400GE technológia, új protokollok, hardverek és technológiai irányok

Nagy Tibor

Cisco Systems
TSA, CCIE#8982
tinagy@cisco.com

Special thanks to Errol Roberts and Mark Nowell

Ethernet Evolution

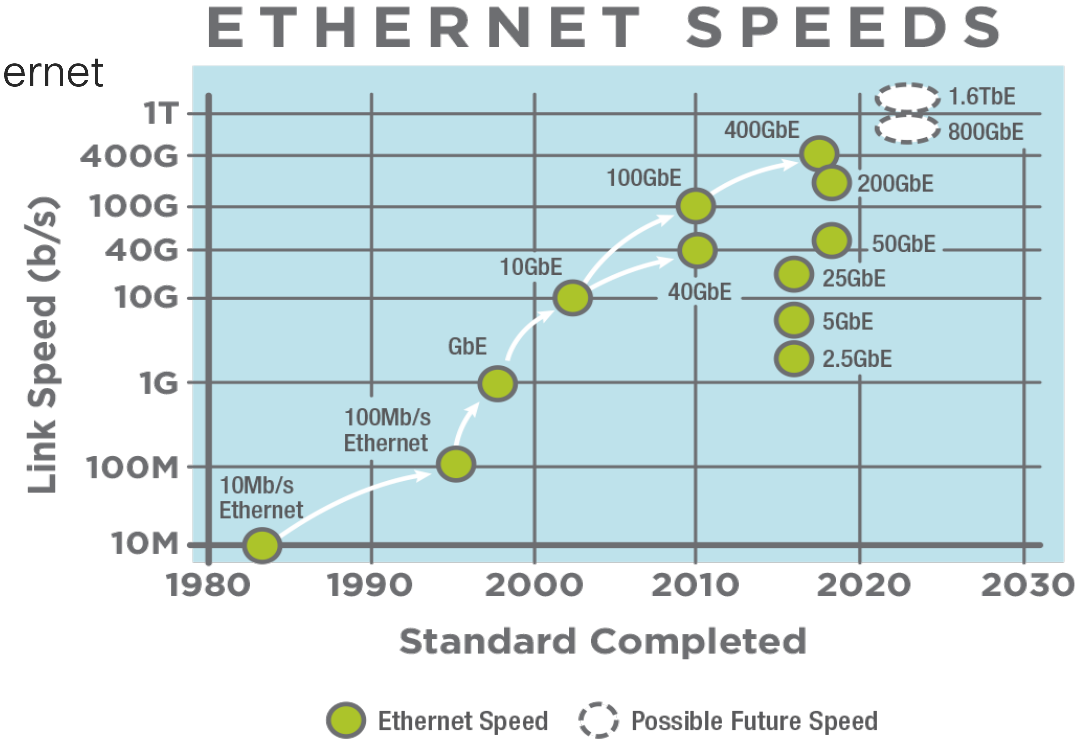
Ethernet Roadmap

6 new speeds in 1st 35 years of Ethernet

6 new speeds in a 2 year span

- 2.5 GbE – 2016
- 5 GbE – 2016
- 25 GbE – 2016
- 50 GbE – late 2018
- 200 GbE – 2017
- 400 GbE – 2017

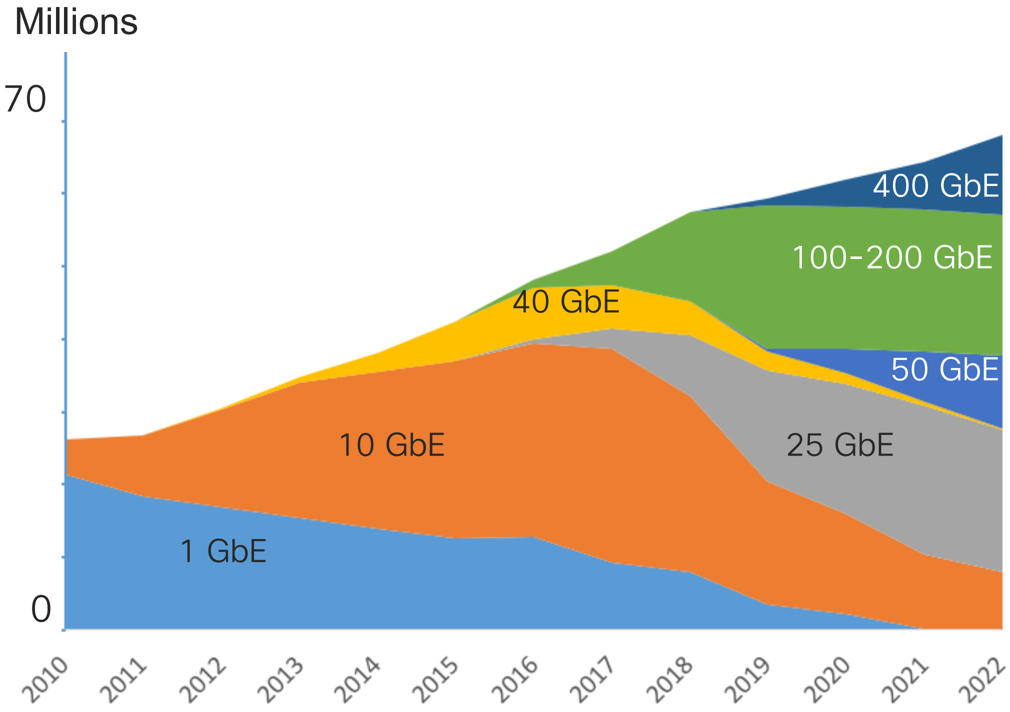
- Proliferation of MSAs/Consortia
  - QSFP-DD, OSFP
  - 100G Lambda
  - COBO



