THE ETHERNET PORTFOLIO FOR HPC

John D'Ambrosia, Huawei
Nathan Tracy, TE Connectivity
Ran Almog, Mellanox
David Rodgers, Teledyne LeCroy

November 16, 2017

SC17
Denver, CO | hpc connects.

NEXT ETHERNET ERA
ethernet alliance

Regarding the Views Expressed

The views expressed on IEEE standards and related products should NOT be considered the position, explanation, or interpretation of the Ethernet Alliance.

Per IEEE-SA Standards Board Bylaws, Dec 2016

"At lectures, symposia, seminars, or educational courses, an individual presenting information on IEEE standards shall make it clear that his or her views should be considered the personal views of that individual rather than the formal position of IEEE. "

Our Mission and Priorities

We are a global community of end users, system vendors, component suppliers and academia

➢ ***Our Mission***

- Promote existing and emerging IEEE 802 Ethernet standards
- Accelerate industry adoption
- Demonstrate multi-vendor interoperability

➢ ***2017 Strategic Priorities***

- Support Existing Technology Deployment
- Support IEEE 802 Standards Development
- Marketing & Education

The Voice of Ethernet

Agenda

The State of Ethernet	John D'Ambrosia, Huawei
Enabling Future Ethernet Connectivity	Nathan Tracy, TE Connectivity
Maximizing Ethernet Performance for Most Demanding Workloads	Ran Almog, Mellanox
Test and Measurement Considerations for Ethernet Applications	David Rodgers, Teledyne LeCroy
Q & A	

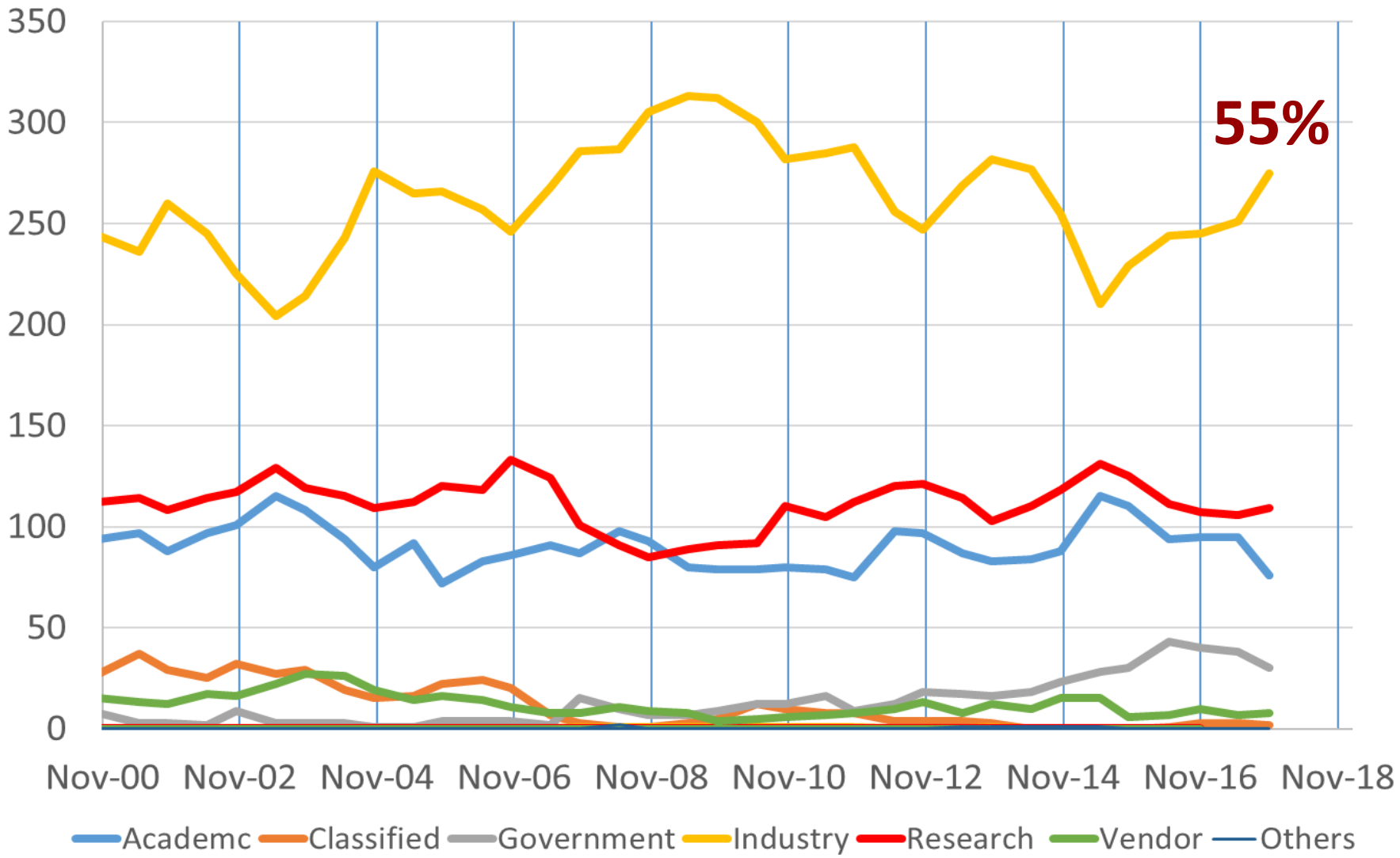
THE STATE OF ETHERNET

John D'Ambrosia

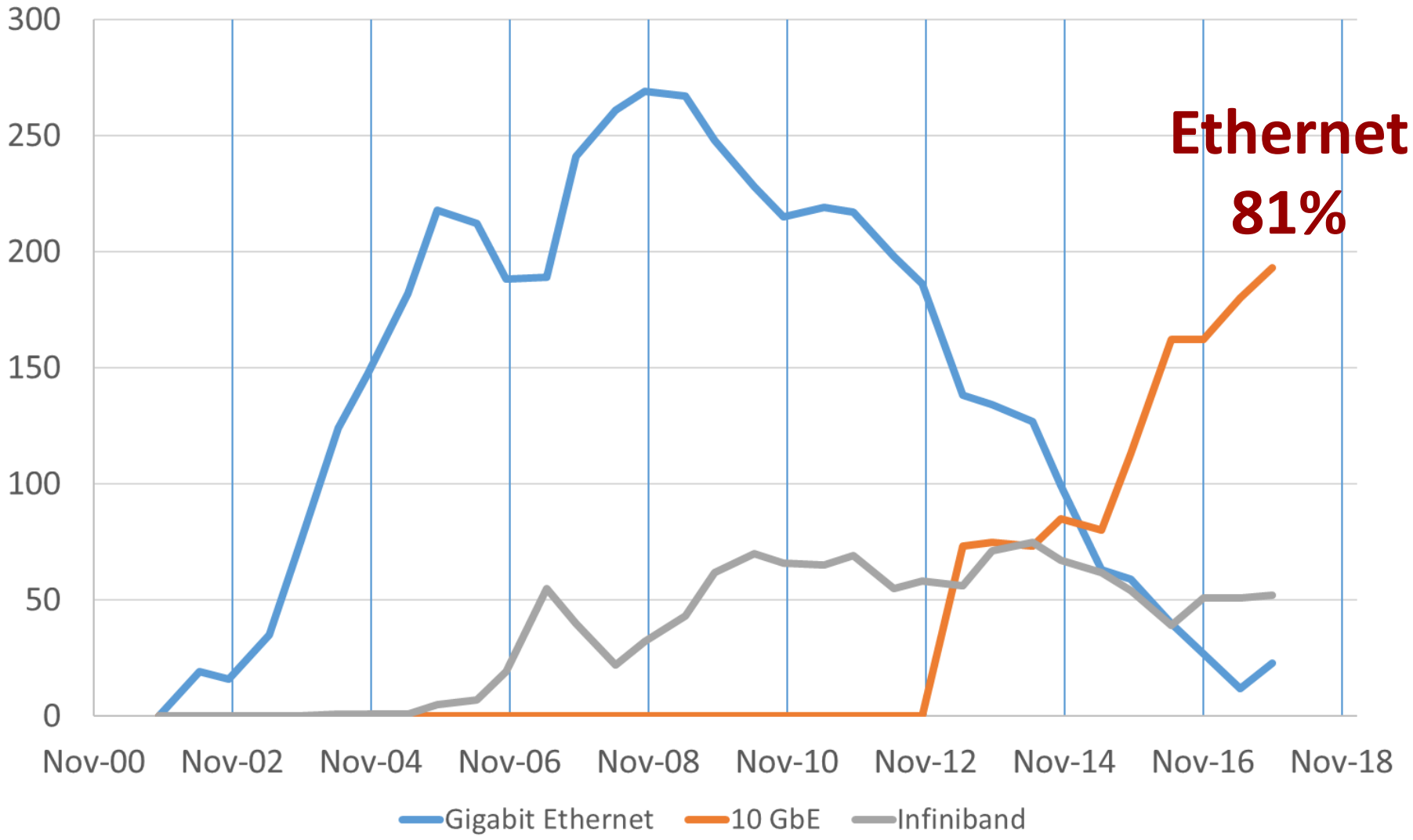
Huawei

November 16 ,2017

Top500 "Segments"



