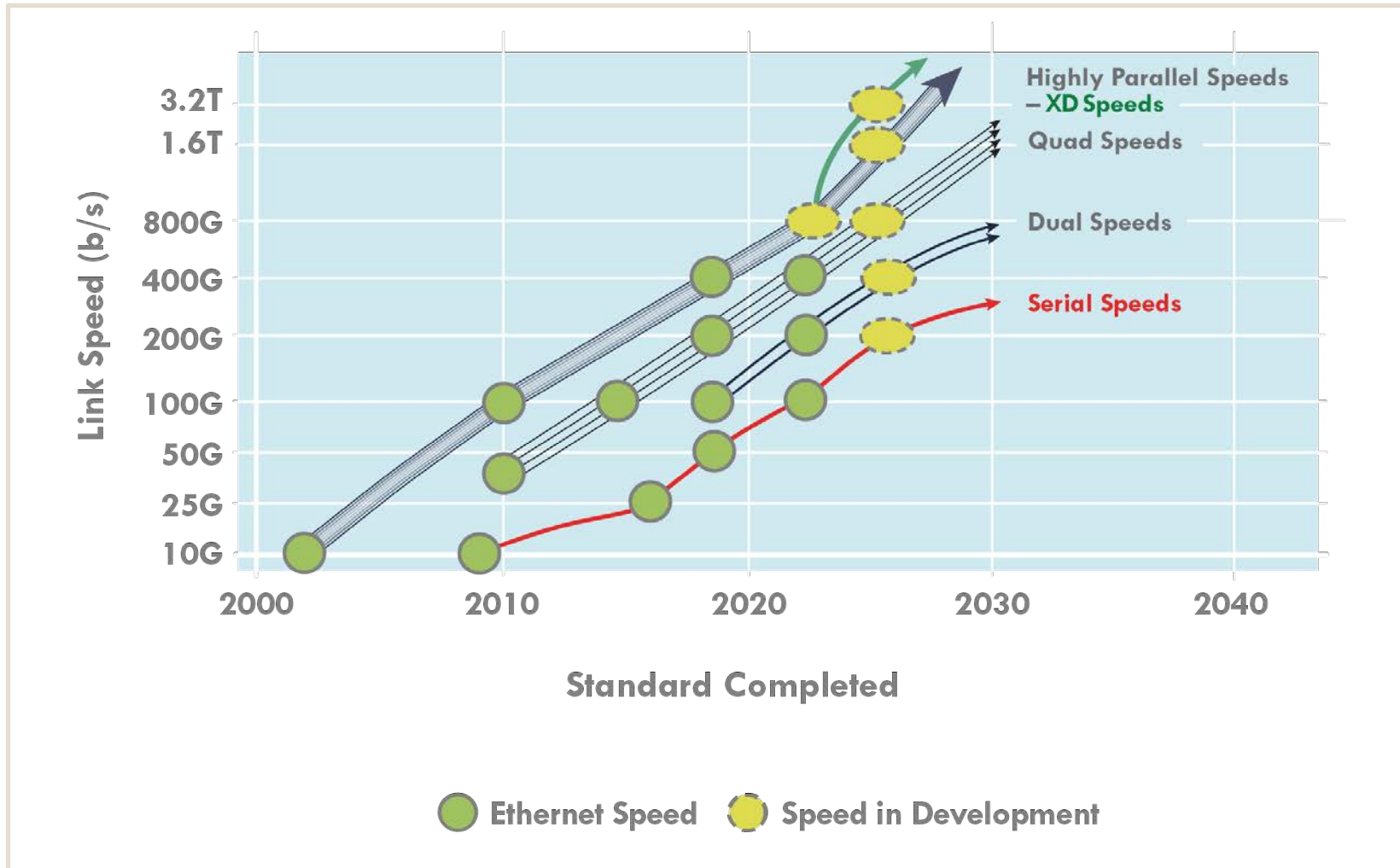
Fatter Pipes, Higher Data Rates



Fatter Pipes, Higher Data Rates



Ethernet Speeds



PAM4 Overview

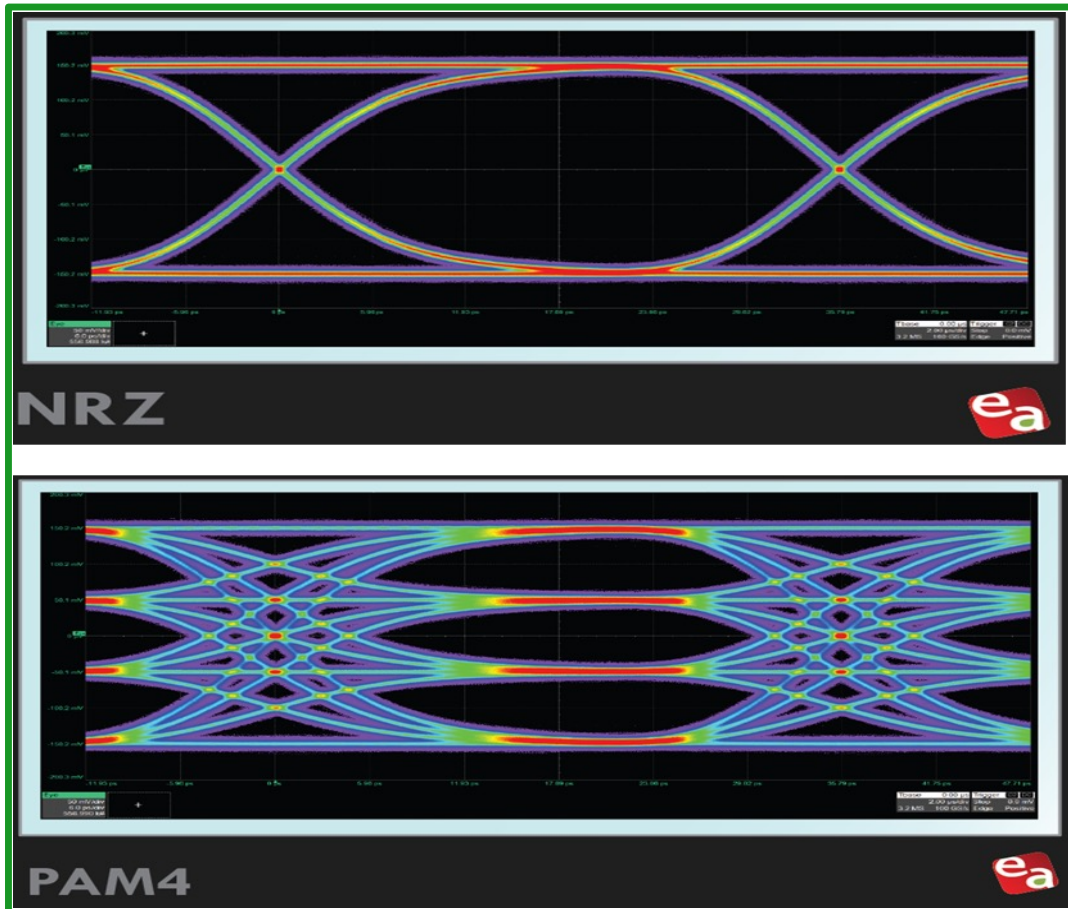
Ryan Harris, Sales and Market Manager, High-Speed Cable Assemblies,
Siemon Company



ethernet alliance

www.ethernetalliance.org

NRZ to PAM4 – More Symbols



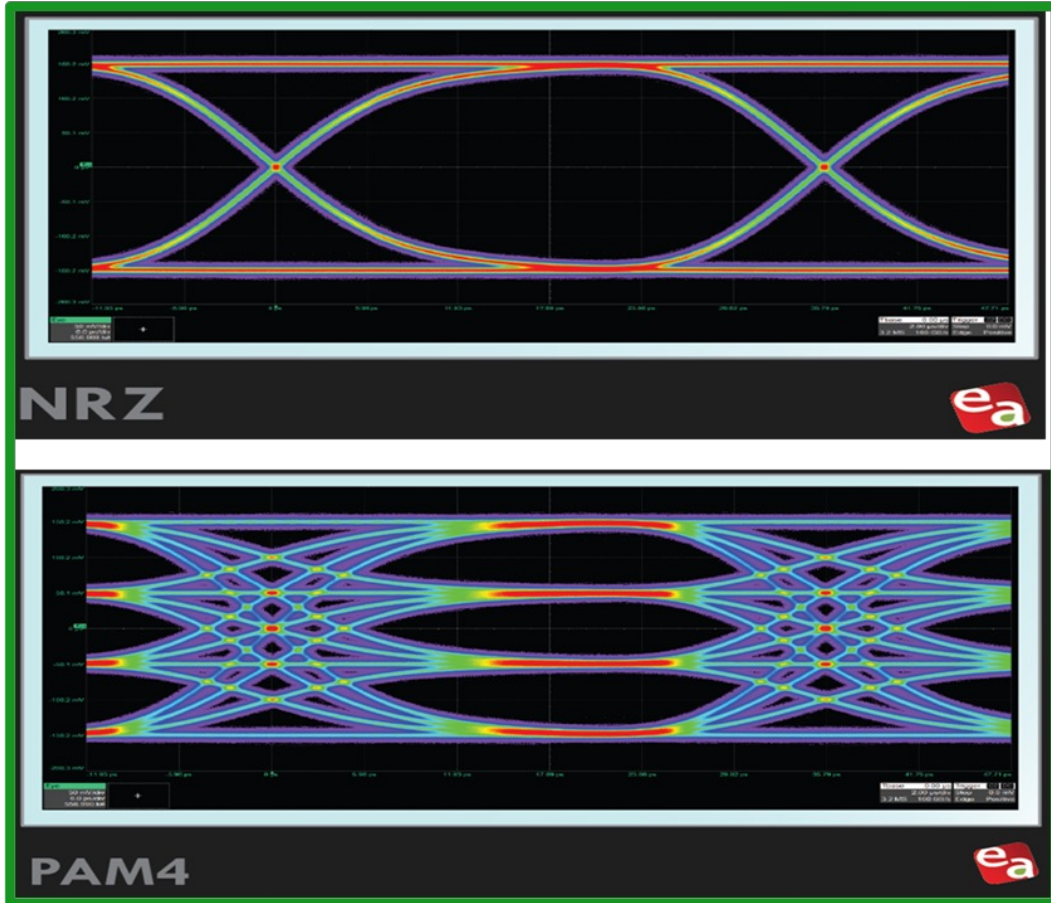
NRZ

- Non-Return to Zero
- uses **0** or **1** as bit symbols for information

PAM4

- Four-level Pulse Amplitude Modulation
- uses **00**, **01**, **10**, **11** as bit symbols for information
- PAM4 symbols enable 2x the Gbps

NRZ to PAM4 – Gbaud Frequency



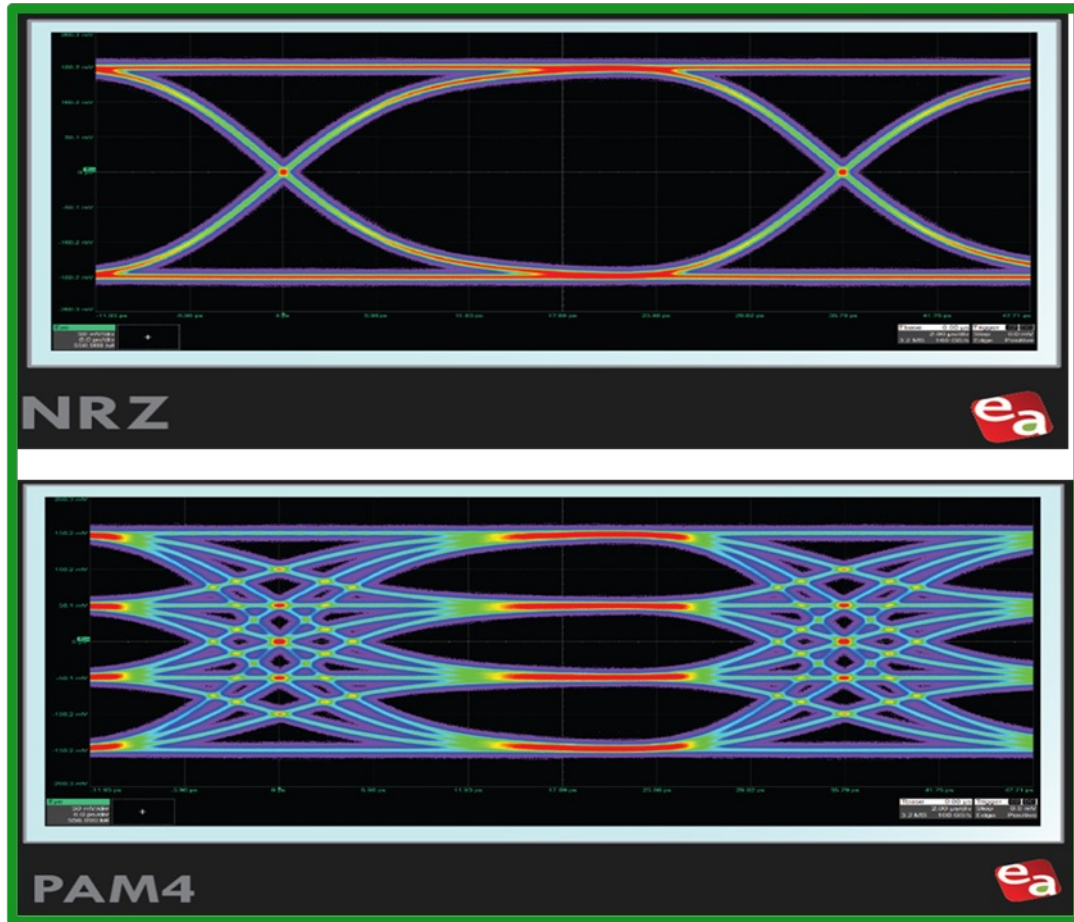
NRZ

- **one** baud carries **one** bit symbol
- 25GAUI is a 25.78125 GBd
- carries 25 Gbps/GbE

