cisco Live!

Let's go

The bridge to possible

Towards High Performance 400G, 800G Data Center

Errol Roberts, Distinguished Engineer
@errolfroberts

BRKDCN-2677

Cisco Webex App

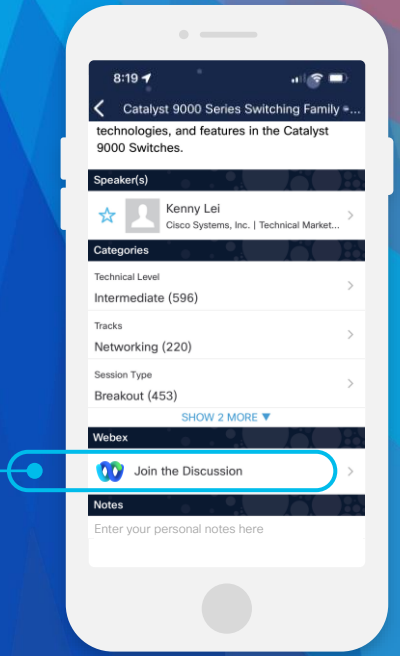
Questions?

Use Cisco Webex App to chat
with the speaker after the session

How

1. Find this session in the Cisco Live Mobile App
2. Click “Join the Discussion”
3. Install the Webex App or go directly to the Webex space
4. Enter messages/questions in the Webex space

Webex spaces will be moderated by the speaker until June 9, 2023.



https://ciscolive.ciscoevents.com/ciscolivebot/#BRKDCN-2677

Agenda

- Market Dynamics
- Building Blocks for 400/800G
- Network Architecture Consideration
- Network Architecture Adoption
- Conclusion

Market Dynamics

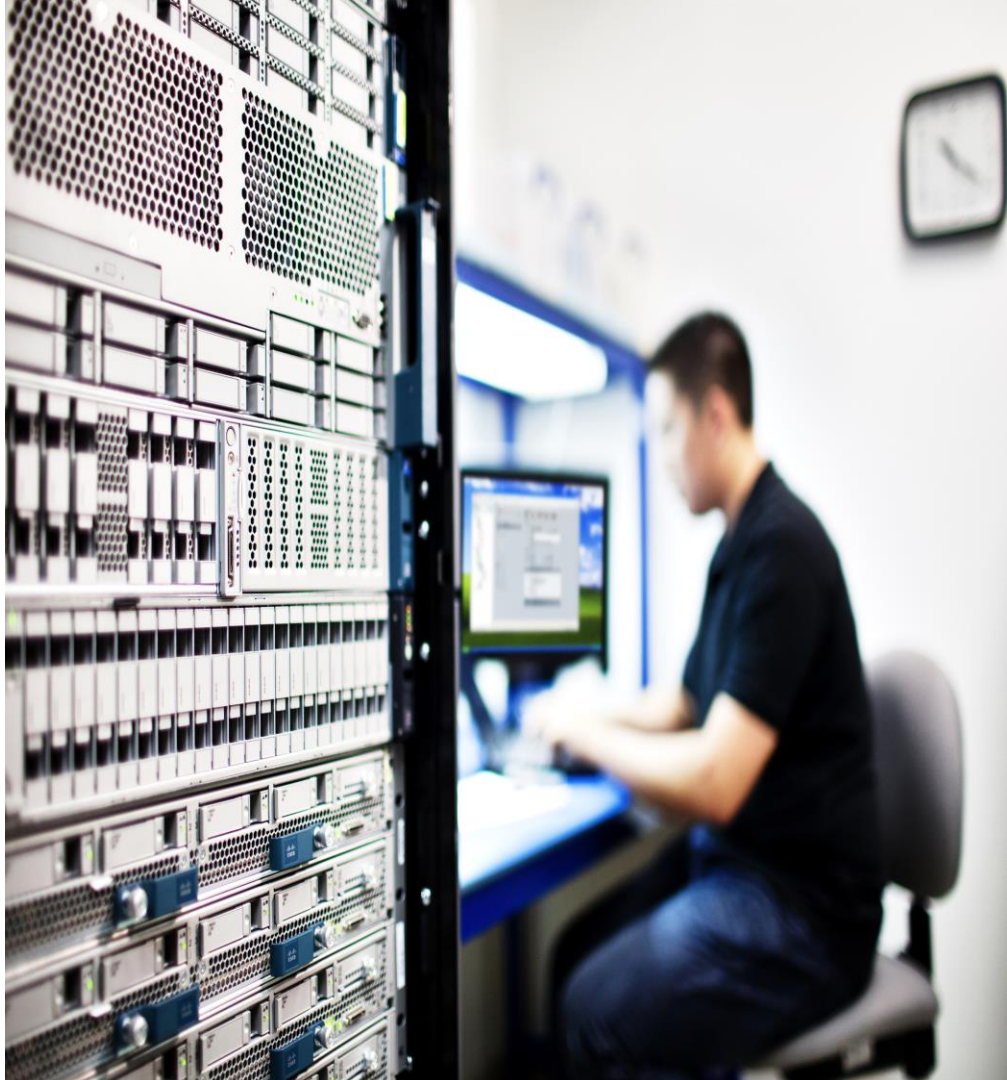
Data Center operator top of mind

Increasing data center capacity and sustainability

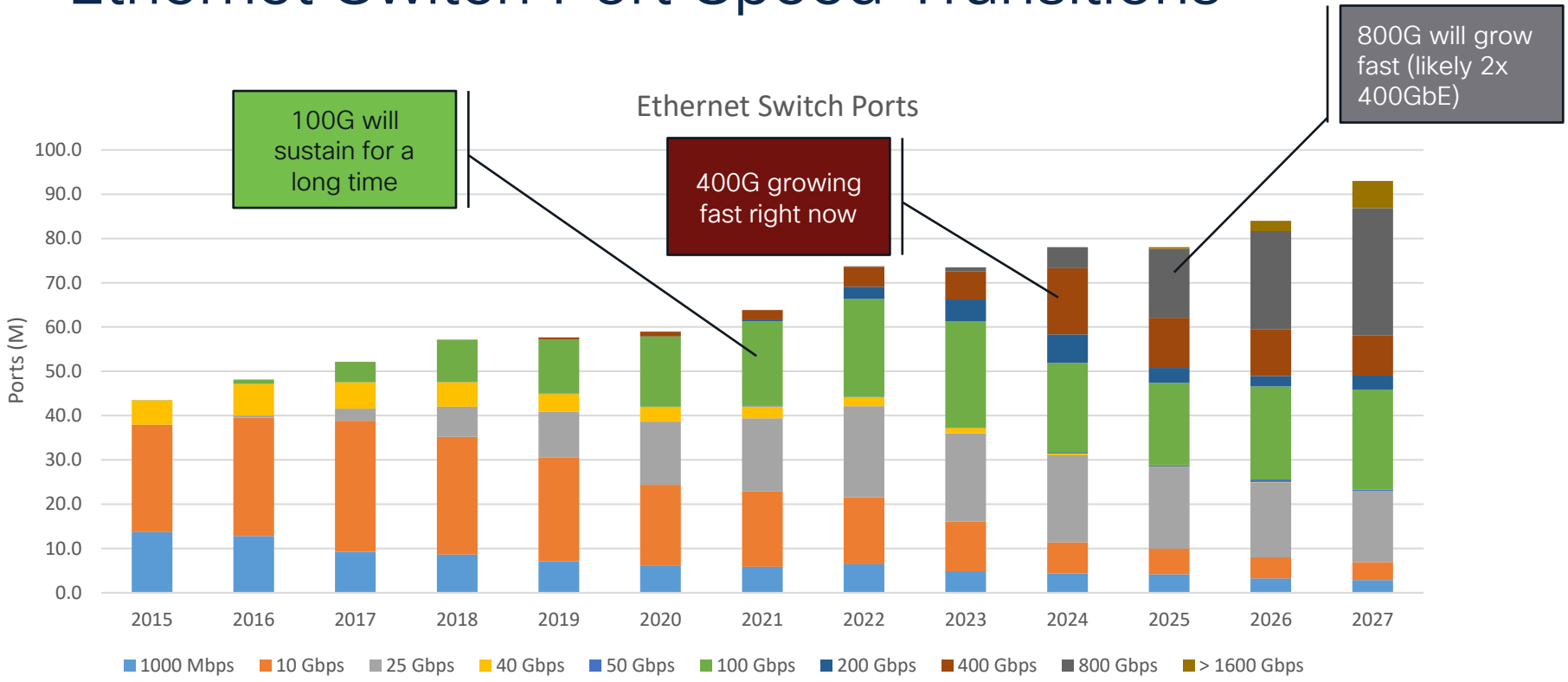
Preserve investments in existing optics infrastructure and cabling

Simplify operations and management of optical links

Preparing for capacity expansion



Ethernet Switch Port Speed Transitions



Source: Dell'Oro's Ethernet Switch – Data Center 5 Year Forecast Report 2023-2027

Market Adoption – Higher Speeds

400G/800G

Hyperscalers

100G/400G/800G fabrics

AI/ML compute clusters

Disaggregation

Webscalers

Scale-out fabrics

25/50/100G server NICs

Vendor NOS supporting open, API-based automation

100G/400G

Enterprise

High performance IO

AI/ML compute clusters

Automation/ Monitoring

Media providers

Fabric for Media (IPFM)

8K uncompressed video driving 100G endpoints

Need for 400G uplinks

Telco service providers

100G/400G fabrics

Space constrained SP DC and edge locations

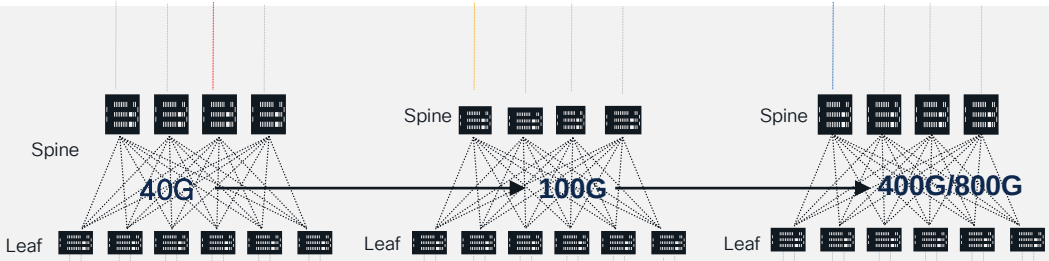
Ready for NFV/5G adoption cycle

Speed evolution in the data center

Inter-Datacenter

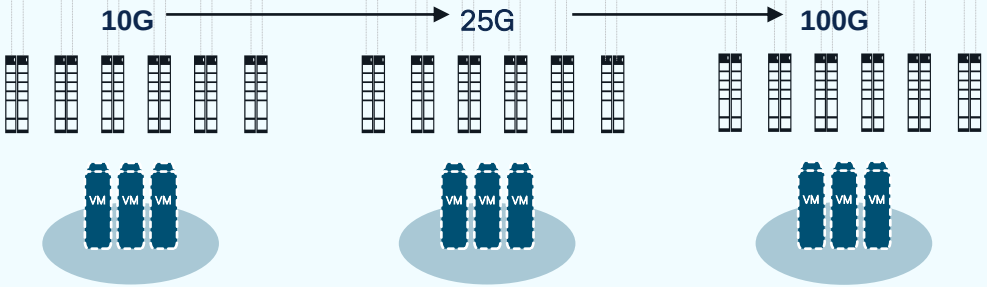
- Transitioning to pluggable DCI (DWDM coherent)
- Open Line System

