SNIA NSF | NETWORKING STORAGE

# New Landscape of Network Speeds

**Live Webcast**
**May 21, 2019**
**10:00 am PT**

# Today's Presenters

**David Chalupsky**
**Intel**

**Craig Carlson**
**Marvell**

**Peter Onufryck**
**Microchip**

**John Kim**
**SNIA NSF Chair**
**Mellanox**

# SNIA-At-A-Glance

SNIA NSF | NETWORKING STORAGE

## SNIA-at-a-Glance

**185**
industry leading organizations

**2,000**
active contributing members

**50,000**
IT end users & storage pros worldwide

Learn more: **snia.org/technical** **@SNIA**

# SNIA Legal Notice

- The material contained in this presentation is copyrighted by the SNIA unless otherwise noted.
- Member companies and individual members may use this material in presentations and literature under the following conditions:
  - Any slide or slides used must be reproduced in their entirety without modification
  - The SNIA must be acknowledged as the source of any material used in the body of any document containing material from these presentations.
- This presentation is a project of the SNIA.
- Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.
- The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.

NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.

# Agenda

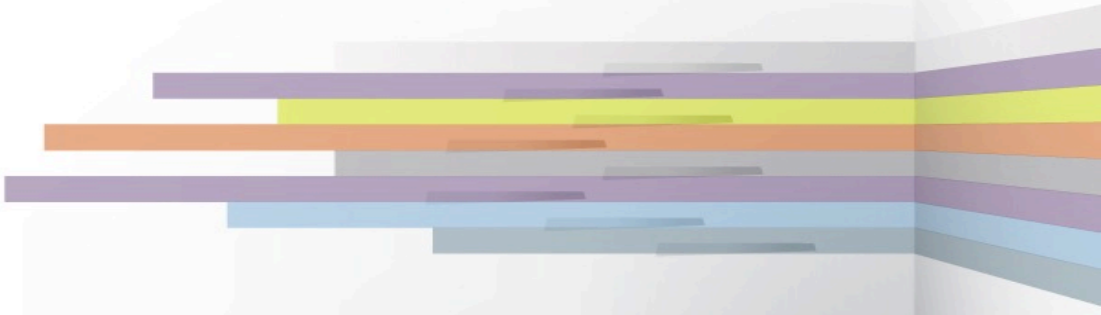
SNIA NSF | NETWORKING STORAGE

- Trends in networking speeds
  - Ethernet
  - Fibre Channel
  - PCI Express
  - InfiniBand
  - What it all means

# The Need for Speed

- Why do customers want faster networking?
  - Faster servers with more CPUs/GPUs/FPGAs
  - Faster storage devices and protocols, like NVMe-oF
  - 4K/8K video
- Big datacenters want wider pipes
  - Spine/superspine links across rows, buildings
  - Fatter uplinks ➔ more efficient fabrics, fewer switches

SNIA NSF | NETWORKING STORAGE



# Ethernet

# Ethernet Agenda

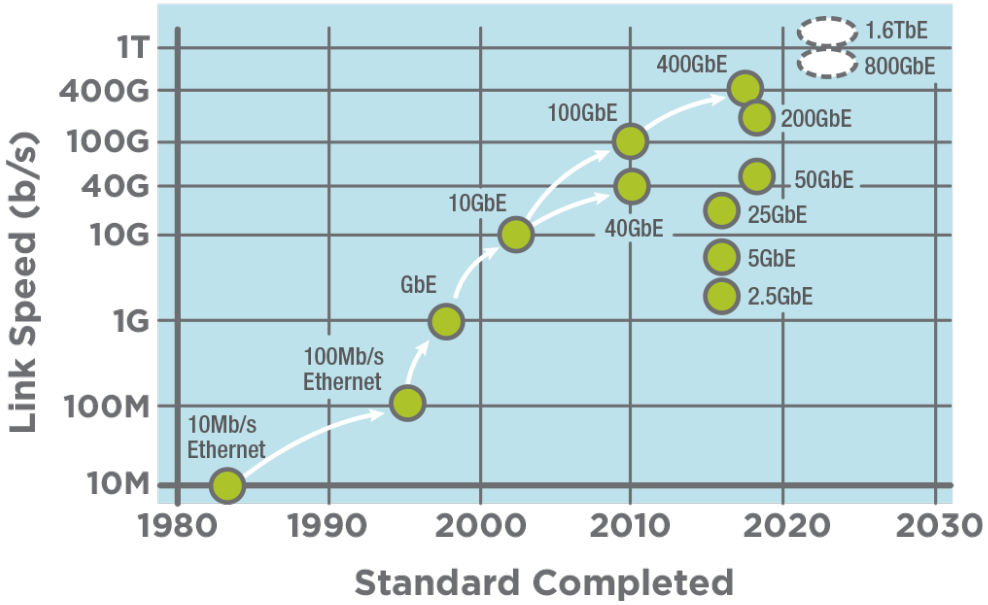
- The Ethernet Roadmap
  - Speeds
  - Applications

Graphics courtesy of the Ethernet Alliance

https://ethernetalliance.org/the-2019-ethernet-roadmap/

# The Ethernet Speed Roadmap

- After creating 6 speeds in 30 years, we[1] developed the next 6 speeds all at once
- Why?



