Looking Beyond 400G

A System Vendor Perspective

Beyond 400 Gb/s Ethernet Study Group

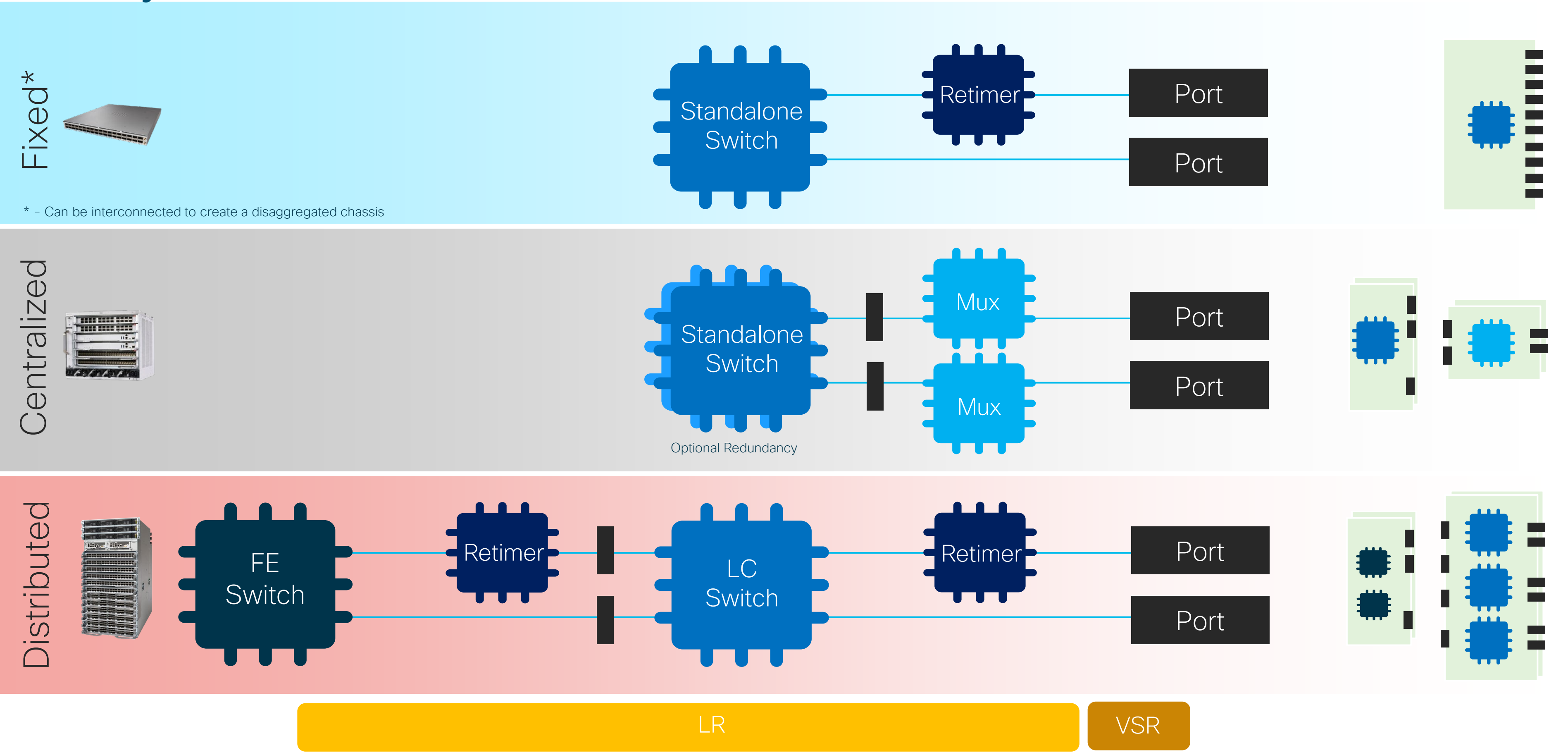
Rakesh Chopra
Cisco Fellow

February 8, 2020

rakchopr@cisco.com
www.linkedin.com/in/rakesh-chopra/
@Rakesh_Chopra1

… Many thanks to Cisco Engineers and Insightful Customers …

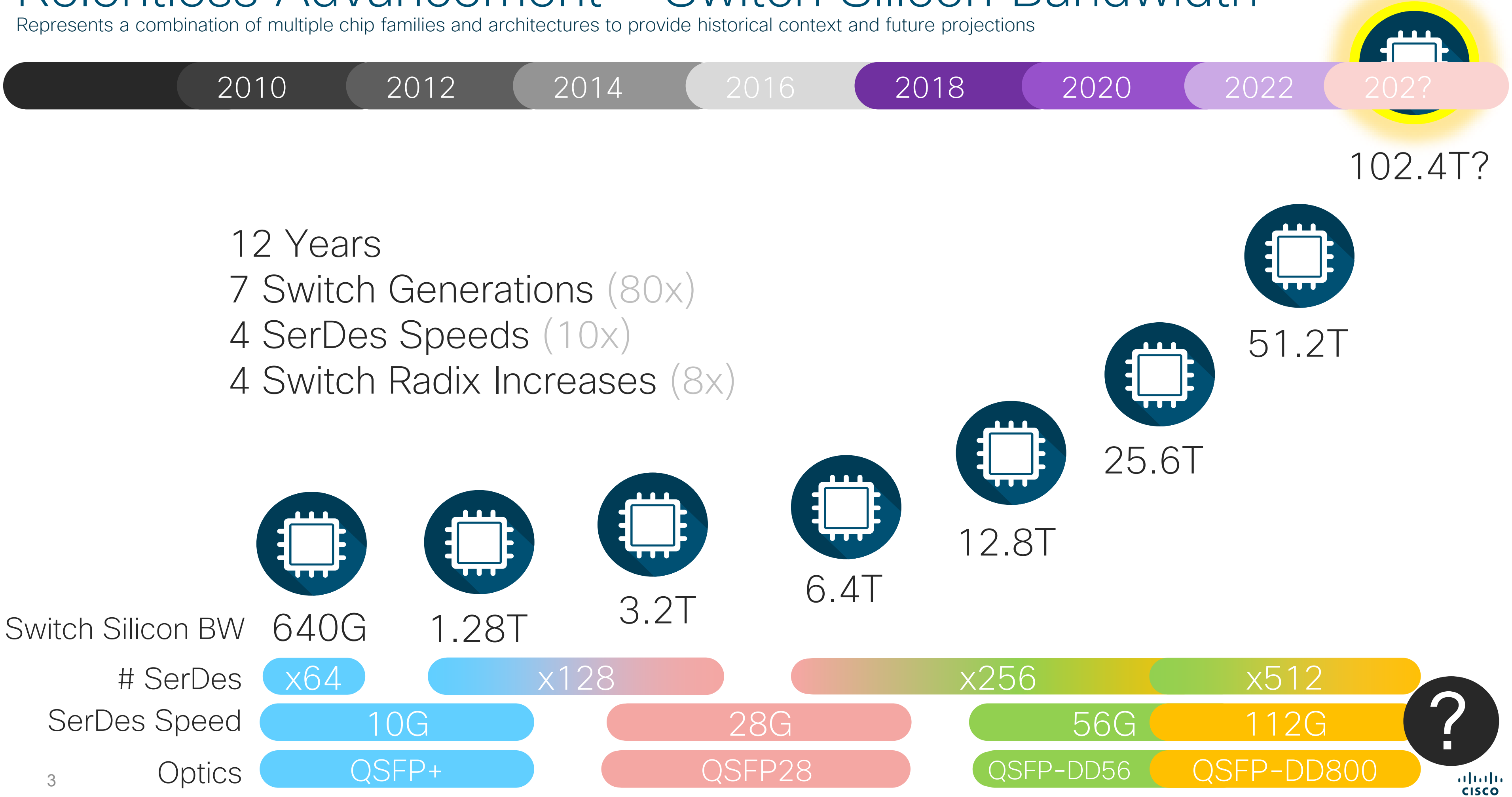
System Architectures



* - Can be interconnected to create a disaggregated chassis

Relentless Advancement – Switch Silicon Bandwidth

Represents a combination of multiple chip families and architectures to provide historical context and future projections



2010 2012 2014 2016 2018 2020 2022 202?

12 Years
7 Switch Generations (80x)
4 SerDes Speeds (10x)
4 Switch Radix Increases (8x)

Switch Silicon BW: 640G, 1.28T, 3.2T, 6.4T, 12.8T, 25.6T, 51.2T, 102.4T?