PAM4

- **one** baud carries **two** bit symbols
- 50GAUI-1 is a 26.5625 GBd
- carries 50 Gbps/GbE

NRZ to PAM4 – Signal Noise & FEC



Short Reach DAC Forward Error Correction (FEC) Settings

Cable type	SFP+	SFP28	SFP56
<i>typical FEC</i>	10 GbE NRZ	25 GbE NRZ	50 GbE PAM4
2-meter DAC	no FEC	no FEC	KP1 FEC
3-meter DAC	no FEC	FC-FEC	KP1 FEC
5-meter DAC	no FEC	RS-FEC	
7-meter DAC	no FEC		

- PAM4 is more noise sensitive over 3 signal eyes
- FEC is determined by cable performance metrics
- Passive DAC with FEC can only support 3-meters
- PAM4 only KP1 FEC setting option
- FEC increases latency to achieve signal health

Cabling Nomenclature

Ryan Harris, Sales and Market Manager, High-Speed Cable Assemblies,
Siemon Company



ethernet alliance

www.ethernetalliance.org

One Lane 50G Ports

SFP56 (1-lane 50 GbE = 50GBASE)

- Using same SFP28 - SFF-8402 specification
- Backwards compatible SFP+/SFP28 cages

Side A (pluggable)	Signal Type	Side A Config.	Cable Type	Side B (pluggable)
SFP56	50 GbE PAM4	1x 50GbE	Straight Through	SFP56
Side A (pluggable)	Signal Type	Side A Config.	Cable Type	Side B (pluggable)
SFP28	25 GbE NRZ	1x 25GbE	Straight Through	SFP28



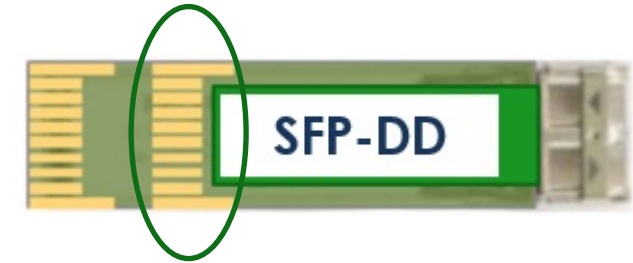
Single Lane Transceiver		
SFP+	SFP28	SFP56
10G NRZ	25G NRZ	50G PAM4



Two Lane 50G Ports

SFP-DD (2-lane 50 GbE = 100GBASE)

- Second row of pins on card edge PCB
- Backwards compatible with SFP+/SFP28



Side A (pluggable)	Signal Type	Side A Config.	Cable Type	Side B (pluggable)
SFPDD	50 GbE PAM4	1x 100GbE	Straight Through	SFPDD
SFPDD	50 GbE PAM4	2x 50GbE	2x Breakout	SFP56

**QSFP56 (2-lane 50 GbE = 100GBASE)

- Uses only 2 electrical lanes and 2 lanes not used
- Backwards compatible with QSFP28 cages
- Typically used on the breakout ends



2 lanes

Four Lane 50G Ports

QSFP56 (4 lane 50 GbE = 200GBASE)

- Using same QSFP28 - SFF-8665 specification
- Backwards compatible with QSFP+/QSFP28

Side A (pluggable)	Signal Type	Side A Config.	Cable Type	Side B (pluggable)
QSFP56	50 GbE PAM4	1x 200GbE	Straight Through	QSFP56
QSFP56	50 GbE PAM4	2x 100GbE	2x Breakout	SFPDD/**QSFP56
QSFP56	50 GbE PAM4	4x 50GbE	4x Breakout	SFP56
Side A (pluggable)	Signal Type	Side A Config	Cable Type	Side B (pluggable)
QSFP28	25 GbE NRZ	1x 100GbE	Straight Through	QSFP28
QSFP28	25 GbE NRZ	2x 50GbE	2x Breakout	**QSFP28
QSFP28	25 GbE NRZ	4x 25GbE	4x Breakout	SFP28

- **QSFP half loaded, only 2 of the 4 lanes used



2 lanes

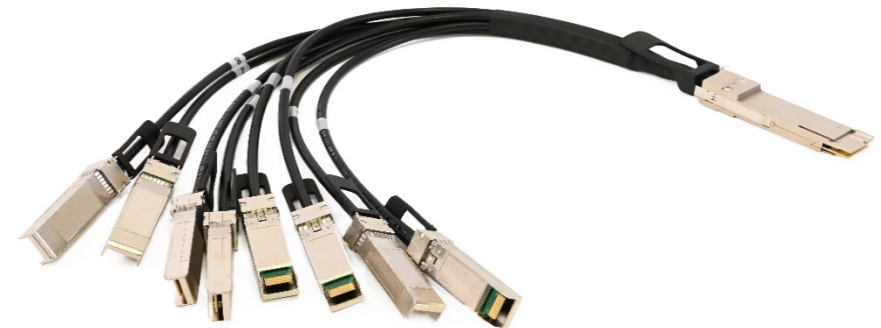
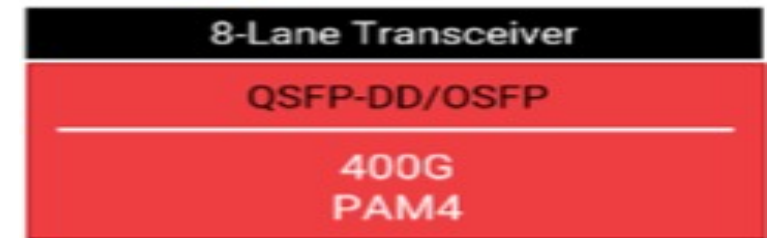
Eight Lane 50G Ports

QSFP-DD (8 lane 50 GbE = 400GBASE)

- Long and shorter signal pins
- Backwards compatible with QSFP+/QSFP28

OSFP (8 lane 50 GbE = 400GBASE)

Side A (pluggable)	Signal Type	Side A Config.	Cable Type	Side B (pluggable)
QSFPDD/OSFP	50 GbE PAM4	1x 400GbE	Straight Through	QSFPDD/OSFP
QSFPDD/OSFP	50 GbE PAM4	2x 200GbE	2x Breakout	QSFP56
QSFPDD/OSFP	50 GbE PAM4	4x 100GbE	4x Breakout	SFPDD/**QSFP56
QSFPDD/OSFP	50 GbE PAM4	8x 50GbE	8x Breakout	SFP56
Side A (pluggable)	Signal Type	Side A Config.	Cable Type	Side B (pluggable)
QSFPDD/OSFP	25 GbE NRZ	1x 200GbE	Straight Through	QSFPDD/OSFP
QSFPDD/OSFP	25 GbE NRZ	2x 100GbE	2x Breakout	QSFP28
QSFPDD/OSFP	25 GbE NRZ	4x 50GbE	4x Breakout	**QSFP28
QSFPDD/OSFP	25 GbE NRZ	8x 25GbE	8x Breakout	SFP28



Ecosystem Deployments

Ryan Harris, Sales and Market Manager, High-Speed Cable Assemblies,
Siemon Company



ethernet alliance

www.ethernetalliance.org

50G per lane – PAM4 Cabling Options



Side A

1



DAC
Direct Attach Copper

DAC where used:

- ToR (Top of Rack)
- 0.5 meters to 3 meters (PAM4)

2



AOC
Active Optical Cable

AOC where used:

- ToR (Top of Rack)
- MoR/EoR (Middle/End of Row)
- 1 meter up to 30 meters

3



Transceiver module
with structured patch cabling

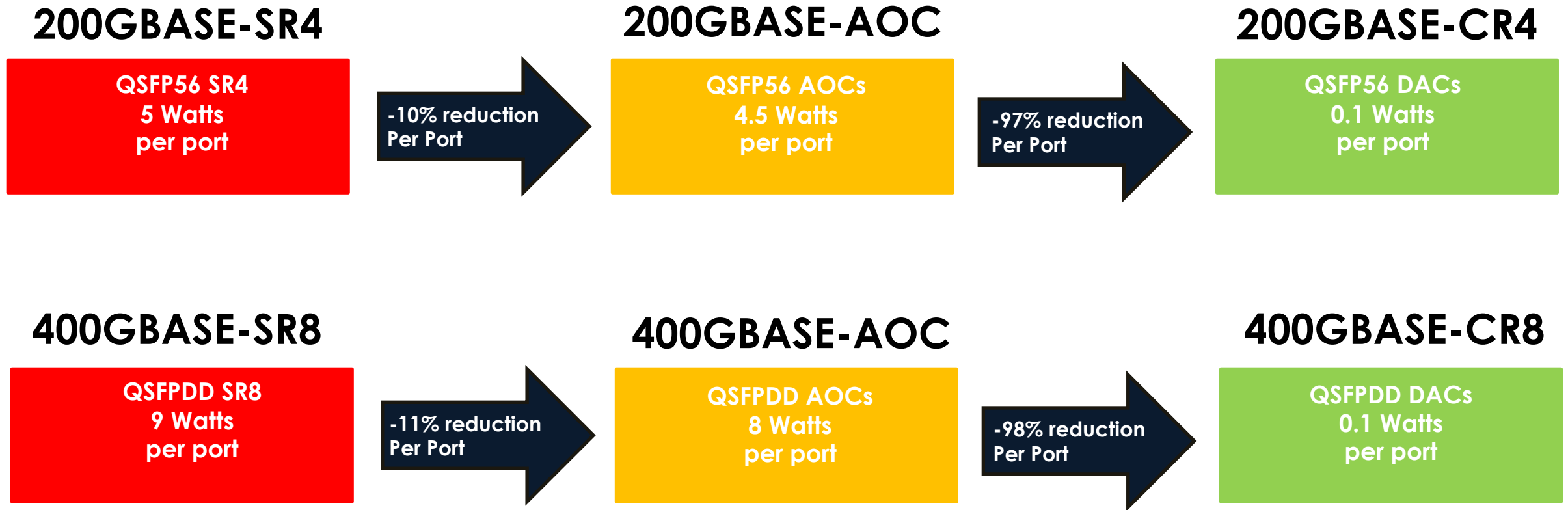
Transceiver where used:

- MoR/EoR (Middle/End of Row)
- (or) Row to Row
- (or) Room to Room
- 1 meter up to 100 meters



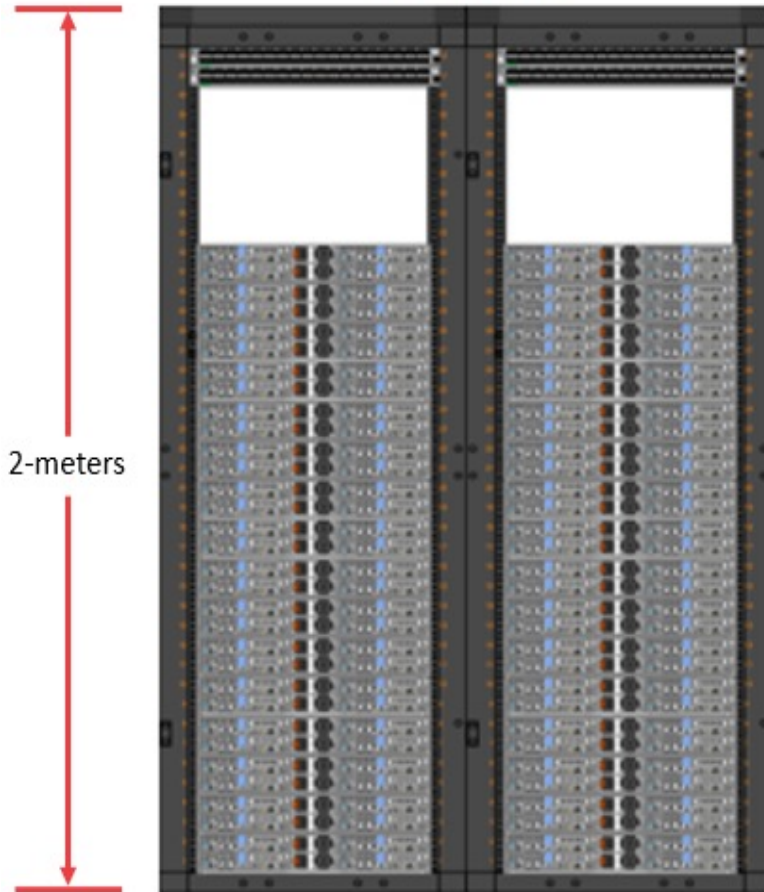
Side B

50G per lane - Power Considerations



50G per Lane – ToR and DAC at the Edge

42U Cabinet



Direct Attach Copper (Passive)

- PAM4 DAC supports up to 3-meters (9.8 ft)

Side A (pluggable)	Signal Type	Side A Config.	Cable Type	Side B (pluggable)
SFP56	50 GbE PAM4	1x 50GbE	Straight Through	SFP56
SFPDD	50 GbE PAM4	1x 100GbE	Straight Through	SFP56DD
SFPDD	50 GbE PAM4	2x 50GbE	2x Breakout	SFP56
QSFP56	50 GbE PAM4	1x 200GbE	Straight Through	QSFP56
QSFP56	50 GbE PAM4	2x 100GbE	2x Breakout	SFPDD/**QSFP56
QSFP56	50 GbE PAM4	4x 50GbE	4x Breakout	SFP56
QSFPDD/OSFP	50 GbE PAM4	1x 400GbE	Straight Through	QSFPDD/OSFP
QSFPDD/OSFP	50 GbE PAM4	2x 200GbE	2x Breakout	QSFP56
QSFPDD/OSFP	50 GbE PAM4	4x 100GbE	4x Breakout	SFPDD/**QSFP56
QSFPDD/OSFP	50 GbE PAM4	8x 50GbE	8x Breakout	SFP56