Top500 "Industry" Segment



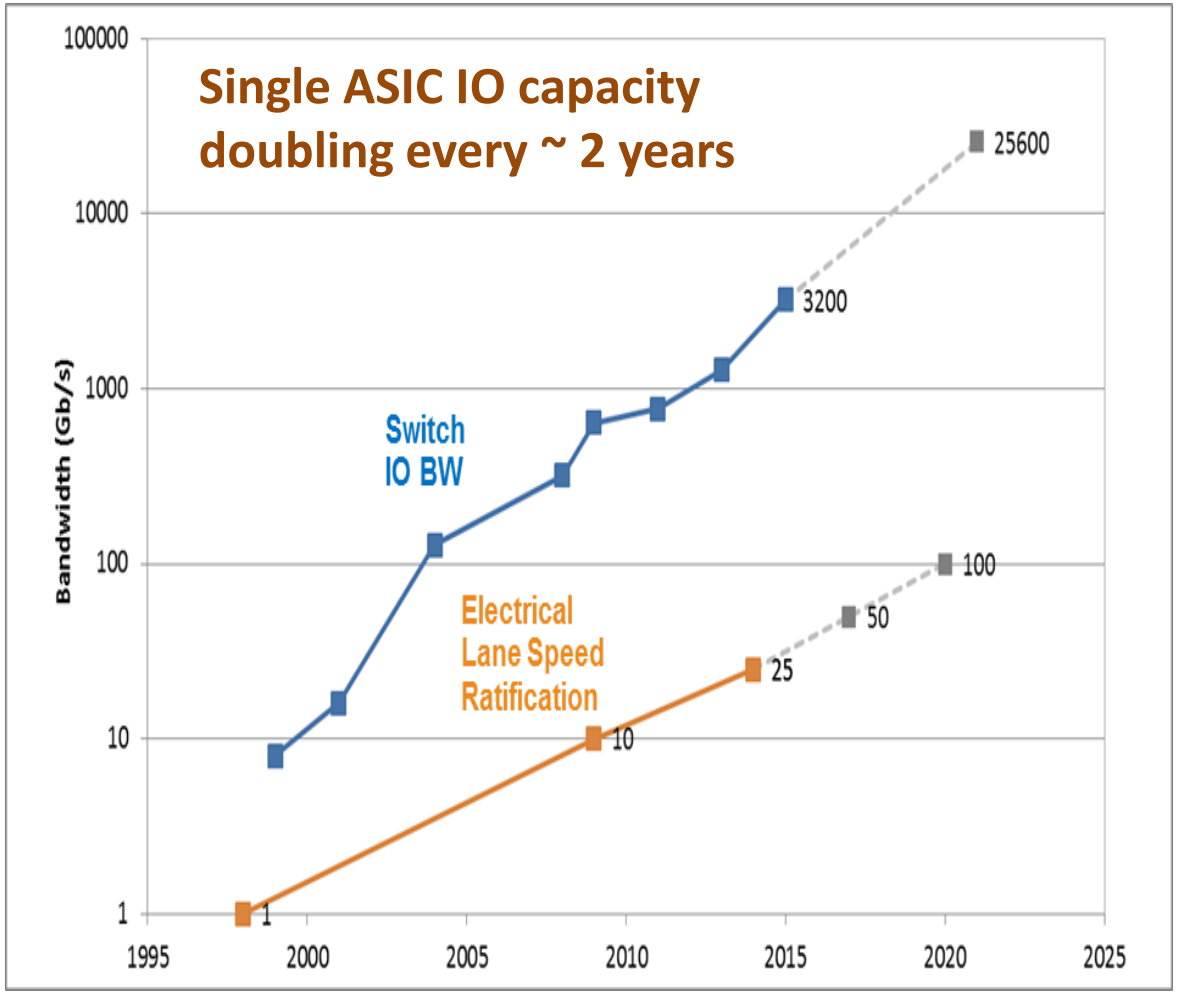
ETHERNET INTERFACES AND NOMENCLATURE

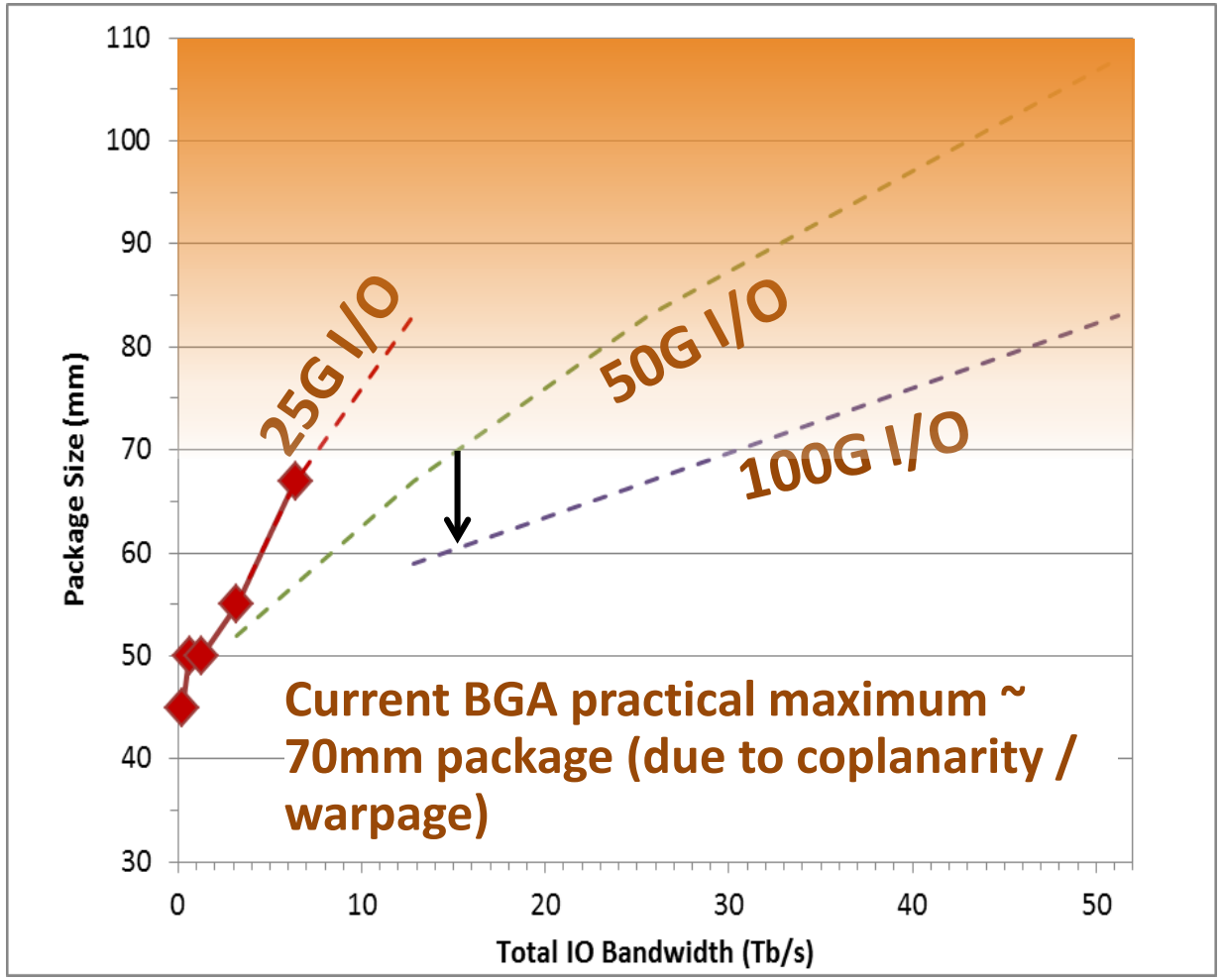
	Electrical Interface	Backplane	Twin-ax	BASE-T (4 Pair)	MMF	500m SMF	2km SMF	10km SMF	40km SMF
10GBASE-	XSBI, XAUI, XFI, SFI	KX4 KR	CX4	T	SR			LR	ER
25GBASE-	25GAUI	KR	CR/CR-S	T	SR			LR	ER
40GBASE-	XLAUI	KR4	CR4	T	SR4/eSR4	PSM4	FR	LR4	ER4
50GBASE-	LAUI-2/50GAUI-2 50GAUI-1	KR	CR		SR		FR	LR	ER?
100GBASE-	CAUI-10 CAUI-4/100GAUI-4 100GAUI-2 CAUI-1?	KR4 KR2 KR?	CR10, CR4, CR2 CR?		SR10 SR4 SR2 SR?	PSM4/DR4 DR	10X10 CWDM4/CLR4	LR4	ER4
200GBASE-	200GAUI-8 200GAUI-4 200GAUI-2?	KR4 KR2?	CR4 CR2?		SR4 SR2?	DR4	FR4	LR4	ER4?
400GBASE-	400GAUI-16 400GAUI-8 400GAUI-4?	KR4?	CR4?		SR16 SR8? SR4?	DR4	FR8 FR4?	LR8	ER8?