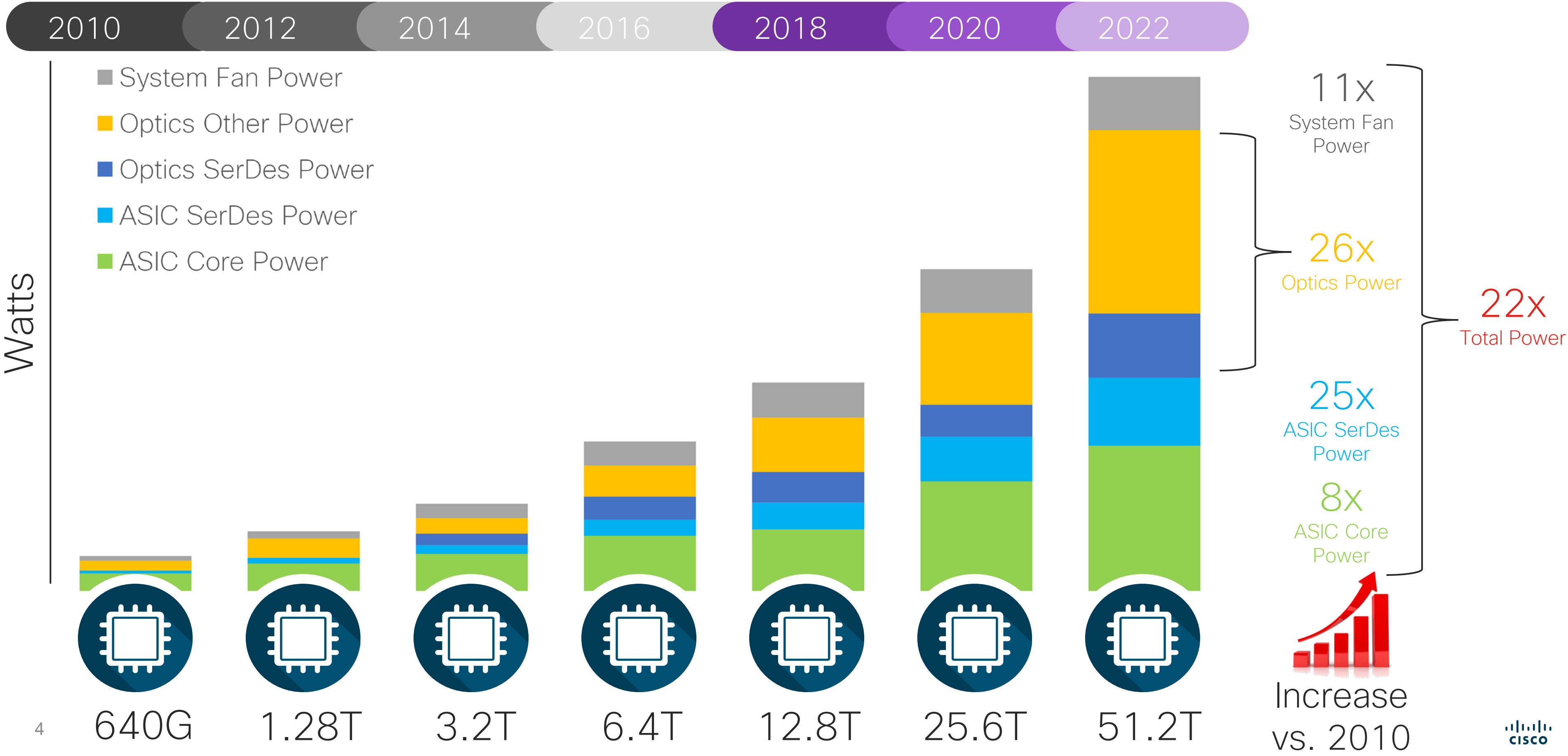
SerDes: x64, x128, x256, x512

SerDes Speed: 10G, 28G, 56G, 112G, ?