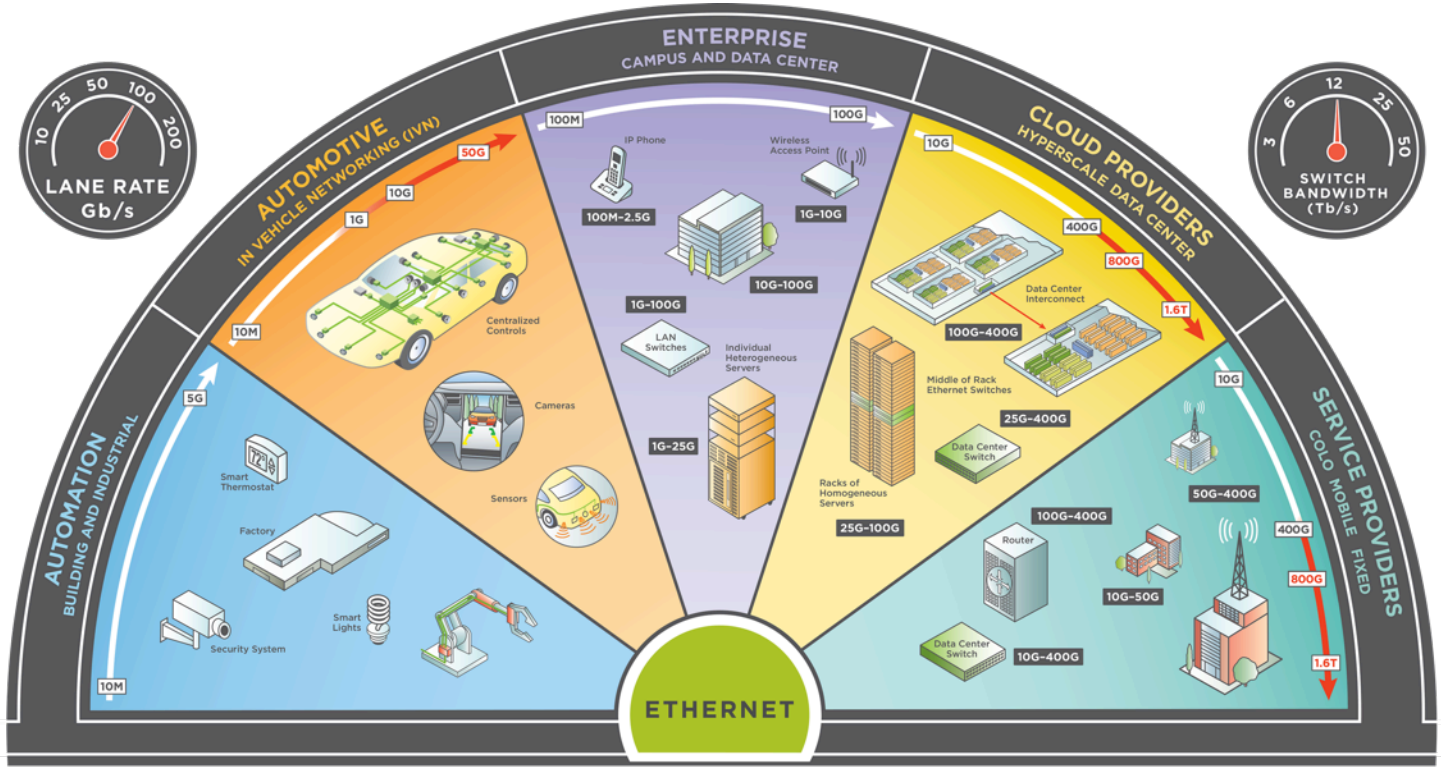

1 – "We" being the IEEE 802.3 working group, the source of Ethernet standards

# Application Diversity

SNIA NSF | NETWORKING STORAGE

- What started in the 1970's as a means to connect minicomputers to printers became so much more
  - The ubiquitous LAN technology
  - Wireless backhaul
  - Corporate backbone
  - Data center interconnect
  - Storage fabric
- Let's look at some application areas in depth

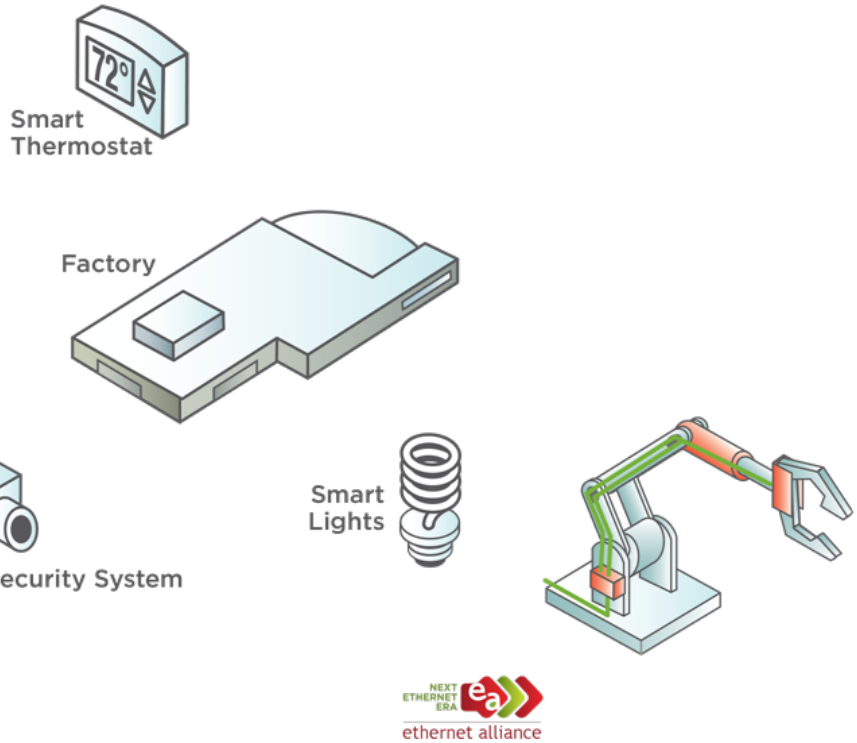
# Ethernet Application Areas



# Automation: Building and Industrial

SNIA NSF | NETWORKING STORAGE

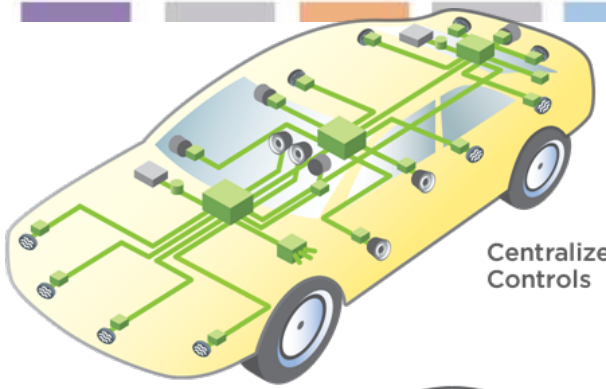
- Reaches from a few meters to 1km
- Speeds from 10Mb/s to 10Gb/s
- Ability to carry power over the data lines
- Rugged
- Low cost



# Automotive In-Vehicle Networking

SNIA NSF | NETWORKING STORAGE

- Reaches up to 15m
- Speeds from 10Mb/s to >10Gb/s
- Power over data lines
- Rugged, EMC resilient
- Light weight
  - Single pair copper
- Fast startup
- Low cost