https://ethernetalliance.org/the-2018-ethernet-roadmap/

NEXT ETHERNET ERA
ethernet alliance

Ethernet Port Speed Transitions



Source: DELL'ORO GROUP

Deployments

10 GbE → 25 GbE

- DC servers, Campus backbone

40 GbE → 100 GbE

10 GbE → 100 GbE

2019 – 2021 early adopters

25 GbE → 50 GbE → 100 GbE

- DC servers

100 GbE → 400 GbE

DC Uplinks, SP Edge/Core

400GE Use Cases

Webscale

Scale-out fabrics

Transition from 10/40G to 25/50/100G server NICs

Lower power per Gigabit

Enterprise Deployments

High performance IO

Increased need to support AI/ML applications/workloads at scale

Enhanced flow level visibility

Service Providers

100G/ 400G fabrics for space constraint environments in SP DC & edge locations

Ready for NFV/ 5G adoption cycle

Choice and Flexibility

400GE technology considerations:

Client and Line/Transport optics

Higher Order Modulation

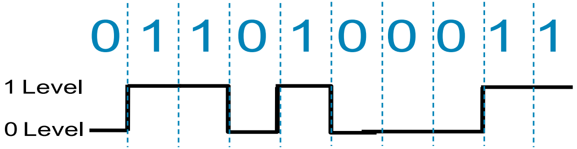
Same Data and Data Rate; but lower frequency (baud rate).

PAM-2

1-bit Symbols

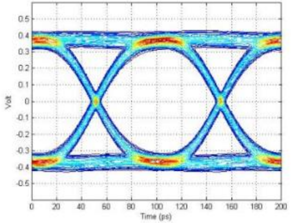
(aka NRZ)

- 1 (1 level)
- 0 (0 level)



PAM-2

(1-bit per symbol)

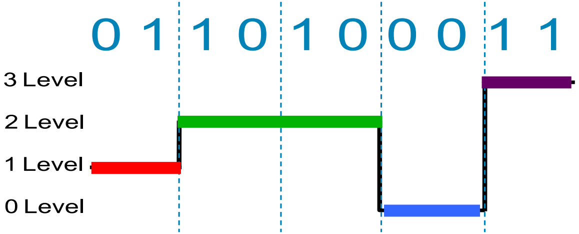


PAM-4

2-bit Symbols

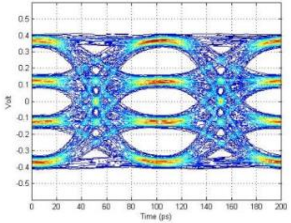
(But 4 levels)

- 1 1 (3 level)
- 1 0 (2 level)
- 0 1 (1 level)
- 0 0 (0 level)



PAM-4

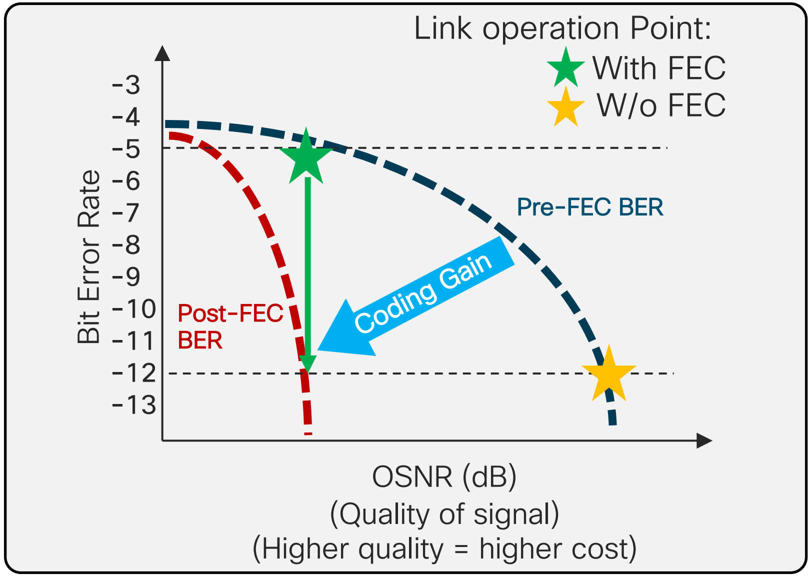
(2-bit per symbol)



Impact of higher order modulation

- Twice the data capacity for same “speed” components
- Enables lower bandwidth components and materials
- Reduces wavelengths & fibers compared to NRZ
- More complex transmitters and receivers

Forward Error Correction
Benefits far out-weigh the drawbacks



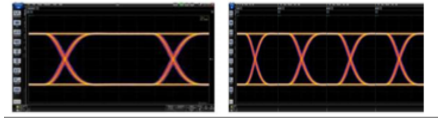
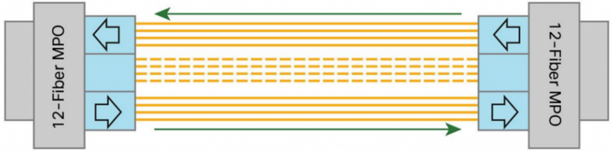
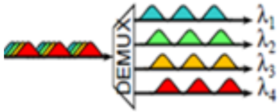
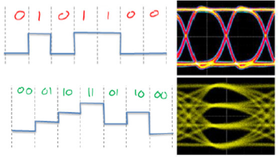
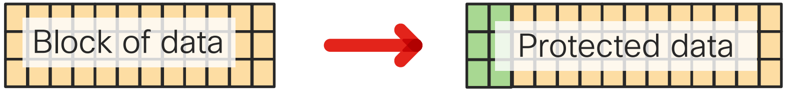
Usage of lower quality optical specifications **significantly reduces** cost and power of solutions