Gray Text = IEEE Standard Red Text = In Standardization Green Text = Future Possible Standard

Blue Text = Non-IEEE standard but complies to IEEE electrical interfaces

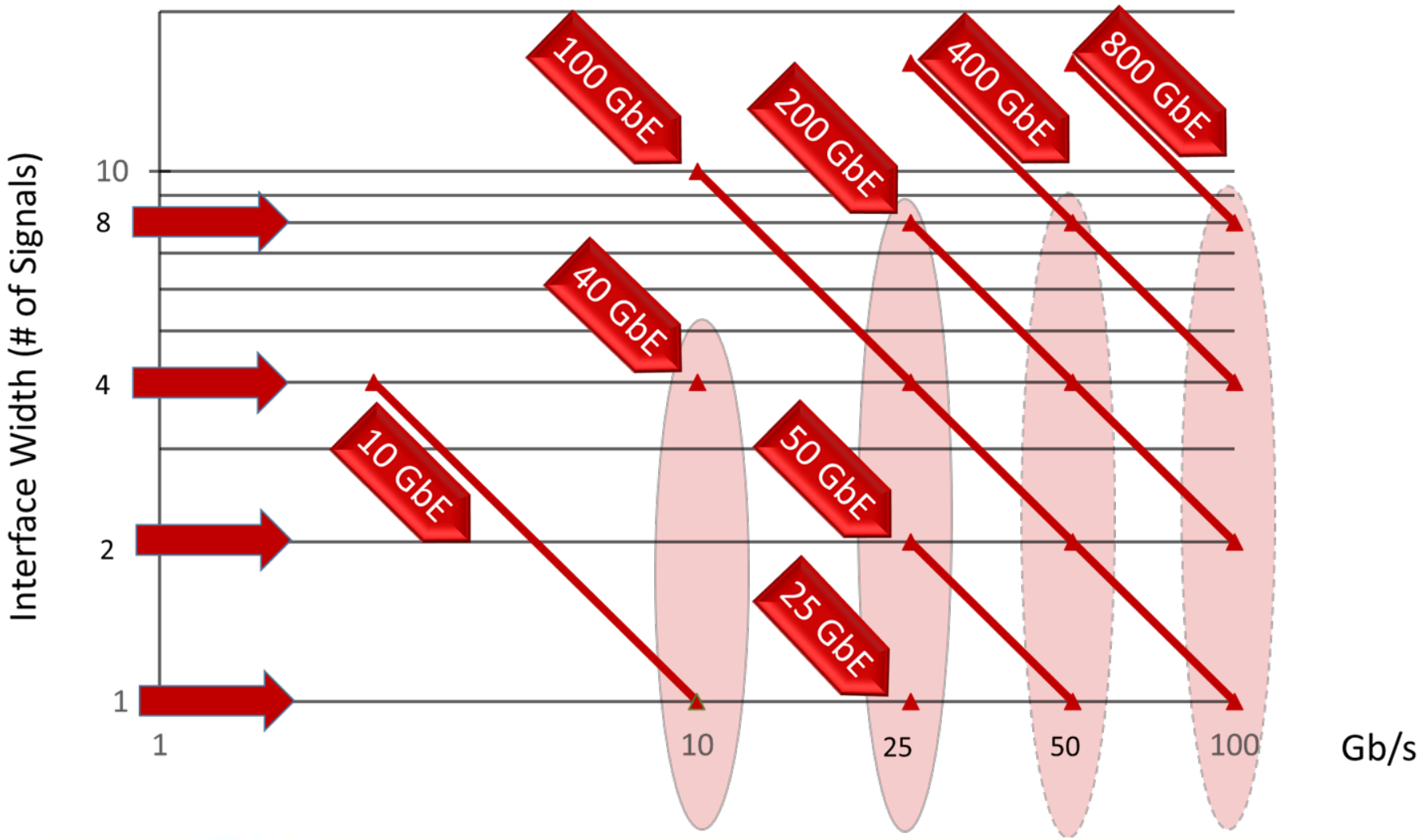
I/O Escape Forcing Transition to Higher Lane Speeds





Source: http://www.ieee802.org/3/ad_hoc/ngrates/public/17_03/goergen_nea_01a_0317.pdf

The New Ethernet Paradigm: Follow the SerDes



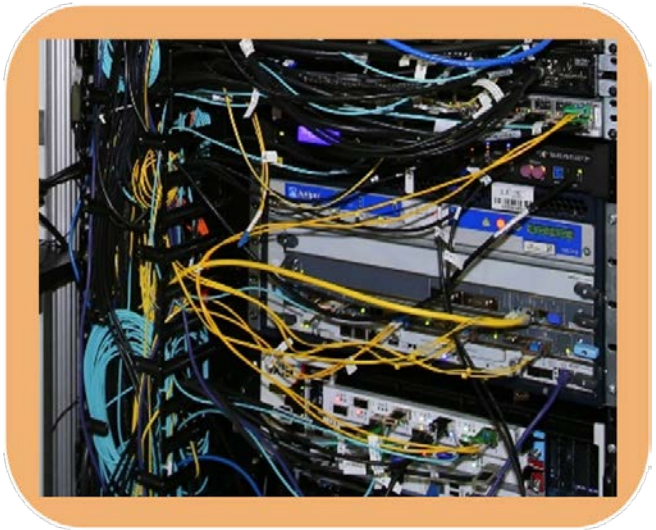
IEEE 802.3 Standards Activities

Project	Description	Schedule
IEEE p802.3bs	200 Gb/s and 400Gb/s Ethernet (electrical interfaces / optical PHYs)	Dec 2017
IEEE p802.3bt	4 Pair Power-Over-Ethernet	Sept 2018
IEEE p802.3ca	25Gb/s, 50 Gb/s, 1000Gb/s EPON	Apr 2020
IEEE p802.3cb	2.5Gb/s and 5Gb/s Backplane	Jun 2018
IEEE p802.3cc	25 Gb/s Ethernet over SMF (10 / 40 km)	Dec 2017
IEEE p802.3cd	50Gb/s, 100 Gb/s ,200Gb/s Ethernet (electrical interfaces, Copper PHYs, Optical PHYs)	Sept 2018
IEEE p802.3.2	YANG Data Models	June 2018
IEEE p802.3cg	10 Mb/s Single Twisted Pair	June 2019
IEEE p802.3ch	Multi-gig Automotive Ethernet	Undefined
Study Group	Beyond 10km Optical PHYs (50Gb/s, 100Gb/s, 200Gb/s, and 400Gb/s Ethernet)	
	100 Gb/s Electrical Interfaces and Electrical PHYs	
	10 Mb/s Backplane Ethernet	
	Next-gen 200G & 400G PHYs for MMF	

The Importance of Multi-vendor Interoperability

Industry investment

- On-going Work
  - PoE (802.3af / 802.3at)
  - 2.5G / 5G / 10G BASE-T
  - 25GbE
  - 100GbE
- Future
  - 4 Pair PoE
  - 25 GbE (10 km / 40 km)
  - 50 GbE
  - 200 GbE / 400 GbE
  - New Signaling
  - New Optical Form Factors