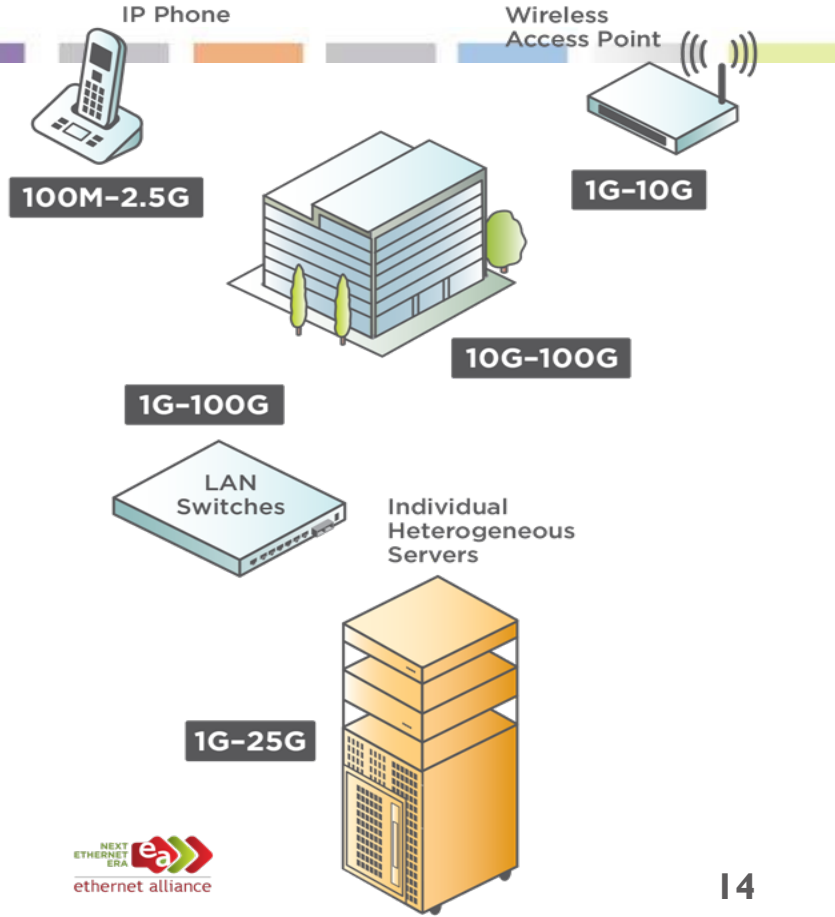





NEXT ETHERNET ERA ethernet alliance

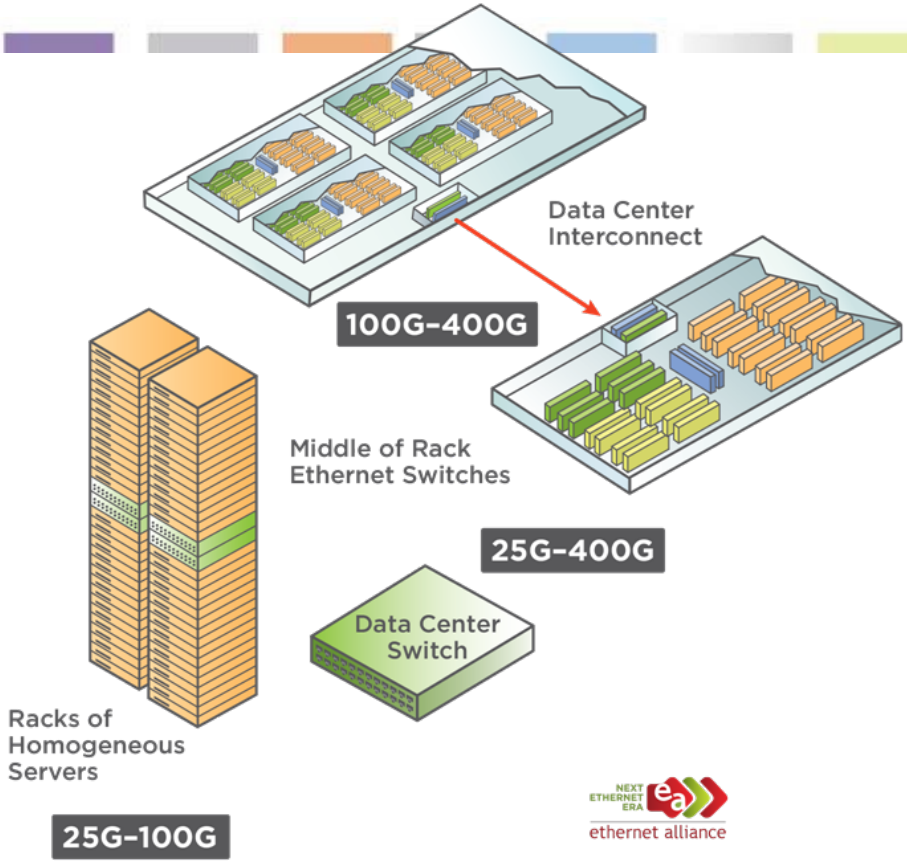
# Enterprise Campus and Data Center

- Reaches 10's to 100's of meters
- Speeds from 100Mb/s to 100Gb/s
- Power over Ethernet
- Ease of use
- Multi-generation infrastructure re-use



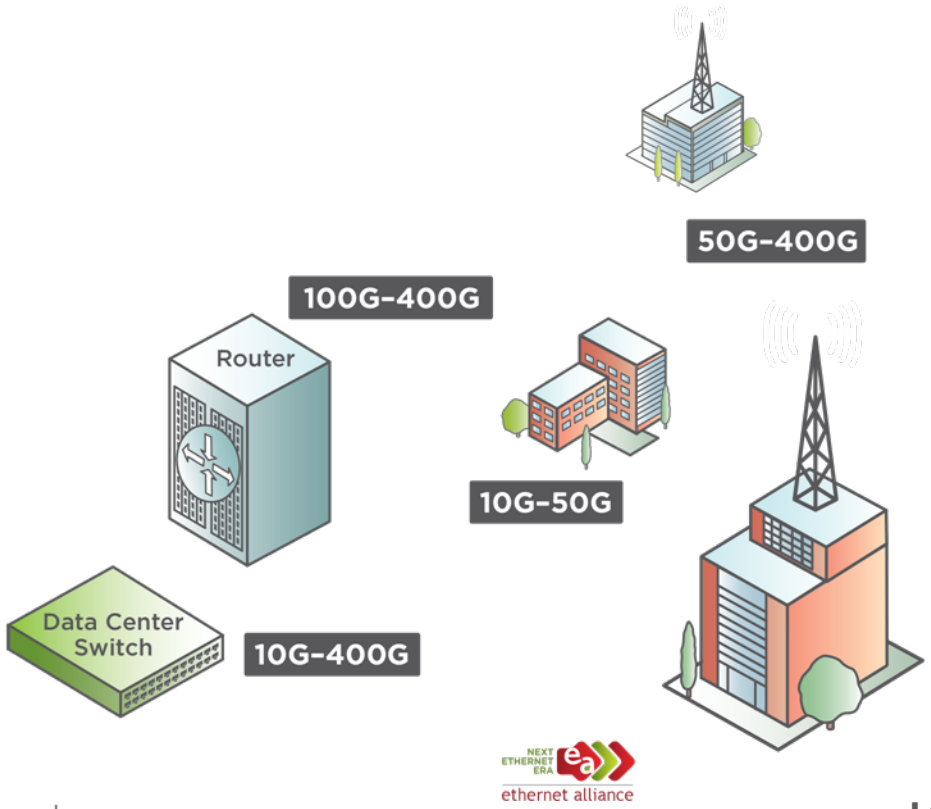
# Cloud Hyperscale Data Center

- Reaches 2m to 500m inside, 2km to 80km outside
- Speeds from 10Gb/s to 400Gb/s
- High volume, homogeneous deployments



