FINISAR

Latest Trends in Optical Interconnects

NANOG 66

San Diego – February 2016

Christian Urricariet

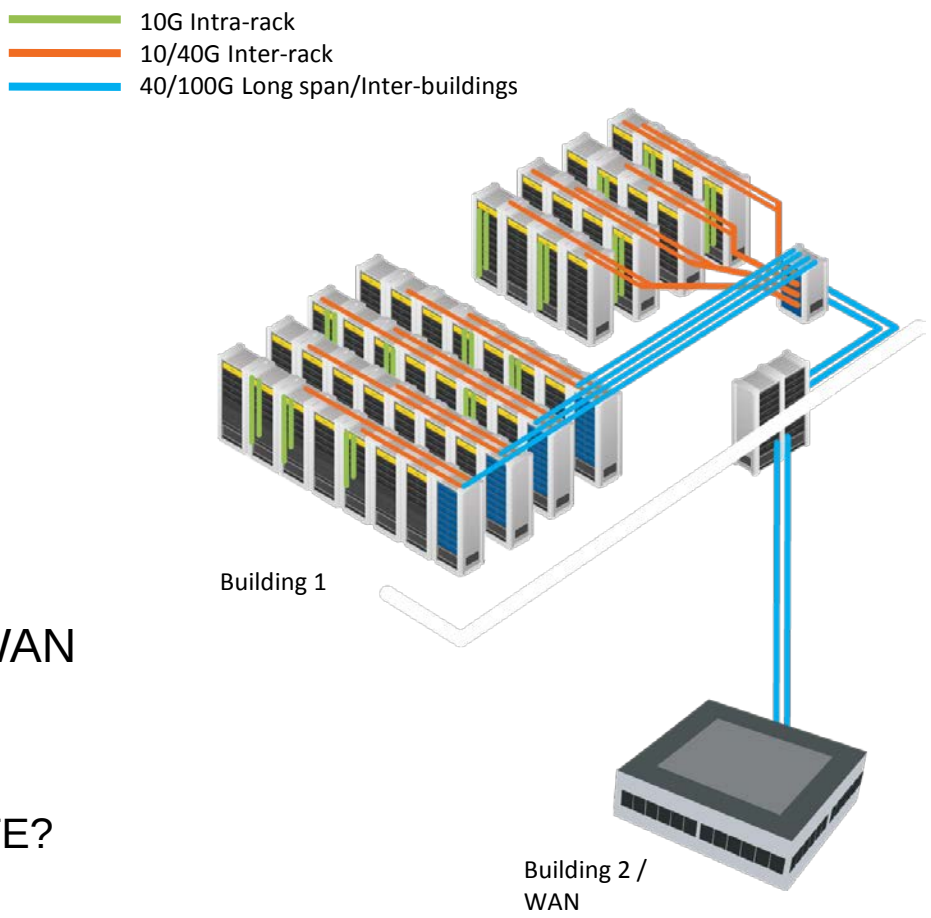
Data Center Connections are Changing

- Data Center connections are moving from 10G/40G, to 25G/100G

- Within the Data Center Rack
  - 10GE being deployed now
  - 25GE to be deployed soon
  - 50GE to the server will follow

- Between Data Center Racks
  - 40GE being deployed now
  - 100GE to be deployed soon
  - What follows? 200GE or 400GE?

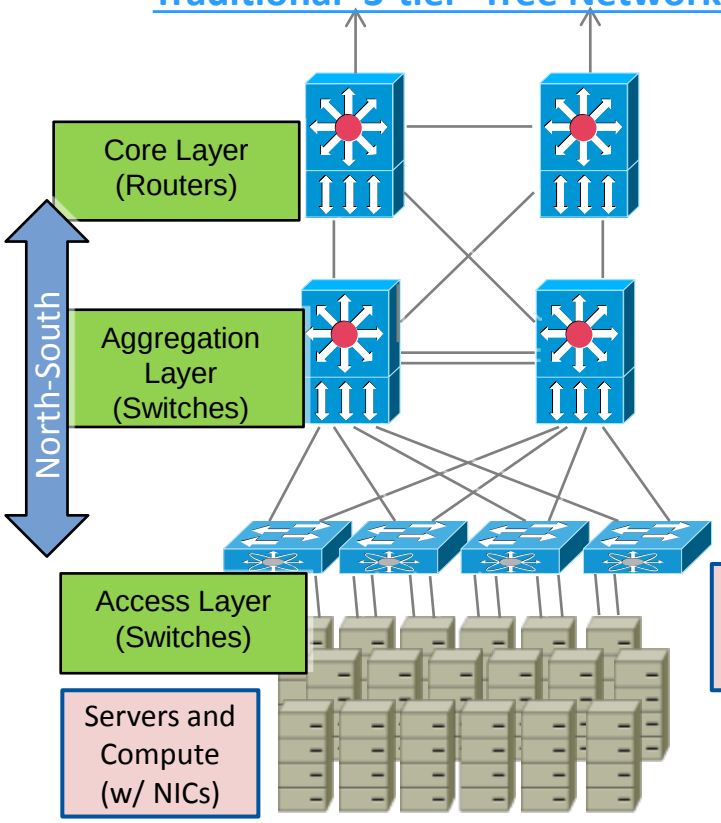
- Long Spans/Inter-Data Centers & WAN
  - 100GE being deployed until now
  - 400GE being standardized now
  - What follows? 800GE, 1TE or 1.6TE?