Switch fabric



- Switch silicon bandwidth growing due to higher Radix and faster Serdes speeds
- Switch ASIC throughput growing: 6.4 Tbps to 12.8 Tbps to 25.6 Tbps to 51.2 Tbps (future)
- Optics increasing from 40Gbps to 100G Gbps to 400Gbps to 800Gbps

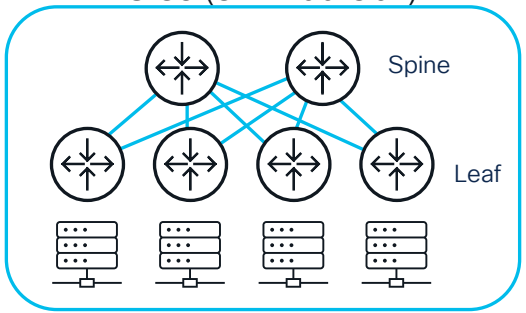
Servers



- Server network connectivity evolves with server processor upgrade cycles as data center traffic grows
- Server port speed is transitioning from 1/10 Gbps to 25 Gbps to 100 Gbps

Why move to higher speeds?
400G → 800G example (same is true for 100G→ 400G)

25.6T user capacity using multiple switches with 12.8T ASICs (32x 400 GbE)



50 Gb/s ASIC IO (SerDes)
32 ports of 400GbE
(128 ports of 100 GbE)

~3000 Watts
26,280 kWh/year

=

25.6T user capacity using single switch with 25.6T ASIC (32x 800 GbE)

100 Gb/s ASIC IO (SerDes)
32 ports of 800G
(64 ports of 400 GbE
256ports of 100 GbE)

~400 Watts
3,504 kWh/year

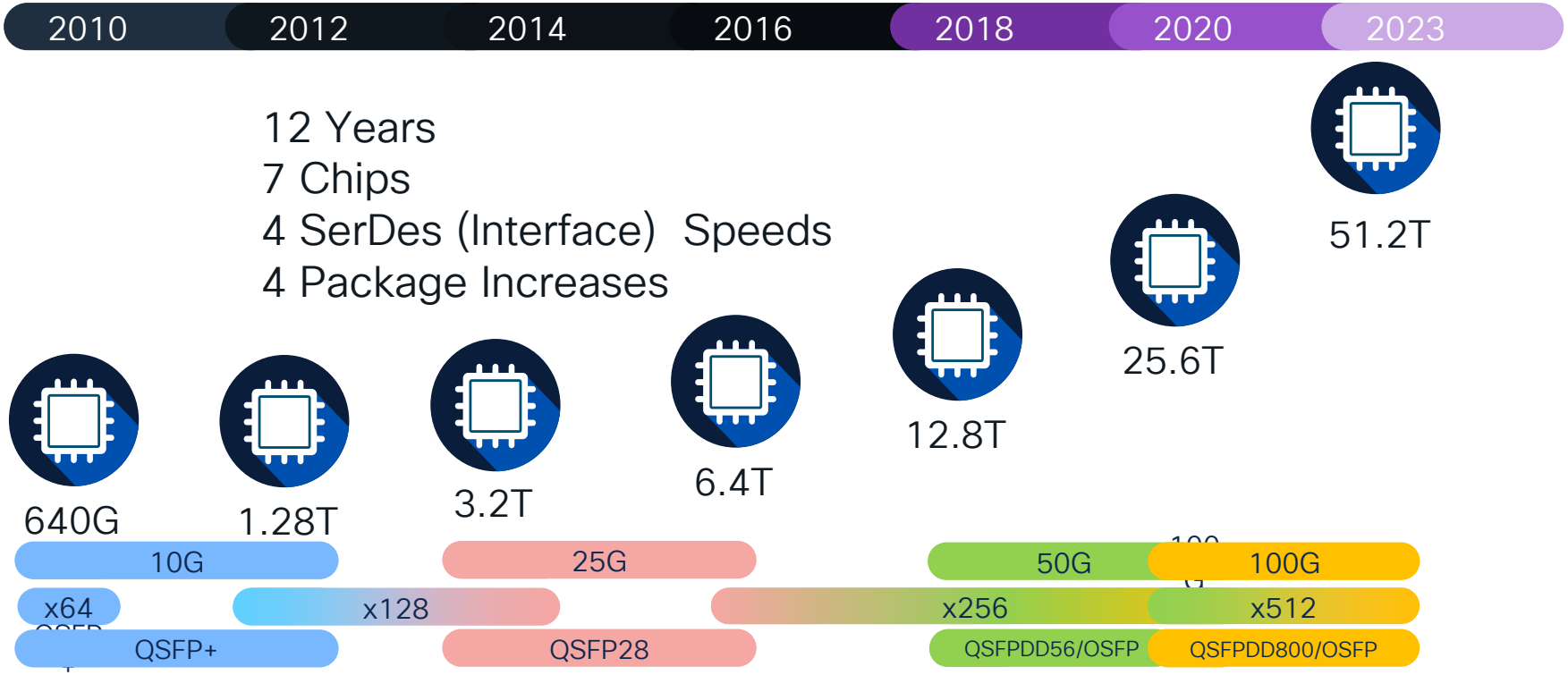
=

Up to 87% Energy Savings

83% less space/fans

Building blocks needed for 400G/800G based Data Center design

ASIC capacity drives importance of optics efficiency



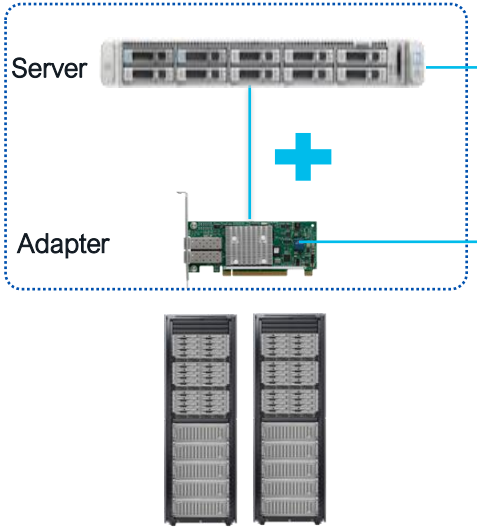
Product and System Flexibility example w/ 25.6T ASIC

Cisco Nexus 9232E Switch

- Compact 1RU 25.6T Switch | 32 800G capable ports
  Up to 64 line rate 400G ports (2x400G breakout)
- 25.6T G100 ASIC (7nm) | 112G SERDES
  108MB fully shared packet buffer
- QSFP-DD800 Ports—backward Compatible
  with QSFP-DD, QSFP28, QSFP+
- Quad Core x86 CPU | 32GB RAM | 128GB SSD
- Cisco NX-OS leaf/spine Capable

Evolution in NIC and server performance

PCIe bandwidth expansion driving higher Ethernet port speeds in the NIC



- Faster CPU
- More and faster memory speeds
- Higher Power

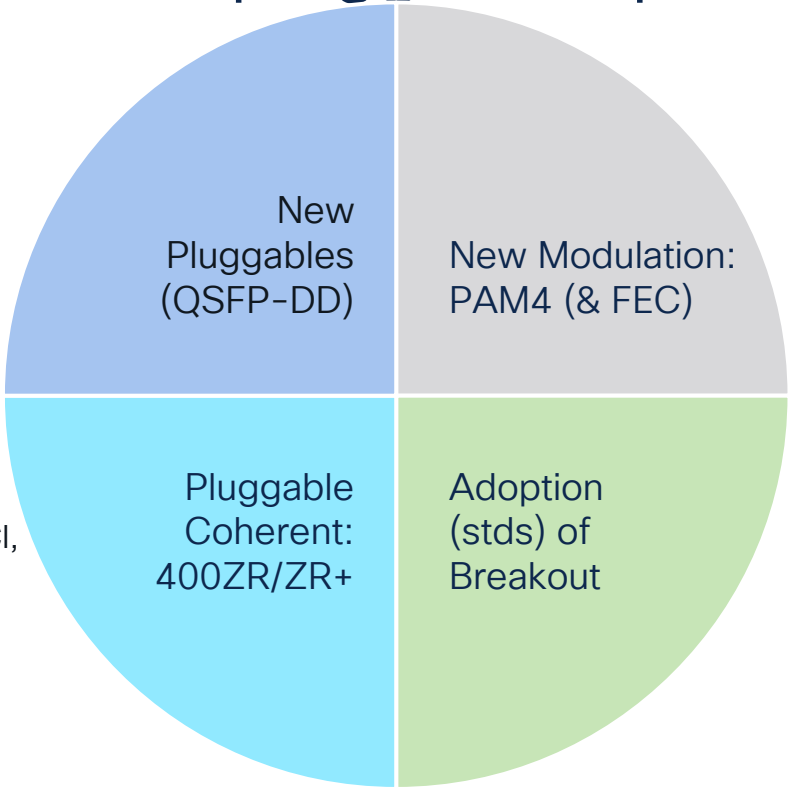
- Increase PCIe Lane performance (GEN 3.0/4.0 and future Gen5.0)
- Increase PCIe Lanes from x8 to x16 and x32
- Increase uplink interface speeds from 10G to 25/100G/200G and 400G
- SmartNIC

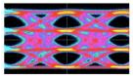
Ports	Modulation	BW per Lane (Single direction)	x8	x16	x32
PCIe Gen2	NRZ	5 Gbps	40 Gbps	80 Gbps	160 Gbps
PCIe Gen3	NRZ	8 Gbps	64 Gbps	128 Gbps	256 Gbps
PCIe Gen4	NRZ	16 Gbps	128 Gbps	256 Gbps	512 Gbps
PCIe Gen5	NRZ	32 Gbps	256 Gbps	512 Gbps	1,024 Gbps
PCIe Gen6	PAM-4	64 Gbps	512 Gbps	1,024Gbps	2,048 Gbps

Innovations in 400G pluggable optics

New pluggable required to support 400G ports (8-wide)
Same faceplate w/2nd row of contacts
Backwards compatibility-
plug QSFP+, QSPF28, QSFP56 into QSPP-DD ports

Higher speed interfaces adopted PAM4 modulation.
Ubiquitous use of FEC.