Ethernet Alliance PoE Certification Program

- According to Dell'Oro
  - 750M PoE Enabled Switch Ports over Next 5 Years
  - Hundreds of Millions of PoE Devices over Next 5 Years
- Distinguishes products based on IEEE 802.3 standards in the market
- Open to general industry
- https://ethernetalliance.org/poecert/

PSE

Example – Class 3 PSE

PD

Example – Class 1 PD

ENABLING FUTURE ETHERNET CONNECTIVITY

Nathan Tracy
TE Connectivity

November 16 ,2017

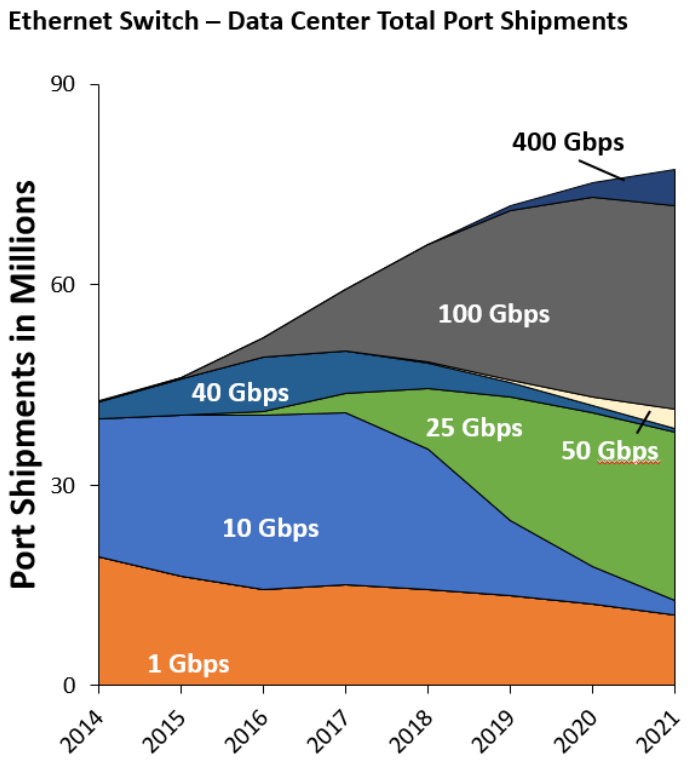
Connectivity Challenges For The Next Generation

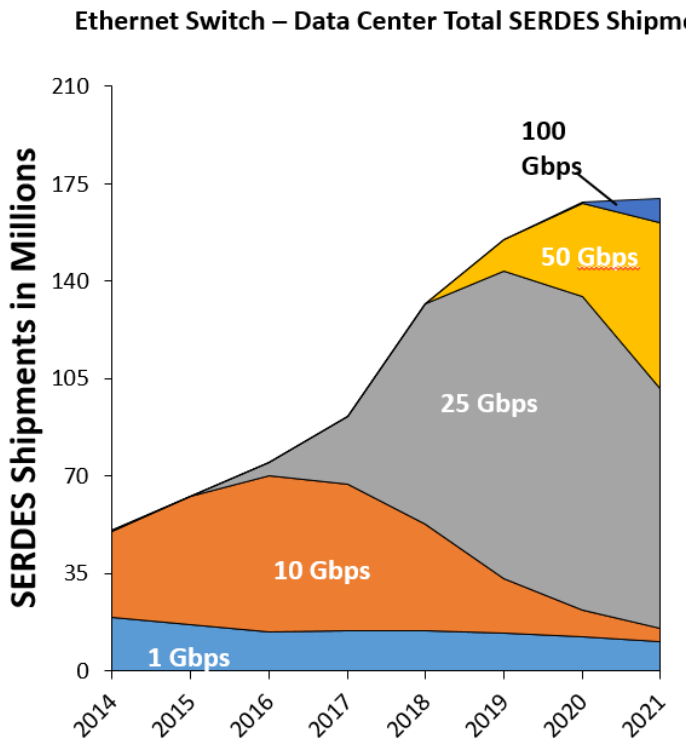
- With 100G Ethernet, we started with 10 x 10Gbps electrical interfaces, then transitioned to 4x25Gbps
  - Enabled a narrower interface (more density)
  - Enabled lower power
- What comes next?
- 50Gbps signaling further reduces the interface width
- But the per port rate needs to get to 400Gbps

Where Are The Data Rates Headed?

Ethernet Switch – Data Center: Total Shipments

650 GROUP
MARKET INTELLIGENCE RESEARCH

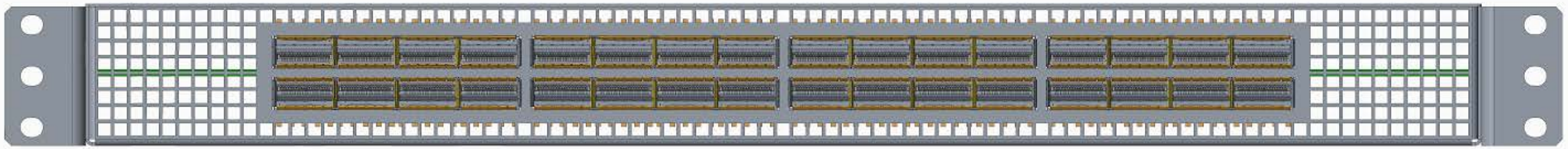




Hyperscale performance is transitioning to 50Gbps

Data provided by 650 GROUP. Further distribution is prohibited

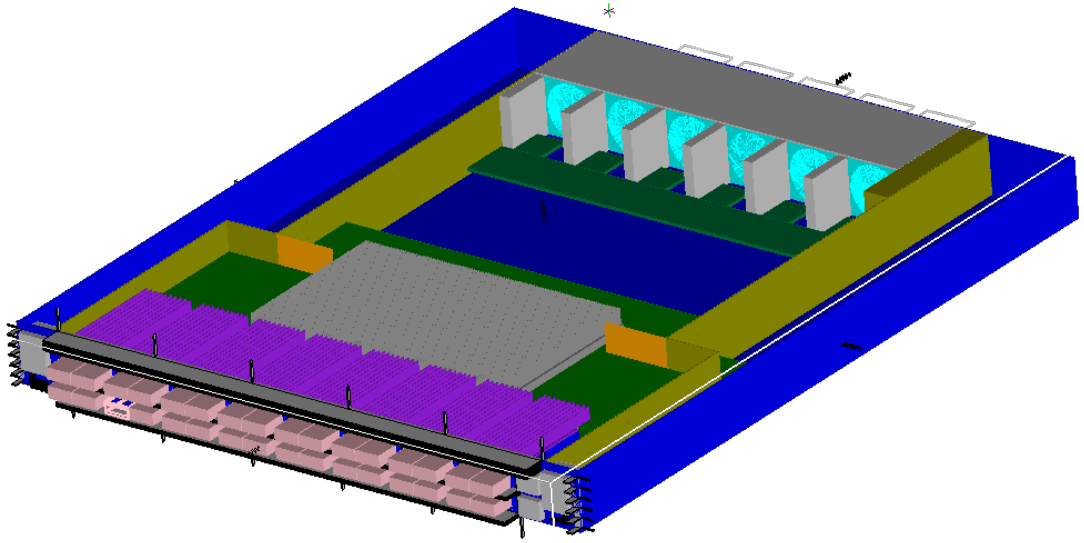
Today → Tomorrow



- Current QSFP Input Output (I/O) port allows 32 to 36 ports per 1RU enclosure
- Each port has 4 electrical channels
- At 25Gbps, enables 3.6Tbps via 100Gbps ports
- At 50Gbps, enables 7.2Tbps via 200Gbps ports
- But we need 400Gbps ports……..