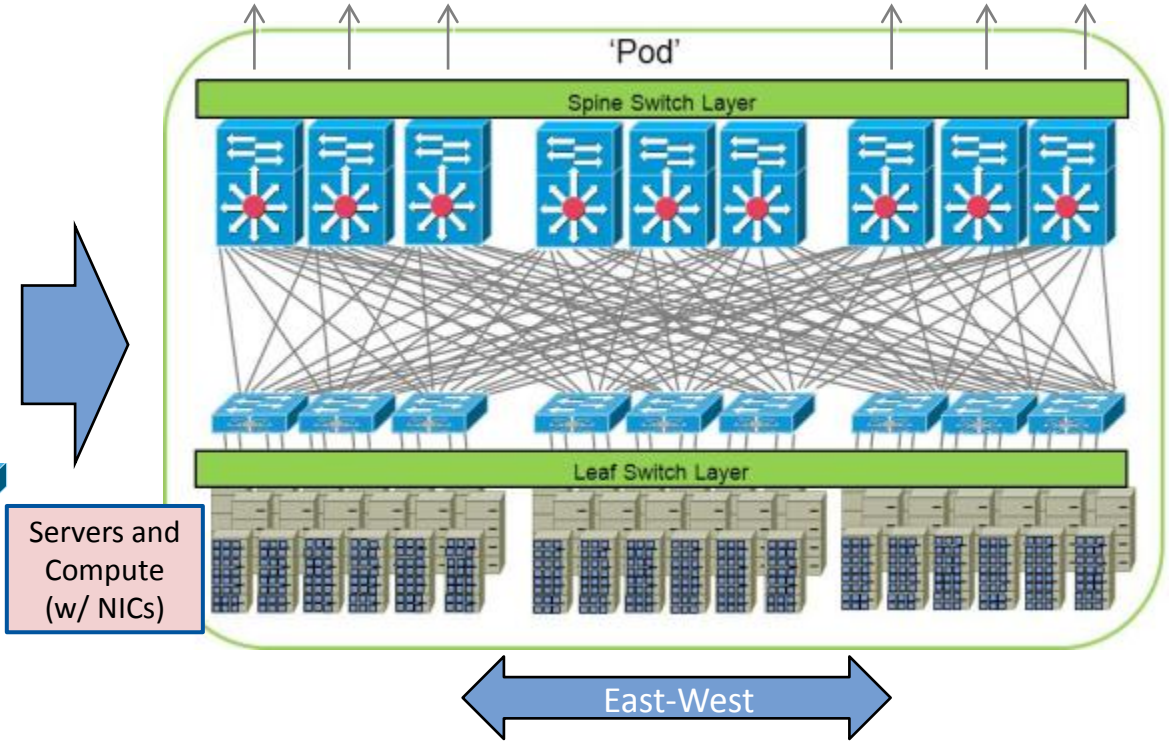
New Architectures in Hyperscale Data Centers

- Most data center networks have been architected on a 3-tier topology
- Cloud data center networks are migrating from traditional 3-tier to flattened 2-tier topology
  - Hyperscale Data Centers becoming larger, more modular, more homogenous
  - Workloads spread across 10s, 100s, sometimes 1000s of VMs and hosts
  - Higher degree of east-west traffic across network (server to server)

Traditional '3-tier' Tree Network

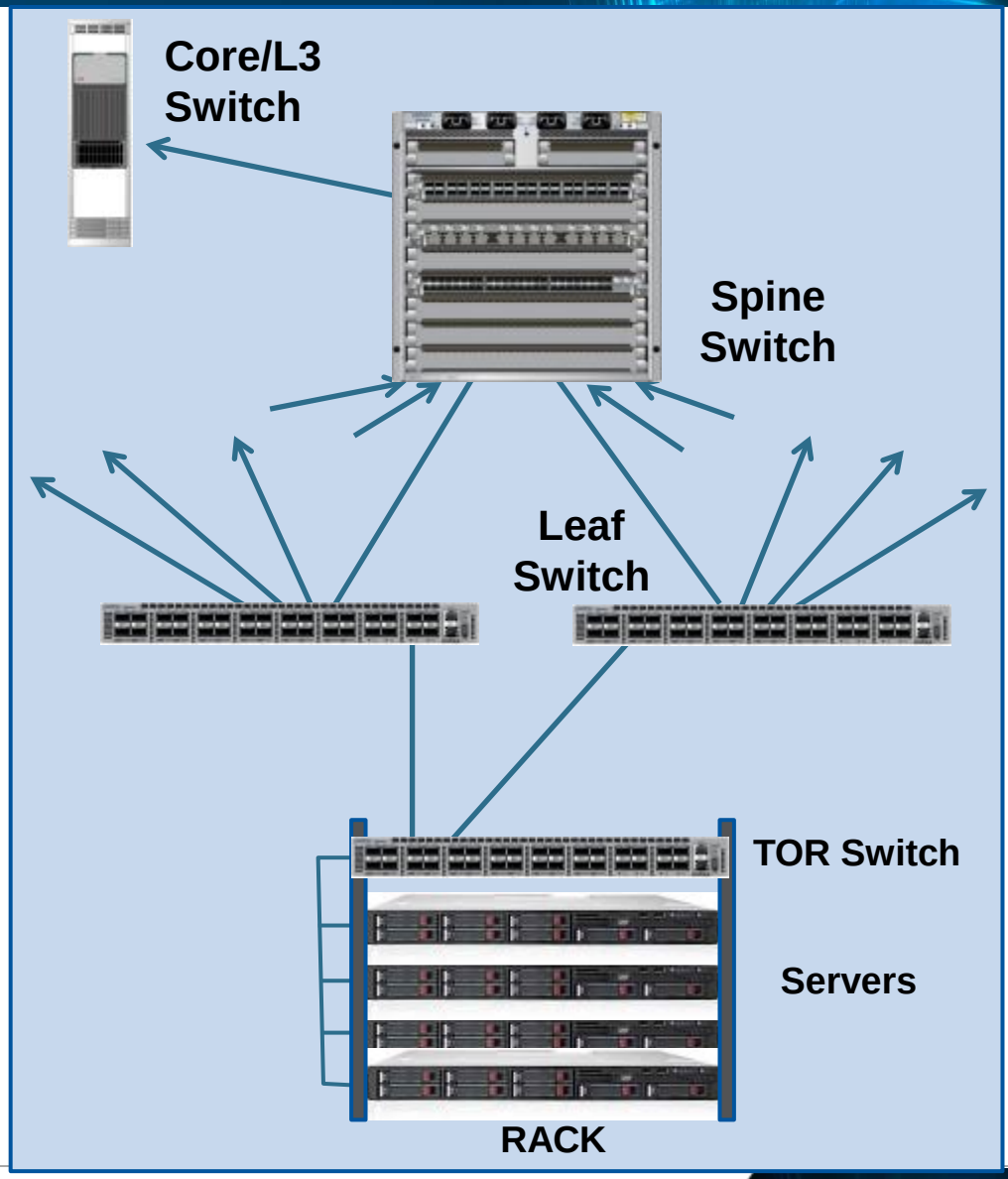


New '2-tier' Leaf-Spine Network



The Hyperscale/Cloud Data Center

- The **RACK** is the minimum building block.
- The goal is to connect as many racks together as possible.
- Heavy ‘East-West’ traffic (server to server).
- Minimum over-subscription.
- Each leaf switch fans out to all spine switches (high radix).