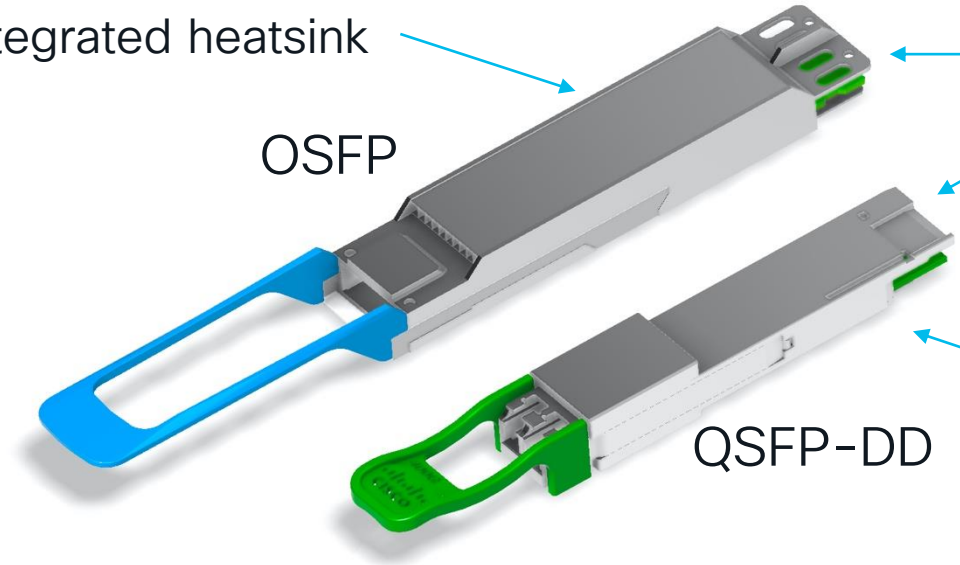
Long reach coherent without system port density reduction DCI, Routed Optical Networking
Thermal efficiency w/ riding heatsink on platform

Pluggable modules supporting multiple lower speed interfaces

400G Optical Modules: QSFP-DD or OSFP

Integrated heatsink

OSFP



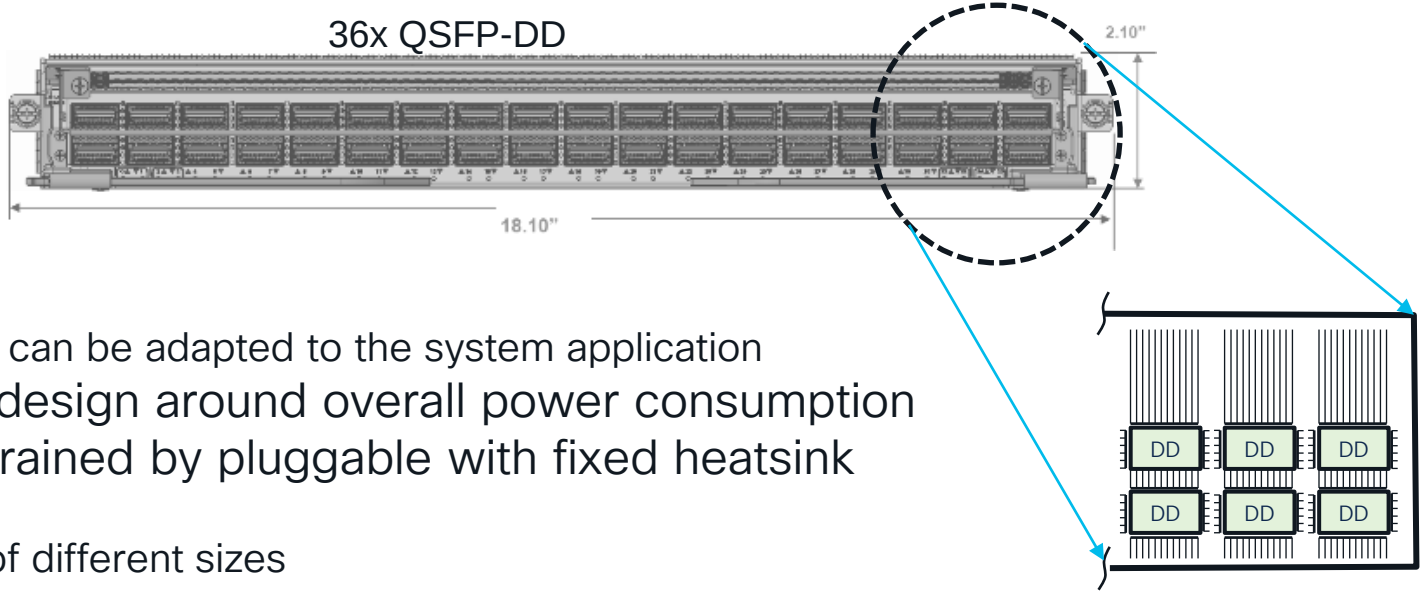
QSFP-DD

8x electrical interface support:

- OSFP – new connector
- QSFP-DD – novel double density approach allowing backwards compatibility with 100G modules (4x interface)

Backwards compatible with 100G QSFP28.
Riding heatsinks inside line card slides on top of the module's flat surface.

Design flexibility with QSFP-DD riding heatsink

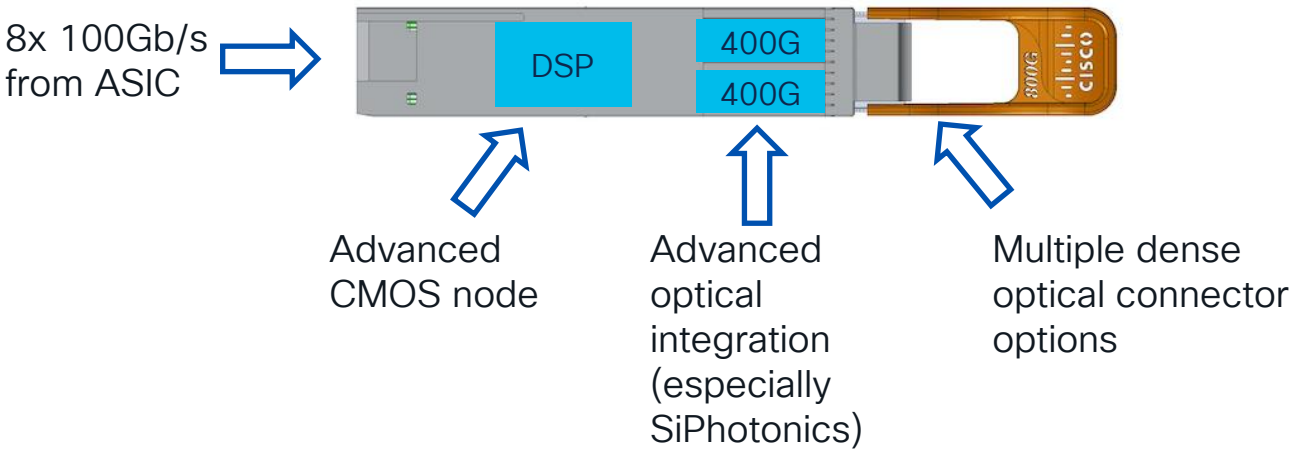


- Thermal design can be adapted to the system application
  - Optimize design around overall power consumption
  - Not constrained by pluggable with fixed heatsink
- Add heatsinks of different sizes
  - Extend heatsinks up into available space
  - High-power row, low-power row

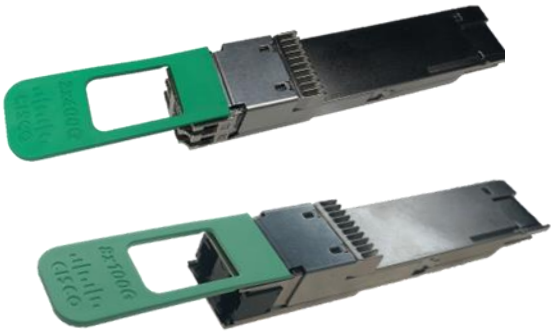
800G supporting dense 400 GbE (aka breakout)

800G form factor enables an economical way to implement 400 GbE

- Maximize the return on investment on the 400 GbE building blocks
- Supports 2x400G / 8x100GE designs



800G Optical Modules: QSFP-DD

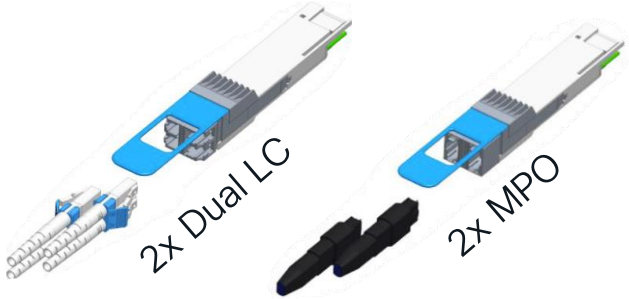


QSFP-DD800

Both variants support all the technical requirements:

- 32 ports in 1 RU
- Electrical signal integrity @ 8x 100 Gb/s
- Thermal cooling capabilities up to 30W

Breakout optical connector options[1]



MPO or LC the connector widely deployed

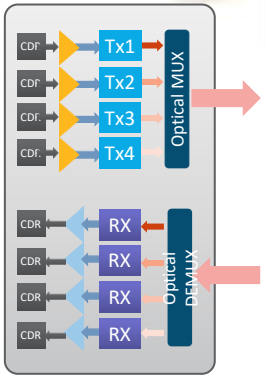
[1] only QSFP-DD shown but similar on OSFP

Single-Wavelength 100G Optics Forward Compatibility

100G Lambda MULTI-SOURCE AGREEMENT

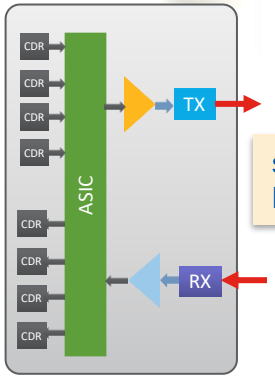
wavelength = lambda = l

Before



QSFP-100G-LR4-S
QSFP-100G-CWDM4-S
QSFP-100G-ER4L-S

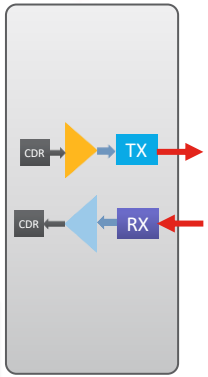
Now



single-lane
PAM4 modulation

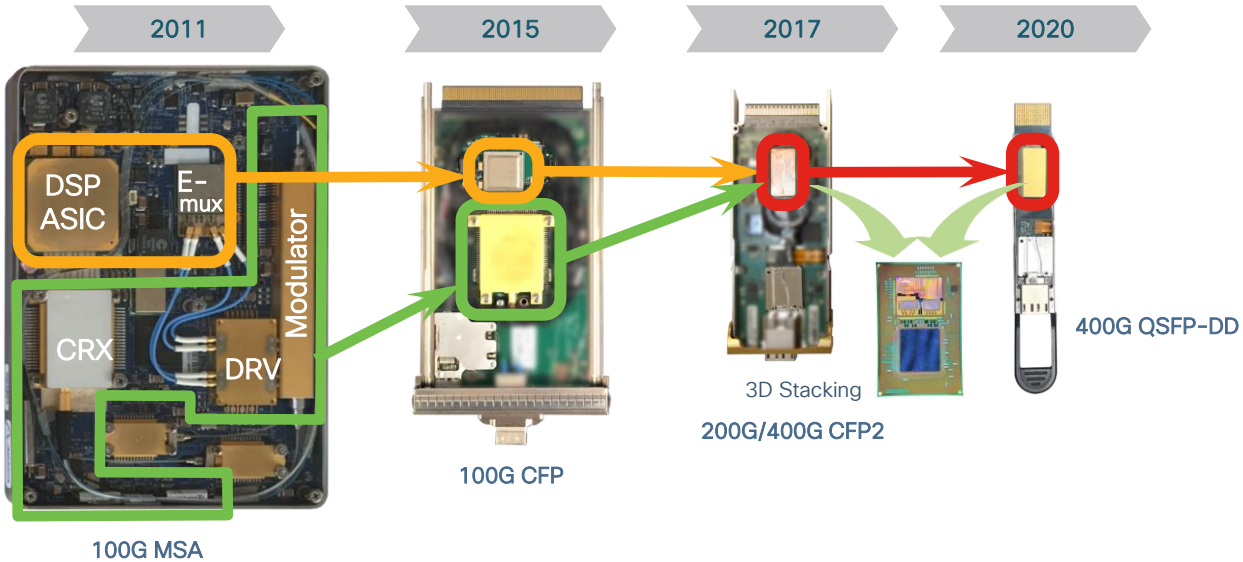
QSFP-100G-DR-S
QSFP-100G-FR-S
QSFP-100G-LR-S
100G ER-Lite