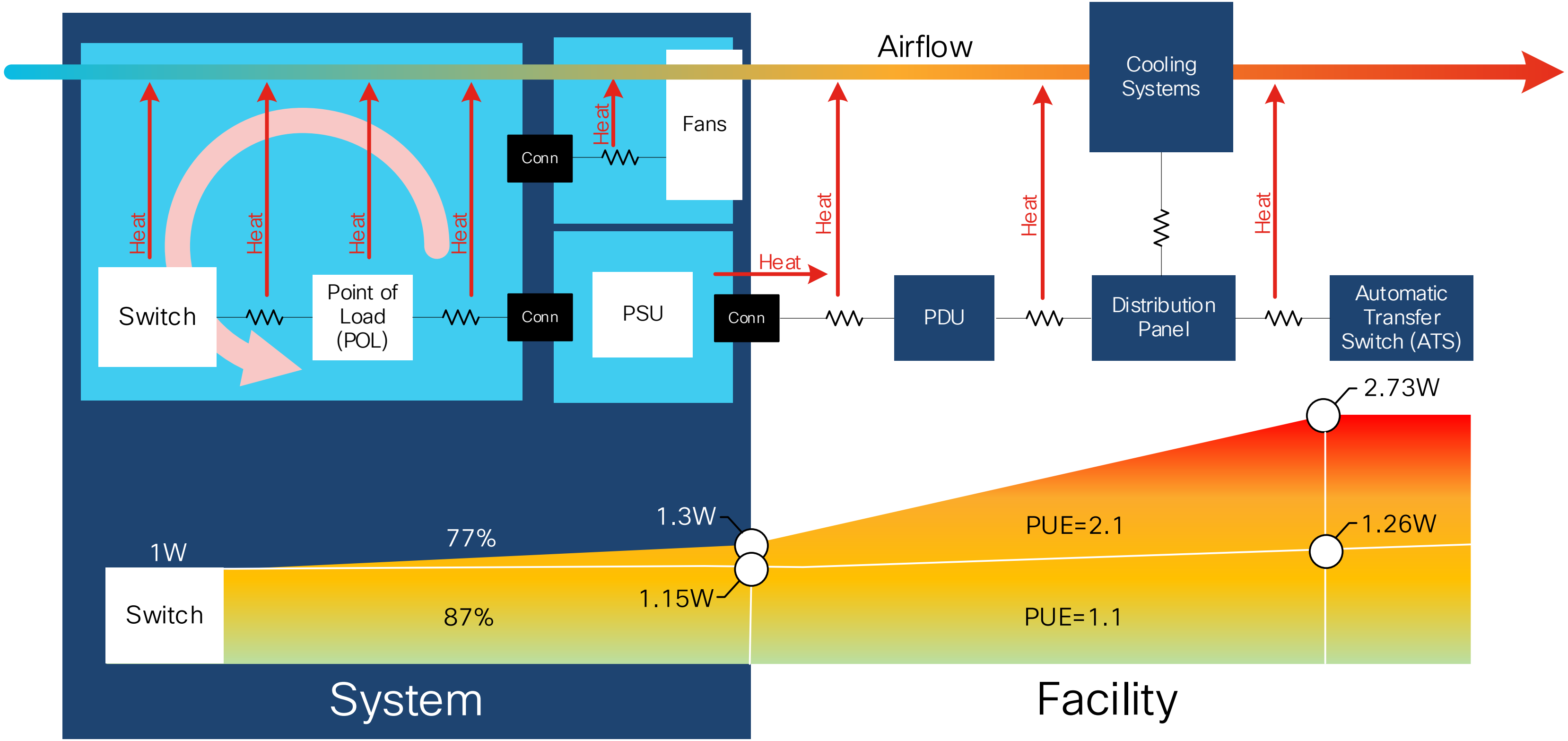
Optics: QSFP+, QSFP28, QSFP-DD56, QSFP-DD800

Relentless Advancement – 80x BW over 12 Years

Represents a combination of multiple chip families and architectures to provide historical context and future projections
Fixed Box Power Breakdown
Retimer Power and other system components not included



The Multiplication Effect of a Watt



Power is THE Problem to Solve

Apollo 13 – Universal Pictures

- ☒ Limits what we can build
- ☒ Limits what can be deployed
- ☒ Limits what our planet can sustain

“Power is Everything”*

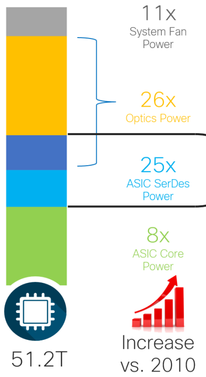
John Aaron- Apollo 13 Flight Controller

Adopt a power first design and deployment methodology

* - Thanks to Kraig Owen for the reference

Co-packaged Optics Is Inevitable

Power savings drives requirement



Must minimize SerDes power

Trends plot from premier CMOS wireline 2018 conference

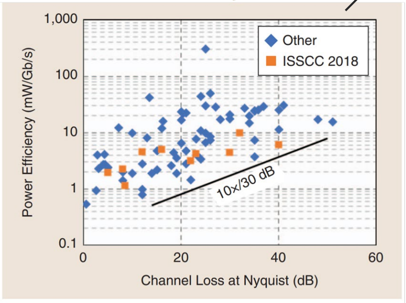
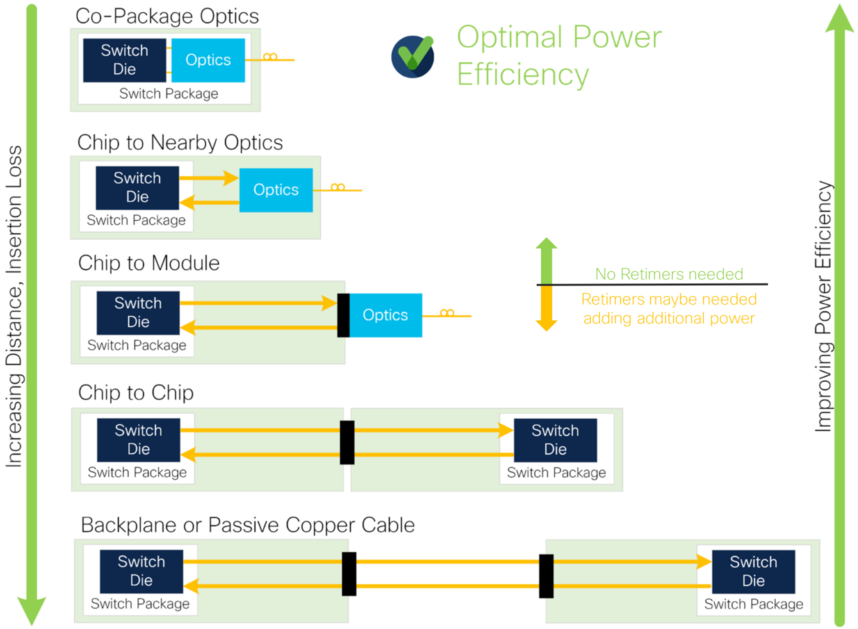