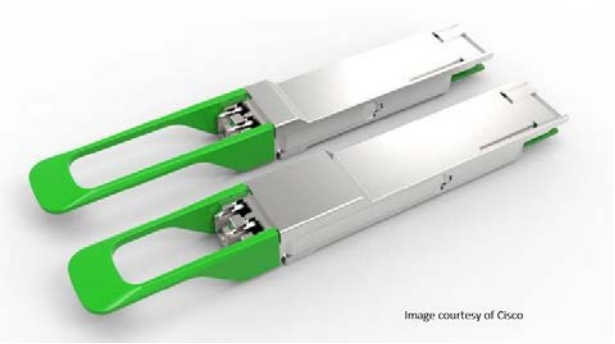
Challenges to 400Gbps Connectivity

- 400Gbps requires 50Gbps signaling and 8 electrical channels
- With a doubling of electrical channels from 4 to 8, how do we fit at least 32 ports in a 1RU enclosure to yield 12Tbps?
- Need a new module/connector form factor solution
  - More lanes
  - Improved electrical performance
  - Improved thermal performance
- The Ethernet Community is up to the task



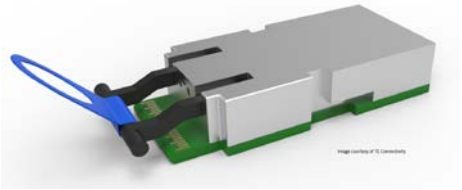
Three New Form Factors Being Developed

- QSFP-DD
  - Goes to 8 channels, maintains ability to accept legacy QSFP ports (backwards capability)



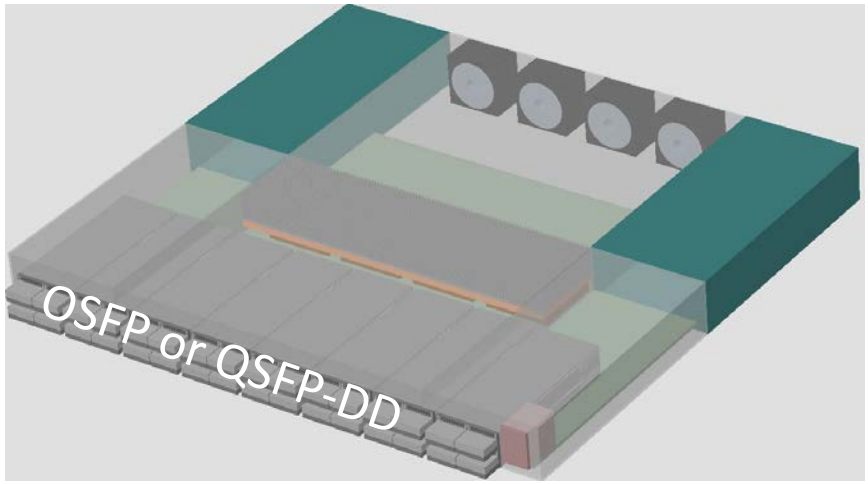
- OSFP
  - New form factor with 8 channels. Includes integrated heat sink for improved thermal performance, uses an adapter for backwards performance

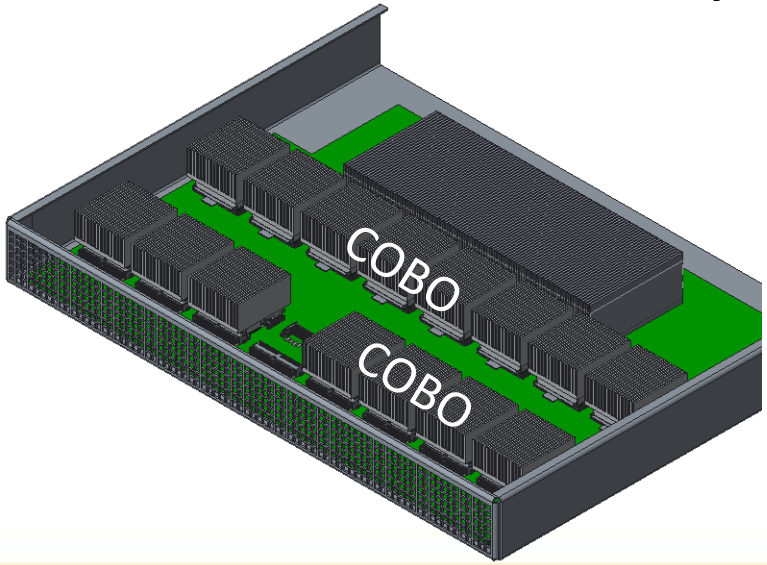
- COBO
  - Defines embedded optics modules. Fits 32 400G modules on a 1RU linecard, allowing more airflow and improved thermal performance



Pluggable or Embedded?

- Pluggable modules are the traditional approach but put all the thermal dissipation of the modules at the face plate and block some airflow
- Embedded optics spread the thermal dissipation, allow larger heat sink and create more area for airflow at the faceplate

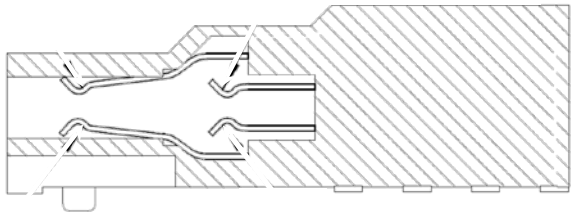




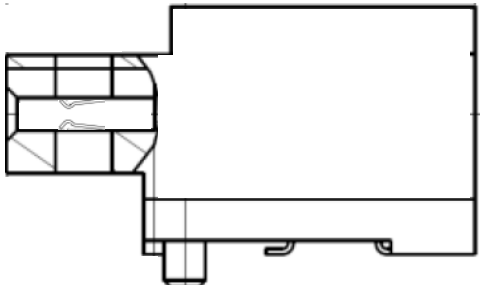
Critical Requirements of a Connector

- Signal Integrity

4 row style connector



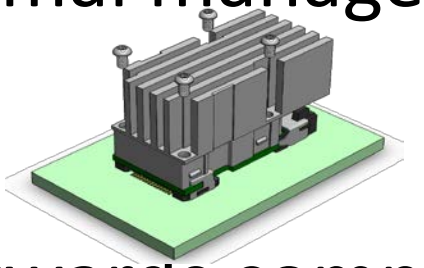
2 row style connector



- 0.6mm contact pitch for COBO and OSFP, 2 row
- 0.8mm contact pitch for QSFP-DD, 4 row
- More channel margin with the simpler 2 row solution
- Copper cable reach is 0.5m longer for OSFP

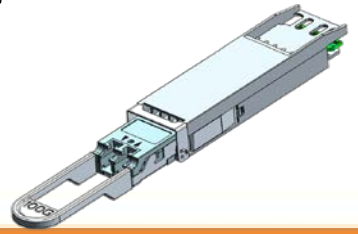
All are designed to the same IEEE electrical requirements (industry consensus)

- Thermal management (sliding, integrated, attached)



- COBO has more room for heat sink, 15+ W
- QSFP-DD sliding heat sink, 12W?
- OSFP integrated heat sink, 15W

- Backwards compatible



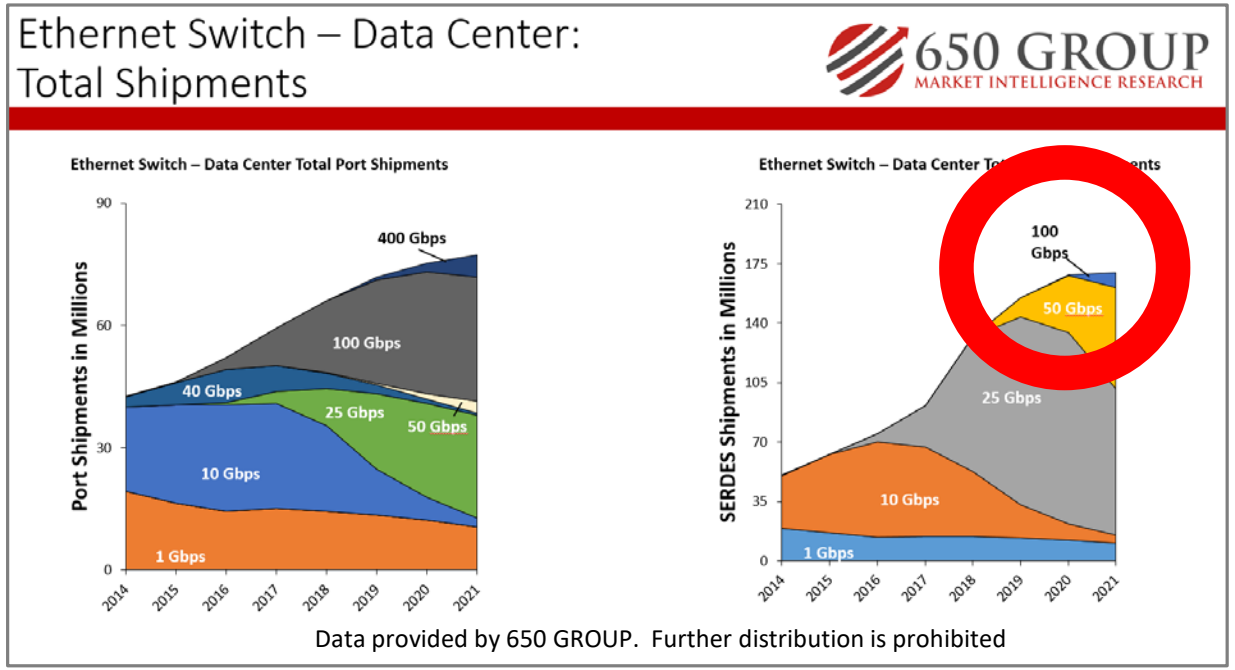
- COBO is disruptive, no backward capability
- QSFP-DD does accept QSFP, i.e. 40G, 100G
- OSFP with adapter accepts QSFP

400G Summary

- There will be choices for 400Gbps implementation
- This discussion has focused on IO, but backplane solutions are available in the market as well
- Important decisions must be made in the connector selection regarding future rates, design margin, thermal performance and cable reach
- Ethernet's silicon, optic, connector and cable suppliers keep on innovating …….