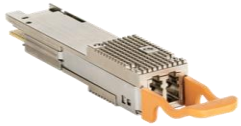
The Goal



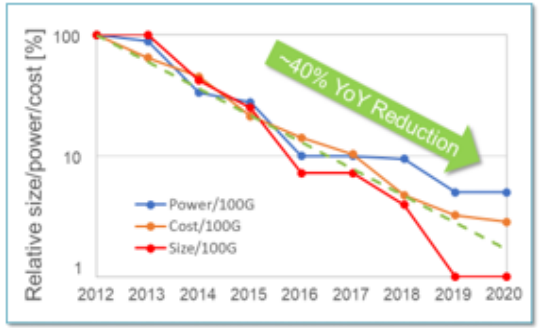
Start with single-lambda today,
stay with single-lambda tomorrow.

400G coherent optics





Reducing Power, Size & Cost through Siliconization



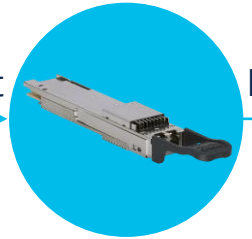
Increased efficiency with siliconization of 400G coherent optics
DWDM interface directly off switch / router
800G and 1.6T coherent will be used for links longer that 2km

Standardization Drives Efficiency

Combines the best of two standardization efforts

OIF 400ZR

Simplified functionality
Edge optimized FEC (C-FEC)

400GE → Client  Line → *400ZR – 120km*

Open ROADM

Flexible Client Mapping +
High Performance FEC (oFEC)

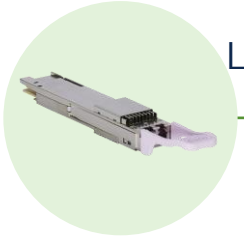
100G
200G
300G
400G
Ethernet
OTN
FlexO

→ Client  Line →

100G
200G
300G
400G
OpenROADM

OpenZR+ MULTI-SOURCE AGREEMENT

100GE
200GE
400GE
Nx 100GE
(Muxponder Mode)

→ Client  Line →

100G
200G
300G
400G

Simplified functionality +
High performance FEC (oFEC)

High performance pluggable modules
Multi-vendor interoperability

Cisco Blog: OpenZR+ Delivers 400G Coherent Optics for Multi-Haul Networks

Cisco and 400 GbE Industry Activities

✓ Complete

Cisco-led

Standards	IEEE 802.3bs ✓ Cisco IEEE 802.3cd ✓ Cisco IEEE 802.3cm ✓ IEEE 802.3cn ✓ IEEE 802.3ct ✓ IEEE 802.3cu ✓ Cisco IEEE 802.3ck ✓ Cisco IEEE 802.3db ✓	400 GbE & 200 GbE MAC & Initial Interfaces 50 GbE MAC & Interfaces (also 100 GbE & 200 GbE PMDs) 400 GbE MMF (BiDi and SR8) Extended reach (40km) 50 GbE, 200 GbE, 400 GbE 100GbE Coherent 80km 100G-FR, 100G-LR, 400G-FR4, 400G-LR4-6 100GE serdes 100/200/400GE MMF (100Gb/s short wavelength)
	OIF400ZR/802.3cw Cisco	400 GbE Coherent 120km / 400 GbE Coherent 80km
	802.3df Cisco	200G/400G/, 800G Ethernet Task Force @ 100Gb/s per lane
	802.3dj Cisco	200G/400G/800G/1.6T Ethernet Task Force @ 200Gb/s per lane
	802.3dk	Greater than 50 Gb/s Bidirectional Optical Access PHYs Task Force.
MSAs*	100G Lambda MSA	100G-FR, 100G-LR, 400G-FR4, 400G-LR4
	QSFP-DD MSA Cisco	400G Form factor
	OSFP MSA Cisco	400G/800G/1.6T Form factor
	SFP-DD MSA	100G Form factor
	DSFP MSA	Alternative 100G Form Factor (Mobile)
	400G-BiDi MSA Cisco	400 GbE MMF BiDi
	QSFP-DD800/1600 MSA Cisco	800G / 1.6T Form Factor

* Multi-Source Agreements – new ones all the time. Not all get wide industry adoption

400/800 Building Block summary

- ASIC Capacity drives systems systems design – Fixed, modular
- 400G pluggable technology is mature and evolving to higher speeds.
- Cisco leading industry standards and MSA solutions
- 400G pluggable brought a lot of innovation that will be extended into next gen
  - New QSFP-DD form factor(s) capable of supporting high density at all reaches. Backwards compatible with 100G QSFP28
  - High-speed PAM4 optics. Higher integration, lower cost
  - Coherent pluggable: 400ZR and 400ZR+ Enables DCI, Routed Optical Network architectures
  - Mainstream adoption of breakout
  - 400G module as 4x100G (SMF) or 8x 50G (MMF). DAC too.
- Support for MM and SM fiber

Network Architecture Considerations

Strategic decisions for efficient data center architectures

How do I optimize data center design flexibility to support 400G?

- Reach, operational flexibility, manageability

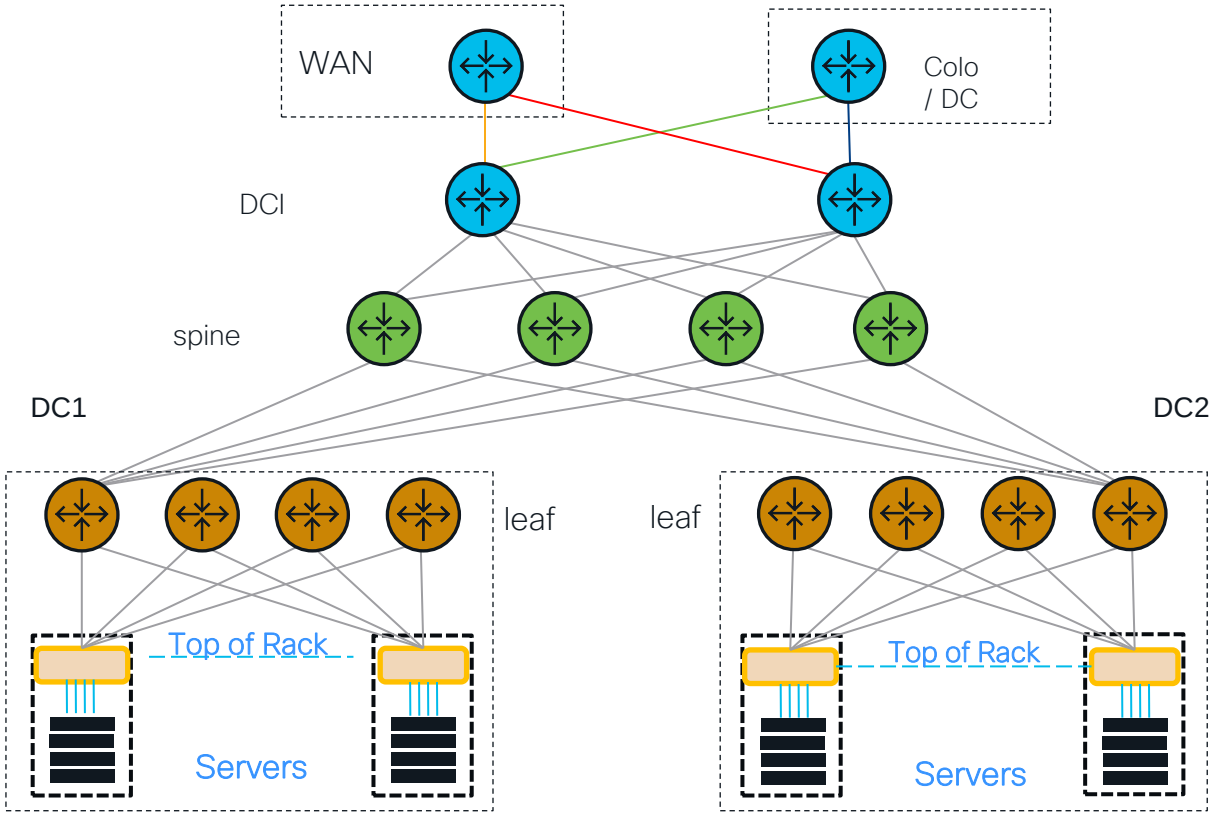
How do I migrate from existing platforms and links to 400G?

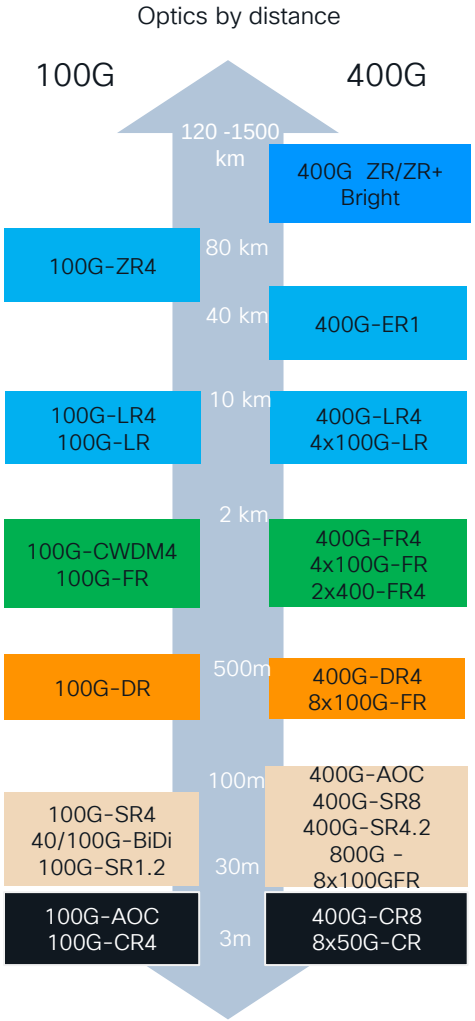
- Operationalizing breakout

How do I ensure network reliability?