FIGURE 15: Transceiver power efficiency versus channel loss.

SerDes power increases with distance

Daly, Denis C., Laura C. Fujino, and Kenneth C. Smith. "Through the Looking Glass-The 2018 Edition: Trends in Solid-State Circuits from the 65th ISSCC." *IEEE Solid-State Circuits Magazine* 10.1 (2018): 30-46.

Architectural Approach to Power Optimization



Co-Package Optics

Optimal Power Efficiency

Chip to Nearby Optics

Chip to Module

No Retimers needed

Retimers maybe needed adding additional power

Chip to Chip

Backplane or Passive Copper Cable

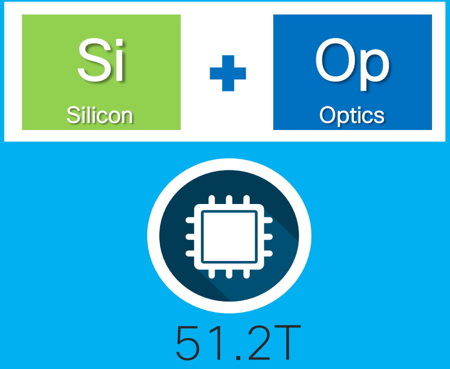
Switch Die
Switch Package
Switch Die
Switch Package
Increasing Distance, Insertion Loss
Improving Power Efficiency