What's Next in Connectivity?

- Remember this slide?



- 100Gbps per differential pair developments have started!
- IEEE has a Study Group
- OIF has a project
- Component suppliers have development programs

Further innovations to address the challenges of signal integrity, reach, thermal, and density will continue!

MAXIMIZING ETHERNET PERFORMANCE FOR MOST DEMANDING WORKLOADS

Ran Almog
Mellanox Technologies

November 16 ,2017

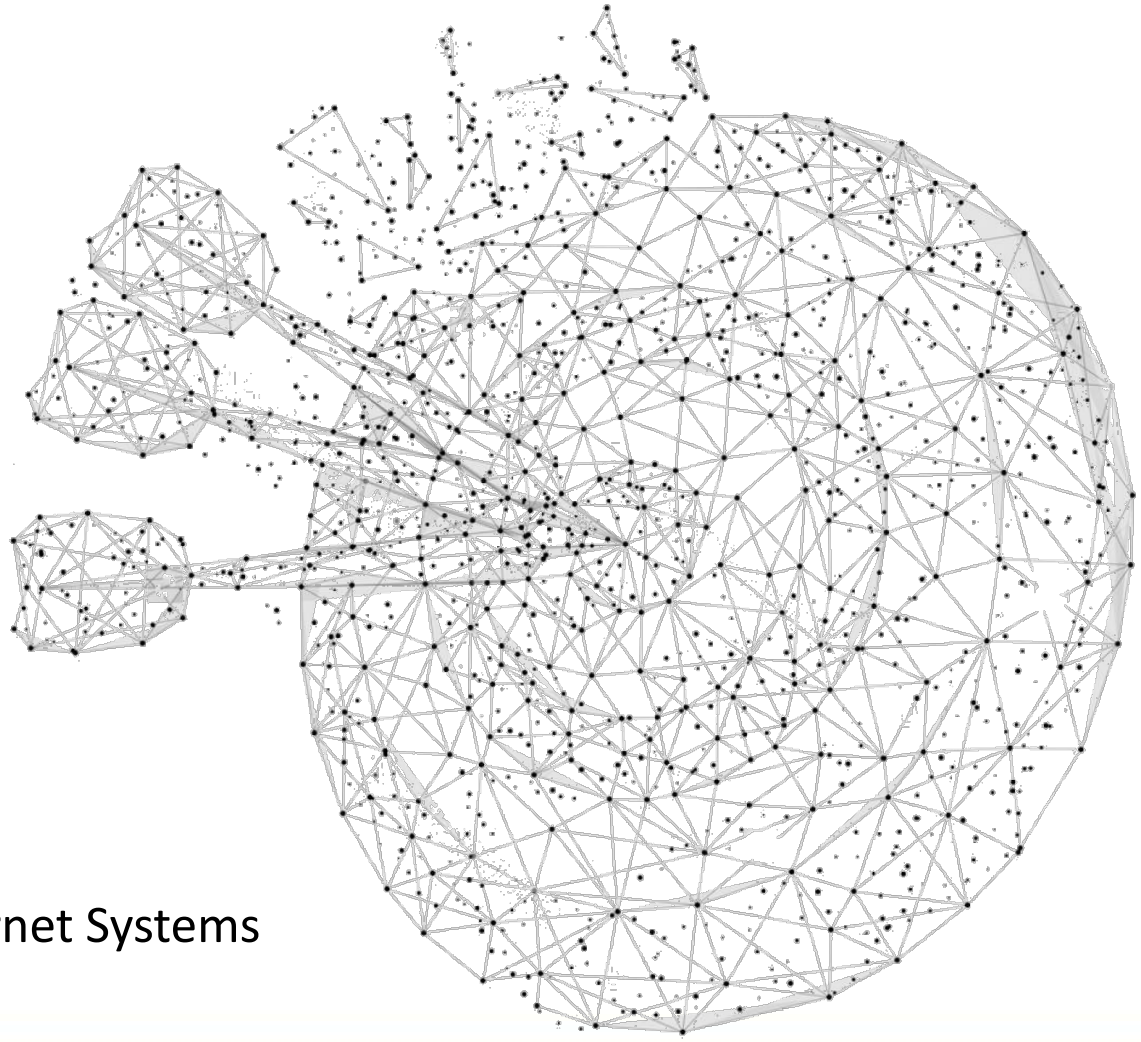
Delivering Highest DC Return on Investment

RoCE™

90% of the Top 10 **Oil and Gas Companies**

60% of Top 5 **Pharmaceutical Companies**

100% of Top 10 **Automotive Manufacturers**

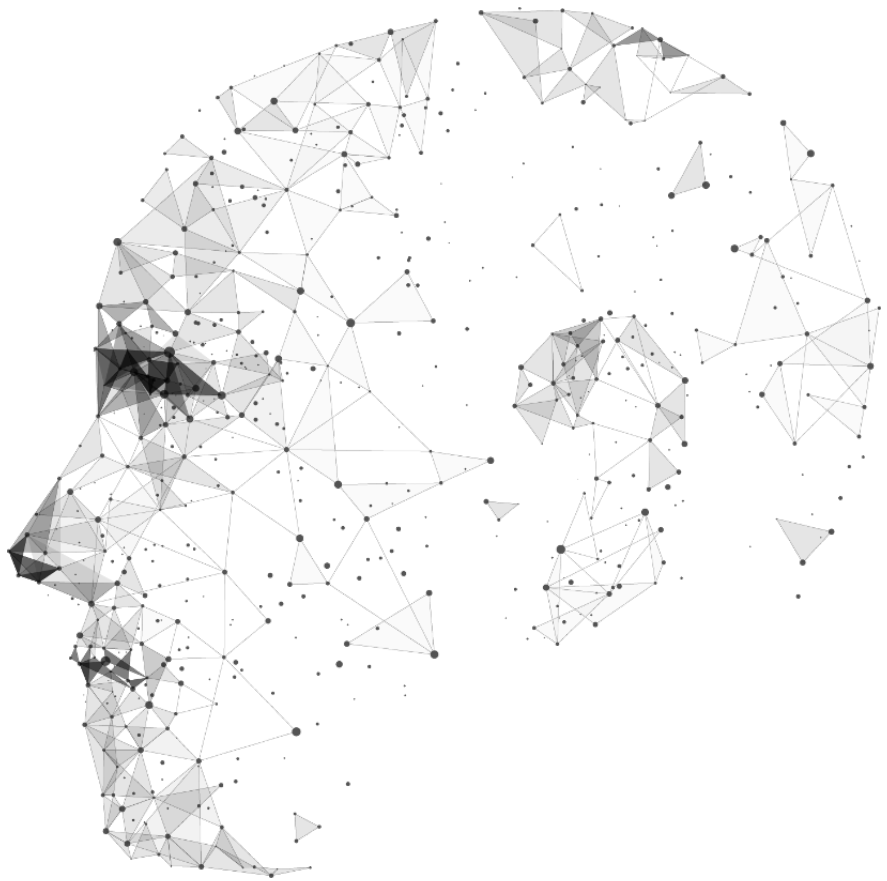


Ethernet Leadership

Connects All of 40G Ethernet Systems

Enabling Most Efficient AI Platforms

- OCP Big Sur Artificial Intelligence Platform
- Real Time Fraud Detection
- Machine Learning System with 400Gb/s
- 18X Speedup For Image Recognition
- 4X Speedup For Image Recognition
- Data Analytics Image Recognition
- World Record For Data Sort, 3X Faster