Different FEC algorithms can be used all with different performance properties

- Reed-Solomon: most common in Ethernet
- Higher performance FECs (e.g. used in Coherent optics)

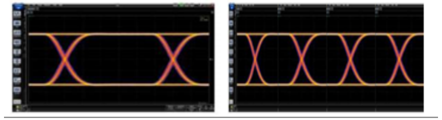
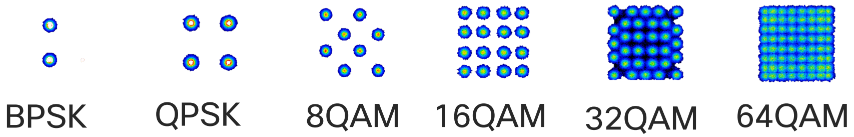
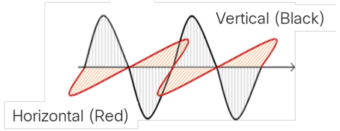
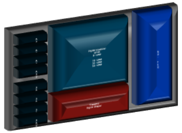
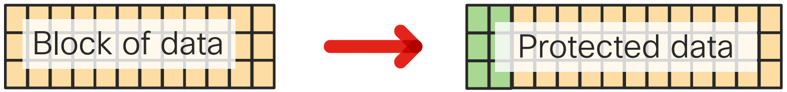
- Incremental latency impact is dependent on implementation and data rate.
- For common Ethernet interfaces latency increase in range of ~50 to 100 ns (equivalent to time of flight over 5-10m of fiber)

Client Optics – Technology summary

Increase baud rate (e.g. 10G to 25G)		SR, LR, etc.
Increase number of fibers (Parallel)	 	SR4, PSM4, DR4,
Increase number of wavelengths (WDM)	 DEMUX λ_1 λ_2 λ_3 λ_4	LR4, ER4, BiDi, CWDM4, FR4
Change modulation format (e.g. NRZ to PAM4)	 Eye Diagram NRZ Each time slot can take on one of two values (0,1) Eye Diagram PAM4 Each time slot can take on one of 4 Discrete values (00,01,10,11)	100G-FR, 100G BiDi, 400G-DR4
Enhance Bit Error Rate with Forward Error Correction (FEC)	 Block of data Protected data	Everything above 40G (except 100G-LR4)

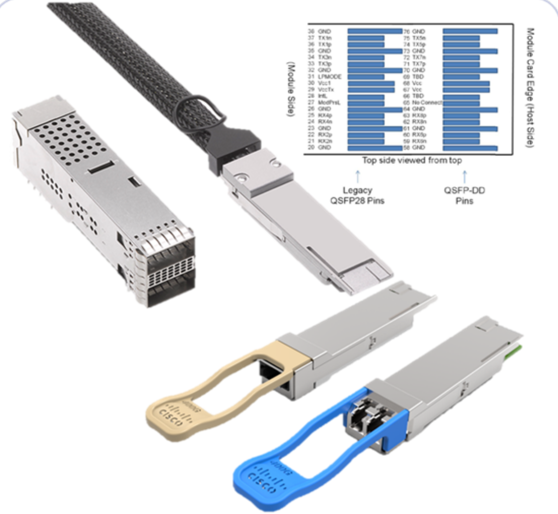
Coherent Optics (Transport/Line)

Technology summary

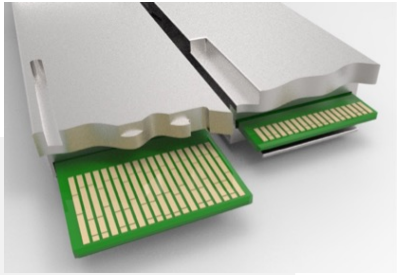
Increase baud rate (e.g. 32 GBaud to 72 GBaud)	
Advanced Modulation (and flexible/hybrid)	
Dual Polarization Modulation	
Tunable Lasers	
Advanced Equalization and Digital Signal Processing	
Enhance Bit Error Rate with Forward Error Correction (FEC)	Block of data Protected data