# Service Providers

- Reaches 2m to 500m inside, 2km to 80km outside
- Speeds from 10Gb/s to 400Gb/s
- CoLo
- Wireless backhaul
- Telecom
- Internet Exchange

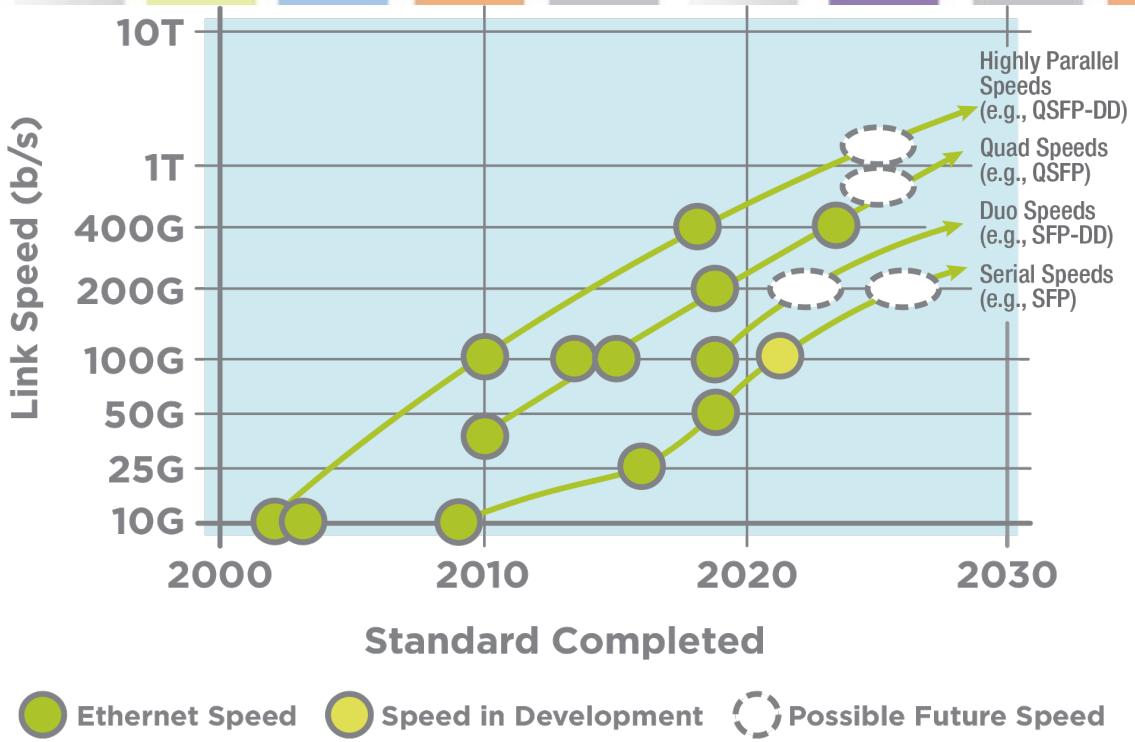


# What About Storage Over Ethernet?

SNIA NSF | NETWORKING STORAGE

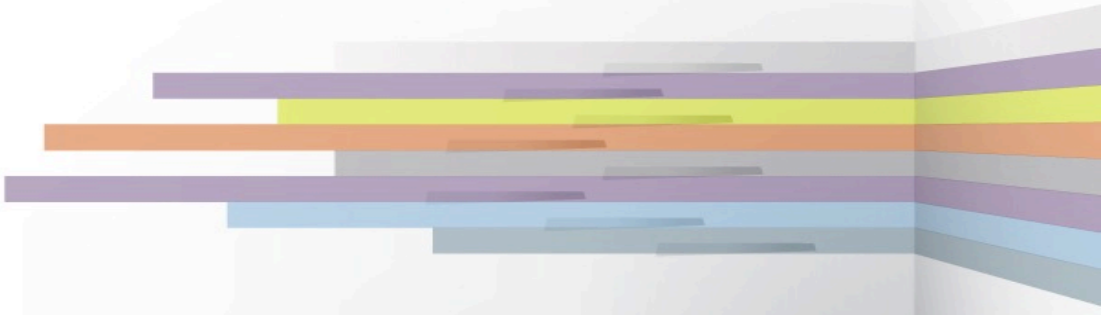
- Ethernet: Layer 1-2 infrastructure for storage protocols past, preset, future
  - Suitable for Block, File, Object oriented applications
  - Network Attached Storage – NFS and SMB
  - iSCSI, iSER, FCoE
  - NVMe over Fabrics, including NVMe over TCP

# TO TERABIT SPEEDS




- What's Next?
- 100Gb/s signaling in definition now, IEEE P802.3ck
- Is 200Gb/s signaling next?

SNIA NSF | NETWORKING STORAGE



# Fibre Channel

# Fibre Channel Physical Standards

- Fibre Channel physical layers defined in FC-PI series of Standards
  - Encoding and Protocol layers defined based on FC-FS series of standards
- Most Recent Standard ratified is FC-PI-7 – 64GFC
- Development has started on FC-PI-8 – 128GFC

| Document | Represents… |
|---|---|
| FC-PI | 1GFC<br>2GFC<br>4GFC |
| FC-PI-2 | 4GFC |
| FC-PI-4 | 8GFC |
| FC-PI-5 | 16GFC |
| FC-PI-6<br>FC-PI-6P | 32GFC<br>128GFC (parallel) |
| FC-PI-7<br>FC-PI-7P | 64GFC<br>256GFC (parallel) |
| FC-PI-8<br>FC-PI-8P | 128GFC<br>512GFC (parallel) |