Ethernet Leadership

The First 100G Ethernet System on The TOP500 List

Maximizing Ethernet Performance with RoCE

High Performance Network

- Highest throughput
- Lowest latency

Advanced Congestion Control

- Early detection and prevention
- Reliable and predictable

Efficient Network Utilization

- Higher server productivity, cost and power savings
- Higher availability of CPU resources to the application

Extensive Visibility

- Optimize network behavior
- Detect, prevent and troubleshoot

RoCE™

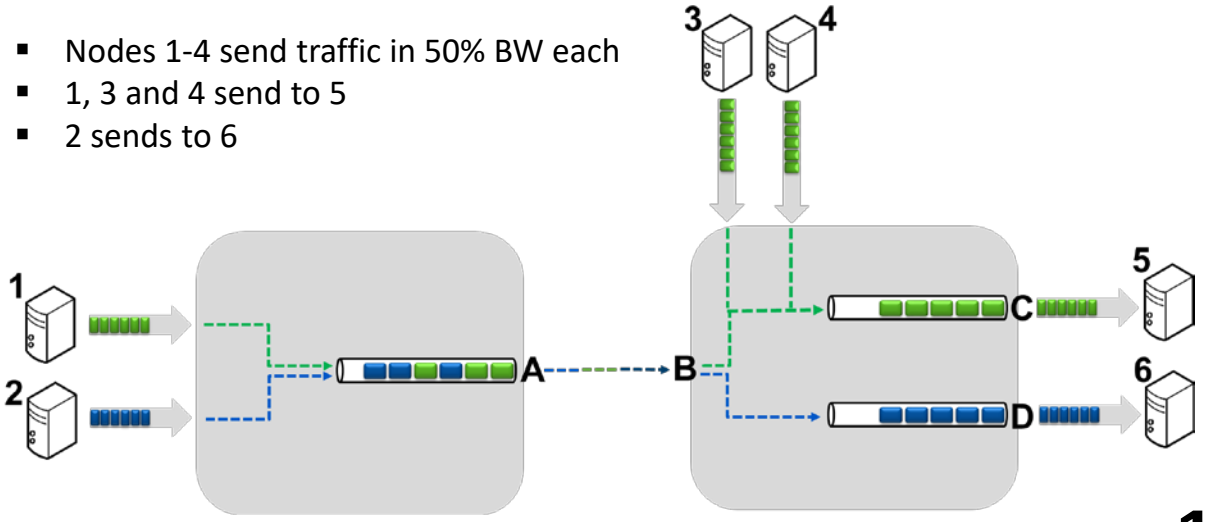
RDMA over Ethernet

RoCE Simplified

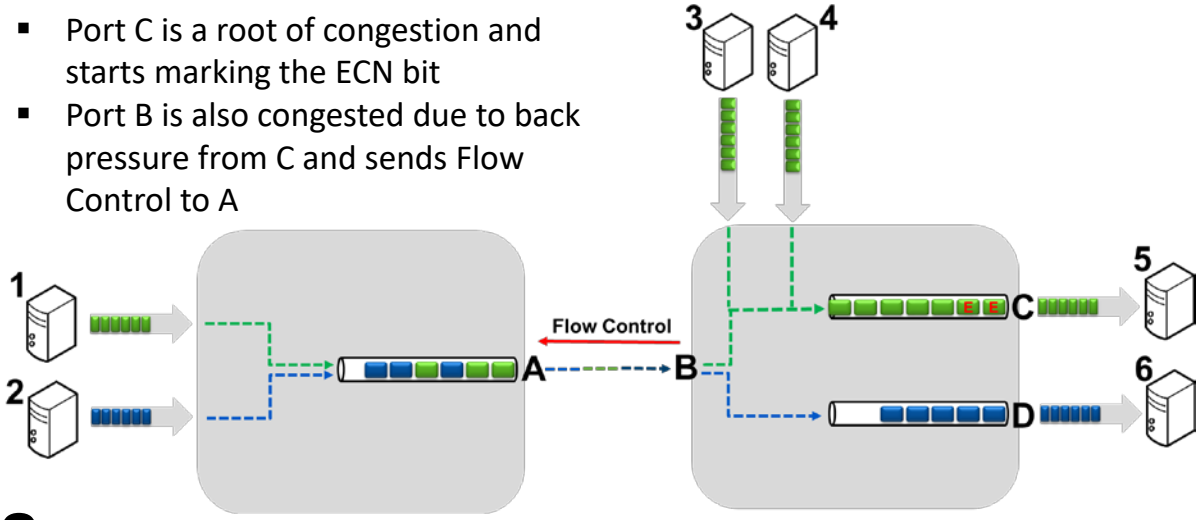
- Soft RoCE
- RoCE for Lossy networks
- Out of the box experience
- ECN capable network
- Automatic PFC configurations

Protect Against Congestion Victims

- Nodes 1-4 send traffic in 50% BW each
- 1, 3 and 4 send to 5
- 2 sends to 6



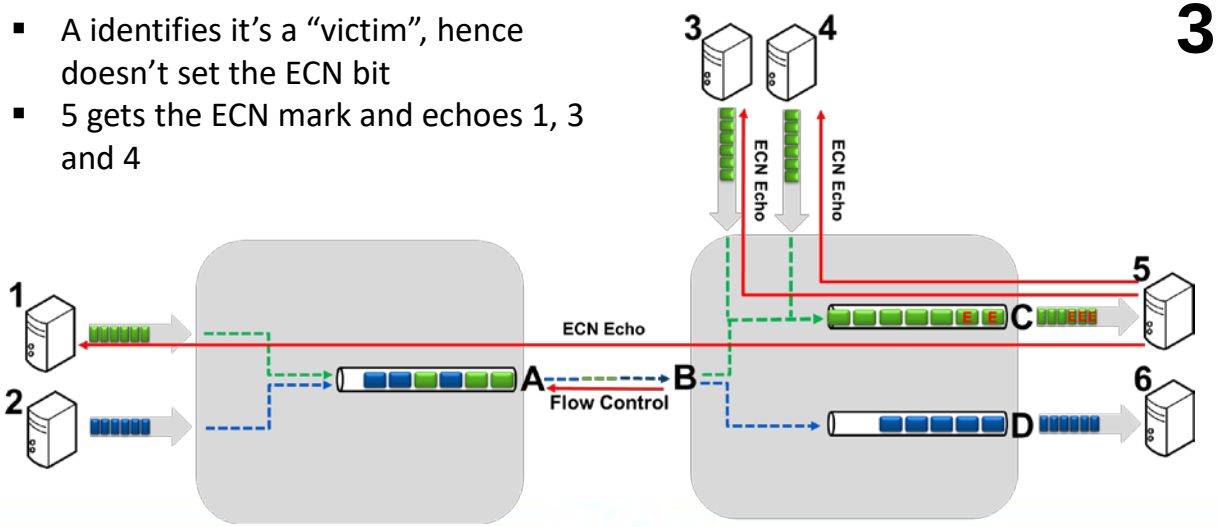
- Port C is a root of congestion and starts marking the ECN bit
- Port B is also congested due to back pressure from C and sends Flow Control to A



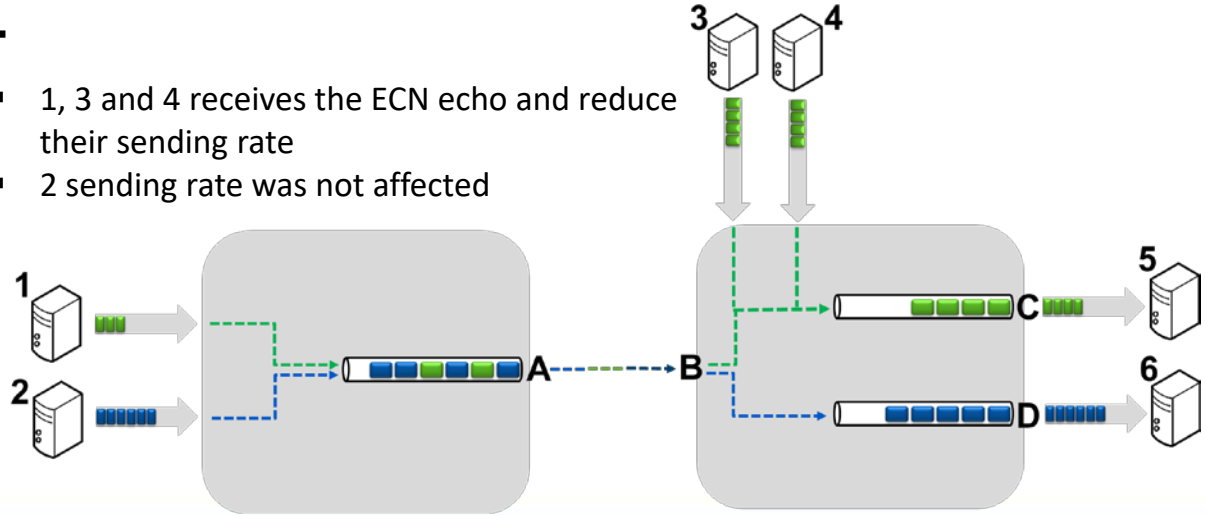
1 2

3 4

- A identifies it's a "victim", hence doesn't set the ECN bit
- 5 gets the ECN mark and echoes 1, 3 and 4