50GPAM4 AN/LT

Craig Foster, Product Line Manager, Storage and Networking, Teledyne LeCroy



ethernet alliance

www.ethernetalliance.org

Evolution of HSN Auto Negotiation and Link Training

- 10Gbs/lane NRZ
- 25Gbs/lane NRZ
- 50Gbs/lane PAM-4
- 100Gbs/lane PAM-4

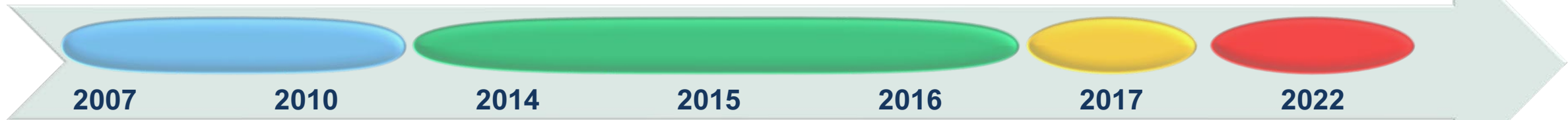
IEEE 802.3
Clause 73
Auto Negotiation

10GBASE-KR
IEEE Cl. 72

100GBASE-
KR4/CR4
IEEE Cl. 92/93

25GBASE-
CR/CR-S and
KR/KR-S
IEEE Cl. 110/111

100G/200G/400G
CR/KR PAM-4
IEEE Cl. 162/163



40GBASE-
KR4/CR4
IEEE Cl. 84/85

ETC
25G CR1/KR1
50G CR2/KR2

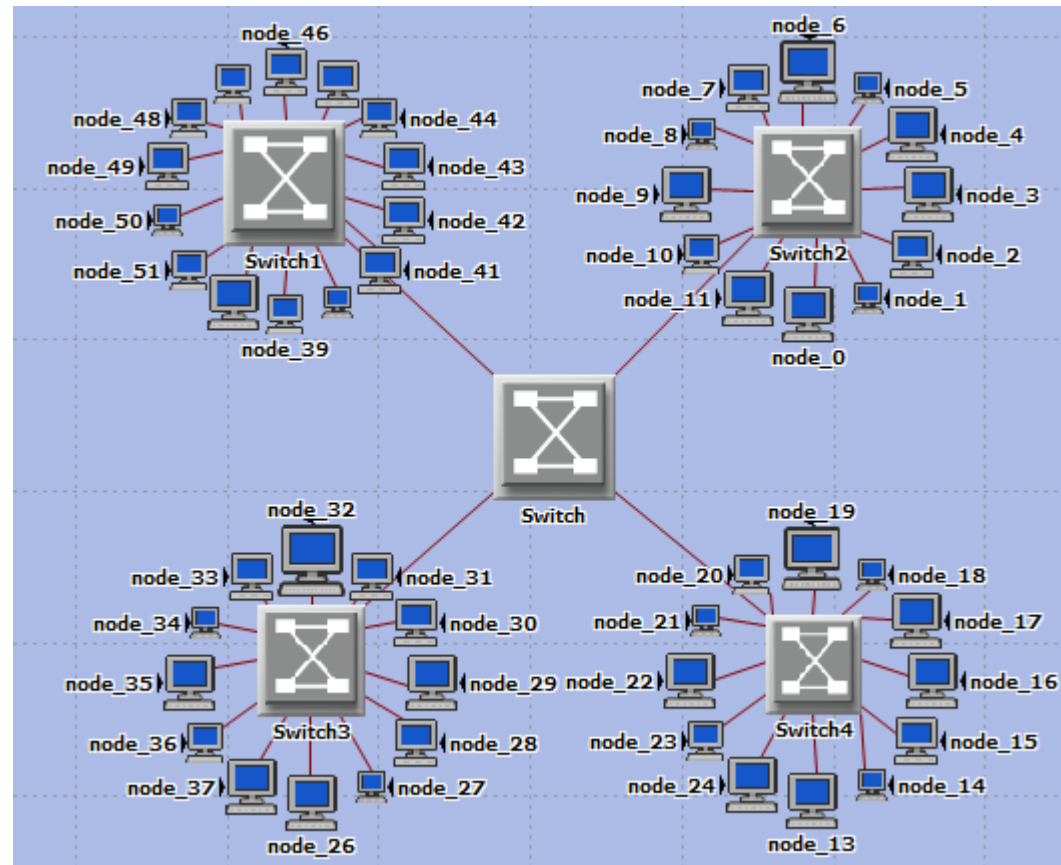
50G/100G/200G
CR/KR PAM-4
IEEE Cl. 136/137

Why is Auto Negotiation important?

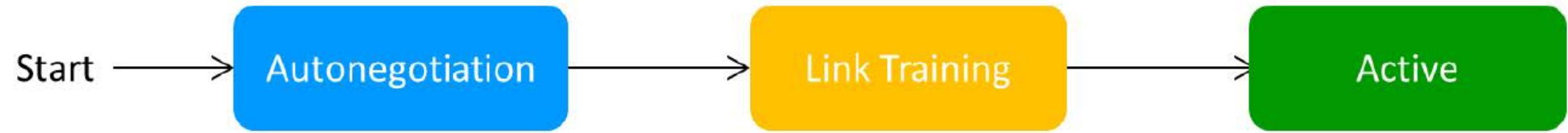
Previous configurations required managers to manually configure each port for speed and port configuration.

With more speeds and configurations available this is becoming less feasible.

Auto Negotiation allows devices to determine the best supported link configuration upon connection without manual interference.



L1



How does link Auto Negotiation work?

Spreadsheet View

No.	Start Time	Port	Speed	Port No	Frame	Frame	Summary
2269	20.999 976 780(s)	P5-Rx	AN	P6		32896 - Auto-Negotiation	0x10:IEEE Std 802.3; Acknowledge=0x0; Next Page=0x0
2270	21.011 135 116(s)	P5-Rx	AN	P6		32896 - Auto-Negotiation	0x10:IEEE Std 802.3; Acknowledge=0x0; Next Page=0x0
2271	21.022 293 457(s)	P5-Rx	AN	P6		32896 - Auto-Negotiation	0x10:IEEE Std 802.3; Acknowledge=0x0; Next Page=0x0
2272	21.033 451 797(s)	P5-Rx	AN	P6		32896 - Auto-Negotiation	0x10:IEEE Std 802.3; Acknowledge=0x0; Next Page=0x0
2273	21.040 280 436(s)	P1-Tx	AN	P1	3 - Auto-Negotiation		0x10:IEEE Std 802.3; Acknowledge=0x0; Next Page=0x1
2274	21.040 281 458(s)	P1-Tx	AN	P1	4 - Auto-Negotiation		0x10:IEEE Std 802.3; Acknowledge=0x1; Next Page=0x1
2275	21.040 282 815(s)	P1-Tx	-	P1	0x00:Loss of Sync		

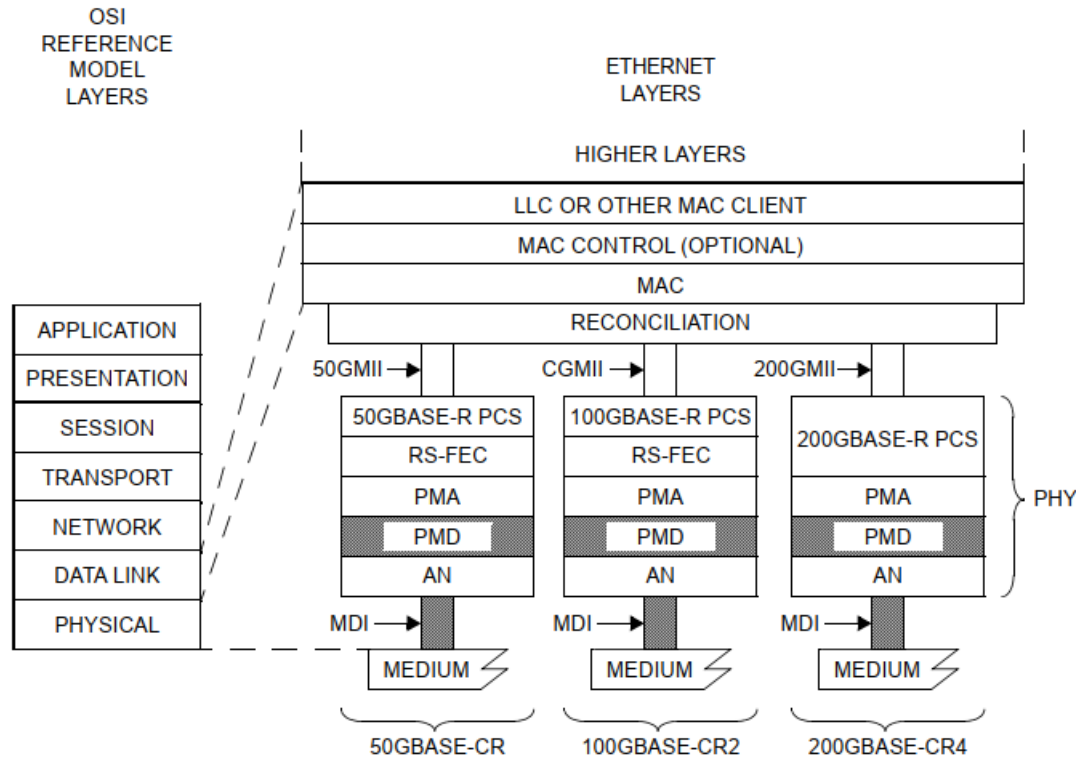
Frame Inspector View

Length: N/A Hide Reserved Fields Marker: Name Description

Index	Data	Field	Value
0001	80 00 38 00	Auto-Negotiation	0x80003800 2008
0002	20 08	Selector Field(S_0:4)	0x10 : IEEE Std 802.3
		Echoed Nonce Field(E_0:4)	0x00
		Pause Ability (C0:C2)	0x0
		C0: PAUSE	0x0
		C1: ASM_DIR	0x0
		Remote Fault	0x0
		Acknowledge	0x0
		Next Page	0x0
		Transmitted Nonce Field(T_0:4)	0x07
		Technology Ability Field(A_0:24)	0x000200
		A0 1000BASE-KX	0x0
		A1 10GBASE-KX4	0x0
		A2 10GBASE-KR	0x0
		A3 40GBASE-KR4	0x0
		A4 40GBASE-CR4	0x0
		A5 100GBASE-CR10	0x0
		A6 100GBASE-KP4	0x0
		A7 100GBASE-KR4	0x0
		A8 100GBASE-CR4	0x0
		A9 25GBASE-KR-S or 25GBASE-CR-S	0x0
		A10 25GBASE-KR or 25GBASE-CR	0x0
		A11 2_5GBASE-KX	0x0
		A12 5GBASE-KR	0x0
		A13 50GBASE-KR Or 50GBASE-CR	0x1
		A14 100GBASE-KR2 Or 100GBASE-CR2	0x0
		A15 200GBASE-KR4 Or 200GBASE-CR4	0x0
		A16 100GBASE-KR1 Or 100GBASE-CR1	0x0
		A17 200GBASE-KR2 Or 200GBASE-CR2	0x0
		A18 400GBASE-KR4 Or 400GBASE-CR4	0x0
		FEC Capability (F0:F3)	0x8
		F2: 25G R5-FEC Requested	0x1
		F3: 25G BASE-R FEC Requested	0x0
		F0: 10 Gb/s per lane FEC Ability	0x0
		F1: 10 Gb/s per lane FEC Requested	0x0

Ethernet Layers

Table 73-5—Priority Resolution



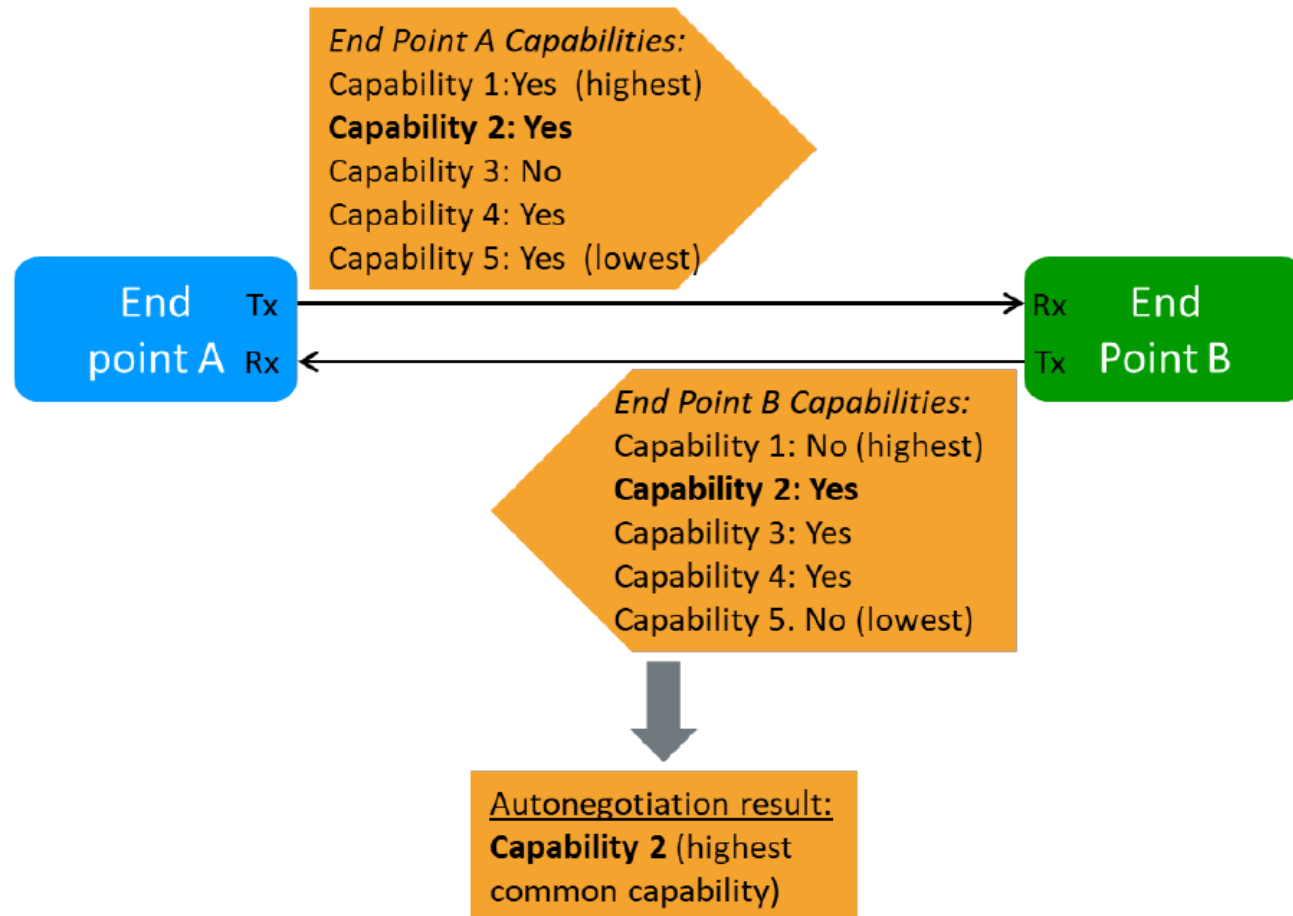
200GMII = 200 Gb/s MEDIA INDEPENDENT INTERFACE
 50GMII = 50 Gb/s MEDIA INDEPENDENT INTERFACE
 AN = AUTO-NEGOTIATION
 CGMII = 100 Gb/s MEDIA INDEPENDENT INTERFACE
 LLC = LOGICAL LINK CONTROL
 MAC = MEDIA ACCESS CONTROL
 MDI = MEDIUM DEPENDENT INTERFACE