Introducing QSFP DD

QSFP-DD



QSFP-DD
400G Module & Cage

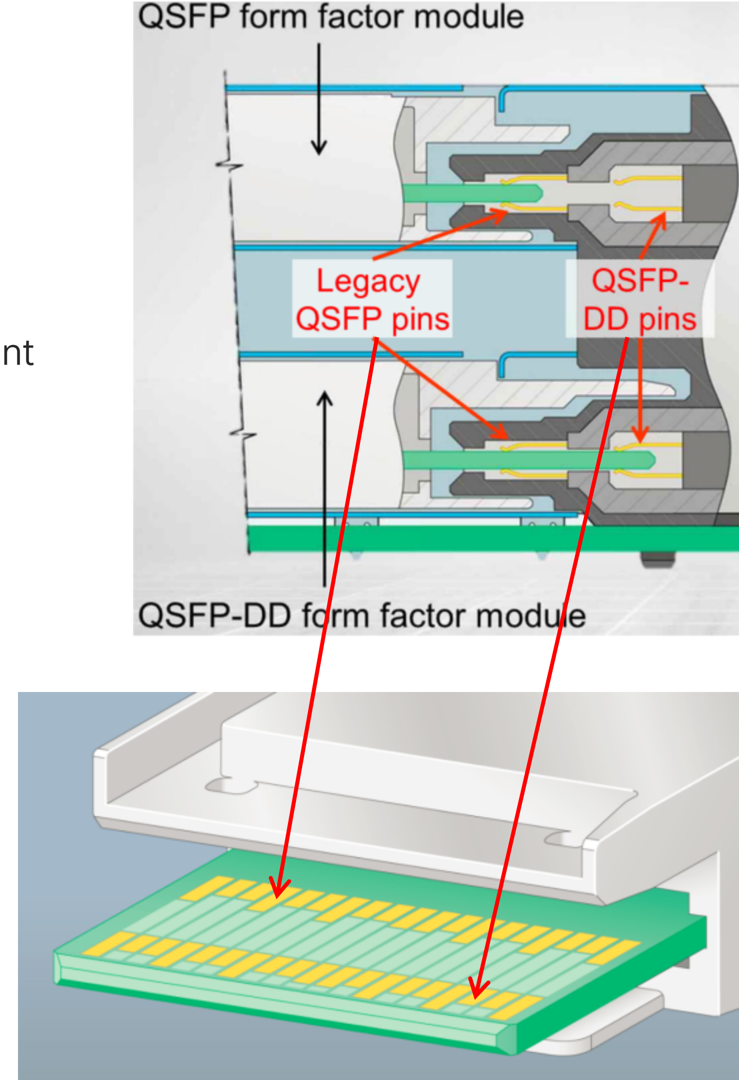


- QSFP-DD MSA has very broad industry support
- MSA has 66 member companies
- Port is backward compatible to QSFP+, QSFP28, QSFP56
- Leverages industry cost structure and production capability of QSFP
  - Nearly 30M QSFP modules deployed to date
  - More than 66M QSFP ports will be deployed by end of '19
- Support 400 GbE and 2/4x 100 GbE designs
- QSFP-DD will support up to 20W of power dissipation
- Broad product offering from copper cable to 400G-ZR

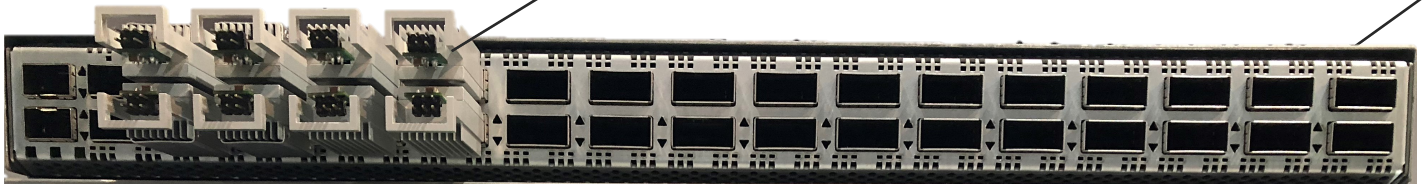
- Supports ASIC interfaces - 400G AUI-8 (8x 50G PAM4)
- Support network requirements for system density: 32 & 36 ports
- Support necessary thermal/SI for implementations (all optical and copper reaches

Optics Innovation – QSFP-DD

- QSFP plus a 2nd row of pins
  - Drop-in upgrade for 100G networks – same port count
  - Maintains 36 ports per RU w/ backward compatibility
- Same faceplate, slightly deeper
- QSFP56-DD for 400G
  - 8 electrical lanes at 50G (56 w/ overhead)
- QSFP28-DD for 200G or 2x 100G
  - 8 electrical lanes at 25G (28 w/ overhead)
- Can support breakouts



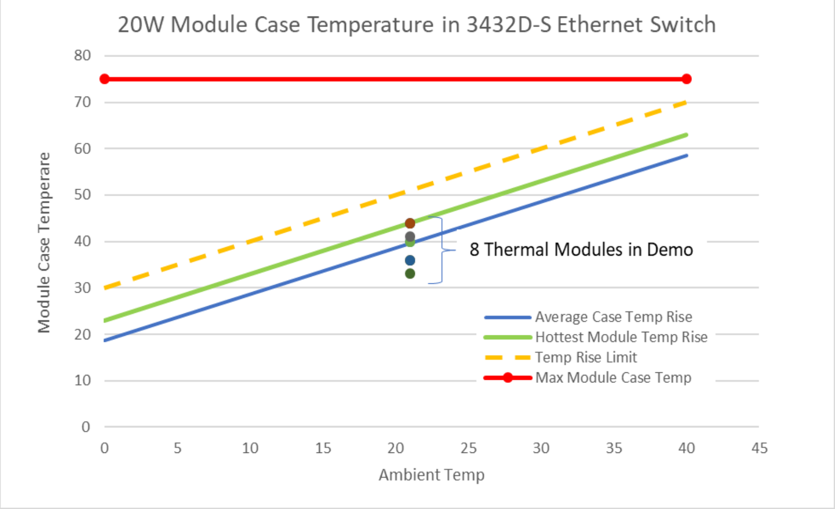
OFC2019 QSFP-DD Thermal Demo



Thermal Modules
21.7W each

Nexus N3K-C3432D-S
Ethernet Switch
32 Ports QSFP-DD

- Cisco N3K-C3432D-S Ethernet Switch with 32 Ports of QSFP-DD
- 8 thermal modules each dissipating 20W
- Average Power: 21.6W
- Average temperature rise: 18.3C
- Cisco SP360 blog on the demo
  - https://blogs.cisco.com/sp/cisco-demonstrates-20w-power-dissipation-of-qsfp-dd-at-ofc-2019