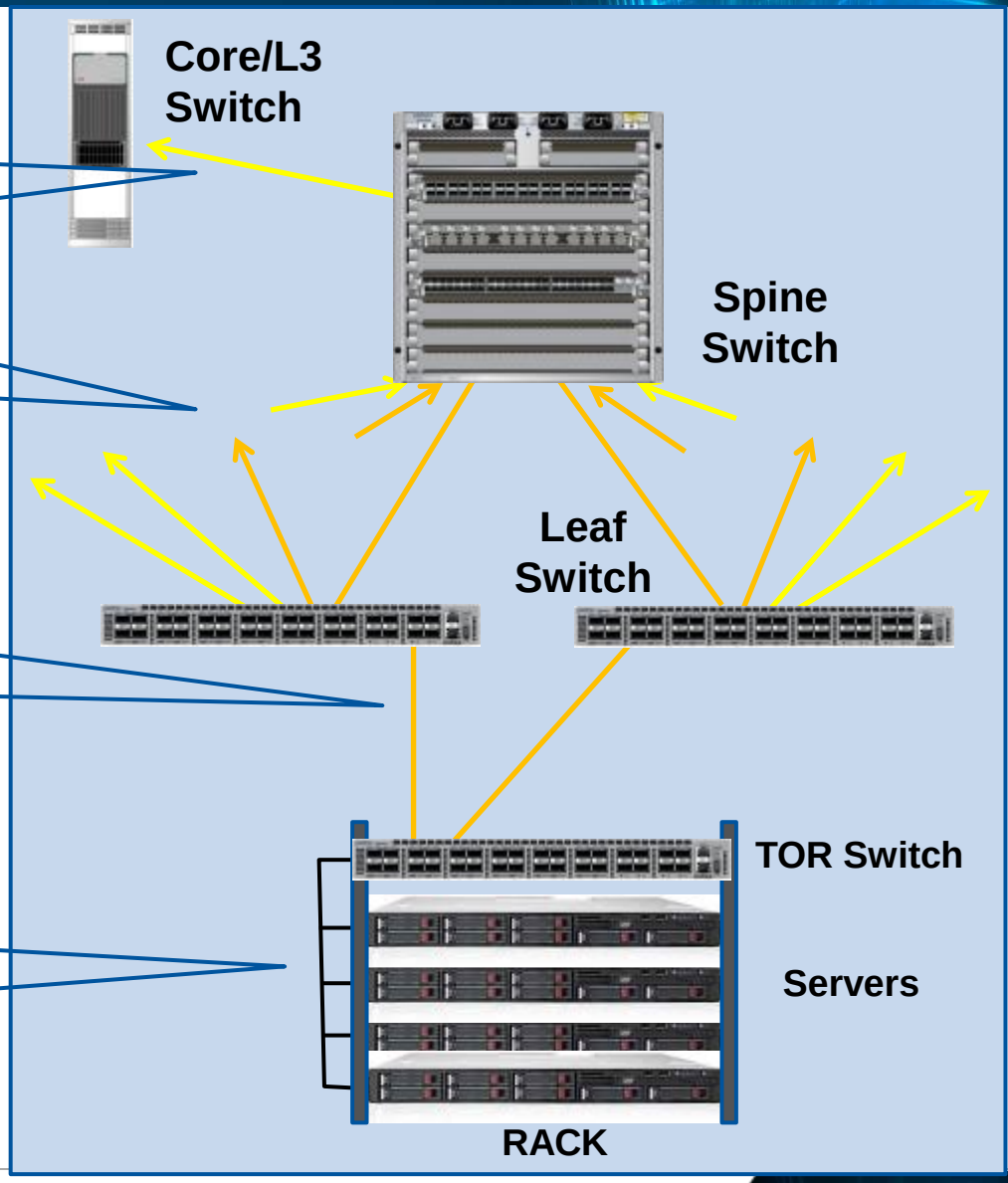
Connections in the Hyperscale/Cloud Data Center

Core Switch/Router to Spine Switch:
Deploying mostly 40GE LR4 today.
Will deploy 100GE CWDM4/LR4 soon.
Roadmap is 200GE or 400GE next.

Spine Switch to Leaf Switch links:
Deploying mostly 40GE SR4/LR4 today.
Will deploy 100GE SR4/CWDM4 soon.
Roadmap may be 200GE SR/LR next.

Leaf Switch to TOR Switch links:
Deploying mostly 40GE SR4 today.
Will deploy 100GE SR4/AOC soon.
Roadmap may be 200GE SR next.

TOR Switch to Server links:
Deploying mostly 10GE SR/DAC today.
Will deploy 25GE SR/AOC soon.
Roadmap is 50GE SR/AOC next.

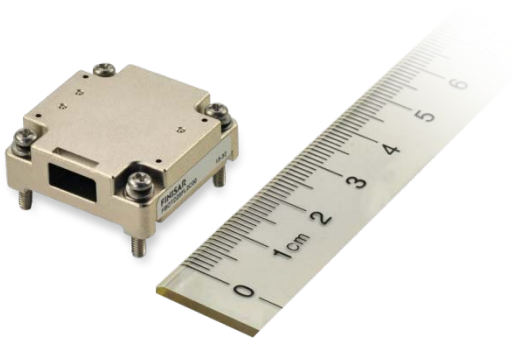


Interconnect Trends in the Data Center Market

- Significant increase in 100G and 25G port density

Interconnect Trends in the Data Center Market

- Significant increase in 100G and 25G port density
  - Smaller form factors, e.g., QSFP28 modules
  - Power dissipation <3.5W
  - Active Optical Cables
  - On-board optics for very high port density

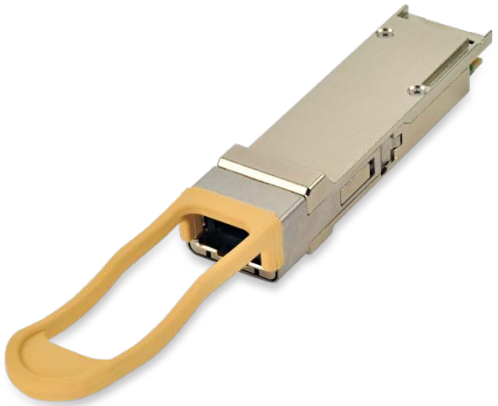


100G Optical Module Form Factor Evolution

CFP	CFP2	CFP4	QSFP28
4 ports/chassis	8-10 ports/chassis	16-18 ports/chassis	18-20 ports/chassis
24W	8W	5W	3.5W

Deployments until today → *time*

100G QSFP28 Module



QSFP28 is the 100GE module form factor of choice for new data center switches

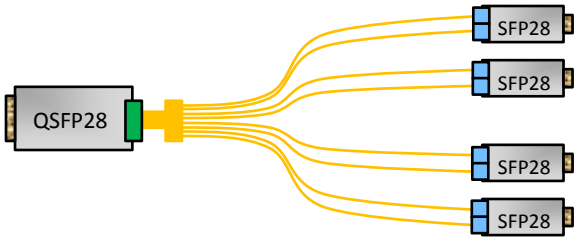
- **100GE optical transceivers**
  - QSFP28 is standardized by SFF-8665 (SFF Committee)
  - It has a 4-lane, retimed 25G I/O electrical interface (CAUI-4)
  - Supports up to 3.5W power dissipation with standard cooling
  - Also used for 4x 25GE applications
- **100GE active optical cables (no optical connector)**