MAC = MEDIA ACCESS CONTROL
 PCS = PHYSICAL CODING SUBLAYER
 PHY = PHYSICAL LAYER DEVICE
 PMA = PHYSICAL MEDIUM ATTACHMENT
 PMD = PHYSICAL MEDIUM DEPENDENT
 RS-FEC = REED-SOLOMON FORWARD ERROR CORRECTION

Figure 136-1—50GBASE-CR, 100GBASE-CR2, and 200GBASE-CR4 relationship to the ISO/IEC Open Systems Interconnection (OSI) reference model and the IEEE 802.3 Ethernet model

Priority	Technology	Capability
1	400GBASE-KR4 or 400GBASE-CR4	400 Gb/s 4 lane, highest priority
2	200GBASE-KR2 or 200GBASE-CR2	200 Gb/s 2 lane
3+4	200GBASE-KR4 or 200GBASE-CR4	200 Gb/s 4 lane, highest priority
4	100GBASE-KR1 or 100GBASE-CR1	100 Gb/s 1 lane
5+2	100GBASE-KR2 or 100GBASE-CR2	100 Gb/s 2 lane
6+3	100GBASE-CR4	100 Gb/s 4 lane
7+4	100GBASE-KR4	100 Gb/s 4 lane
8+5	100GBASE-KP4	100 Gb/s 4 lane
9+6	100GBASE-CR10	100 Gb/s 10 lane
10+7	50GBASE-KR or 50GBASE-CR	50 Gb/s 1 lane
11+8	40GBASE-CR4	40 Gb/s 4 lane
12+9	40GBASE-KR4	40 Gb/s 4 lane
13+10	25GBASE-KR or 25GBASE-CR	25 Gb/s 1 lane
14+11	25GBASE-KR-S or 25GBASE-CR-S	25 Gb/s 1 lane, short reach
15+12	10GBASE-KR	10 Gb/s 1 lane
16+13	10GBASE-KX4	10 Gb/s 4 lane
17+14	5GBASE-KR	5 Gb/s 1 lane
18+15	2.5GBASE-KX	2.5 Gb/s 1 lane
19+16	1000BASE-KX	1 Gb/s 1 lane, lowest priority

How does link Auto Negotiation work?



Base Pages

Auto-Negotiation	0x80003800 2008
Selector Field(S_0:4)	0x10 : IEEE Std 802.3
Echoed Nonce Field(E_0:4)	0x00
Pause Ability (C0:C2)	0x0
C0: PAUSE	0x0
C1: ASM_DIR	0x0
Remote Fault	0x0
Acknowledge	0x0
Next Page	0x0
Transmitted Nonce Field(T_0:4)	0x07
Technology Ability Field(A_0:24)	0x000200
A0 100BASE-KX	0x0
A1 10GBASE-KX4	0x0
A2 10GBASE-KR	0x0
A3 40GBASE-KR4	0x0
A4 40GBASE-CR4	0x0
A5 100GBASE-CR10	0x0
A6 100GBASE-KP4	0x0
A7 100GBASE-KR4	0x0
A8 100GBASE-CR4	0x0
A9 25GBASE-KR-5 or 25GBASE-CR-5	0x0
A10 25GBASE-KR or 25GBASE-CR	0x0
A11 2_5GBASE-KX	0x0
A12 5GBASE-KR	0x0
A13 50GBASE-KR Or 50GBASE-CR	0x1
A14 100GBASE-KR2 Or 100GBASE-CR2	0x0
A15 200GBASE-KR4 Or 200GBASE-CR4	0x0
A16 100GBASE-KR1 Or 100GBASE-CR1	0x0
A17 200GBASE-KR2 Or 200GBASE-CR2	0x0
A18 400GBASE-KR4 Or 400GBASE-CR4	0x0
FEC Capability (F0:F3)	0x8
F2: 25G R5-FEC Requested	0x1
F3: 25G BASE-R FEC Requested	0x0
F0: 10 Gb/s per lane FEC Ability	0x0
F1: 10 Gb/s per lane FEC Requested	0x0

Each device begins Auto negotiation by sending a Base Page that describes what speeds, signaling, and FEC are supported by the device.

The devices repeatedly send these pages until the ACK bit is set – indicating that the Base Page has been received.

Next Pages

Message Page indicates a code number that describes what pages are coming next

These are sent until the ACK bit is set

Field	Value
Auto-Negotiation	0xA015CAC0 FB20
Auto-negotiation OUI Extended Next page	0xA015CAC0 FB20
Message Code	0x005 : Organizationally Unique Identifier (OUI) tag code
ACK1	0x1
ACK2	0x0
MP	0x1
ACK	0x0
Next Page	0x1
OUI_12:2	0x656
OUI_23:13	0x7D9

Additional Pages

More pages are sent (Extended Technology Ability Field, Remote Fault message, OUI, Phy identify tag code)

Field	Value
Auto-Negotiation	0xC0400040 0000
Auto-negotiation OUI Extended Next page	0xC0400040 0000
Extended Technology Abilities	0x3
D_8:2	0x00
OUI_1:0	0x2
T Bit	0x0
D_13:12	0x0
ACK	0x0
Next Page	0x0
D_19:16	0x0
25GBASE-KR1	0x0
25GBASE-CR1	0x0
D_23:22	0x0
50GBASE-KR2	0x0
50GBASE-CR2	0x1
D_26	0x0
400GBASE-KR8 / CR8	0x0
LL-RS-FEC Ability	0x0
FEC Control	0x0
F1: Clause 91 FEC ability	0x0
F2: Clause 74 FEC ability	0x0
F3: Clause 91 FEC requested	0x0
F4: Clause 74 FEC requested	0x0
LL-RS-FEC Request	0x0

These are sent until the ACK bit is set

If the Next Page bit is set, then
More pages will be sent in the same way

What to look for in Auto Negotiation

- Did Auto Negotiation occur?
- Was it sent in both directions?
- Were the capability bits set correctly?
- Incorrect transmission or response of parameters in base or extended pages
- Time Outs - Receiver does not acknowledge receipt of a page, or takes too long to ACK
- Did the link progress to Link Training?

How does link training work?

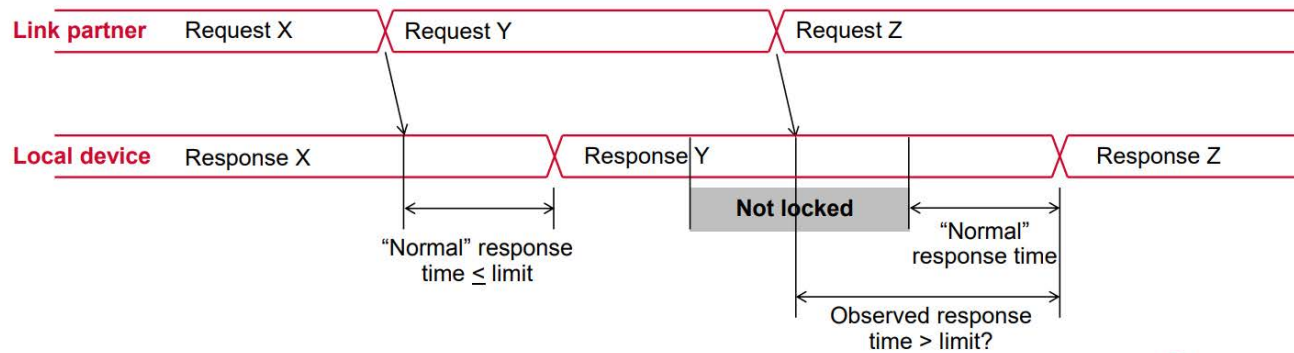


L1 – Transmitter Training

Example



- IEEE Std 802.3-2015, 92.7.12 item b)
 - The start of the period is the frame marker of the training frame with the new request and the end of the period is the frame marker of the training frame with the corresponding response.
 - A new request occurs when the coefficient update field is different from the coefficient field in the preceding frame. The response occurs when the coefficient status report field is updated to indicate that the corresponding action is complete.



18 | IEEE P802.3cd Task Force, July 2016



https://standards.incits.org/apps/group_public/download.php/82832/T11-2013-162v3.pdf

https://www.ieee802.org/3/cd/public/July16/healey_3cd_01a_0716.pdf

L1 – Transmitter Training

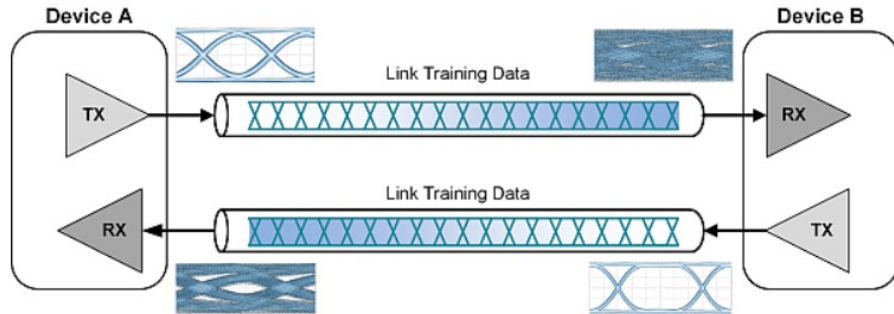
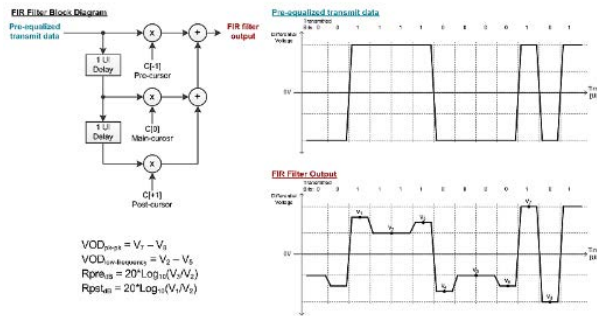


Figure 2 Link training phase



WHITE PAPER

The NRZ and PAM4 Line Code

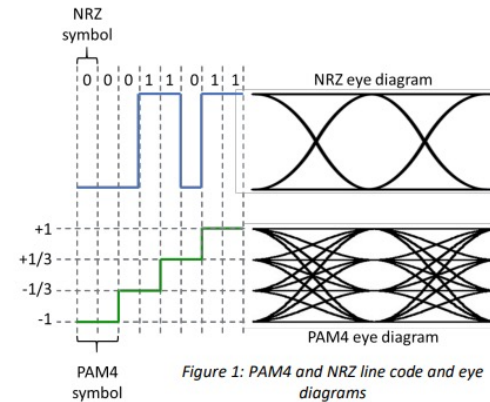
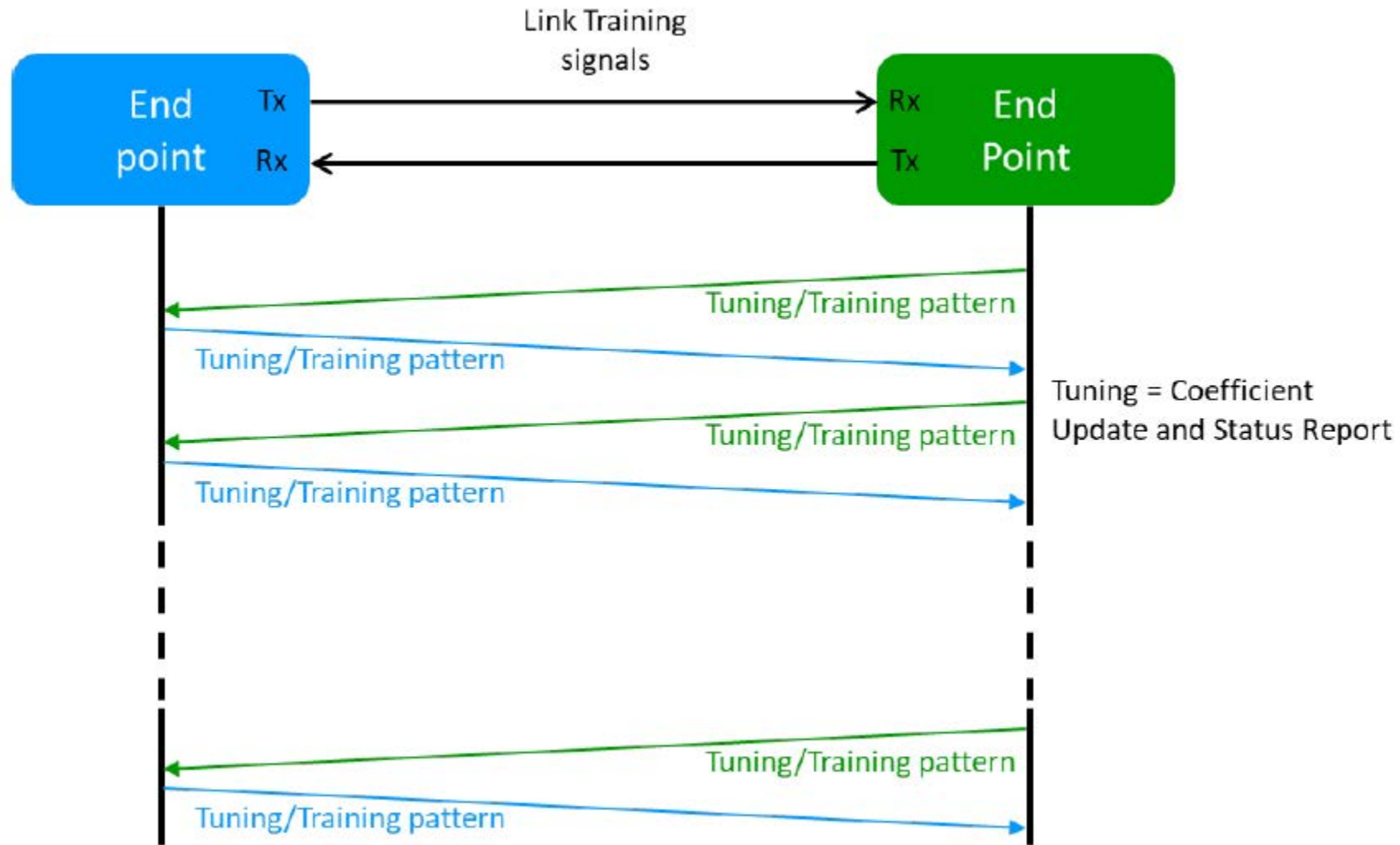