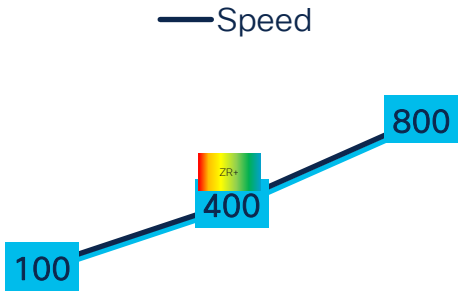
- Component integration improves reliability even at higher speeds

400G optics cover the entire network

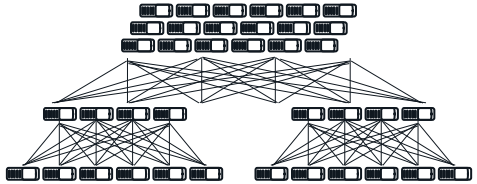




400G DC Fabric



High Scale Leaf/Spine based Designs



Common network architecture between 100G, 400G and 800G

- Same physical port densities, same media reaches
- Continued investment in fiber plant – SM & MM fiber

Flexibility to adopt 400G breakout for high radix 100 GbE design

- Connectivity to 100 GbE equipment

Design flexibility (Switch Platform)

- High bandwidth, high port density platform flexibility w/ fixed, modular
- Link bandwidth distribution
- Port flexibility – non-coherent / coherent use cases, and mixed data rates
- Cabling flexibility

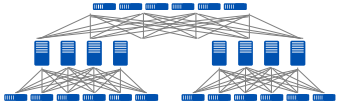
Backward and forward compatibility (QSFP, QD-DD)

- 40G, 100G, 400G, 800G

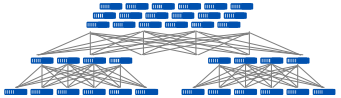
Operational and design flexibility with 400G

2-Tier Leaf/Spine

3-Tier Leaf/Spine

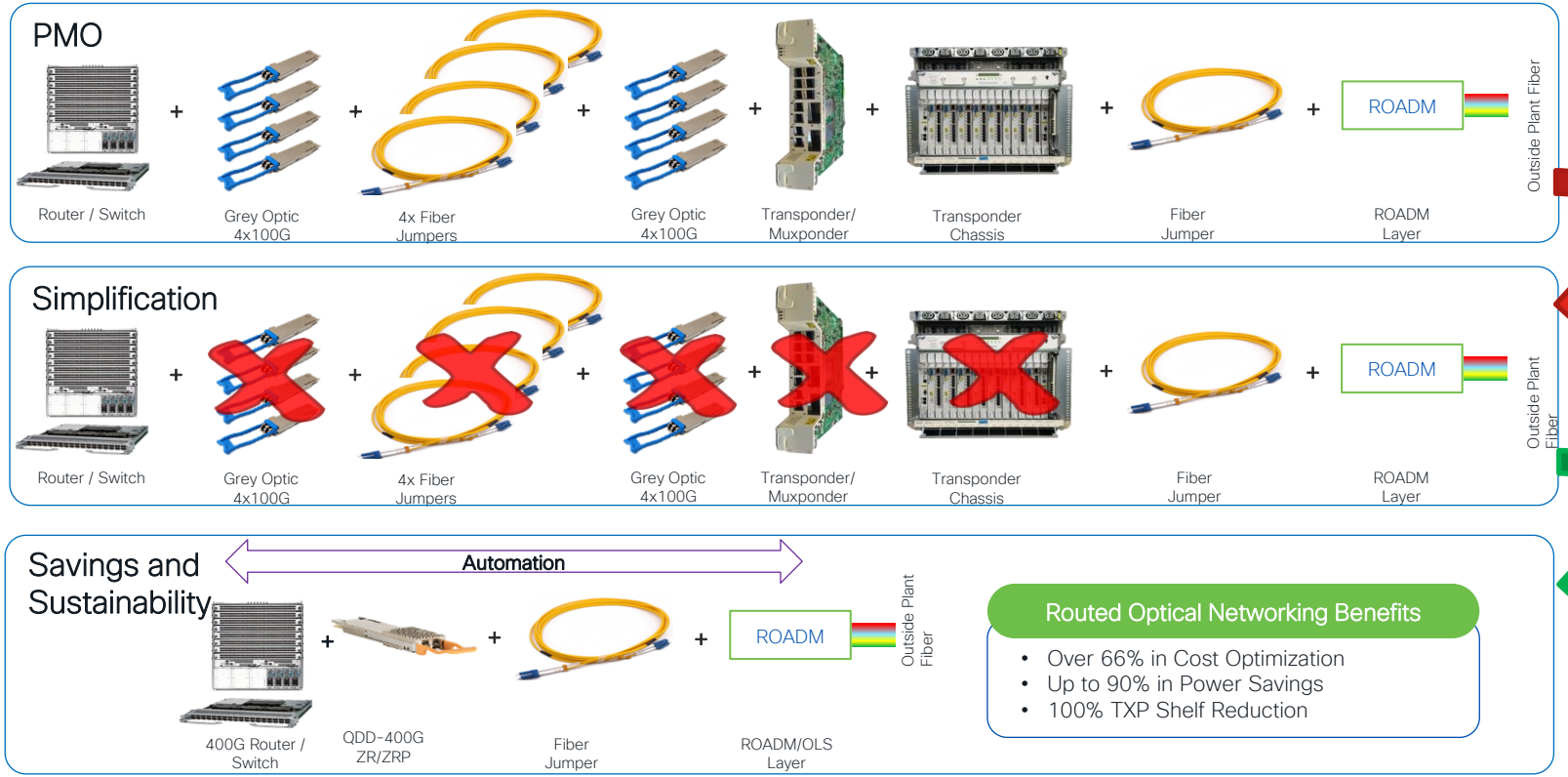


Evolved 3-Tier Leaf/Spine

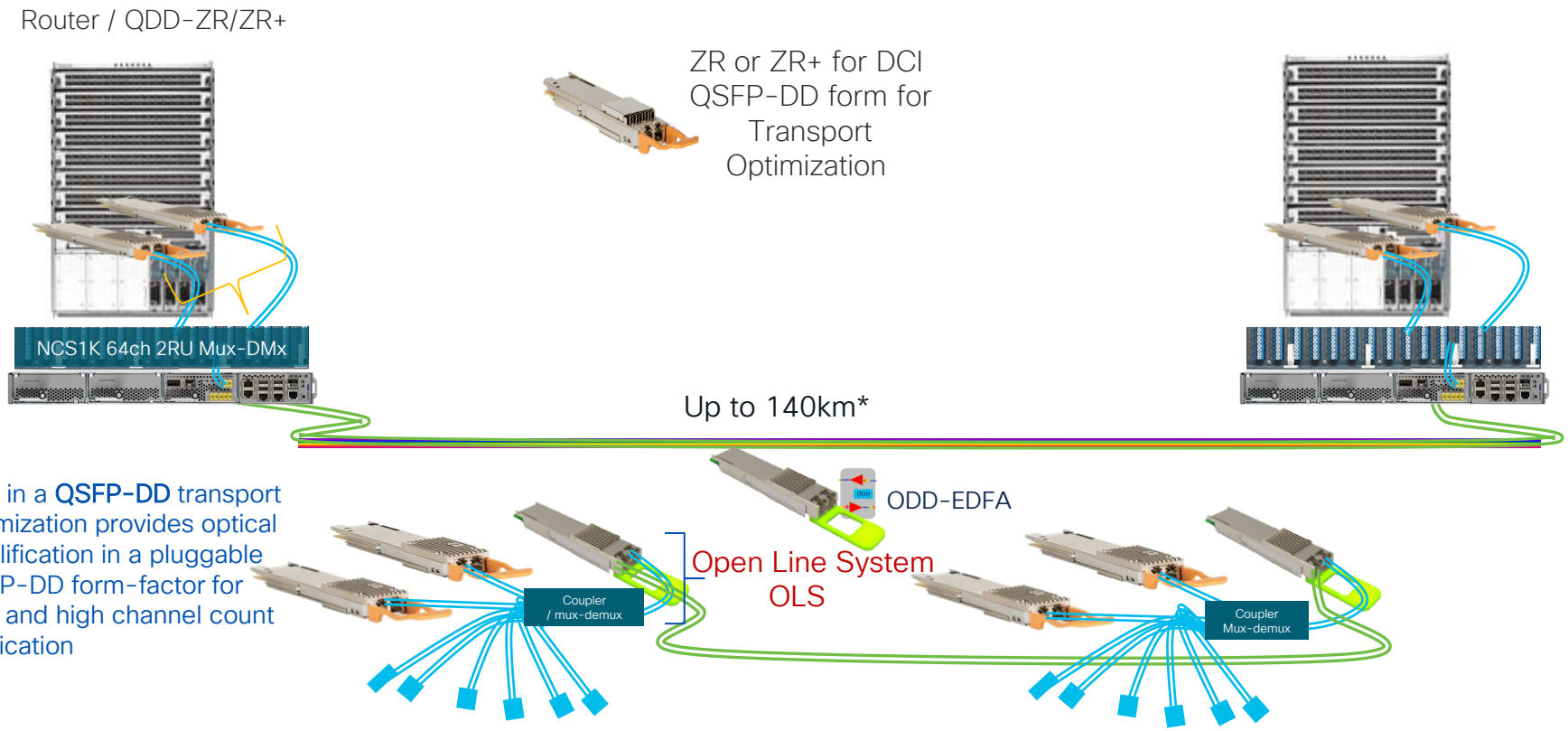


- Increasing scale-out in all tiers
- Improve system capacity with dense 400G/800G platforms
- Cost optimization with lower cost/bit and improved power efficiency
- Design flexibility w/ fix, modular platform – scale, hops, latency
- Latency optimization with single SoC switch for network designs
- Improved application performance – high bandwidth 400G fabric
  - Improved ECMP performance – bigger flows, larger flow buckets
  - Intelligent buffering
- Breakout for leaf spine and server access – improved design flexibility