QSFP28: 100G *and* High-Density 25G

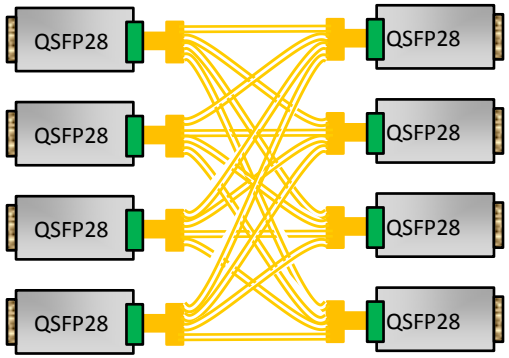
- QSFP28 = Quad SFP28
- QSFP28 is both a 100G ***and*** a high-density 25G form factor

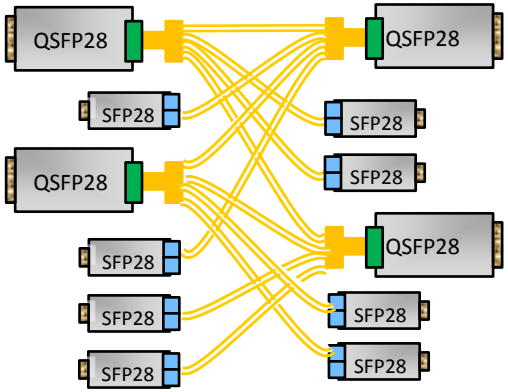
Point-to-Point 100G

4x25G Breakout



4x25G Shuffle





General Case:

Breakout *and* Shuffle

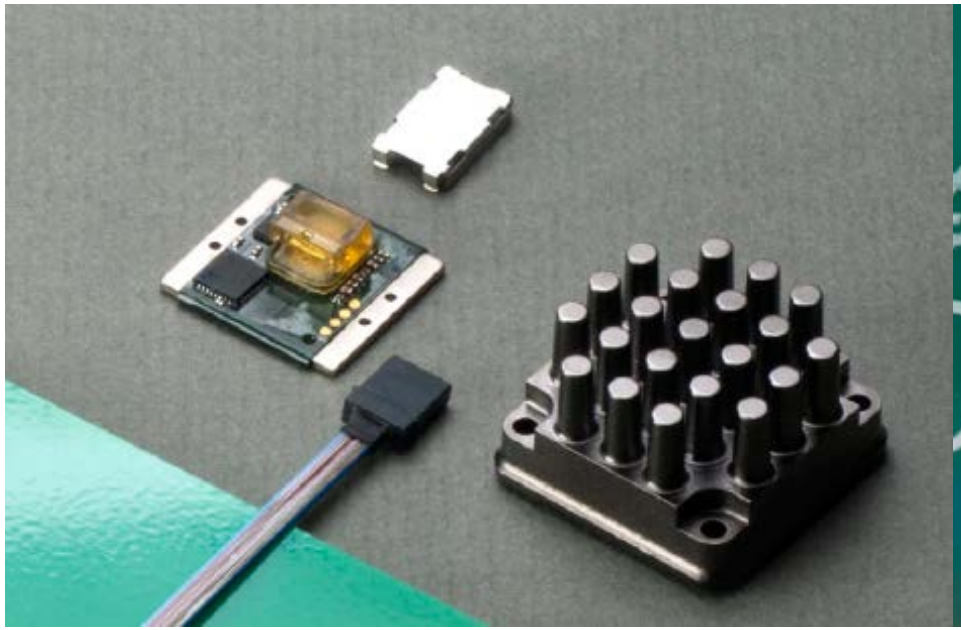
QSFP28 will have very high volumes, because it supports both 100G and 25G links.

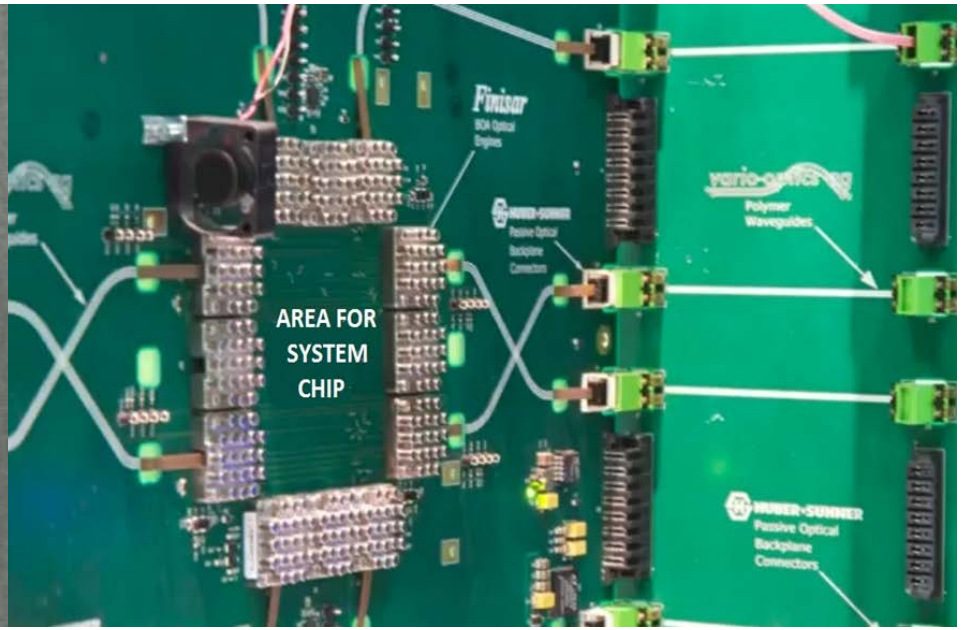
25G SFP28 Module

SFP28 is the 25GE module form factor of choice for new Servers / NICs

- **25GE optical transceivers**
  - SFP28 is standardized by the SFF Committee
  - It has a 1-lane, retimed 25G I/O electrical interface
  - Supports up to 1W power dissipation with standard cooling
  - Used for 25GE ports in server and switches
- **25GE active optical cables**

Board-Mounted Optical Assembly (BOA)