# FCIA Roadmap

SNIA NSF | NETWORKING STORAGE

| Product Naming | Throughput (Mbytes/s) | Line Rate (Gbaud) | T11 Specification Technically Complete (Year)* | Market Availability (Year)* |
|---|---|---|---|---|
| 1GFC | 200 | 1.0625 | 1996 | 1997 |
| 2GFC | 400 | 2.125 | 2000 | 2001 |
| 4GFC | 800 | 4.25 | 2003 | 2005 |
| 8GFC | 1,600 | 8.5 | 2006 | 2008 |
| 16GFC | 3,200 | 14.025 | 2009 | 2011 |
| 32GFC | 6,400 | 28.05 | 2013 | 2016 |
| 128GFC | 25,600 | 4X28.05 | 2014 | 2016 |
| 64GFC | 12,800 | 28.9 PAM-4 (57.8Gb/s) | 2017 | 2019 |
| 256GFC | 51,200 | 4X28.9 PAM-4 (4X57.8Gb/s) | 2017 | 2019 |
| 128GFC | 25,600 | TBD | 2020 | Market Demand |
| 256GFC | 51,200 | TBD | 2023 | Market Demand |
| 512GFC | 102,400 | TBD | 2026 | Market Demand |
| 1TFC | 204,800 | TBD | 2029 | Market Demand |

# Signaling Rate Abbreviations

SNIA NSF | NETWORKING STORAGE

| Abbreviation | Signaling rate | Number of Lanes | Data rate |
|---|---|---|---|
| 1GFC | 1.0625 MBd | 1 | 100 MB/s |
| 2GFC | 2.125 MBd | 1 | 200 MB/s |
| 4GFC | 4.250 MBd | 1 | 400 MB/s |
| 8GFC | 8.500 MBd | 1 | 800 MB/s |
| 16GFC | 14.025 MBd | 1 | 1600 MB/s |
| 32GFC | 28.050 MBd | 1 | 3200 MB/s |
| 64GFC | 28.900 MBd | 1 | 6400 MB/s |
| 128GFC | 112.200 MBd | 1 or 4 | 12800 MB/s |
| 256GFC | 115.600 MBd | 4 | 25600 MB/s |

MB/s = Mega*bytes* per second
MBd = Mega*baud* per second

# Just Completed…

- FC-PI-7 (64GFC) ratified mid 2018
- 64GFC had to be backward compatible to 32GFC and 16GFC.
  - Backward compatibility and “plug and play” to utilize existing infrastructure with new speeds is always a must have for FC development.
- Existing cable assemblies must plug into 64GFC capable products
  - LC (connector) and SFP+ (form factor)
- Reach goals
  - 100 meters for multi-mode short reach optical variant using OM4/OM5 cable plants
    - OM4 optical fibre has a higher optical bandwidth than OM3 fibre which leads to longer reach at a given speed.
  - 10KM for single mode optical variant
  - Electrical variant for backplane applications
- 64GFC will double the throughput of 32GFC
- Corrected  bit-error-rate (BER) target of 1e-15
  - Advanced bit error recovery achieved through FEC

# Forward Error Correction for 64GFC

- Forward Error Correction (FEC) is mandatory for all types of 64GFC links
- Transmitter encodes the data stream in a redundant way using an error correcting code
- 64GFC uses a block code called Reed Solomon.
  - 64GFC uses RS(544,514)
  - Allows correction of single bit errors or burst errors for 15 ten-bit symbols out of 5140 bits sent
- 64GFC uses terms such as uncorrected BER which is the minimum BER to be expected pre-FEC encoding/decoding
  - Uncorrected BER is 1e-04 range or lower
  - FEC-corrected BER is 1e-15 range or lower
  - These numbers help identify the usefulness of FEC in making 64GFC links robust

**Forward Error Correction (FEC)**

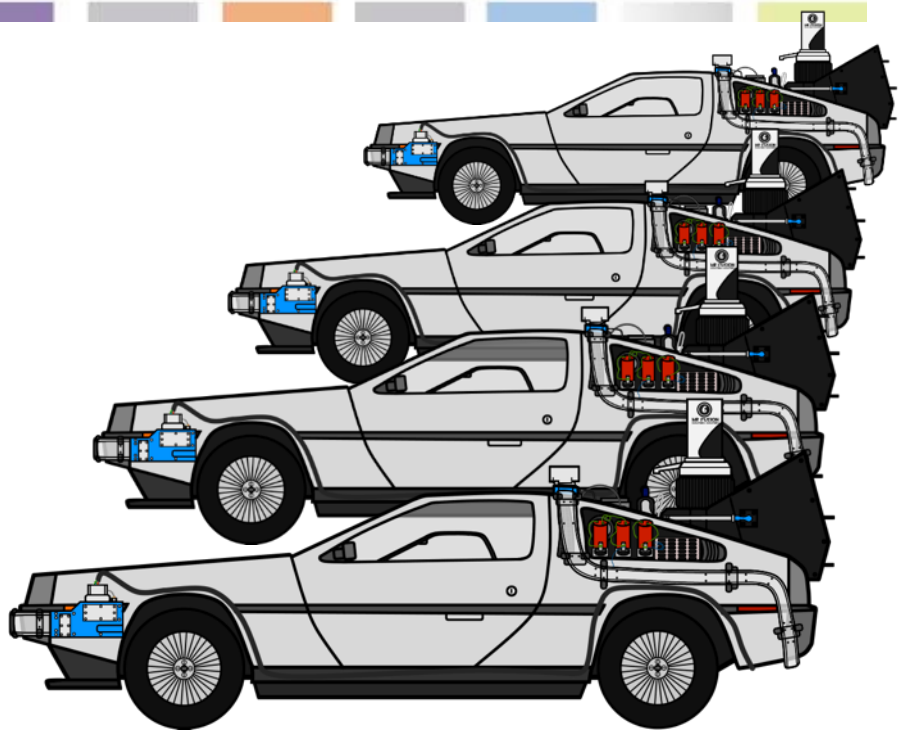
A set of algorithms that perform corrections that allow for recovery of one or more bit errors

- SNIA Dictionary

# 256GFC (Parallel Four Lane)

SNIA NSF | NETWORKING STORAGE

- FC-PI-7P will describe a four lane 64GFC variant that has a throughput of 256GFC (4x64GFC)
  - Standard currently in development
  - Expect first letter ballot to be mid-2019
- Data striped across the four lanes
- MRD requested the following variants
  - 100m on multi-mode cable OM4/OM5
  - 2km single mode variant
- Backward compatibility with 128GFC (4x32GFC) is also a requirement



# What's New

- FC-PI-8 - 128GFC
- FC-NVMe-2 – Advanced error recovery

# 128GFC FC-PI-8 Planned Requirements



- Backward compatible to 64GFC and 32GFC
- Same external connectors as 32/64GFC
- Existing cable assemblies will work with 128GFC
- Multi-mode cable plant reach is 100 meters on OM4/OM5
- Single mode cable plant reach of 10KM
- 128GFC links should double the throughput in MB/sec of 64GFC links
- Corrected BER target of 1e-15
- Reduce latency of 64GFC by up to 20%
- A four lane parallel 512GFC is also planned