DCI – Simplification, Savings and Sustainability



400ZR enables simplified DCI



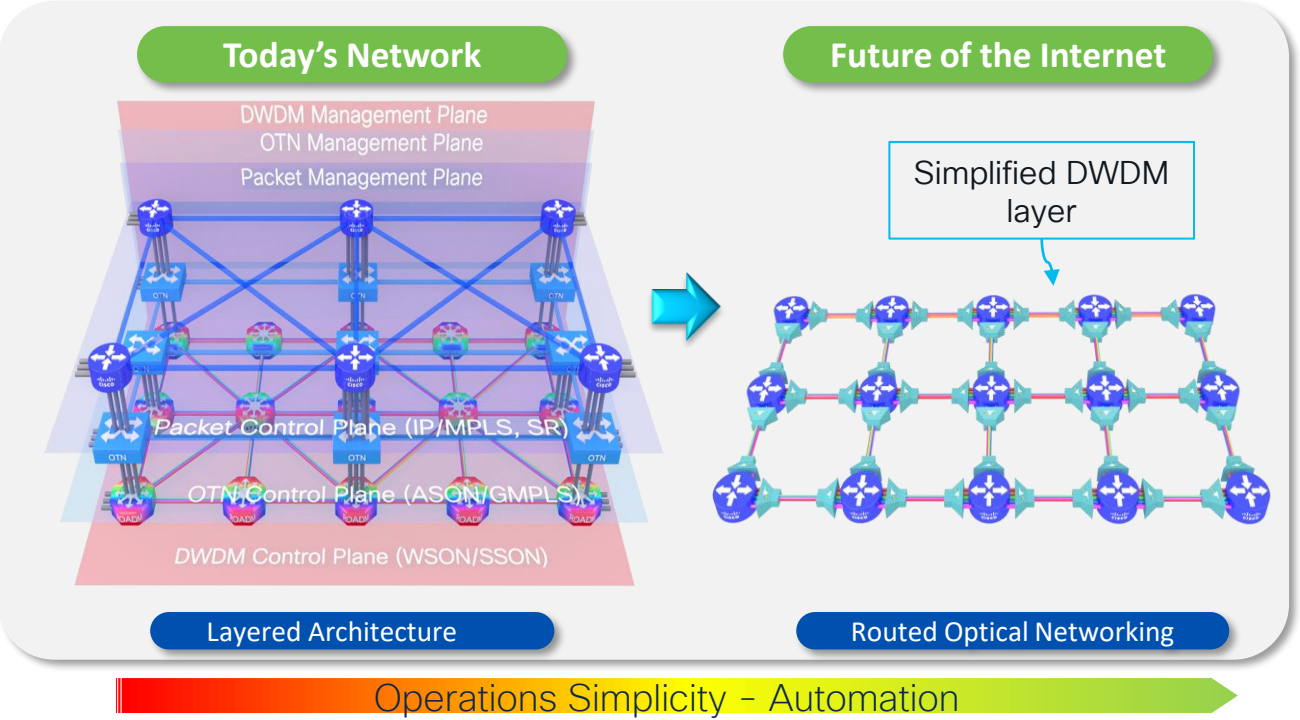
OLS in a **QSFP-DD** transport optimization provides optical amplification in a pluggable QSFP-DD form-factor for Low and high channel count application

400G Coherent pluggable enables Routed Optical Network

DWDM interfaces directly off switch/router with no loss of density

Flattened network architecture

Significantly lower TCO

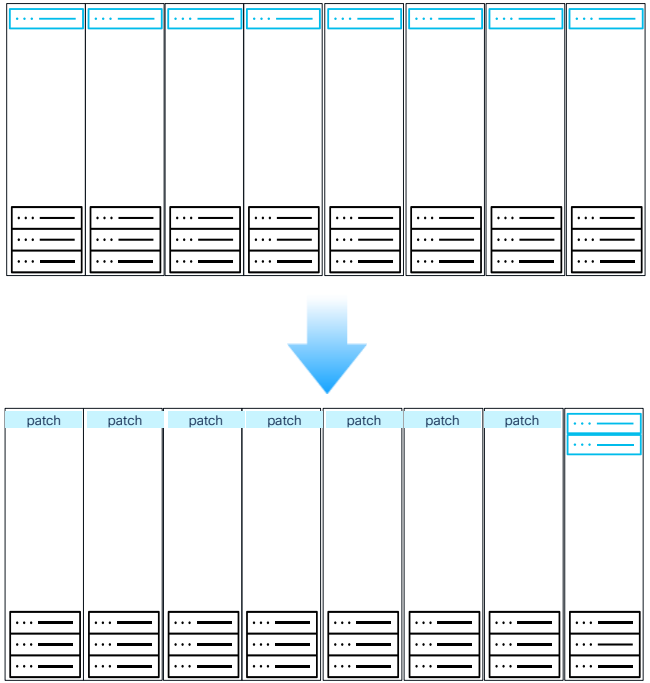


Network Architecture Adoption

space, power, cabling

Improve power efficiency with 400G

DC Server Rack Architecture w/400G: Considerations



Current Architecture

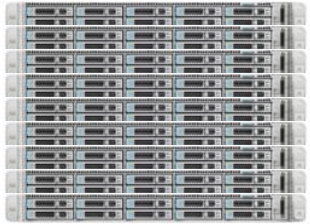
- One ToR Switch (1RU) per Cabinet
- Provides connectivity to 16-32 servers
- One port per server

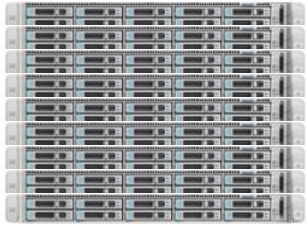
New Architecture

- One or two ToR Switches (2RU) per (8) Cabinets
- Provides connectivity to 128-256 servers
- One port can service 4 or 8 servers with 4 to1 (SR4.2) or 8 to1 (SR8) breakouts

Adopt 400g and save 15% of power saving

400GE Economics – Transceiver Power

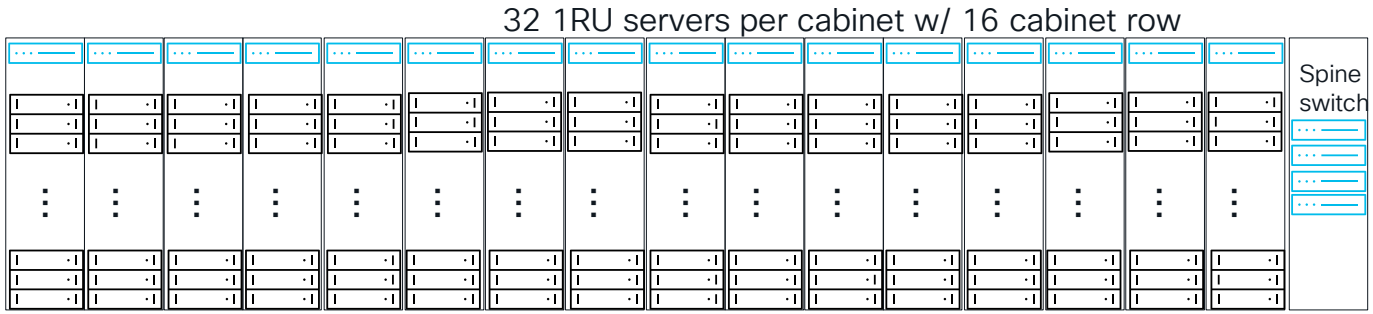


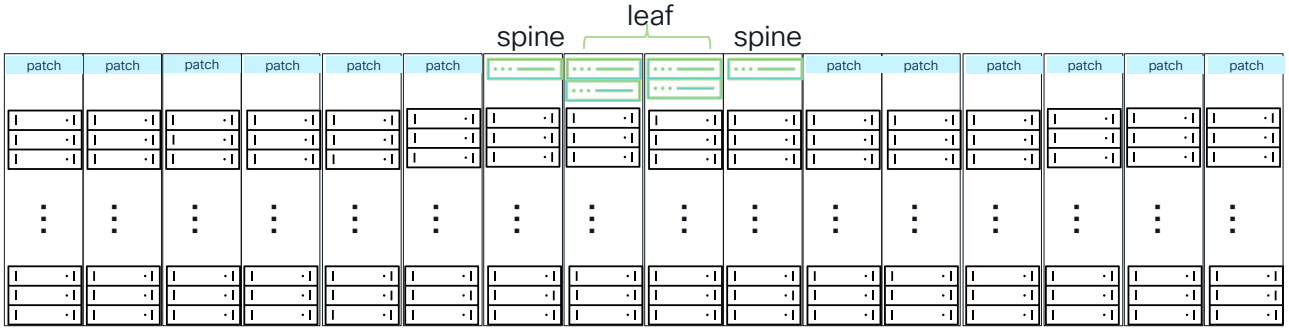


Breaking out 400G to 4x100G modules is a 15% power savings in comparison to equivalent capacity to 8x100G modules

400G: Space, cabling and devices consideration

Example of DC Server Rack Architecture





Current Architecture

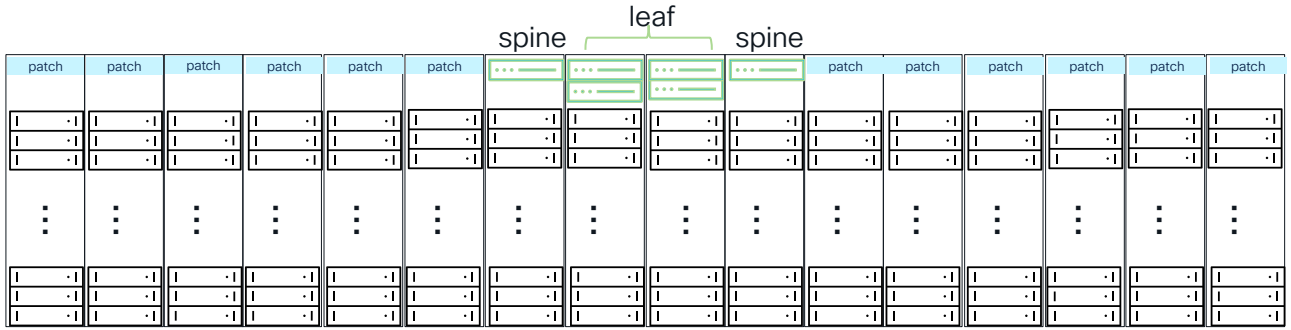
- One 100G Nexus 9364C-GX ToR Switch (2RU) per Cabinet
- One port per server
- Provides connectivity to 32 servers
- 32 downlinks 16 uplinks 2:1 oversubscription

New Architecture

- Two ToR Nexus 9364D-GX2A Switches (2RU) per (8) Cabinets
- One port can service 4 servers with 4 to 1 (SR4.2) breakouts
- Provides connectivity to 256 servers
- 32 ports downlinks 16 uplinks 2:1 oversubscription
- Spine switch in server rack – reduce need for extra switch rack

400G: Space, cabling and devices consideration

Total power savings with switch reduction



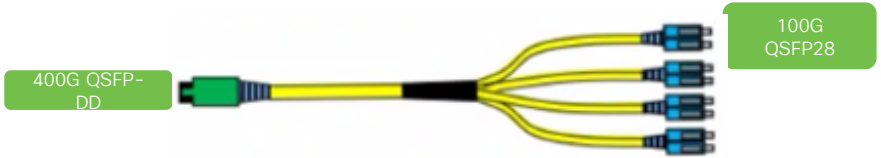
Server Switch reductions - 16 server cabinet row