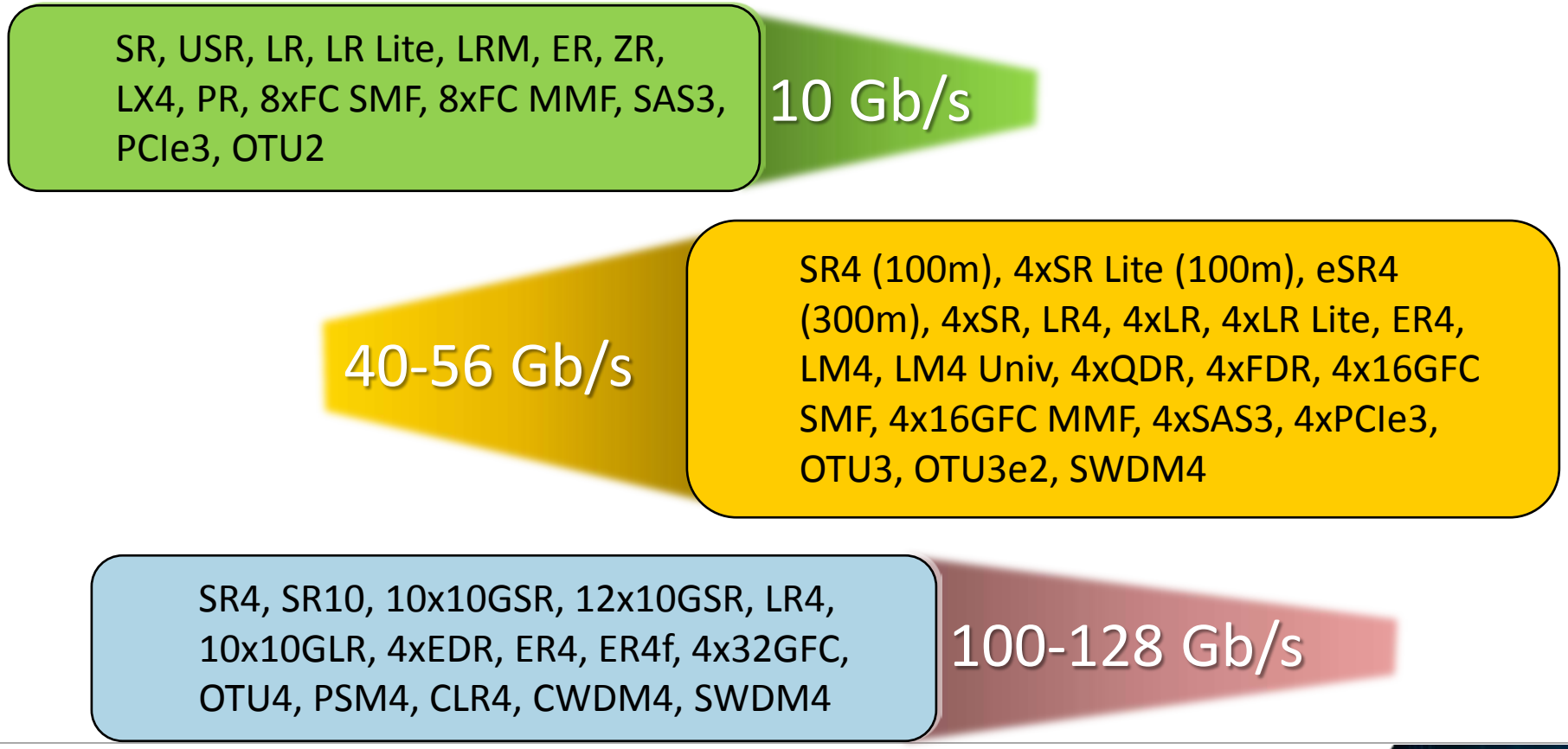
- These optics are not pluggable; they are mounted on the host PCB
- Used today on core routers, supercomputers and some switches
- Very short host PCB traces enable low power dissipation
- Higher bandwidth density can be achieved by:
  - More channels: Up to 12+12 Tx/Rx, or 24Tx and 24Rx
  - Higher data rate per channel: 10G/ch and 25G/ch variants today, 50G/ch in the future

Interconnect Trends in the Data Center Market

- Significant increase in 100G and 25G port density
- Extension of optical links beyond the Standards

Optical Standards Proliferation

- Duplex and parallel optics products continue to proliferate
- This results in a proliferation of standards, *de facto* standards, MSAs, and proprietary codes, each optimized for a particular use case



40G Ethernet QSFP+ Modules

	Parallel (MPO)	Duplex (LC)
Multimode	SR4 • 100/150m eSR4 & 4xSR • 300/400m	A duplex multimode product is required to re-use the same fiber plant used for 10GE LM4 • 140/160m/1km
Single Mode	4xLR • 10km 4xLR Lite • 2km	LR4 • 10km ER4 • 40km

Black = Standardized interfaces

Blue = MSA/Proprietary interfaces

Parallel links ***can*** be broken out to 4 separate 10G connections

Duplex WDM ***cannot*** be broken out to separate 10G connections

Multimode distances refer to OM3/OM4
Single mode distances refer to SMF28

100G Ethernet QSFP28 Modules

	Parallel (MPO)	Duplex (LC)
Multimode	SR4 & 4x25G-SR • 70/100m SR4 without FEC • 30/40m	A duplex multimode product is required to re-use the same fiber plant used for 10GE
Single Mode	PSM4 • 500m	LR4 • 10km CWDM4/CLR4 • 2km

Black = Standardized interfaces

Blue = MSA/Proprietary interfaces

Parallel links ***can*** be broken out to 4 separate 10G connections

Duplex WDM ***cannot*** be broken out to separate 10G connections

Multimode distances refer to OM3/OM4
Single mode distances refer to SMF28

Impact of Latency on 25G/100G Ethernet Optical Links

- Various recent 25G and 100G Ethernet standards and MSAs require the use of **RS-FEC** (aka, “KR4 FEC”) on the host to increase overall link length:

- RS-FEC does not increase the total bit rate, but it introduces an additional ***latency of ~100ns*** in the link.
  - Some applications like HFT have little tolerance for latency.

Standard	Link Length with RS-FEC
IEEE 802.3bm 100GBASE-SR4	100m on OM4 MMF
IEEE P802.3by 25GBASE-SR	100m on OM4 MMF
100G CWDM4 MSA	2km on SMF
100G PSM4 MSA	500m on SMF

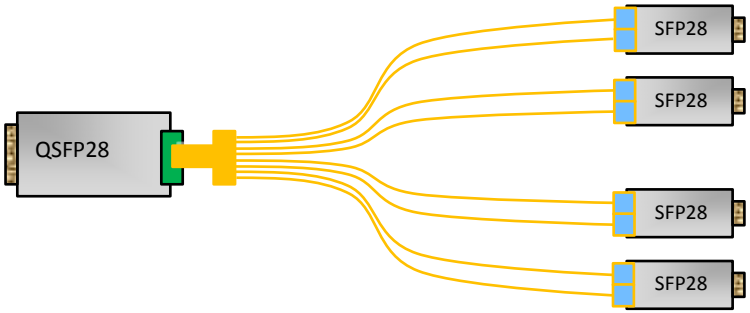
- The fiber propagation time of each bit over 100m of MMF is **~500ns**
  → The amount of additional latency introduced by RS-FEC may be significant for the overall performance of short links <100 meters (see next page).