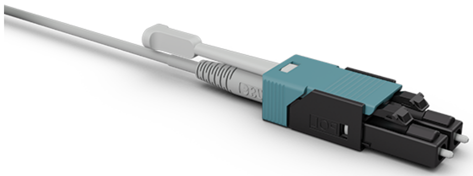
Infrastructure Considerations

Optical Connector Considerations

Two dominant ferrule/connector types

New

LC – single fiber ferrule

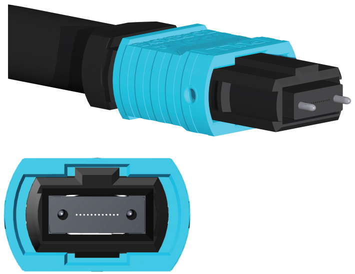


Uni-boot designs are common

Courtesy US CONEC

Used for all duplex fiber applications – SFP, QSFP, QSFP-DD etc

MPO/MTP – multiple fiber ferrule

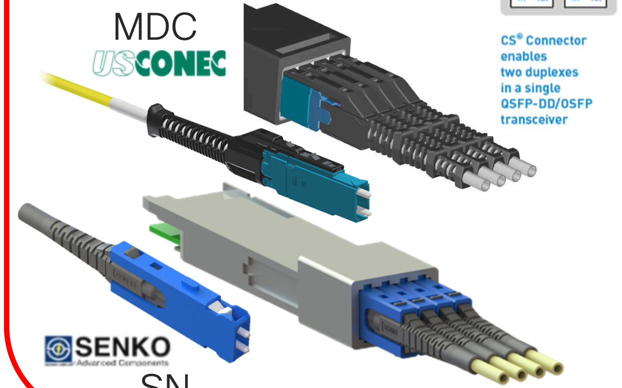


Courtesy US CONEC

Traditionally used for all parallel fiber applications – QSFP, QSFP-DD etc

New Dual ferrule connectors

CS® Duplex



1 2 TX RX
1 2 TX RX
MDC
US CONEC
CS® Connector enables two duplexes in a single QSFP-DD/OSFP transceiver
SENKO Advanced Components
SN

New high-density connectors for breakout applications (market adoption pending…)

400 GbE modules and use cases

Distance	3+ m	100 m	500m-2km	10 km	100+ km
Optics	400G-CR8 8x 50G-CR 400G-AOC(30m)	400G-SR8 400G-SR4.2 400G-DR4	400G-DR4 400G-FR4	400G-LR4 400G-LR8	400ZR 400ZR+
Media	Copper Cables / AOC (Active Optical Cable)	MMF / SMF	SMF	SMF	SMF

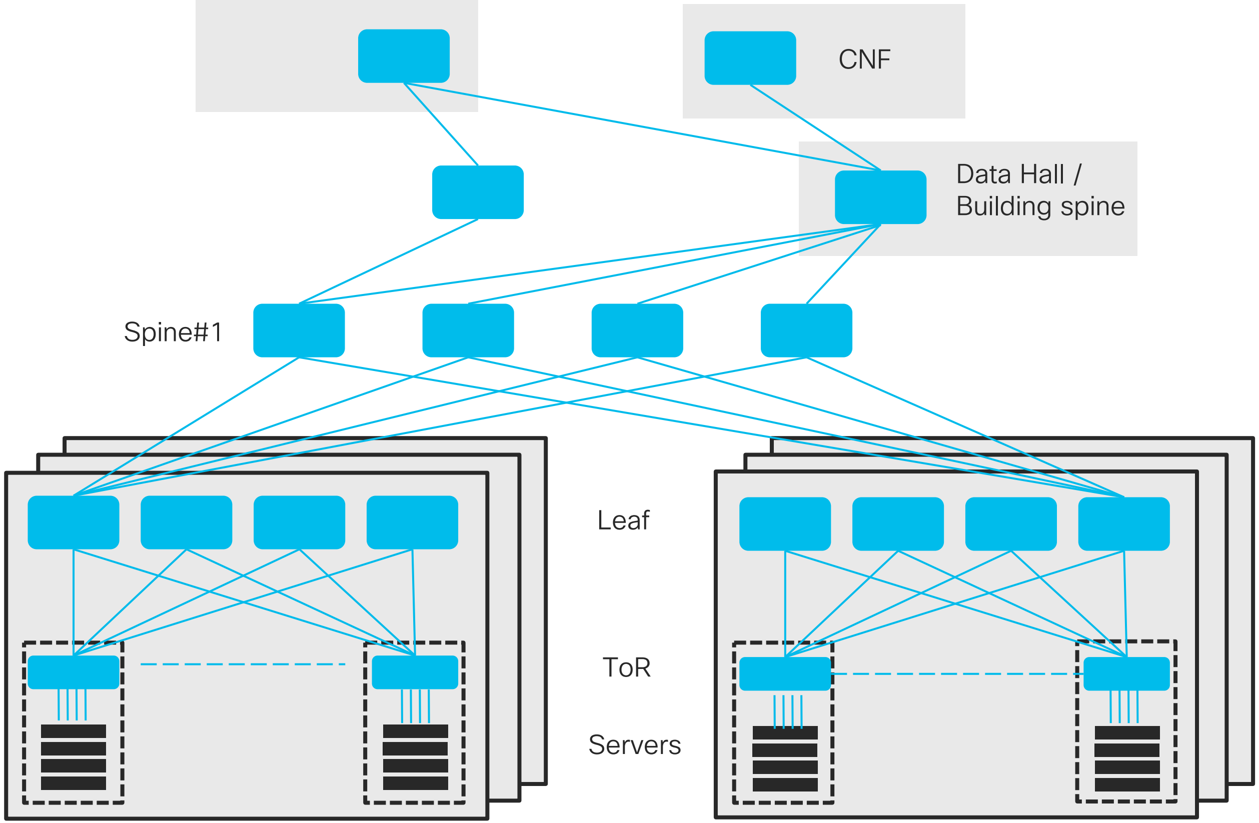
400G Optics available from Cisco in CY2019