- (16) ToR switches to (4) ToR switches per server cabinet row
- Typical 2RU 100G switch (Nexus 9364C-GX) 811W x 16 = 13 kW
- Typical 2RU 400G switch (9364D-GX2A) 1324W x 4 = 5.3 kW
- 7.7kW power savings on switch reductions per 16 cabinet row

400G: Space, cabling and devices consideration

Total power savings with transceiver reductions

- Current Architecture
  - 4.3 W 100G transceivers : (512) = 2.2kW
  - Total Power Used = 2.2kW



- New Architecture
  - 12 W 400G transceivers : (64) in breakout = 768W
  - 4.3 W 100G transceivers : (256) in breakout = 1.1kW
  - Total Power used = 1.87kW
- Current Architecture 2.2kW – New Architecture 1.87kW = Transceiver Total Power Savings 330W per 16 cabinet row 15% power savings
- Total power savings for New Architecture with Switch and Transceivers Reductions = **8kW per 16 cabinet row** **47% reduction in switch power load!**

400G: Space, cabling and devices consideration

Total power savings with switches and transceivers vs. DAC Cables

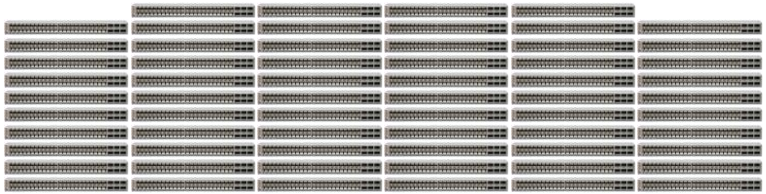
- Current Architecture with DAC Cables
  - .5 W 100G DAC Cables – (512) = 256W
  - Total Power Used = 256kW
- New Architecture
  - 12 W 400G transceivers – (64) in breakout = 768W
  - 4.3 W 100G transceivers – (256) in breakout = 1.1kW
  - Total Power used = 1.87kW
- Current Architecture 256kW – New Architecture 1.87kW = Transceiver Total Power Savings (1.61kW)
- Total power savings for New Architecture with Switch and Transceivers Reductions = 7.7kW – 1.61kW = **6.1kW per 16 cabinet row** **40% reduction in switch power load!**

400G: Space, cabling and devices consideration

Sustainability Use Case: AI/ML 400G

9504

9364D-GX2A



93180YC-FX3

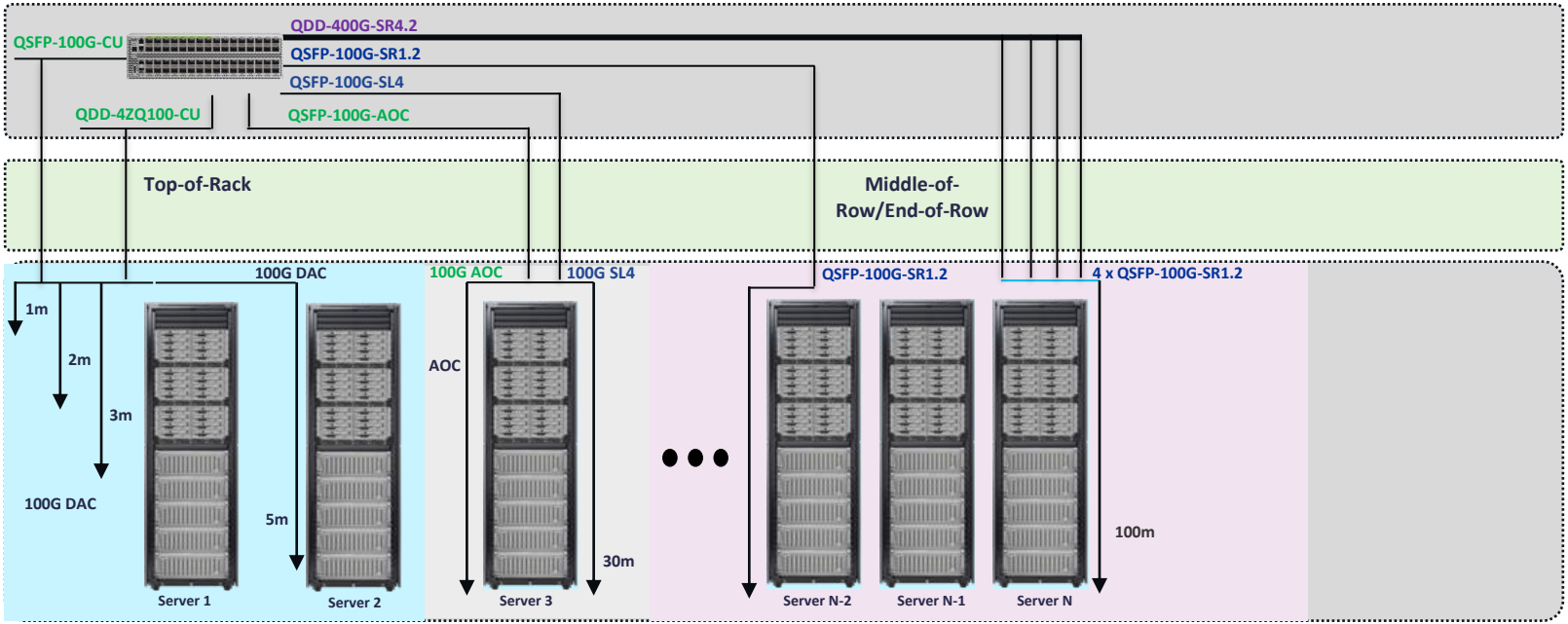
9364-GX2A

- 3072 x 25G Server Ports
- 64 x 9348YC-FX3 Leaf Switches
- 3:1 Oversubscription (25.6Tbps)
- 29.8 kW System Power
- 1.2 Watts/Gbps
- 78 RU

- 3072 x 100G Access Layer Ports
- 16 x 9364D-GX2A
- 3:1 Oversubscription (102.4Tbps)
- 26.4 kW System Power
- .26 Watts/Gbps
- 40 RU

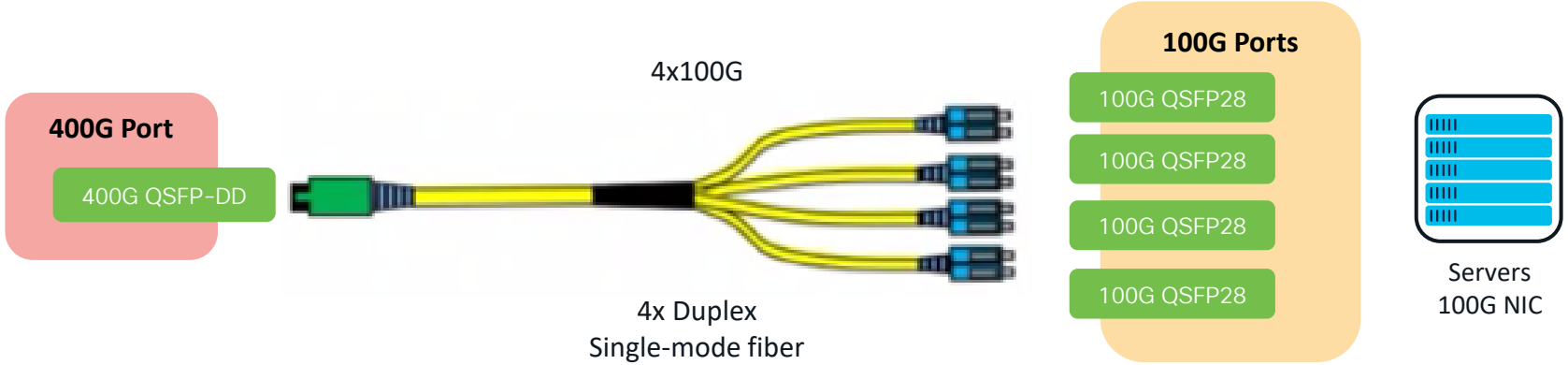
Sever Rack Transceiver Connectivity Options

High Performance 100G Server Connectivity Options

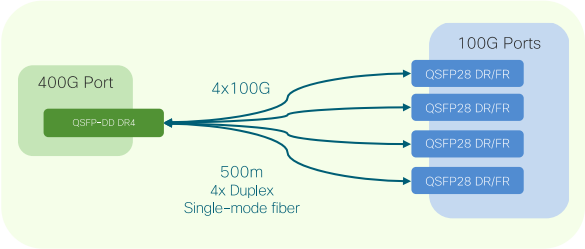


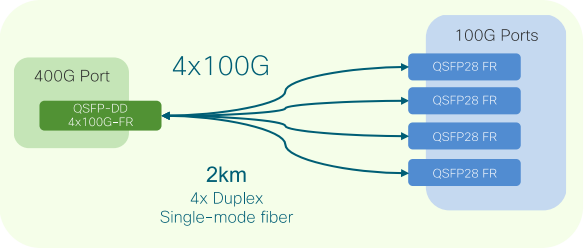
Breakout: 400G to 100G connectivity

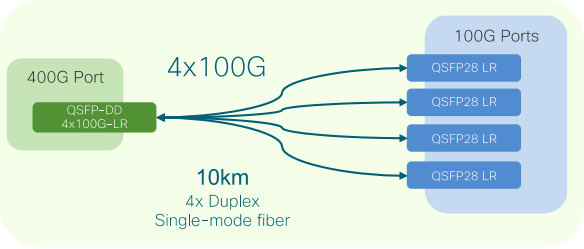
Maximize port efficiency + forward compatibility with 100G single lambda