- But the fiber propagation time of each bit over 500m of SMF is **~2500ns**
  → The amount of latency introduced by RS-FEC is **not** significant for the overall performance of links >500 meters.

Low-Latency QSFP28 SR4 and SFP28 SR without FEC

- Support of 25G/100G Ethernet links ***without FEC***
  - Lower latency
  - Lower host power dissipation

- Standard QSFP28 and SFP28 form factors
- Supports 4:1 fan-out configuration
- Up to 30 meters on OM3 / 40 meters on OM4 MMF



Interconnect Trends in the Data Center Market

- Significant increase in 100G and 25G port density
- Extension of optical links beyond the Standards
- Reutilization of existing 10G fiber plant on 40G and 100G

Why Duplex Multimode Fiber Matters

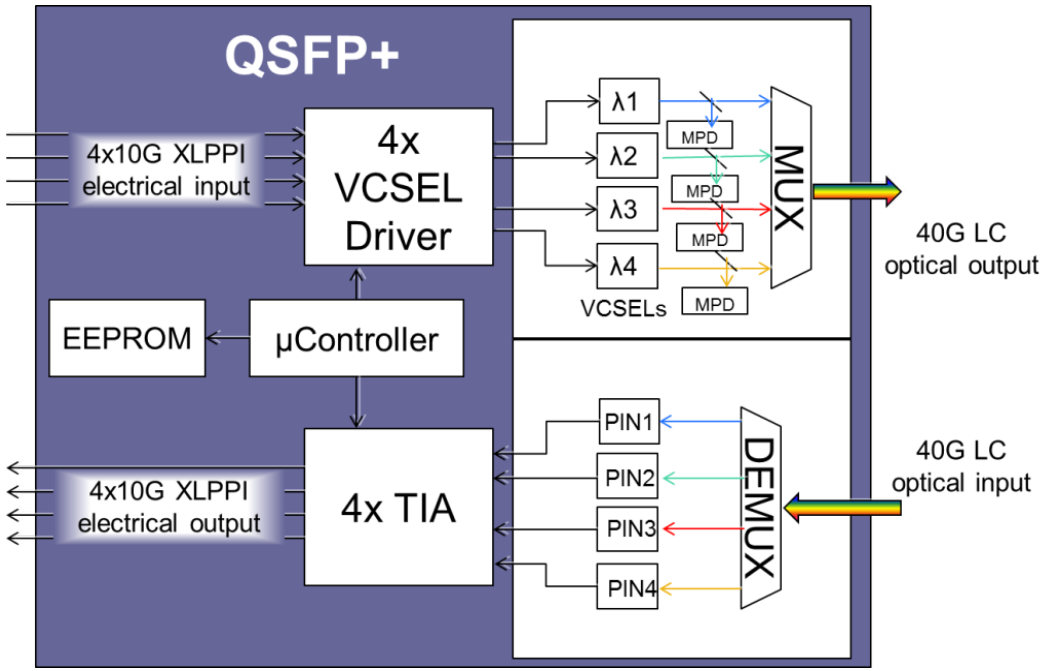
- Data centers today are architected around 10G Ethernet
- Primarily focused on 10GBASE-SR using **duplex MMF (LC)**
- Data center operators are migrating from 10G to 40G or 100G, but want to maintain their existing fiber infrastructure
  - SR4 requires ribbon multimode fiber with an MPO connector
    - ***Not provided by pre-installed fiber plant***
  - LR4 requires single mode fiber
    - ***Not provided by pre-installed fiber plant***

Data centers want to upgrade from 10G to 40 and 100G
without touching the duplex MMF fiber infrastructure

Introducing Shortwave WDM (SWDM)

- SWDM uses 4 different wavelengths in the 850nm region, where MMF is optimized, which are optically multiplexed inside the transceiver.
- SWDM enables the transmission of 40G (4x10G) and 100G (4x25G) over existing duplex multimode fiber, using LC connectors.

Block diagram of a 40G SWDM QSFP+ Transceiver



- Industry group to promote SWDM technology for duplex MMF in data centers.
- Finisar is a founding member of the SWDM Alliance.
- More information at www.swdm.org

swdm4™

40G Ethernet QSFP+ Modules

	Parallel (MPO)	Duplex (LC)
Multimode	SR4 • 100/150m eSR4 & 4xSR • 300/400m	Bi-directional • Limited use SWDM4 • Being tested LM4 • 140/160m/1km
Single Mode	4xLR • 10km 4xLR Lite • 2km	LR4 • 10km ER4 • 40km

Black = Standardized interfaces

Blue = MSA/Proprietary interfaces

Parallel links ***can*** be broken out to 4 separate 10G connections

Duplex WDM ***cannot*** be broken out to separate 10G connections

Multimode distances refer to OM3/OM4
Single mode distances refer to SMF28

100G Ethernet QSFP28 Modules

	Parallel (MPO)	Duplex (LC)
Multimode	SR4 & 4x25G-SR • 70/100m SR4 without FEC • 30/40m	SWDM4 • Being tested
Single Mode	PSM4 • 500m	LR4 • 10km CWDM4/CLR4 • 2km

Black = Standardized interfaces

Blue = MSA/Proprietary interfaces

Parallel links ***can*** be broken out to 4 separate 10G connections

Duplex WDM ***cannot*** be broken out to separate 10G connections

Multimode distances refer to OM3/OM4
Single mode distances refer to SMF28

Interconnect Trends in the Data Center Market

- Significant increase in 100G and 25G port density
- Extension of optical links beyond the Standards
- Reutilization of existing 10G fiber plant on 40G and 100G
- Moving beyond 100G, to 200G and 400G

Distinct 200G/400G Applications in the Market

- **Service Provider Applications**:
  400GE Router-Router and Router-Transport client interfaces
  - Critical requirements are time to market and supporting multiple reaches.
  - Currently deploying tens of thousands of 100GE CFP/CFP2/CFP4.
  - First generation 400GE client module will need to provide a port density advantage with respect to using 4x QSFP28.

- **Data Center and Enterprise**:
  200GE uplinks and 4x100GE fan-outs
  - Critical requirement is high port count/density.
  - Currently planning on deploying 25G SFP28 on the server and 100G QSFP28 on switches starting in CY2016.
  - A 400G “QSFP112” module will take several years to be feasible due to power dissipation and size limitations.
  - A better product for the next generation of switches may be a **200GE QSFP56** module, which could also support 4x50GE fan-out.
  - Servers have a roadmap to 50GE I/O already.