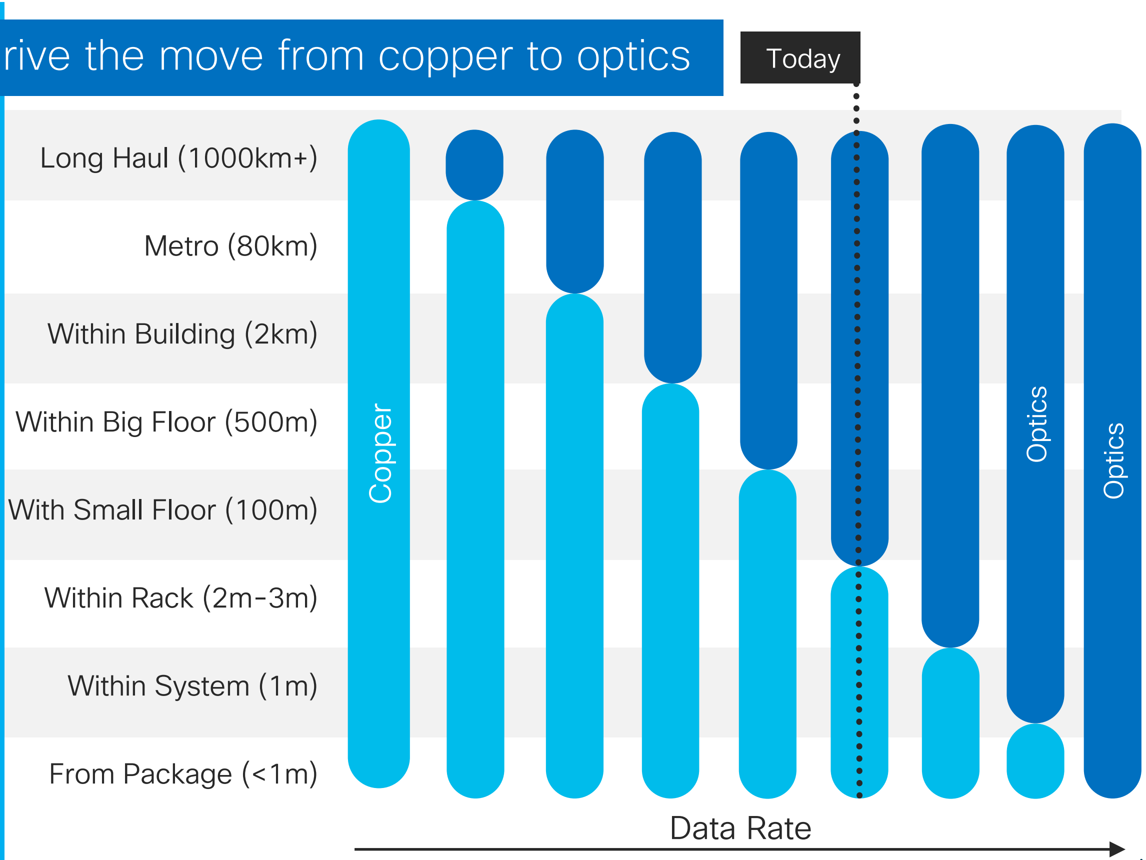
Co-packaged Optics Is Inevitable

and viable in the 51.2T generation

Higher data rates and distance drive the move from copper to optics

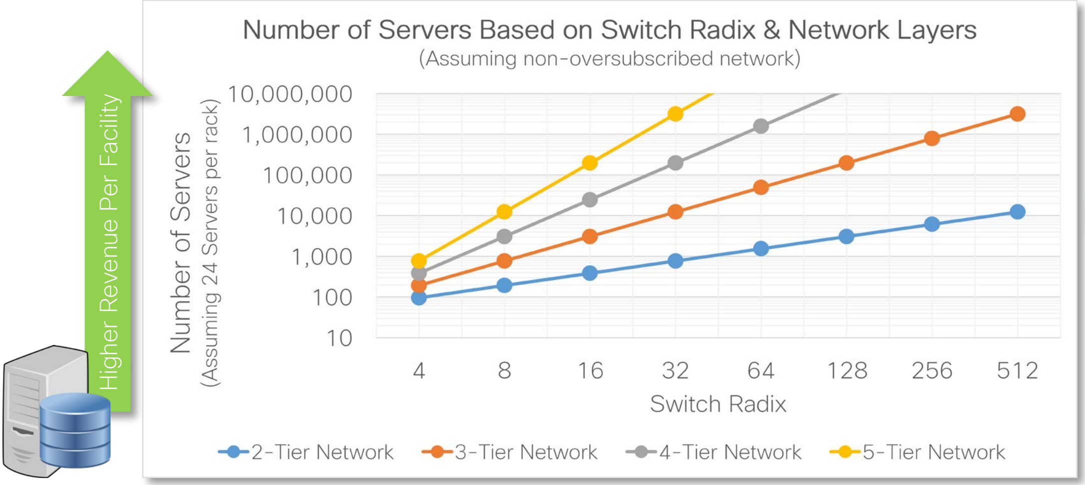
Future innovations will only be possible with silicon and optical integration





Building Your Data Center

Impact of Switch Radix

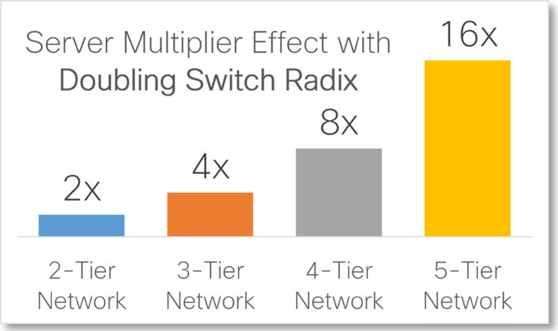


Graph concept leveraged from R. Nagarajan, Ilya Lyubomirsky, "Next-Gen Data Center Interconnects: The Race to 800G"
Adjusted to hold servers per rack constant