# FC-NVMe-2 Being Ratified

- 2nd Revision of FC-NVMe being ratified now
  - Focus is on error Recovery
    - Errors recovered at the transport level
    - Quick error recovery – Errors are recovered before NVMe layer knows anything happened
  - Reliable nature of FC combined with new error recovery makes FC ideal for NVMe and all flash array datacenters

# Customer Use Case

SNIA NSF | NETWORKING STORAGE

- Biggest use case is the all-flash array datacenter
- FC with FC-NVMe is perfectly positioned to support this use
  - ~70% of AFAs connect with Fibre Channel (Source: Gartner)
- FC-NVMe ratified in 2017 (2nd rev. FC-NVMe-2 being ratified now)
  - NVMe over Fabrics on Enterprise-ready and trusted Fibre Channel
  - Support for all of the trusted Fibre Channel services
    - FC Name Server
    - FC State Change notification
    - Zoning
    - Management Services
    - Security

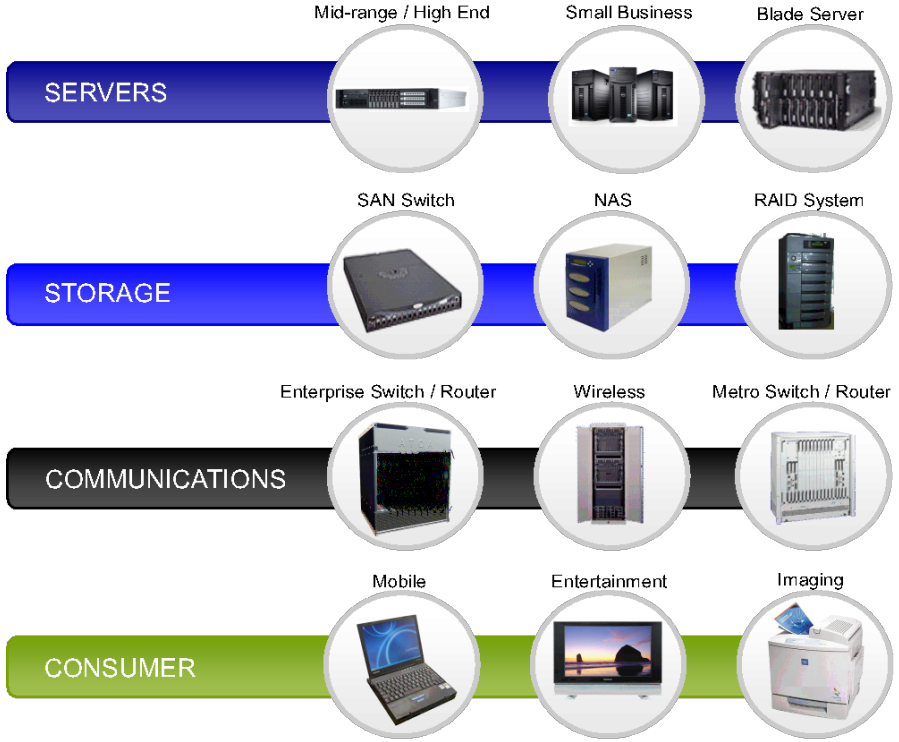
SNIA NSF | NETWORKING STORAGE

# PCI Express

# PCI Express ® (PCIe®)

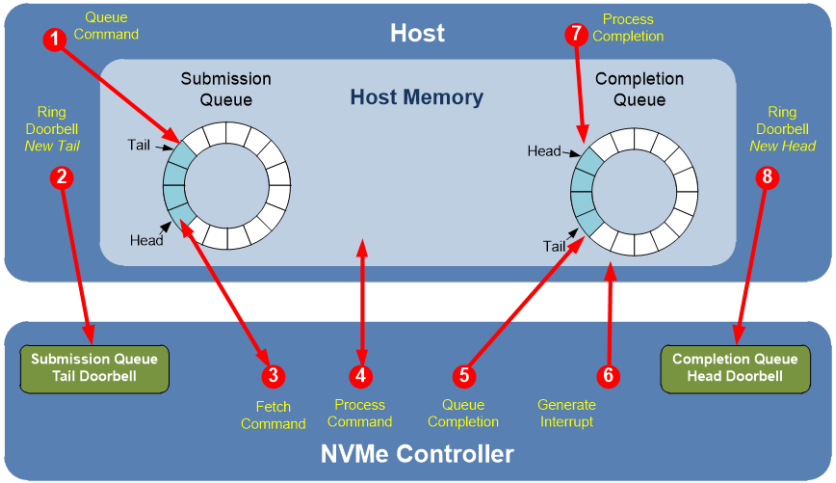
SNIA NSF | NETWORKING STORAGE

- Specification defined by PCI-SIG
  - www.pcisig.com
- Packet-based protocol over serial links
  - Software backward compatible with PCI and PCI-X
  - Reliable, in order packet transfer
- High performance and scalable from consumer to enterprise
- Primary application is as an I/O interconnect