200GE and 400GE Standardization

- The 400GE Standard is already being defined in IEEE P802.3bs.

Interface	Link Distance	Media type	Technology
400GBASE-SR16	100 m	32f Parallel MMF	16x25G NRZ Parallel
400GBASE-DR4	500 m	8f Parallel SMF	4x100G PAM4 Parallel
400GBASE-FR8	2 km	2f Duplex SMF	8x50G PAM4 LAN-WDM
400GBASE-LR8	10 km	2f Duplex SMF	8x50G PAM4 LAN-WDM

  - Electrical I/O: CDAUI-8 8x50G PAM4
    CDAUI-16 16x25G NRZ
  - 400GE Standard is expected to be ratified in December 2017

- 50G and 200G Ethernet standardization by IEEE has started.

- Optics suppliers are already working on components to support these new rates.
  - Based on VCSELs, InP DFB laser and Si Photonics technologies
  - ICs and test platforms that support PAM4 encoding

50G, 200G and Next-Gen 100G Ethernet Standardization

- 200GE PMD objectives to be studied by IEEE 802.3bs:

Interface	Link Distance	Media type	Technology
200GBASE-SR4	100 m	8f Parallel SMF	4x50G PAM4 850nm
200GBASE-FR4	2 km	2f Duplex SMF	4x50G PAM4 CWDM
200GBASE-LR4	10 km	2f Duplex SMF	4x50G PAM4 CWDM

- 50GE PMD objectives to be studied by new IEEE Task Force:

Interface	Link Distance	Media type	Technology
50GBASE-SR	100 m	2f Duplex MMF	50G PAM4 850nm
50GBASE-FR	2 km	2f Duplex SMF	50G PAM4 1300nm window
50GBASE-LR	10 km	2f Duplex SMF	50G PAM4 1300nm window

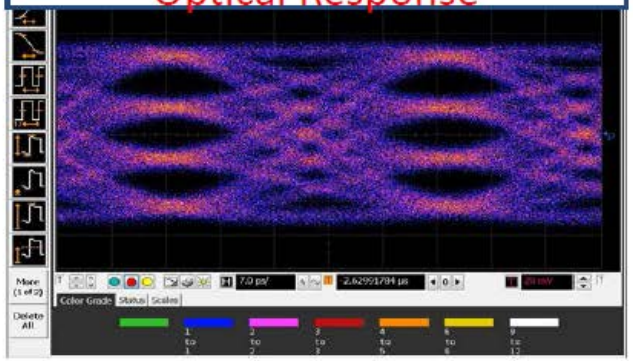
- Next-Gen 100GE PMD objectives to be studied by new IEEE Task Force:

Interface	Link Distance	Media type	Technology
100GBASE-SR2	100 m	2f Duplex MMF	2x50G TBD
100GBASE-xR2	x km	2f Duplex SMF	2x50G TBD

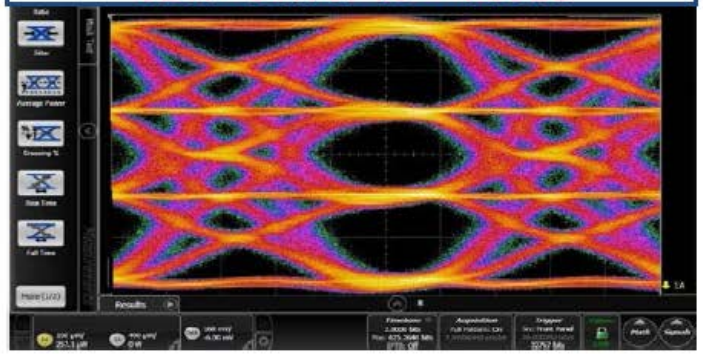
Technical Feasibility: 50 Gb/s PAM4 at Finisar

Bench top PAM4 experiments using 25Gb/s VCSELs

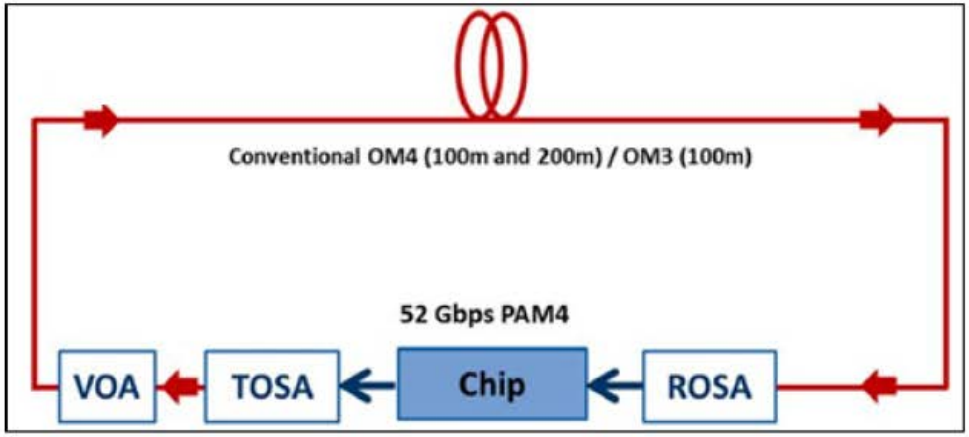
50Gb/s AWG + VCSEL
Optical Response

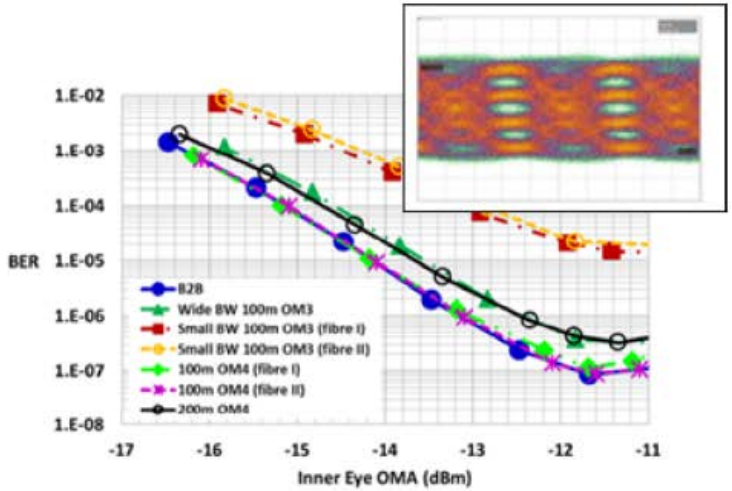


50Gb/s PAM4 Optical Response
CDR+ Driver IC + VCSEL

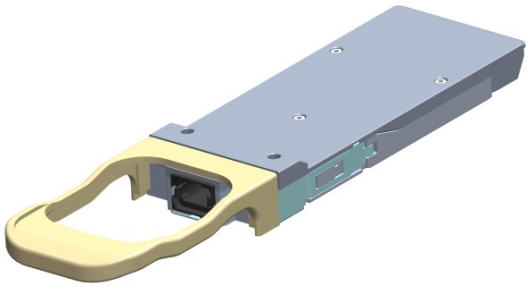


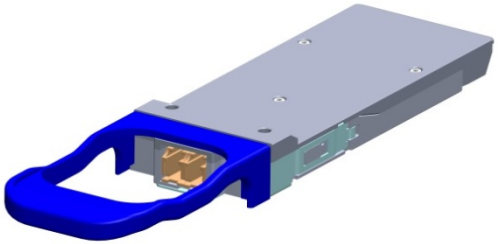
and early PAM4 PHY evaluation….