Breakout Options



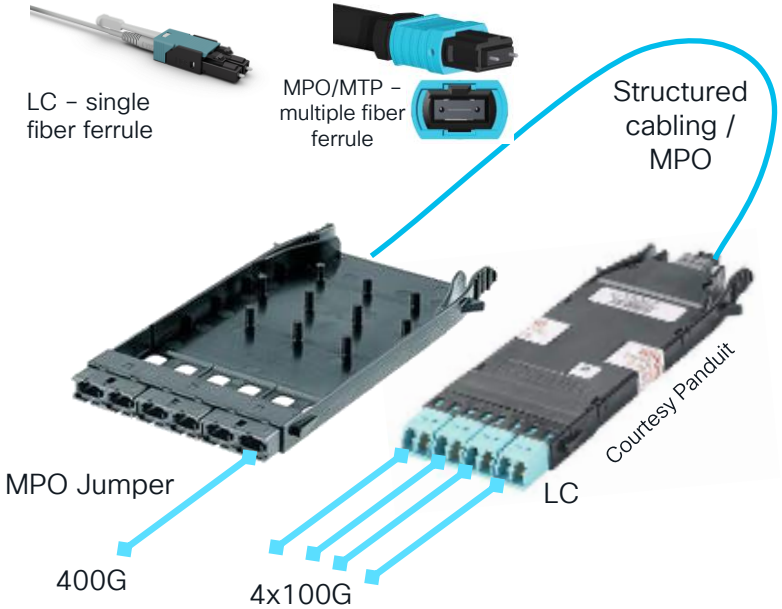




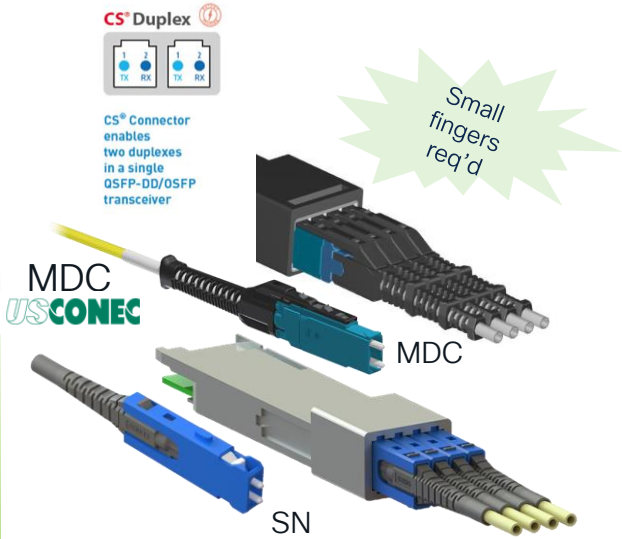
Optical Connector Considerations

Multiple options exist

Breakout cables

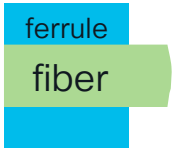


Using structured cabling and breakout cassettes

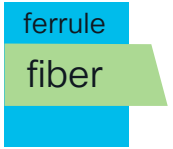


New dense VSFF connectors in module nose

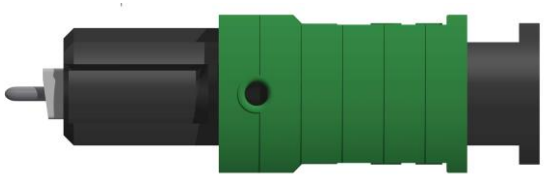
Deployment considerations:
Multi-fiber (MPO) connectors: Angled (APC) vs flat polish (UPC)



Ultra Polish (UPC)
- all single SMF/MMF fiber connectors (LC)
- Vast majority of MMF MPO

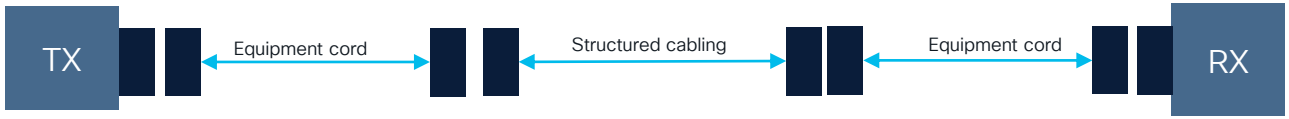


Angled Polish (UPC)
- All SMF MPO
- Some recent introduction for MPO MMF

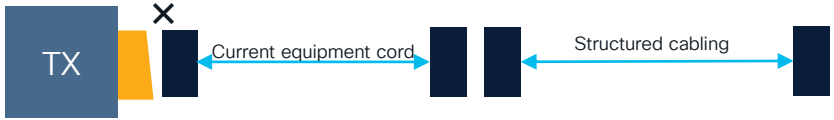


Some recent 400G MMF specs defined use of MPO APC. Awareness will prevent deployment issues

APC deployment usage



Current Structure



Network installer doesn’t realize to change equipment cord

- Additional cost of replacing equipment cords
- Confusion around needing specific equipment cords for specific PMDs
- Incorrect mismatch of PC and APC results in out of spec fiber plant (air gap). Unclear if damage risk exists or not
- Risk of large product returns expected to fiber installer or module manufacturer

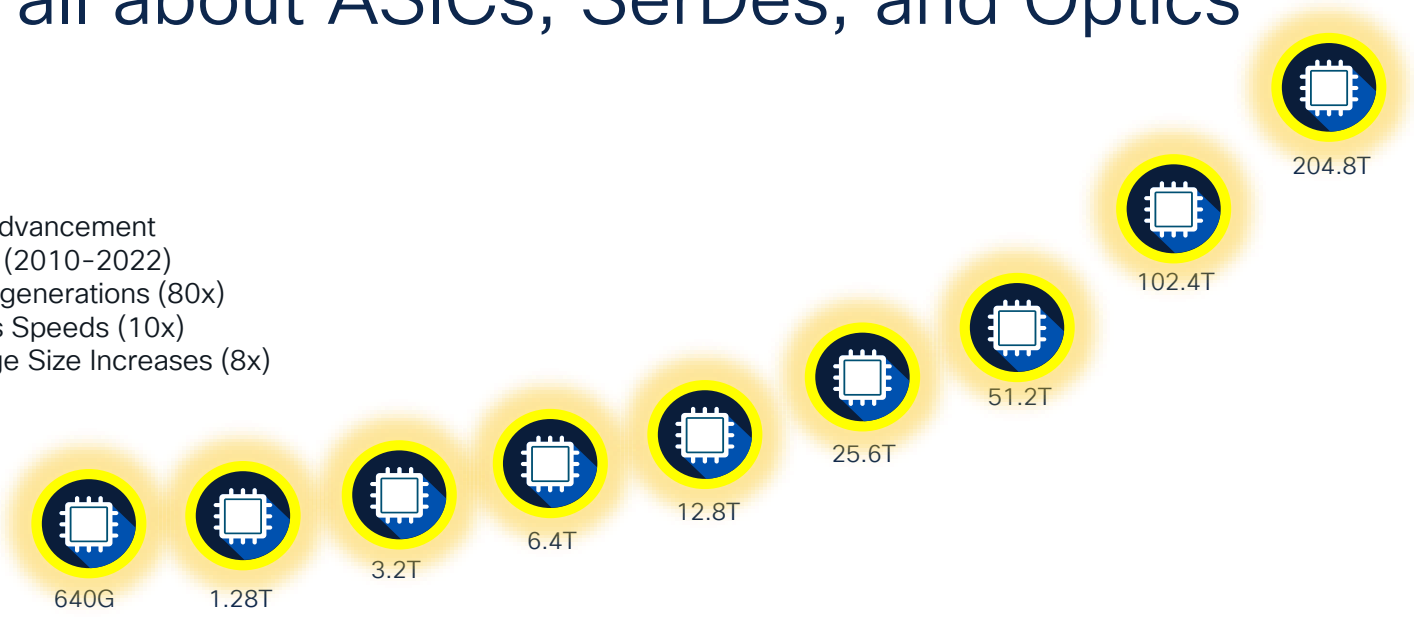
Summary

- 400G/800G switch deployment will impact, space, power and provide efficiency gains –port, network design, sustainability
- High 400G deployment require facilities architecture consideration– cabling, switch placement, use of breakout
- Brownfield deployment – opportunity to use existing connector and structured cabling
- Sustainability improvements

Conclusion

It’s all about ASICs, SerDes, and Optics

Relentless Advancement
- 12 years (2010-2022)
- 7 switch generations (80x)
- 4 SerDes Speeds (10x)
- 4 Package Size Increases (8x)



← 204.8T is 16x the bandwidth of 12.8T

160x increase in switching bandwidth over 14 years

Switching Speeds	640G	1.28T	3.2T	6.4T	12.8T	25.6T	51.2T	102.4T	204.8T	
Year	2010	2012	2014	2016	2018	2020	2022	~2024 ?	~2026 ?	
Serdes GHz	10G		28G		56G		112G			2x in 4 years
SerDes Lanes	64	128		256		512			1024	2x in 4 years
Optics Pkg	QSFP+		QSFP28		QSFP-DD56		QSFP-DD800	QSFP-DD1600		4x in 4 years

ASIC density continues to redefine how products are built.
Gates & GHz. SerDes & Interconnect. Optics & wavelengths.

Credit: adapted from https://blogs.cisco.com/sp/co-packaged-optics-and-an-open-ecosystem

Path ahead

Pluggable optics roadmap continues and extends beyond 800G

QSFP-DD

All QSFP

400G (8x 50G) | 800G (8x 100G) | 1.6T (8x 200G)

OSFP

400G (8x 50G) | 800G (8x 100G) | 1.6T (8x 200G)

Differences exist in system design and support of large current installed base

XD

1.6T (16x 100G)

No compatibility between OSFP & OSFP-XD (despite initial intent)
Market adoption unclear

8x

16x

Host side interface width

Market demand (time)

Summary

- Transition to 400G/800G is well underway, can provide tremendous benefits
  - higher bandwidth and performance
- Optical breakout improves switch port efficiency and power efficiency
- Coherent optics enable cost effective DCI, Routed Optical Networking, maximize switch port efficiency – reduction in footprint of system
- Increase capacity with less footprint
  - Efficient system, architecture and reuse of pluggable modules
  - Sustainability improvements

Fill out your session surveys!

Attendees who fill out a minimum of four session surveys and the overall event survey will get **Cisco Live-branded socks** (while supplies last)!

Attendees will also earn 100 points in the **Cisco Live Challenge** for every survey completed.

These points help you get on the leaderboard and increase your chances of winning daily and grand prizes

Continue your education

- Visit the Cisco Showcase for related demos
- Book your one-on-one Meet the Engineer meeting
- Attend the interactive education with DevNet, Capture the Flag, and Walk-in Labs
- Visit the On-Demand Library for more sessions at www.CiscoLive.com/on-demand

Thank you

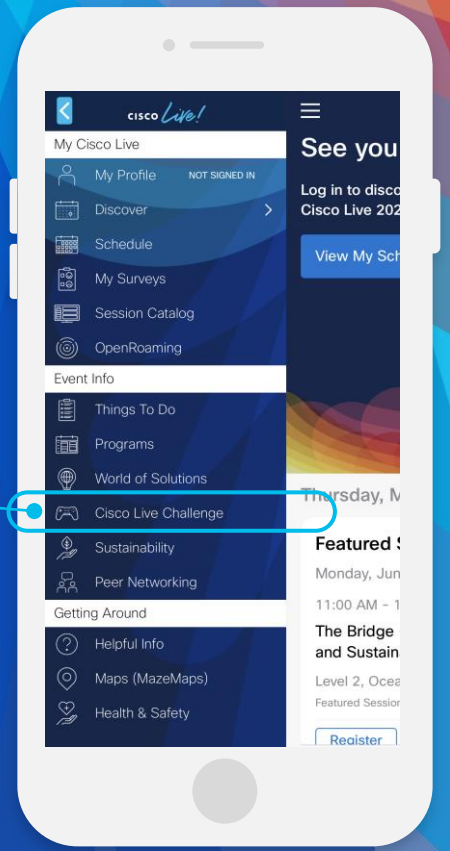
Cisco Live **Challenge**

Gamify your Cisco Live experience!

Get points for attending this session!

How:

1. Open the Cisco Events App.
2. Click on 'Cisco Live Challenge' in the side menu.
3. Click on View Your Badges at the top.
4. Click the + at the bottom of the screen and scan the QR code:



cisco Live!

Let's go