Figure 1: PAM4 and NRZ line code and eye diagrams

Many Ethernet connections have been based on the Non-Return-to-Zero (NRZ) line code, which transfers 1 bit per clock symbol. So far this has been used for SerDes speeds up to 25 Gbps. A 100 Gbps Ethernet connection defined with current standards is aggregated over four 25 Gbps SerDes – and even older standards define 100GbE sent on ten 10 Gbps SerDes. Higher transmission rates could be achieved by sending signals as a higher multiple of more 25 Gbps SerDes, but a desire to reduce the spectral bandwidth and number of SerDes used for the signal has driven a new line code for the high-speed signals: PAM4 (Pulse Amplitude Modulation), which encodes two bits in a single symbol by using 4 signal levels. This provides 53.125 Gbps SerDes with a symbol rate (or baud rate) of 26.5625 GBaud. For simplicity these are normally referred to as 50 Gbps or 25 GBaud SerDes. Figure 1 illustrates the difference between the NRZ and PAM line codes.

<https://www.edn.com/what-is-link-training-and-when-should-i-use-it/>
 Xena and Teledyne LeCroy White Paper

How does link training work?



How are the messages sent?

- The messages are sent with Manchester encoding (see next 2 slides)
- They are sent over and over until a response is given
- The systems do not have to achieve frame lock in order to communicate
- Frames with Manchester errors are/should be ignored

Start Time	Port	Speed	Port No	Frame	Frame	
02.838 117 277 000(s)	← P10	200G PAM4	← P2		Training Sequence	completed ; Lane No= 1
02.838 117 903 000(s)	← P10	200G PAM4	← P2		1684 - Training Sequence	completed ; Lane No= 1
02.838 308 298 000(s)	P9 →	200G PAM4	P1 →	905 - Training Sequence		completed ; Lane No= 1
02.838 876 614 000(s)	P9 →	-	P1 →	0x00:Loss of Sync		Lane No= 1
02.839 175 157 000(s)	← P10	-	← P2		0x00:Loss of Sync	Lane No= 1
02.843 117 735 000(s)	← P10	200G PAM4	← P2		28468 - Training Sequence	completed ; Lane No= 2
02.848 010 854 000(s)	← P10	200G PAM4	← P2		2424 - Training Sequence	completed ; Lane No= 3
02.848 509 181 000(s)	P9 →	200G PAM4	P1 →	910 - Training Sequence		completed ; Lane No= 3
02.849 080 633 000(s)	P9 →	-	P1 →	0x00:Loss of Sync		Lane No= 3
02.849 532 573 000(s)	← P10	-	← P2		0x00:Loss of Sync	Lane No= 3
02.860 505 757 000(s)	P9 →	200G PAM4	P1 →	741 - Training Sequence		Preset 2 ; Lane No= 2
02.860 971 143 000(s)	P9 →	-	P1 →	0x00:Loss of Sync		Lane No= 2

Manchester Encoding

136.8.11.1.1 Frame marker

Training frames are delimited by a specific sequence of PAM4 symbols. The training frame marker is a run of 16 consecutive “3” symbols followed by a run of 16 consecutive “0” symbols. This sequence is not found in the control field, status field, or training pattern and it uniquely identifies the beginning of a training frame.

136.8.11.1.2 Control and status fields

The control field comprises 16 bits with the structure defined in 136.8.11.2. The status field comprises 16 bits with the structure defined in 136.8.11.3.

Each bit of the control and status fields is sent as a differential Manchester encoded (DME) cell, where each cell is eight unit intervals in length. The specific rules for this encoding follow.

- a) A transition from 0 to 3 or from 3 to 0 occurs at the start of each cell.
- b) A transition from 0 to 3 or from 3 to 0 at the midpoint of a cell, i.e., four unit intervals from the transition at the beginning of the cell, corresponds to a logical one.
- c) The absence of a transition at the midpoint of a cell corresponds to a logical zero.

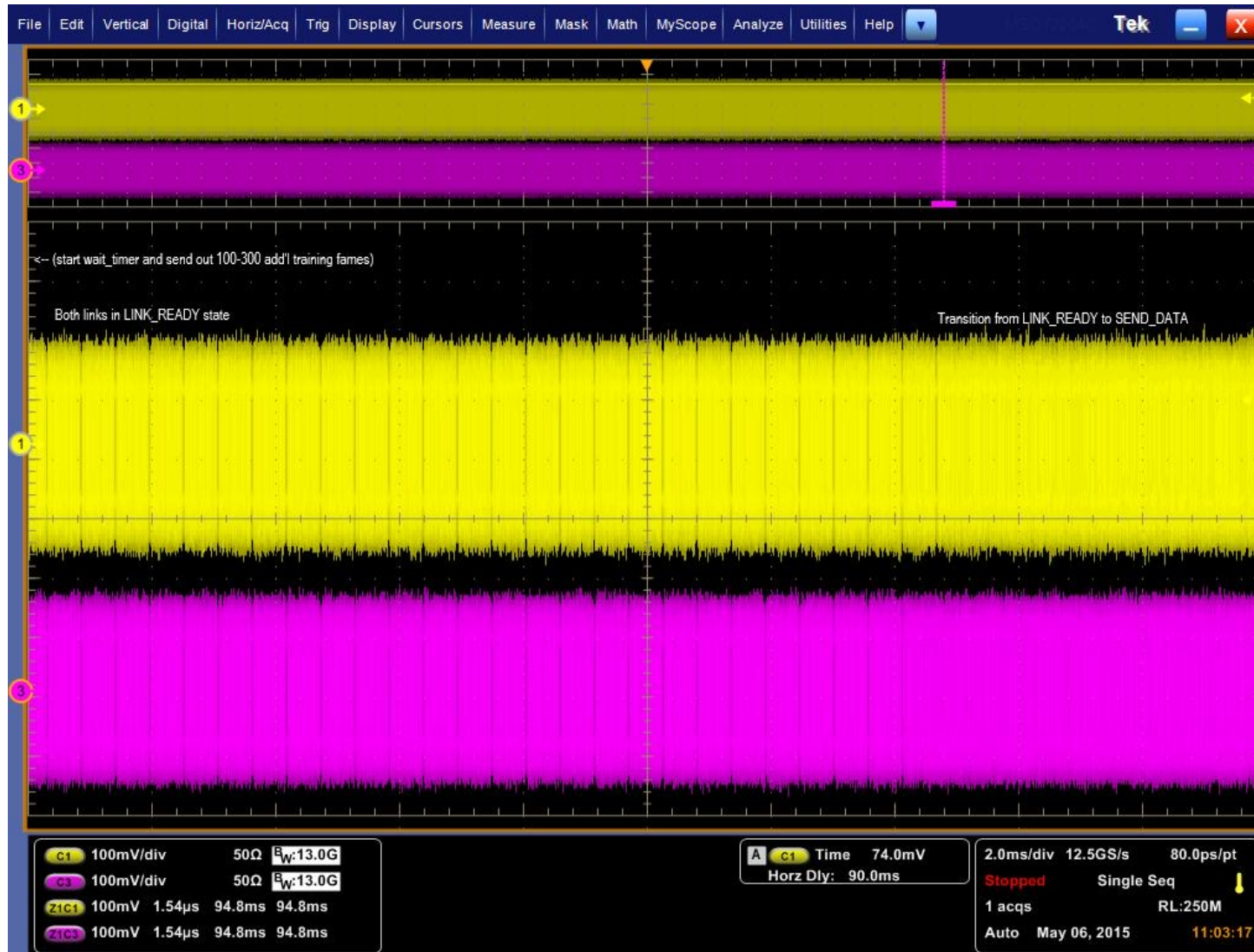
The control field is transmitted immediately after the frame marker. The status field is transmitted immediately after the control field. Within each field, the order of transmission is from bit 15 to bit 0.

When a training frame is received, if a violation of the DME encoding rules is detected within the control field or the status field, the contents of both fields in that frame are ignored.

Alternating Speeds



Ready to Send Data (PRBS to PCS)



Link Training Frame - Control

Table 136-9—Control field structure

Bit(s)	Name	Description
15:14	Reserved	Transmit as 0, ignore on receipt
13:12	Initial condition request	13 12 1 1 = Preset 3 1 0 = Preset 2 0 1 = Preset 1 0 0 = Individual coefficient control
11:10	Reserved	Transmit as 0, ignore on receipt
9:8	Modulation and precoding request	9 8 1 1 = PAM4 with precoding 1 0 = PAM4 0 1 = Reserved 0 0 = PAM2
7:5	Reserved	Transmit as 0, ignore on receipt
4:2	Coefficient select	4 3 2 1 1 0 = $c(-2)$ 1 1 1 = $c(-1)$ 0 0 0 = $c(0)$ 0 0 1 = $c(1)$
1:0	Coefficient request	1 0 1 1 = No equalization 1 0 = Decrement 0 1 = Increment 0 0 = Hold

Link Training Frame - Status

Table 136-10—Status field structure

Bit(s)	Name	Description
15	Receiver ready	1 = Training is complete and the receiver is ready for data 0 = Request for training to continue
14:12	Reserved	Transmit as 0, ignore on receipt
11:10	Modulation and precoding status	11 10 1 1 = PAM4 with precoding 1 0 = PAM4 0 1 = Reserved 0 0 = PAM2
9	Receiver frame lock	1 = Frame boundaries identified 0 = Frame boundaries not identified
8	Initial condition status	1 = Updated 0 = Not updated
7	Parity	Even parity bit
6	Reserved	Transmit as 0, ignore on receipt
5:3	Coefficient select echo	5 4 3 1 1 0 = $c(-2)$ 1 1 1 = $c(-1)$ 0 0 0 = $c(0)$ 0 0 1 = $c(1)$
2:0	Coefficient status	2 1 0 1 1 1 = Reserved 1 1 0 = Coefficient at limit and equalization limit 1 0 1 = Reserved 1 0 0 = Equalization limit 0 1 1 = Coefficient not supported 0 1 0 = Coefficient at limit 0 0 1 = Updated 0 0 0 = Not updated

Link Training Frame - Example

Training Sequence	0x00000000
Control Field	0x0000
Initial condition request	0x0 : Individual coefficient control
Modulation and precoding request	0x0 : PAM2
Coefficient select	0x0 : c(0)
Coefficient request	0x0 : Hold
Status Report Field	0x0000
Receiver Ready	0x0 : Continue
Modulation and precoding status	0x0 : PAM2
Receiver frame lock	0x0 : Frame boundaries not identified
Initial condition status	0x0 : Not updated
Parity	0x0
Coefficient select echo	0x0 : c(0)
Coefficient status	0x0 : Not updated

Previous Coefficients

TXFFE Overview

- 3 tap feed forward equalization
- The output waveform is a summation of the following:
 - Input waveform * Pre-Cursor tap (inverted)
 - Input waveform * Main tap (delayed by 1UI)
 - Input waveform * Post-Cursor tap (inverted, delayed by 2 UI)