PID	Description	Optical Connector
QDD-400-CUxM	400G QSFP-DD to QSFP-DD Passive Copper Cable **1/2/2.5/3m**	N/A
QDD-400G-AOCxxM	400G QSFP-DD to QSFP-DD Active Optical Cable, **1/2/3/5/7/10/15/20/25/30m**	N/A
QDD-400G-DR4-S	400G QSFP-DD Transceiver, 400GBASE-DR4, MPO-12, 500m parallel SMF, can be used as 4X 100G breakout to QSFP-100G-FR-S	MPO-12 SMF APC
QDD-400G-FR4-S	400G QSFP-DD Transceiver, 400G-FR4, Duplex LC, 2km Duplex SMF	Duplex SMF LC
QDD-400G-LR8-S	400G QSFP-DD Transceiver, 400GBASE-LR8, Duplex LC, 10km Duplex SMF	Duplex SMF LC

Architectural and Deployment Considerations

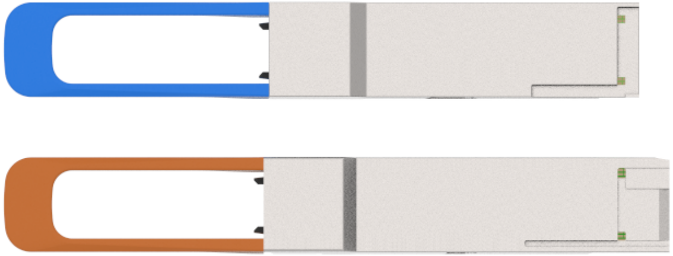
DC Topology – Scale

100G-DCO	400G-ZR/ZR+	100's km
100G-LR4	400G-LR4 400G-FR4	< 10km
100G-CWDM4 100G-PSM4	400G-FR4 400G-DR4	100's m
100G-AOC 100G-BiDi	400G-AOC 400G-SR8 400G-SR4.2	10's m
100G-CR4 4x25G-CR	400G-CR8 8x50G-CR	<3m



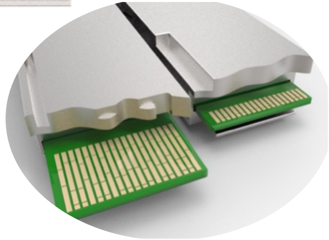
400 GbE to 100 GbE backwards compatibility