# PCIe and NVMe™

- PCIe is the interface used for direct attached NVMe SSDs




## NVMe Operation

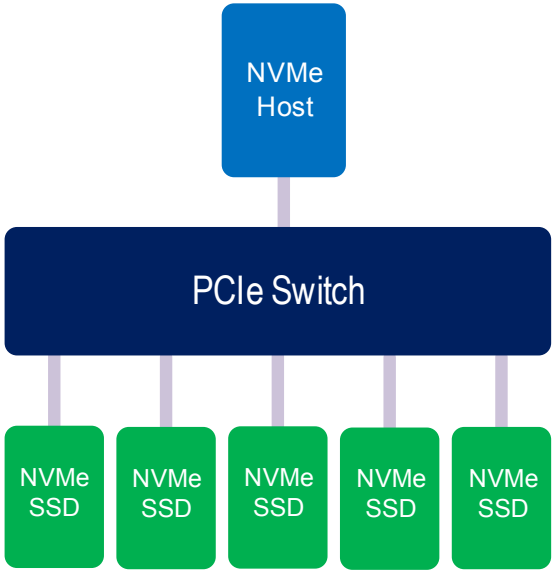
PCIe Card


M.2


U.2

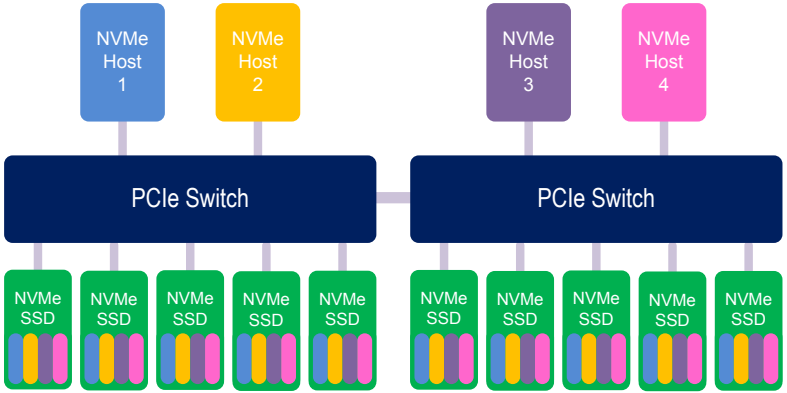
## NVMe SSD Form Factors

# PCIe NVMe JBOF

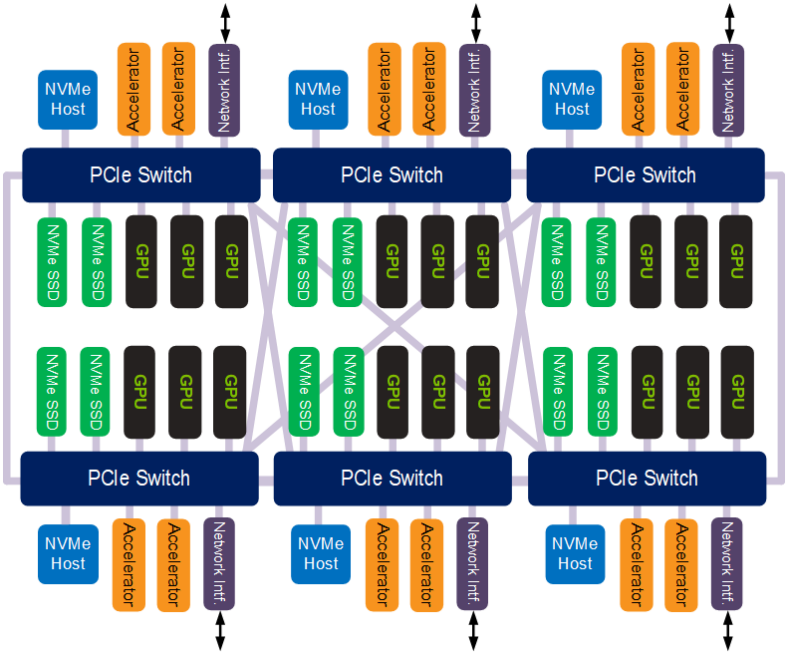



**Facebook Lightning PCIe NVMe JBOF**

# PCIe Fabrics

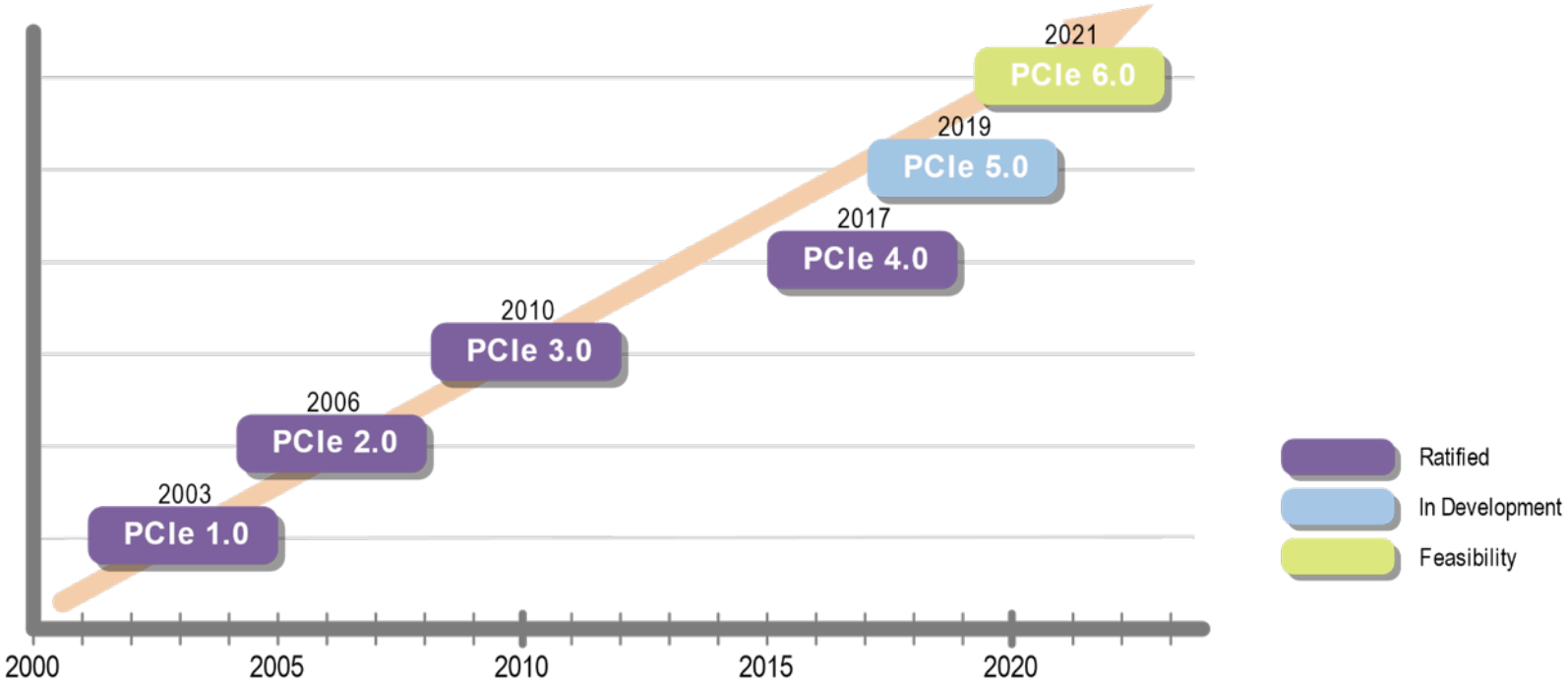



**Small PCIe Fabric with Shared NVMe Storage**




**PCIe Fabric with storage, accelerators, general purpose processors and networking**

# PCIe Roadmap



# PCIe Characteristics

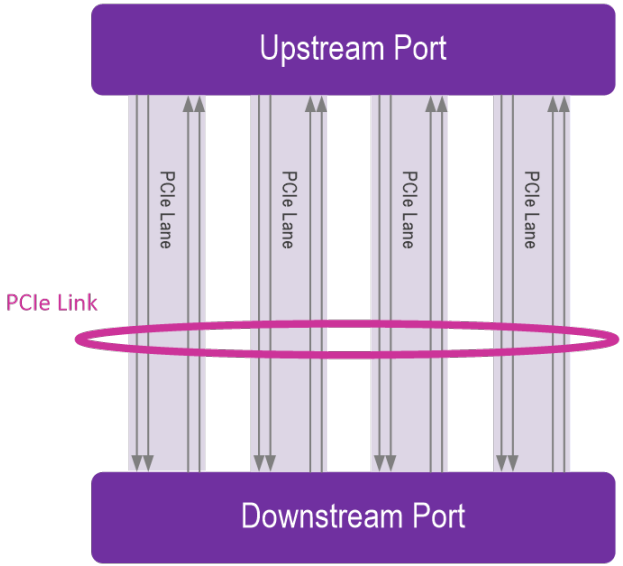
SNIA NSF | NETWORKING STORAGE

## Scalable Speed

| Specification | Generation | Raw Bit Rate | Encoding | Bandwidth Per Lane Each Direction | Total x16 Link Bandwidth |
|---|---|---|---|---|---|
| PCIe 1.0 | Gen 1* | 2.5 GT/s | 8b10b | ~ 250 MB/s | ~ 8 GB/s |
| PCIe 2.0 | Gen 2* | 5.0 GT/s | 8b10b | ~500 MB/s | ~16 GB/s |
| PCIe 3.0 | Gen 3* | 8 GT/s | 128b/130b | ~ 1 GB/s | ~ 32 GB/s |
| PCIe 4.0 | Gen 4 | 16 GT/s | 128b/130b | ~ 2 GB/s | ~ 64 GB/s |
| PCIe 5.0 | Gen 5 | 32 GT/s | 128b/130b | ~4 GB/s | ~128 GB/s |
| PCIe 6.0 | Gen 6 | 64 GT/s | 128b/130b | ~8 GB/s | ~ 256 GB/s |

** Source – PCI-SIG PCI Express 3.0 FAQ*

## Scalable Width




x1, x2, x4, x8, x12, x16, x32

# Speed Costs, How Fast Do You Want to Go?

SNIA NSF | NETWORKING STORAGE

- As speed increases, the distance signals travel in a given board material decreases
- Ways to increase signal distance
  - More expensive board material
  - PCIe retimers (or PCIe switches)
  - Cables
- Applications that require higher performance will move to faster speeds
  - High speed networking
  - GPUs and accelerators
  - NVMe SSDs based on next gen NVM

| Board Material | Board Material Cost Increase | Trace Length | |
|---|---|---|---|
| | | Gen4 | Gen5 |
| Meg2 | Baseline | ~ 15.5 in | ~ 13 in |
| Meg4 | 15% to 20% | ~ 19 in | ~ 14.5 in |
| Meg6 | 25% or more | ~ 24.2 in | ~ 16.25 in |

SNIA NSF | NETWORKING STORAGE

# InfiniBand

# InfiniBand and Use Cases

- Low-latency interconnect fabric
  - Supports RDMA & other offloads
- Top use cases
  - High Performance Computing
  - Artificial Intelligence / ML
  - Media & Entertainment
  - Specialized cloud
  - Data storage for above use cases

# Current InfiniBand Speeds

SNIA NSF | NETWORKING STORAGE

- FDR (56Gb/s) and EDR (100GB/s) InfiniBand
- Usually 4 lanes – 4x14Gb/s or 4x25Gb/s
- EDR similarities to 100GbE and128G FC
  - Bandwidth, cabling, signaling
  - Some FDR similarities to 64G FC
- FDR/EDR designed for PCIe Gen3