Doubling Radix adds 2x-16x more servers

Scale Out

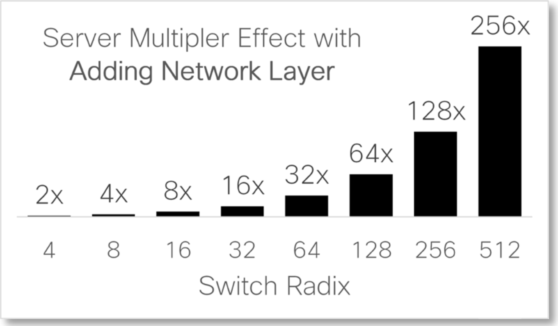
Wider Radix



Adding a layer adds 2x-256x more servers

Scale Up

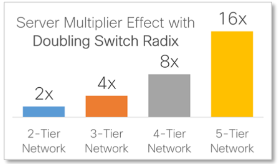
More Layers



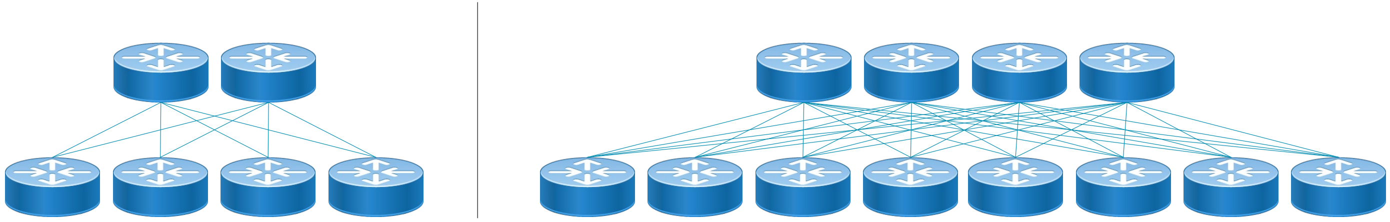
Impact of Switch Radix

Case against increasing Switch Radix

Doubling Radix adds 2x-16x more servers depending on the layers in the network



Increasing radix adds cabling complexity, cost and weight

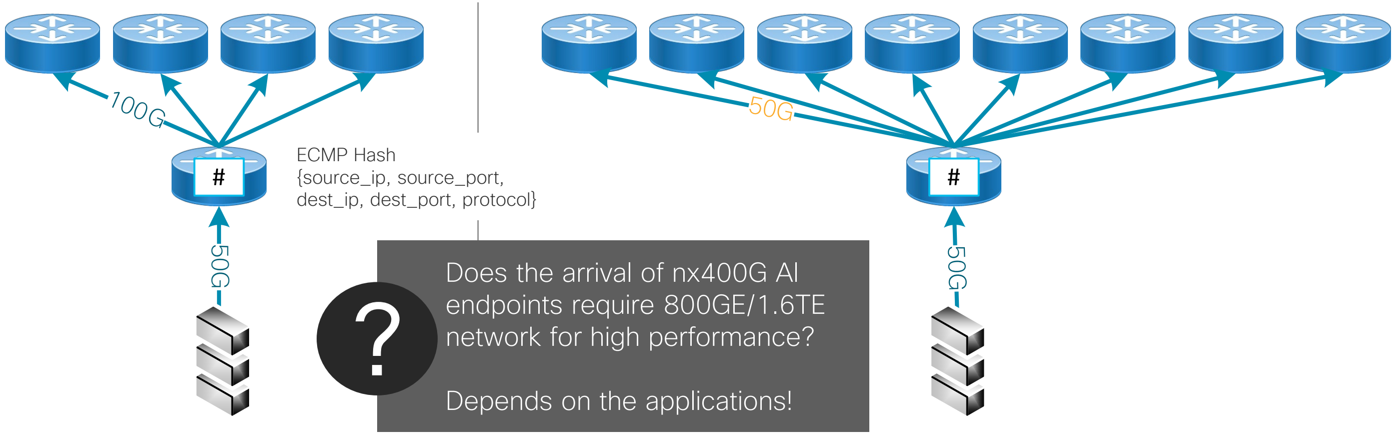


☒ Complicated Cabling

Increasing radix decreases link (mac) speed for the same switch bandwidth

As the flow speed approaches the link speed link utilization decreases

12.8T	
x32	400GE
x64	200GE
x128	100GE
x256	50GE



ECMP Hash
{source_ip, source_port,
dest_ip, dest_port, protocol}

?