400 GbE QSFP-DD ports are backwards compatible with 100 GbE QSFP28 modules



QSFP28

QSFP-DD

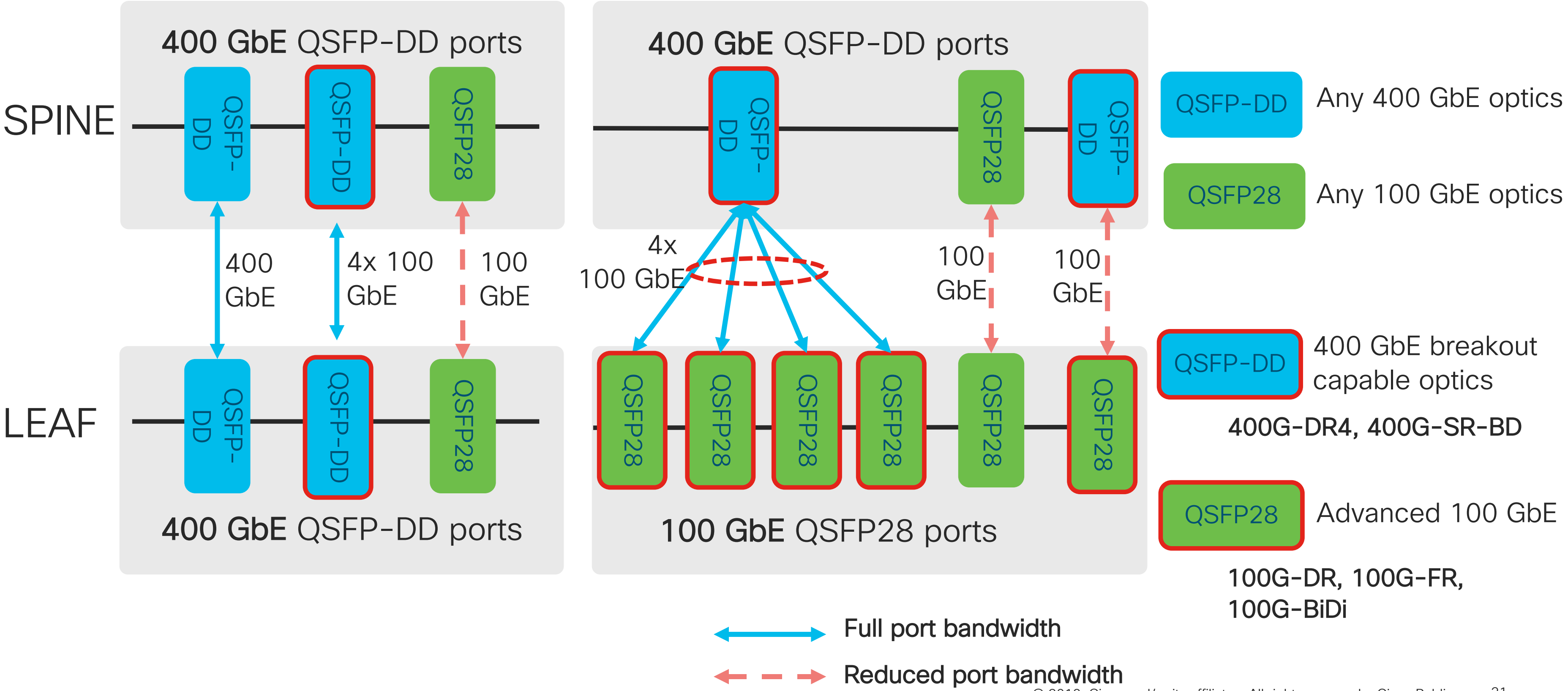


- Ease of migration to next higher speed
- Connecting to legacy equipment
- Mix and match optics for cost reasons

- 400 GbE is based on 8x 50Gb/s interfaces
- ASICs with 50 Gb/s interfaces can down-rate to 25 Gb/s (or 10 Gb/s)
  - Compatible with QSFP28 (4x 25 Gb/s)