  - x8 lanes for FDR, x16 lanes for EDR

# EDR InfiniBand—Arriving Now

SNIA NSF | NETWORKING STORAGE

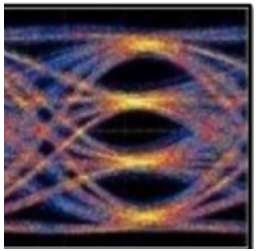
- HDR = 200Gb/s InfiniBand
  - 4x50Gb/s (EDR) or 2x50Gb (HDR100) lanes
- Equipment shipping in beta
  - Adapters, switches, cables and transceivers
- EDR designed for PCIe Gen4
  - x16 lanes PCIe Gen4
  - Or 2 slots each with x16 lanes of PCIe Gen3

# InfiniBand Characteristics

- Lossless network
- Credit-based flow control
- Forward Error Correction (FEC) not used
  - Very reliable cables with extra-low BER
  - Means lower latency
- Ethernet also offers no-FEC option
  - With higher-quality materials on short cables (2m max.)

# 200Gb/s Signaling

- Change from NRZ to PAM-4 Signaling
  - 25GHz clock but 2 bits per tick instead of 1
  - 4x25Gb/s = 100Gb/s
  - 4x50Gb/s = 200Gb/s
  - 2x50Gb/s = 100Gb/s on 2 lanes
- Same change for multiple fabrics
  - 200Gb Ethernet, 256G FC, HDR InfiniBand



# 200Gb/s Cabling



- Cabling and Form Factors
  - QSFP56 is 4x50; SFP56 is 1x50
  - QSFP-DD & OSFP allow 8x50Gb/s
  - Shorter copper cables: 3m (not 5m)
- Coming next: 100Gb/s data rate
  - 50GHz signaling, PAM4 (2 bits per tick)
  - 4x100Gb/s = 400Gb/s

**QSFP28**
4x25G

**QSFP56**
4x50G

**QSFP-DD**
8x50G

**SFP28**
**1**x25G NRZ

**SFP56**
**1**x50G PAM4

**MPO-12**

**LC Duplex**

SNIA NSF | NETWORKING STORAGE

# What Does It All Mean?

# Faster Networks are Here



- Get more bandwidth when you need it
  - Support faster servers/storage
  - More efficient large datacenter fabrics
- Fastest speeds only deployed where needed
  - May use only for fastest servers or switch uplinks
- Match network speed to PCIe bus bandwidth
  - Fastest new network speeds need PCIe Gen4
  - Vendors working on PCIe Gen5

# More Webcasts

- More SNIA NSF Webcasts Available on-demand at:
  - https://www.snia.org/forums/esf/knowledge/webcasts-topics
- Next Live Webcast:
- *Intro to Incast, Head of Line Blocking and Congestion Management*
  - June 18, 2019, 10:00 am PT
  - Register at: https://www.brighttalk.com/webcast/663/356343

# After This Webcast

SNIA NSF | NETWORKING STORAGE

- Please rate this webcast and provide us with feedback
- This webcast and a PDF of the slides will be posted to the SNIA Networking Storage Forum (NSF) website and available on-demand at www.snia.org/forums/nsf/knowledge/webcasts
- A full Q&A from this webcast, including answers to questions we couldn't get to today, will be posted to the SNIA-NSF blog: sniansfblog.org
- Follow us on Twitter @SNIANSF

# Thank You