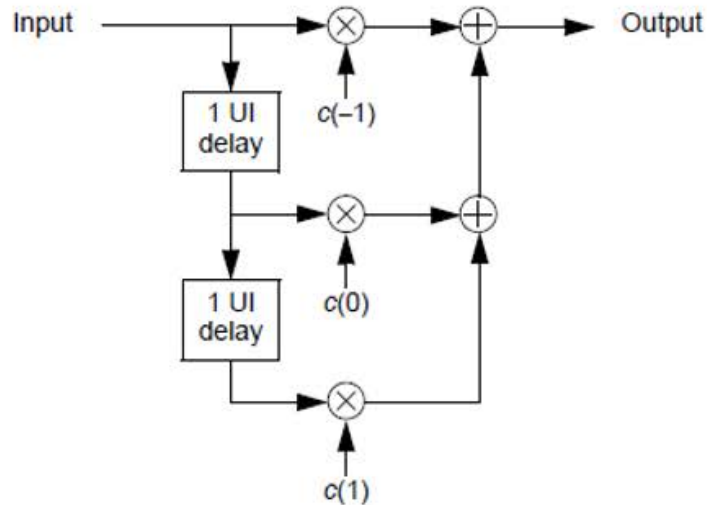
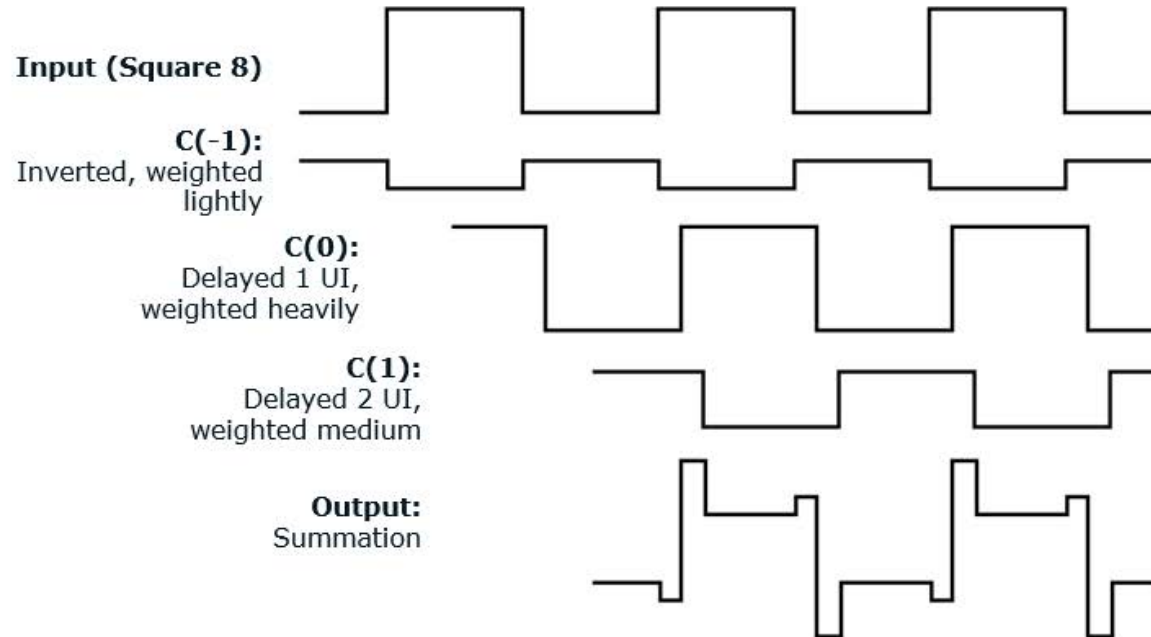


Figure 72-11—Transmit equalizer example



Coefficients

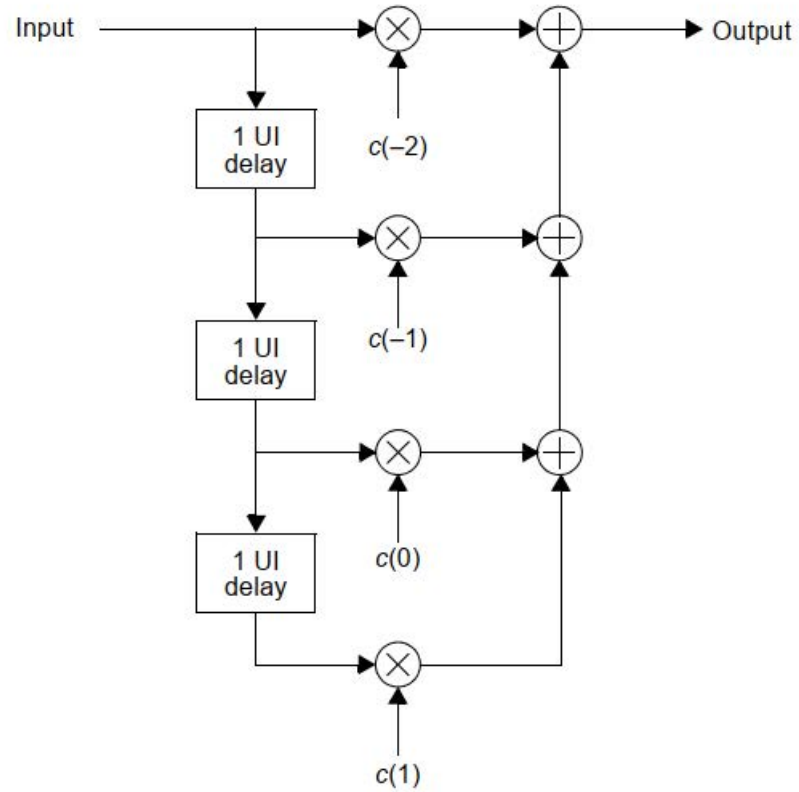
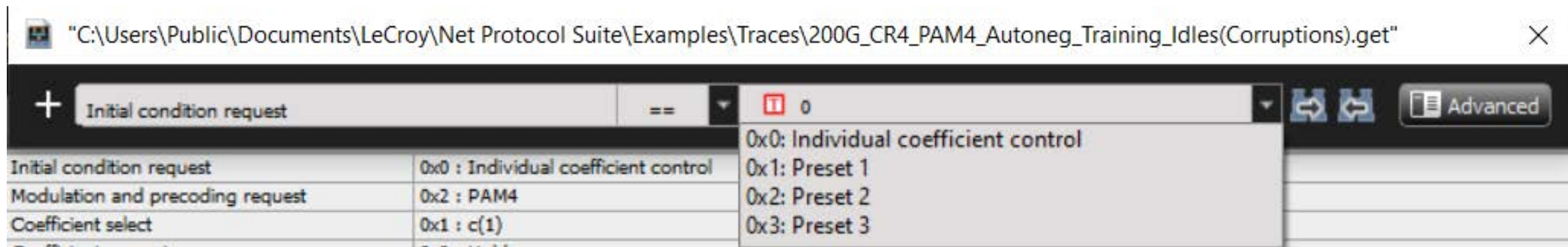


Figure 136–10—Transmit equalizer functional model

Link Training Control – Initial Condition Request



ICR	c(-2)	c(-1)	c(0)	c(1)	Comment
Preset 1	0	0	1	0	Default
Preset 2	0	-0.15	.75	-0.1	Most common
Preset 3	0	-0.25	.75	0	

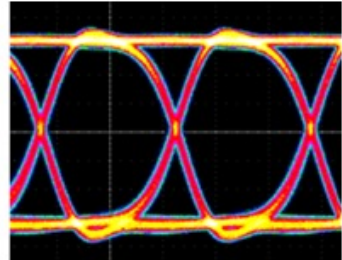
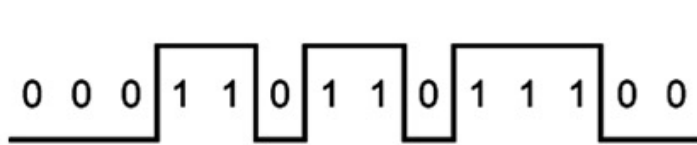
Link Training Control – Modulation and Precoding Request

"C:\Users\Public\Documents\LeCroy\Net Protocol Suite\Examples\Traces\200G_CR4_PAM4_Autoneg_Training_Idles(Corruptions).get"

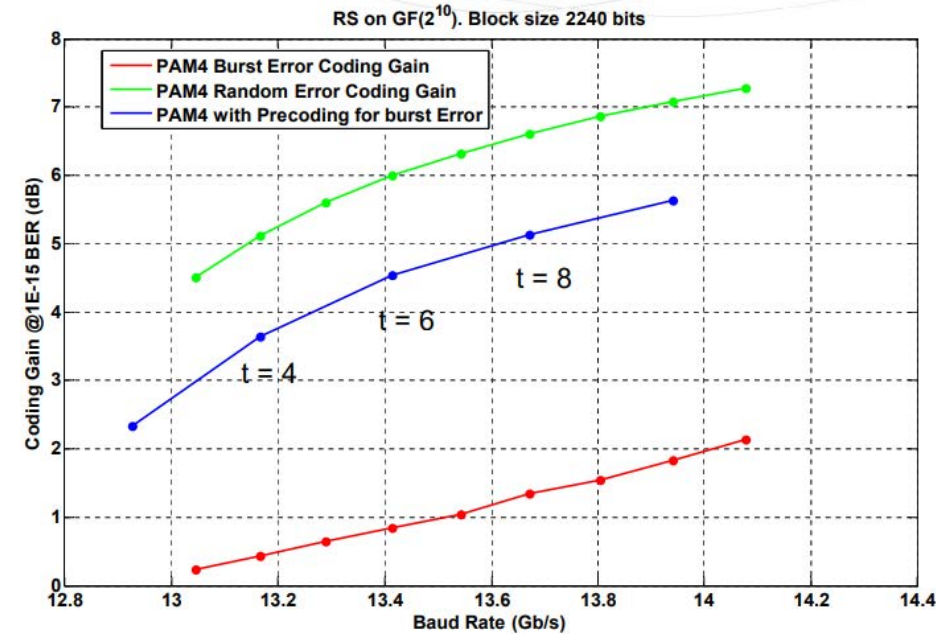
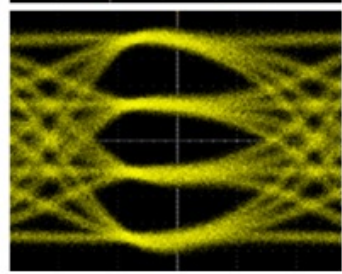
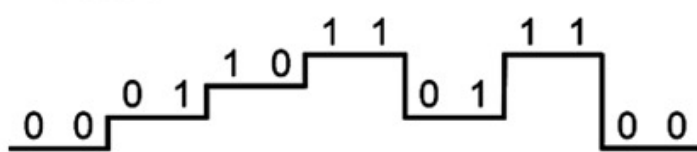
Modulation and precoding request == d

- 0x0: PAM2
- 0x2: PAM4
- 0x3: PAM4 with precoding
- 0x1: Reserved

PAM2-NRZ



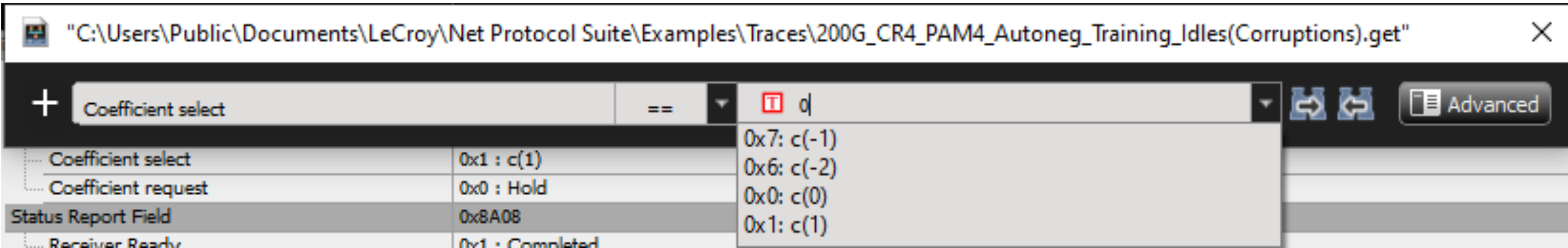
PAM4



[Precoding proposal for PAM4 modulation](#)

[High-speed Signal Interconnection Technology of Next-generation Data Center-PAM4](#)

Link Training Control – Coefficient Select



The screenshot shows a network protocol analyzer window with the following details:

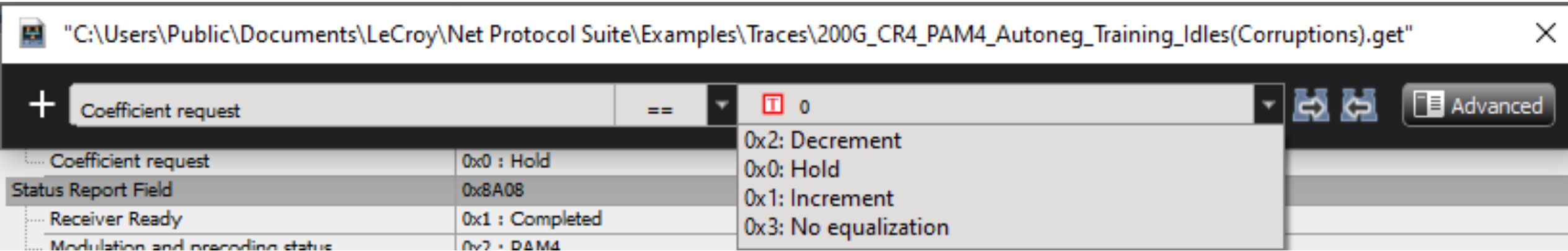
Field	Value
Coefficient select	0x1 : c(1)
Coefficient request	0x0 : Hold
Status Report Field	0x8A08
Receiver Ready	0x1 : Completed

The packet details pane shows a 'Coefficient select' packet with a value of '0x0'. A dropdown menu is open, showing the following options:

- 0x7: c(-1)
- 0x6: c(-2)
- 0x0: c(0)
- 0x1: c(1)

- Specifies the coefficient that is to be modified as defined in the request
- Presets are primarily sent to c(0) which updates all coefficients

Link Training Control – Coefficient Request



The screenshot shows a network protocol analyzer window with the title "C:\Users\Public\Documents\LeCroy\Net Protocol Suite\Examples\Traces\200G_CR4_PAM4_Autoneg_Training_Idles(Corruptions).get". The main display area shows a table of fields with a dropdown menu open over the "Coefficient request" field. The dropdown menu lists four options: "0x2: Decrement", "0x0: Hold", "0x1: Increment", and "0x3: No equalization". The "Coefficient request" field in the table is currently set to "0x0 : Hold".