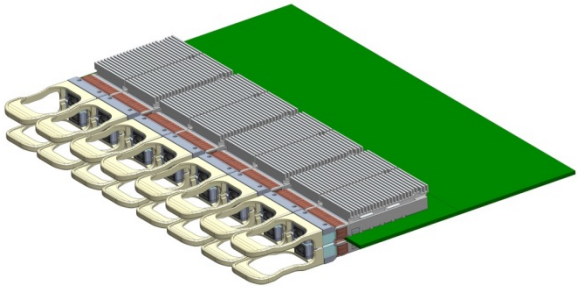




400GE CFP8 Optical Transceiver Module

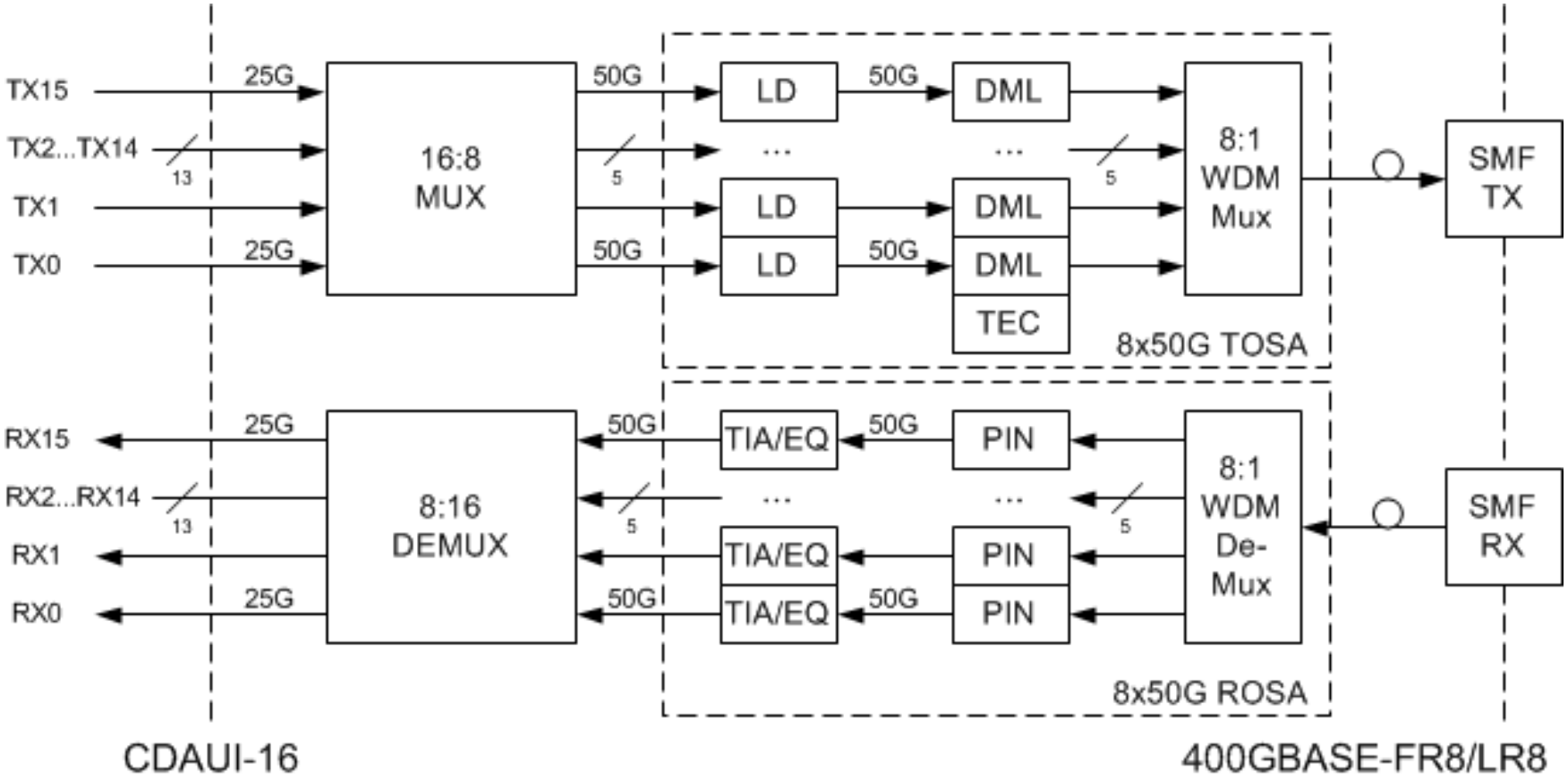






- **CFP8** is a proposed first-generation 400GE form factor.
- Module dimensions are **similar to CFP2**.
- Enables **6.4 Tb/s** per host board (8x2 modules in a 1RU configuration).
  - Supported ports: 16x400G, 64x100G, 128x50G, 256x25G
- Supports standard IEEE 400G **multimode and single mode** interfaces
- Supports either **CDAUI-16** (16x25G) or **CDAUI-8** (8x50G) electrical I/O.
- It is being standardized by the **CFP MSA**

400GBASE-FR8/LR8 CFP8 Generic Block Diagram



- 8x50G PAM4 optical modulation
- 16x25G NRZ electrical interface to the host

Summary

- Large growth in web content and applications is driving:
  - Growth in bandwidth and changes in data center architectures
  - Subsequent growth in number of optical links
  - Large increase in power requirements
- 25G, 40G and 100G optics support this growth today with:
  - Smaller module form factors for higher port density
  - Lower power consumption and cost per bit
  - Increased performance to leverage existing fiber infrastructure
- New Ethernet speeds are being standardized: 50G, 200G, 400G
- Questions?
- Contact Us
  - E-mail: christian.urricariet@finisar.com
  - www.finisar.com

FINISAR®

Thank You