- 1, 3 and 4 receives the ECN echo and reduce their sending rate
- 2 sending rate was not affected



Open Composable Networks

Microsoft SONiC

MLNX-OS Mellanox OS

Customer-OS

Cumulus® Linux® Network OS

Metaswitch Networks

Network OS
Choice

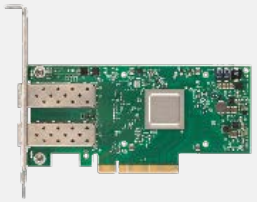
sai SDK switchdev

Open APIs

Spectrum™

NEO

Automation



End-to-End Interconnect

TEST AND MEASUREMENT CONSIDERATIONS FOR ETHERNET APPLICATIONS

David J. Rodgers

Teledyne LeCroy PSG

November 16 ,2017

ethernet alliance

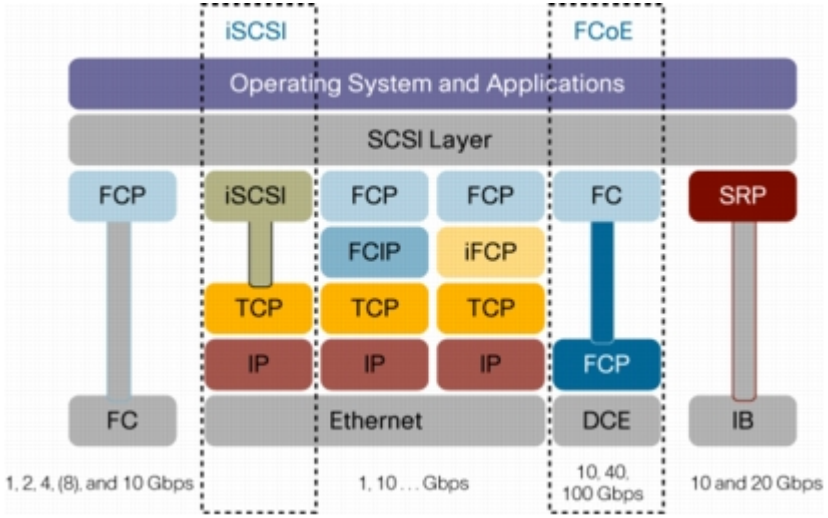
Basic Considerations!

- **Testing and Validation Needs to Keep Up**
- **Integrating Higher Speeds in the SAN**
  - 10/40GbE, now 25/100GbE and 50/200GbE right around the corner
  - Closing in on the “Holy Grail” of 100GbE
- **Ethernet Fabrics Fueling Storage Explosion**
  - Speed and Optimization meeting QOS Expectations
    - iSCSI, FCoE, NVMf, NFS, IBXoE, FCIP, iSER, iWARP, RoCE, Routable RoCE (v2)
- **Conforming to Standards**
- **Keep on Budget and Keep Users Happy!**

TELEDYNE LECROY
Everywhereyoulook™

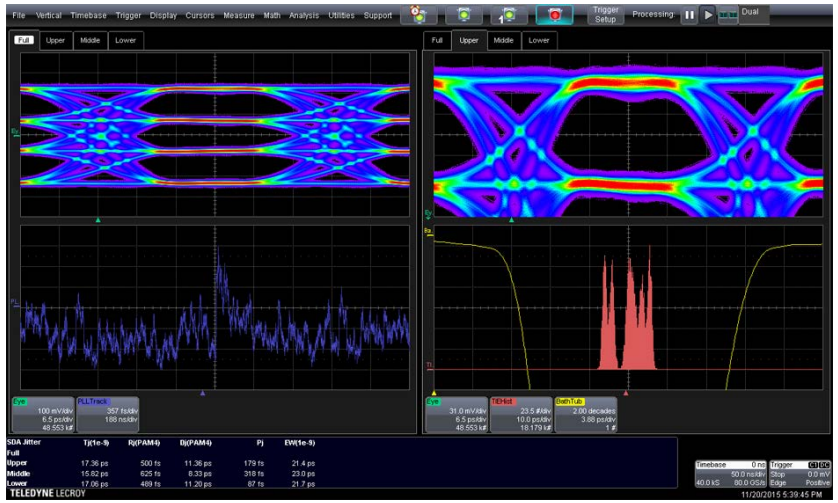
Conformance to Standards

- **Ethernet Standards Evolving at Breakneck Pace**
  - Automotive
  - 25GbE to 100GbE, now 50GbE to 200GbE
  - Soon, 100GbE to 400GbE
- **Storage Solutoins Leveraging Speed**
  - FCoE, iSCSI
  - NVMf
- **Standards beget Interoperability?**
  - Interpretation and implementation differences abound
    - Increase in speed has added complexities



Interoperability in the Real World

- **No two vendors implementations are identical**
- **There is a "protocol" to the phy**
  - Auto-negotiation
  - Link Training
  - FEC
- **New Speeds adding new complexities**
  - NRZ vs PAM4 signaling
- **Testing needs to be "standardized" and repeatable**
  - Interop PlugFests, 3rd party testing services



TELEDYNE LECROY
Everywhereyoulook™

Key Interoperability Challenges

- **Identifying Participants**
  - Characterizing Functionality of All Ecosystem Players
- **Determining Root Cause**
- **Crafting the Solution**
- **Remediation Validation**
  - Test the fix
- ***Timely Resolution!***

TELEDYNE LECROY
Everywhereyoulook™

NEXT ETHERNET ERA
ethernet alliance

Effective Observation

Fabric Management
Utility/Hypervisor

Traffic Tap and
DPI (Wire Shark)

Line Rate
Analysis

Adding Line Rate Analysis

- ***Purpose Built Protocol Tools!***
  - Compliment to, not replacement for Traditional Tools
- **Invisible to the Fabric Under Test**
  - Unbiased traffic capture of all layers
- **Agnostic to Participants, Traffic Type, and Transport Media**
  - Real Time Triggering, Post Capture Data Analysis
- **Traffic Modification/Error Injection Functions**
  - *Determine Root Cause!*
  - Proof Remediation before deployment

TELEDYNE LECROY
Everywhereyoulook™

Testing the Fix

- **Once remediation is applied, does it work?**
- **Lab Recreation of the offending condition(s)**
  - Proof of concept on the bench – Analyzer/Jammer
  - Observation on the link - Analyzer
- **Test out additional corner case scenarios**
  - Prescreen pending releases
  - Reuse profiles/test cases

TELEDYNE LECROY
Everywhereyoulook™

NEXT ETHERNET ERA
ethernet alliance

Investigative Challenge

"Often times (the problem)requires the *recreation of a given fault in a lab environment*, which is problematic without an appropriate toolset. For this specific purpose an in-band protocol analyzer and *error injection utility* is now an integral part of my troubleshooting arsenal."

"The error injection capability of these tools is of even higher importance to me however.

Reference excerpted from Teledyne LeCroy user case study.

TELEDYNE LECROY
Everywhereyoulook™

Error Injectors

- Provide effective, programmatic recreation of faulty conditions/events
- Drop commands, responses
- Insert Errors at all levels
- Counters and Timers for true-to-traffic conditions
- Best when tightly integrated to the analysis tools

TELEDYNE LECROY
Everywhereyoulook™

NEXT ETHERNET ERA
ethernet alliance

ReCap

- Ethernet is a Juggernaut
- Content delivery and Storage Demands are High
- Consistent and predictable interoperation is mandatory
- Speed adds exponential Influences on the EcoSystem
- Testing, Testing, Testing

<u>Tool Sets and Methodologies Must Evolve</u>

TELEDYNE LECROY
Everywhereyoulook™

NEXT ETHERNET ERA
ethernet alliance

QUESTIONS?

ethernet alliance