Field Name	Value
Coefficient request	0x0 : Hold
Status Report Field	0x8A08
Receiver Ready	0x1 : Completed
Modulation and precoding status	0x2 : PAM4

- Hold – Don't change the selected coefficient
- Increment – Increment the selected coefficient by:
 - between 0.005 and 0.05 for -1, 0, 1 coefficients
 - between 0.005 and 0.025 for -2 coefficient
- Decrement – Decrement the selected coefficient by:
 - between 0.005 and 0.05 for -1, 0, 1 coefficients
 - between 0.005 and 0.025 for -2 coefficient
- No equalization – Preset 1 or set to zero

Link Training Status – Receiver Ready

The screenshot shows a network protocol suite trace window. The title bar indicates the file path: "C:\Users\Public\Documents\LeCroy\Net Protocol Suite\Examples\Traces\200G_CR4_PAM4_Autoneg_Training_Idles(Corruptions).get". The main area displays a table of events. The first event is "Receiver Ready" with a value of "0x1 : Completed". The second event is "Modulation and precoding status" with a value of "0x2 : PAM4". A search filter is applied to the "Receiver Ready" event, showing a dropdown menu with options "0x1: Completed" and "0x0: Continue". The search filter is currently set to "0".

Event	Value
Receiver Ready	0x1 : Completed
Modulation and precoding status	0x2 : PAM4

- Link training is complete for this lane and direction

Link Training Status – Parity

The image shows two screenshots of a network analysis tool's filter interface. The top screenshot shows a filter rule: "Parity == 00". Below the filter bar, a table shows the field "Parity" with the value "0x0". The bottom screenshot shows a filter rule: "Parity == 01". Below the filter bar, a table shows the field "Parity" with the value "0x1". Both screenshots include a search bar with the filter expression, a search button, a refresh button, and an "Advanced" button.

Even parity bit:

“The parity bit is calculated based on the other bits in the control field and status field to create even parity for these fields. Even parity ensures that the transmitted control and status fields (see 136.8.11.1.2) are DC balanced. This field is ignored on receipt. “

Link Training Status – Coefficient Select Echo

The screenshot shows a network protocol analyzer window with the following details:

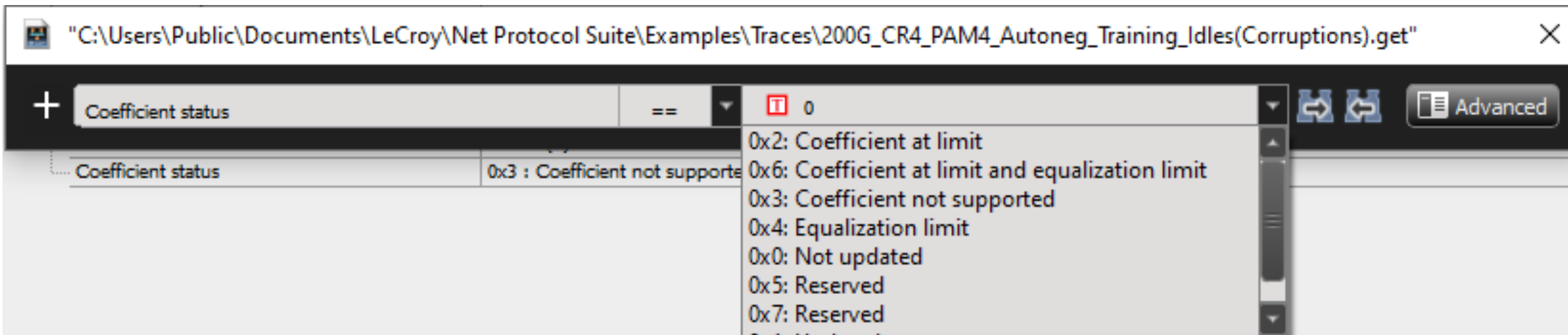
Field Name	Value
Coefficient select echo	0x0 : c(0)
Coefficient status	0x3 : Coefficient not supported

The dropdown menu for the Coefficient Status field contains the following options:

- 0x7: c(-1)
- 0x6: c(-2)
- 0x0: c(0)
- 0x1: c(1)

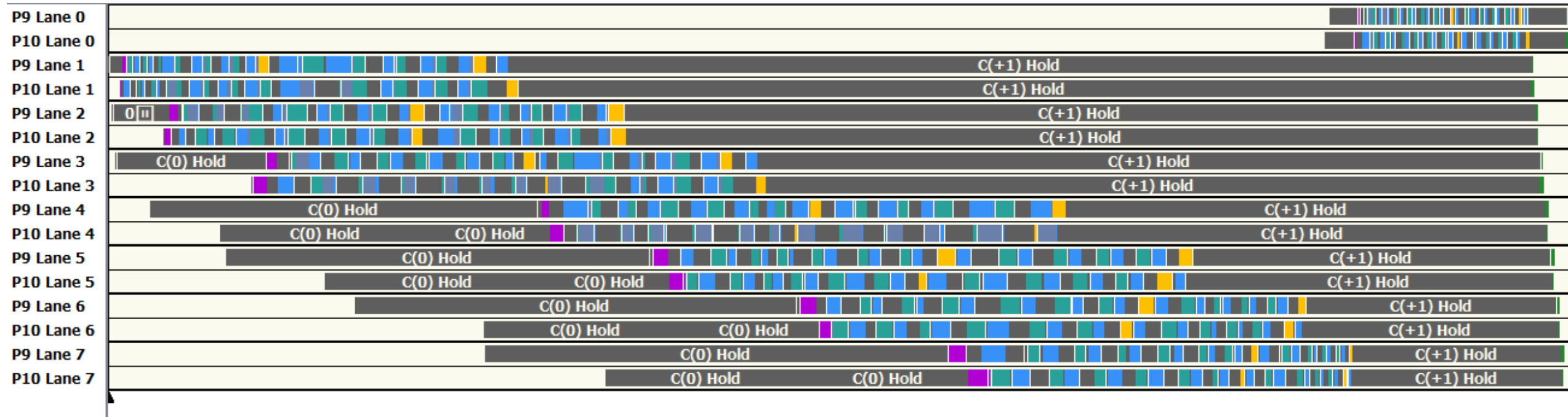
Specifies which coefficient the Coefficient Status pertains to.

Link Training Status – Coefficient Status



- Not updated – the coefficient has not been updated to the requested value
- Updated (0x1) – the coefficient has been updated to the requested value
- Coefficient at limit – A minimum (decrement) or a maximum (increment) has been reached. The requested won't action will not be applied
- Coefficient not supported – the device does not support the selected coefficient (note I have seen coefficient requests for unknown coefficients)
- Equalization limit – Adjusting the coefficient would exceed the transmitters voltage limits. The coefficient has not yet reached its limit so if other coefficients are changed then it could be updated
- Coefficient at limit and equalization at limit – Adjusting the coefficient would exceed the transmitters voltage limits and the coefficient's limits as well.

What it looks like over time



AN/LT to PCS

The screenshot displays a simulation interface with a top status bar and a block diagram below. The status bar shows a time of 04.371 037 618 280(s), a port P9 connected to a 200G PAM4 source, and a port P1 with a 'Remote Fault' error highlighted in a red box. Below the status bar is a control panel with a 'Go to Iteration: 2 of 2' field and various navigation buttons including play, stop, zoom, and level selection (L0-L3). The block diagram at the bottom shows a sequence of blocks: AN (green), TS (green), AN (white), TS (white), PCS (white), TS (white), and PCS (white). A red arrow points from the bottom right towards the PCS block in the diagram.

Link Layer Validation

The screenshot displays the Teledyne LeCroy Net Protocol Suite interface, showing various views for link layer validation. The top view is the Exchange View, which is a table of network events. Below it are the Link State View and Timeline View. The bottom section contains the Listing/State Diagram View and the Navigation View.

No.	Marker	Start Time	Duration	Port No	Lane No	Speed	Modulation	an(Coefficients	Protocol	Frame	Nominal FC_Port RTT (ns)	Frame	Summary
5		0:00:00.035 (hr)	30992...	P9	0	400G PAM4	0x0:PAM2	0x0:c(0)	Ethernet	4937901 - Training Sequ...			Lane No=0
6		0:00:00.039 (hr)	20025...	P9	1	400G PAM4	0x0:PAM2	0x0:c(0)	Ethernet	31906 - Training Sequen...			Lane No=1
7		0:00:00.045 (hr)	18031...	P9	2	400G PAM4	0x0:PAM2	0x0:c(0)	Ethernet	28729 - Training Sequen...			Lane No=2
8		0:00:00.046 (hr)	16349...	P9	0	400G PAM4	0x0:PAM2	0x0:c(0)	Ethernet	26049 - Training Sequen...			Lane No=0
9		0:00:00.049 (hr)	20050...	P9	3	400G PAM4	0x0:PAM2	0x0:c(0)	Ethernet	31945 - Training Sequen...			Lane No=3

The Listing/State Diagram View shows a table of events and a corresponding state transition diagram.

Port	Event	Start	Stop
P9	Auto Negotiation	0:00:00.035 (hr)	0:00:00.046 (hr)
P9	Auto Negotiation Enable	0:00:00.035 (hr)	0:00:00.035 (hr)
P9	AN Good Check Transition to "AN Good Check" occurred while in state "" Reason: Training Sequence detected	0:00:00.035 (hr)	0:00:00.035 (hr)
P9	Transmit Disable	0:00:00.035 (hr)	0:00:00.035 (hr)
P9	Ability Detect	0:00:00.035 (hr)	0:00:00.046 (hr)
P9	AN Good Check Transition to "AN Good Check" occurred while in state "Ability Detect" Reason: Training Sequence detected	0:00:00.046 (hr)	0:00:00.046 (hr)
P9	AN Good	0:00:00.046 (hr)	0:00:00.046 (hr)
P9	Transmitter Training	0:00:00.046 (hr)	0:00:00.249 (hr)
P9	Train Init	0:00:00.046 (hr)	0:00:00.046 (hr)
P9	Send Training	0:00:00.046 (hr)	0:00:00.046 (hr)

The state diagram shows the following states and transitions:

- Auto Negotiation Enable** (0:00:00.035 (hr) to 0:00:00.035 (hr), Duration: NA)
- AN Good Check** (0:00:00.046 (hr) to 0:00:00.046 (hr), Duration: NA)
- AN Good** (0:00:00.046 (hr) to 0:00:00.046 (hr), Duration: NA)
- Next Page Wait** (State Never Entered)

The Navigation View shows a sequence of states: AN, TS, AN, TS.

What to look for in Link Training

- Did Link Training occur?
- Was it sent in both directions?
- Did Link Training end with the 'Complete' on all lanes and directions?
- Did the link progress to PCS (Physical Coding Sublayer)?
- Are there 'Remote Faults'?
- Make sure there aren't multiple iterations of AN/LT (unless they were specifically caused for testing)

Consistent Link Training?

Completed Comparison

Port 9

Lane	c(-2)	c(-1)	c(0)	c(1)	Presets and limits
0	0.0000(13)	-0.0275(1) 0.0000(12)	0.9725(2) 0.6150(1) 1.0000(10)	0.0000(13)	preset 1 on c(0) (12) N/A(1)
1	0.0000(13)	0.0000(13)	1.0000(13)	0.0000(13)	preset 1 on c(0) (12) N/A(1)
2	0.0000(13)	0.0000(13)	1.0000(13)	-0.0275(1) 0.0000(12)	preset 1 on c(0) (10) N/A(3)
3	0.0000(13)	-0.0275(1) 0.0000(12)	1.0000(13)	-0.0275(1) 0.0000(12)	preset 1 on c(0) (12) N/A(1)

Port 10

Lane	c(-2)	c(-1)	c(0)	c(1)	Presets and limits
0	0.0000(13)	0.0000(13)	1.0000(13)	0.0000(13)	preset 1 on c(0) (12) N/A(1)
1	0.0000(13)	0.0000(13)	0.9725(1) 1.0000(12)	0.0000(13)	preset 1 on c(0) (12) N/A(1)
2	0.0000(13)	-0.0275(3) 0.0000(10)	0.9450(1) 1.0000(12)	0.0000(13)	preset 1 on c(0) (12) N/A(1)
3	0.0000(13)	0.0000(13)	0.9725(1) 1.0000(12)	-0.0275(1) 0.0000(12)	preset 1 on c(0) (12) N/A(1)

Summary

- Reliable Link Training is key to making communication reliable over copper
- Testing with proper visibility is needed to ensure the training is working as expected
- It is important that all frames with errors are ignored
- Even after the link is trained, it is important to make sure it remains stable

Link Establishment

Sam Johnson, HSN Co-Subcommittee Chair, Intel



ethernet alliance

www.ethernetalliance.org

Link Establishment

Link can only be established if both end points are in a compatible configuration

- Auto-negotiation resolves PHY parameters including link speed, lane count, and FEC mode
- Static configurations must match exactly

Proprietary link establishment methodologies

- Automatic media-based static configuration
- Loop of supported configurations
- Received signal detection
- Manual configuration

Port Configuration Options: Breakout modes



Assessing Link Health

Specification requirement for PAM-4 link health: Frame Loss Ratio (FLR)

- Previous standards used BER requirements, usually 1×10^{-12}
- With every link protected by FEC and ~ 6 dB coding gain, standards change to FLR
- 6.2×10^{-10} for 64-octet frames with min IPG for 50/100GbE
- 6.2×10^{-11} for 200/400GbE
- FLR requirement expected to be met with raw BER of 2.4×10^{-4} for 50Gb per lane.