Leaf-Spine Deployment considerations

Some examples



SPINE

LEAF

400 GbE QSFP-DD ports

QSFP-DD | QSFP-DD | QSFP28

400 GbE | 4x 100 GbE | 100 GbE

QSFP-DD | QSFP-DD | QSFP28

400 GbE QSFP-DD ports

400 GbE QSFP-DD ports

QSFP-DD | QSFP28 | QSFP-DD

4x 100 GbE | 100 GbE | 100 GbE

QSFP28 | QSFP28 | QSFP28 | QSFP28 | QSFP28 | QSFP28

100 GbE QSFP28 ports

QSFP-DD Any 400 GbE optics

QSFP28 Any 100 GbE optics

QSFP-DD 400 GbE breakout capable optics

400G-DR4, 400G-SR-BD

QSFP28 Advanced 100 GbE

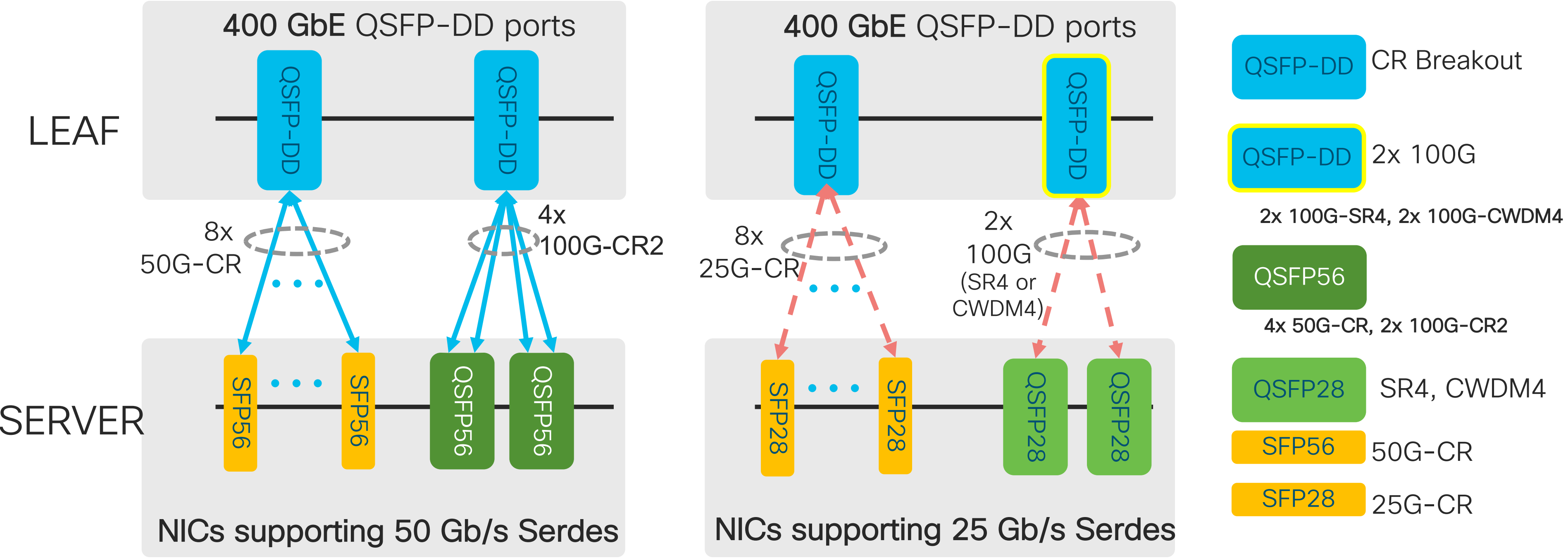
100G-DR, 100G-FR, 100G-BiDi

Full port bandwidth

Reduced port bandwidth

400 GbE Leaf-Server Deployment considerations

Some examples



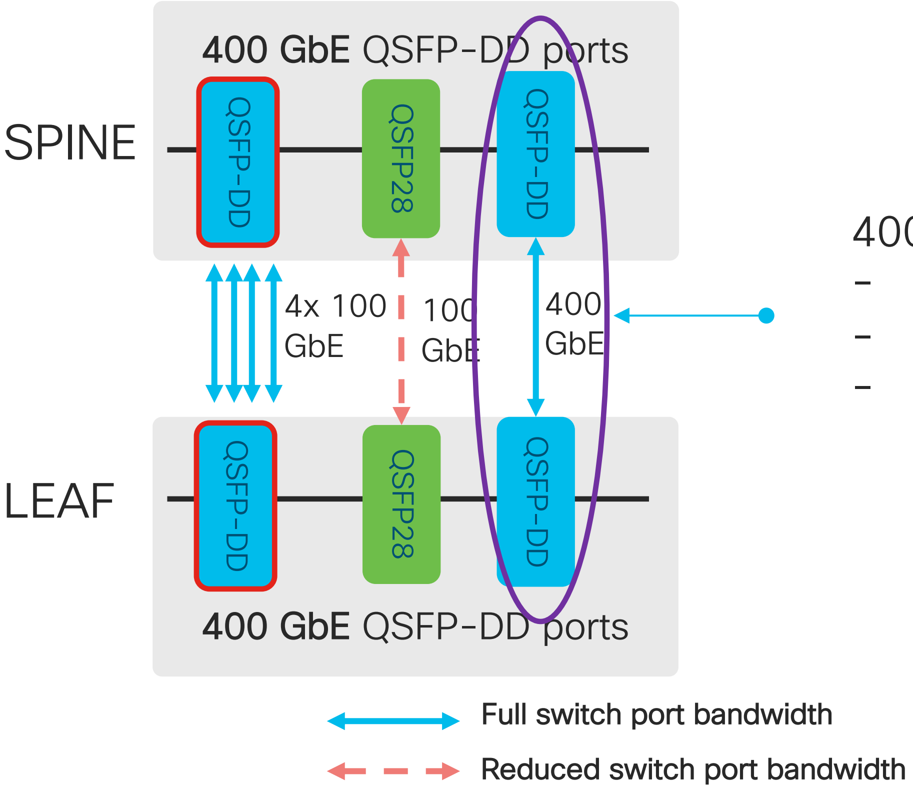
Previous slide's optical options all possible for server IO too

↔ Full switch port bandwidth

⇠ ⇢ Reduced switch port bandwidth

Leaf-Spine Deployment considerations

Some examples



400 GbE

- Better hash efficiency
- Less ports manage vs n x 100 GbE links
- Improved system performance

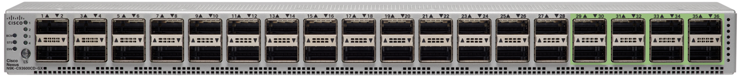
400GE DC switches available from Cisco

Nexus 9316D-GX – 16p NXOS/ACI Spine

16p 400 QSFP-DD

- 400G ACI/NX-OS Spine
- 400G ACI/NX-OS Leaf (future!)
- Flexible port speeds:
  - 16p of 40/100/400G
  - 64p 100G bandwidth in compact 1RU formfactor
  - Breakout capable to 10/25/50/100/200G
- Flexible TCAM Templates with 80MB Buffer
- Enhanced Telemetry- FT, FTE, SSX, INT transparent & postcard

Nexus 93600CD-GX – NXOS/ACI Leaf



28p 40/100 QSFP-28
+ 8p 400 QSFP-DD

- 400G ACI/NX-OS Leaf
- 400G ACI/NX-OS Spine (future!!)
- High performance for AI/ML workloads
- Flexible port speeds:
  - 28p of 40/100 + 8p of 400
  - Breakout capable to 10/25/50/100/200G*
- Flexible TCAM Templates with 80MB buffer
- Enhanced Telemetry- FT, FTE, SSX, INT transparent & postcard

** ports 25-36*

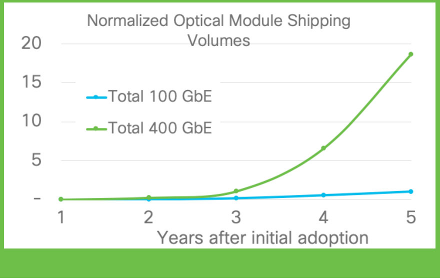
Nexus 3432D-S

Nexus 3432D-S

- 1RU fixed NXOS switch (H-Innovium release)
- 32ports of QSFP-DD
- Flexible connectivity options
  - 1x 400/100/40GE
  - 4x100/50G/25/10GE
  - 8x 50GE
- 70MB Buffer
- Telemetry- INT
- Low Latency

Summary

Demand for 400 GbE is here



Uniquely for 400 GbE, multiple solutions will exist in a common QSFP-DD pluggable form factor

Industry is broadly engaged to deliver 400 GbE now

Cisco is engaged in most 400GE development areas

You make **possible**