Pre-FEC error ratio and FEC codeword bin counters can give assessment of link health and margin

No spec for receiver statistics (eye height, eye opening)

Time to Link heavily impacted by link establishment methodology

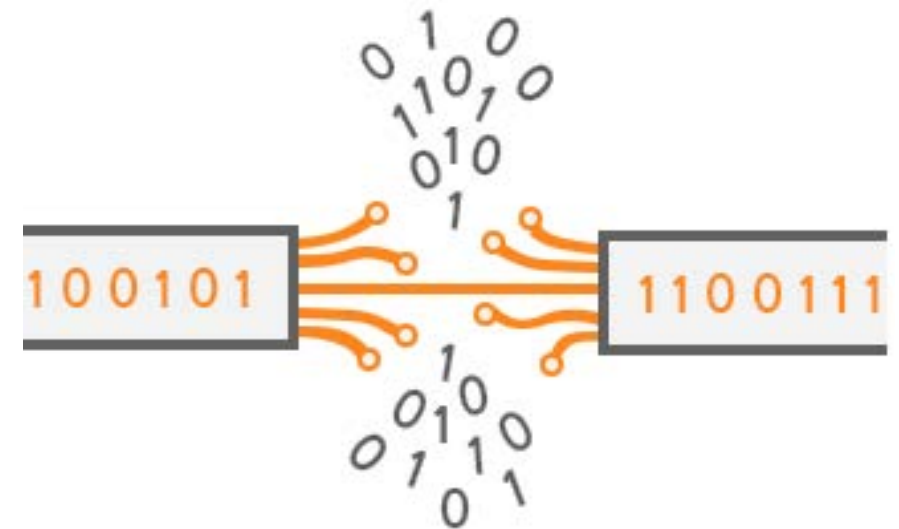


Photo Source: <https://www.vyopta.com/blog/video-conferencing/understanding-packet-loss/>

Interop Challenges and Common Issues

Sam Johnson, HSN Co-Subcommittee Chair, Intel



ethernet alliance

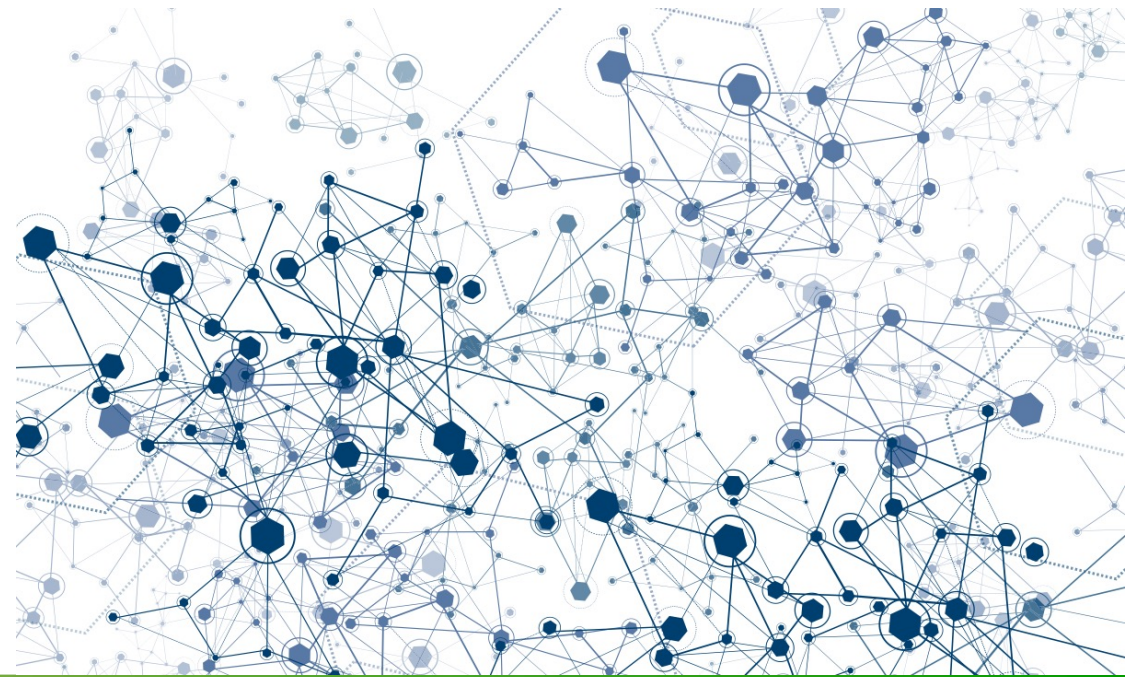
www.ethernetalliance.org

Ethernet Interoperability

Definition: The ability of devices from different vendors to work together seamlessly in order to establish a reliable data connection for the passing of Ethernet traffic

Recipe for success:

- Technology compatibility
- Specification compliance of devices and medium
- Device configuration
- Testing, testing, testing



The Importance of Interoperability Testing

Interop testing **during** product development:

- Set product development direction
- Solve problems before they are found in the field
- Improve product robustness

Interop testing **after** product development:

- New products to market can cause increased issues
- Product use-cases expand over time
- Product updates can result in very different behavior



Interop Challenges: Spec vs. Ecosystem

Legacy speed support

Conflicting link establishment methods

Evolving standards, especially for modules

Consortium vs. Standards compatibility

Prevalence of non-compliant device and configurations in deployment, for example:

- Auto Negotiation disabled on DACs
- Lack of RS-FEC support in legacy devices
- Better-than-spec devices
- Proprietary technologies
- Compliance with pre-ratified standards



Common Issues: System vs. Cables/Modules

Module EEPROM accuracy, lack of compliance enforcement

- Impacts media identification: type, capabilities, power, etc.
- Date code differences

Module control, communication and timing behavior

System to system implementation variation

Module insertion and squelching

Various specifications: SFF vs. CMIS

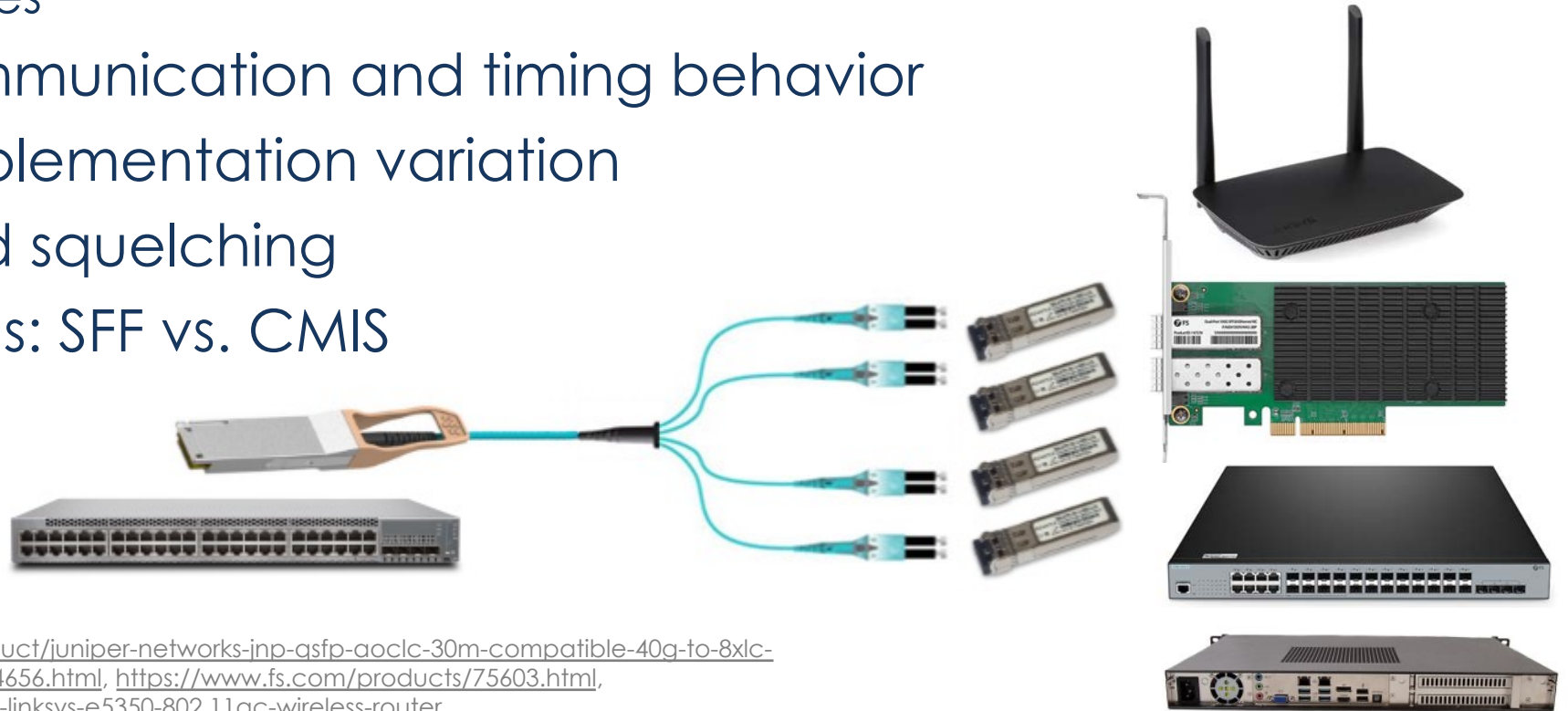


Photo sources: <https://advance.technology/product/juniper-networks-jnp-asfp-aoclc-30m-compatible-40g-to-8xlc-breakout-30m/>, <https://www.fs.com/products/134656.html>, <https://www.fs.com/products/75603.html>, <https://www.sweetwater.com/store/detail/E5350--linksys-e5350-802.11ac-wireless-router>, <https://www.onlogic.com/mk100b-54/>

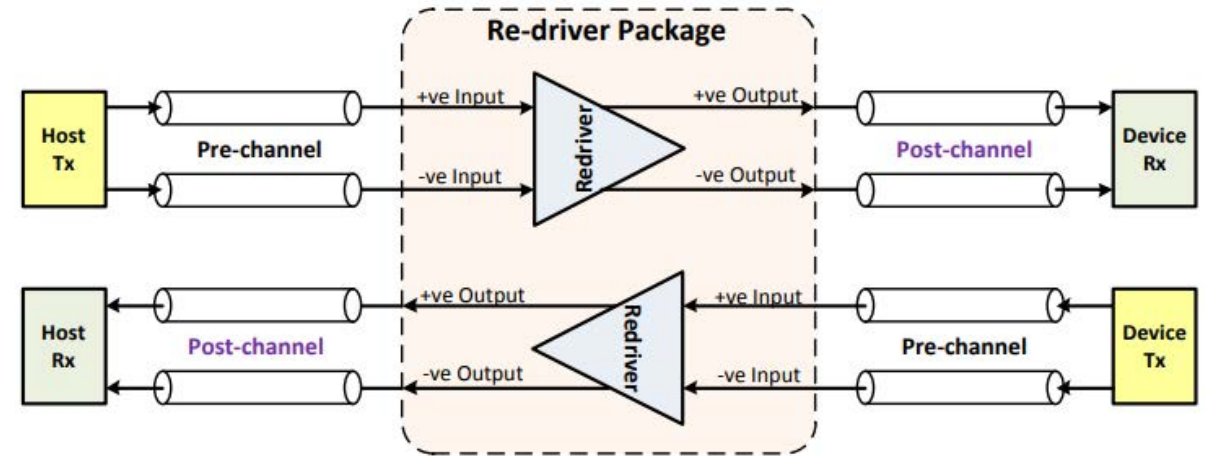
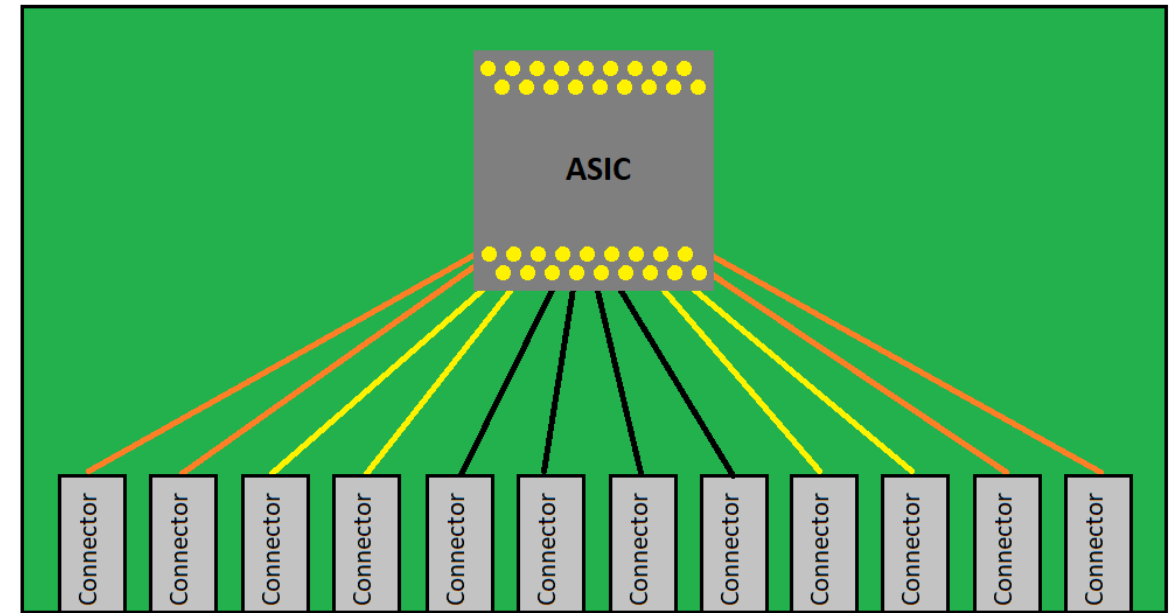
System Design Challenges

Platform form factor and routing implications

- Switches, outside port routes
- ASICs with integrated Ethernet

Repeaters/External PHYs used to extend channel reach

- Redrivers: noise sensitivity, electrical compliance, device location impact
- Retimers: link training interference, protocol awareness requirement



Interop Plugfest Value

Sam Johnson, HSN Subcommittee Co-Chair, Intel



ethernet alliance

www.ethernetalliance.org

Ethernet Alliance Plugfest Opportunity

Multi-vendor event open to EA members to test product interoperability and conformance methodologies

- Participation from System, T&M and component vendors

Emphasize testing on **latest technologies** as well as interop with **legacy speeds** and the established ecosystem

- Opportunity to evaluate and interoperate with pre-release products
- Testing includes system to system and system to module interop testing, evaluating link establishment and link health reliability
- Establish BKMs for test and measurement methodologies
- Allow time for debug and issue resolution

Receive vendor specific and anonymized reports

Next event coming soon, May 1st 2023!



Thank you for watching!

If you have any questions or comments, please email admin@ethernetalliance.org

 **@ethernetalliance**

 **@EthernetAllianc**

 **Ethernet Alliance**