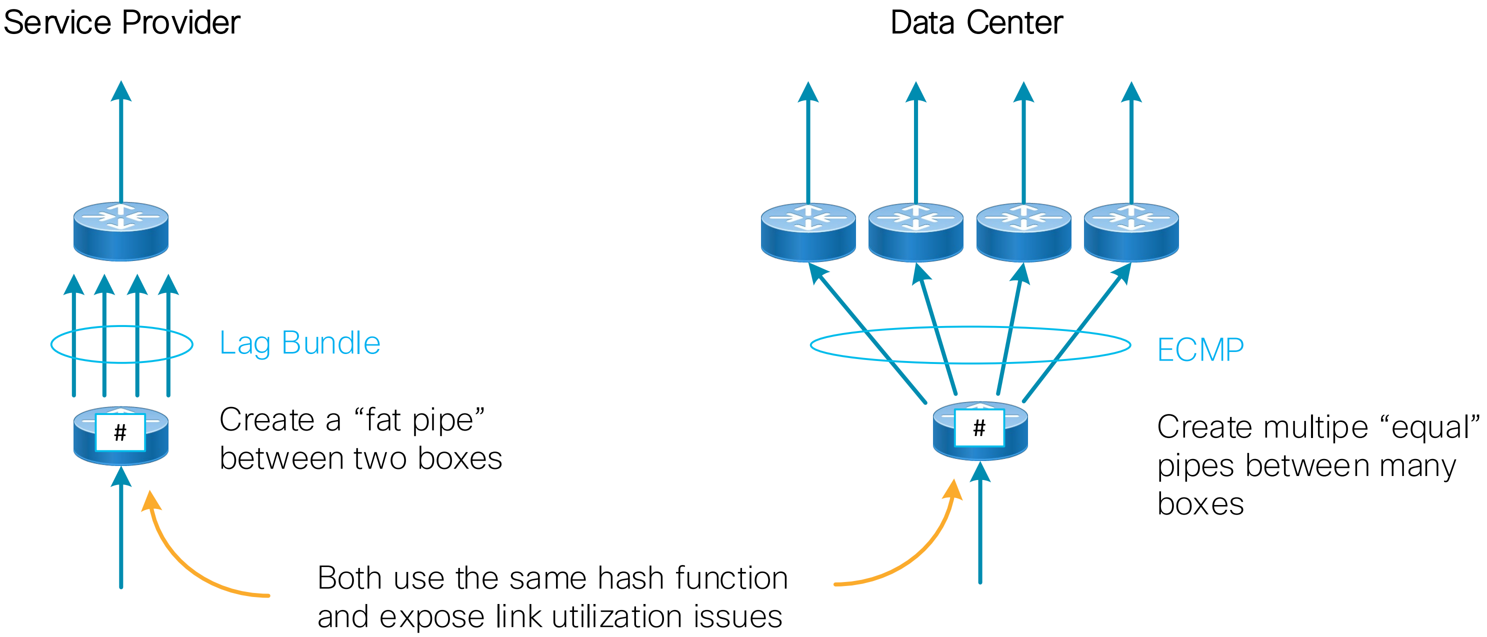
Does the arrival of nx400G AI endpoints require 800GE/1.6TE network for high performance?

Depends on the applications!

☒ Poor link utilization

LAG vs. ECMP

The Basic Topology



Create a “fat pipe” between two boxes

Create multipe “equal” pipes between many boxes

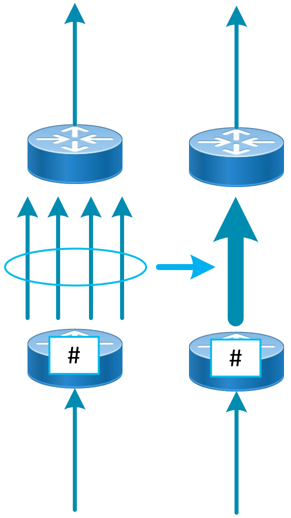
Both use the same hash function and expose link utilization issues

Note : Service Providers use ECMP as well but not in an equivalent fundamental way

LAG vs. ECMP

The advantages of higher speed MACs aren't as clear as they used to be

Service Provider

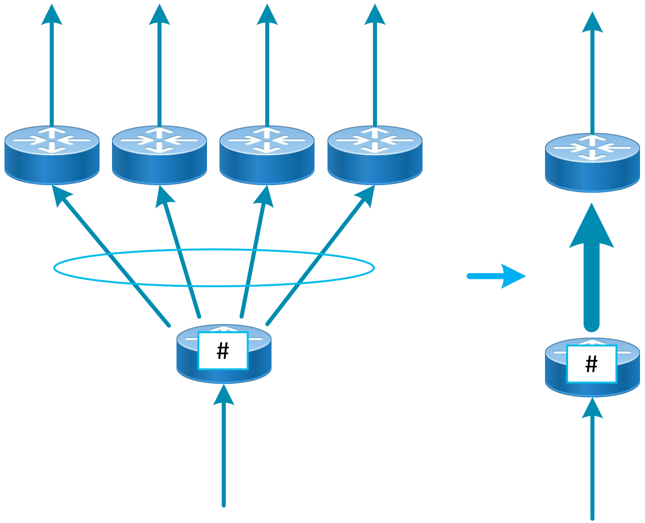


Increase MAC speed

Create a "fat pipe" between two boxes with no hash inefficiencies

No downside to replacing a LAG bundle with a higher speed Ethernet MAC

Data Center



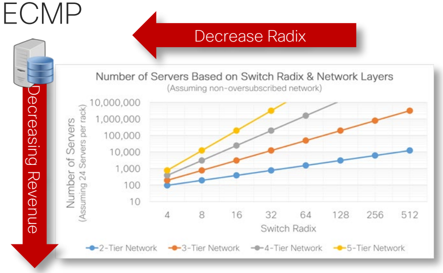
Increase MAC speed

Shrink the number of boxes we can connect to

Downside for higher speed MAC with ECMP

For same speed silicon:

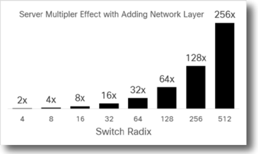
- As you increase MAC speed
- Decrease your radix
- Decreases your switches per DC
- Lower revenue potential



Impact of Switch Radix

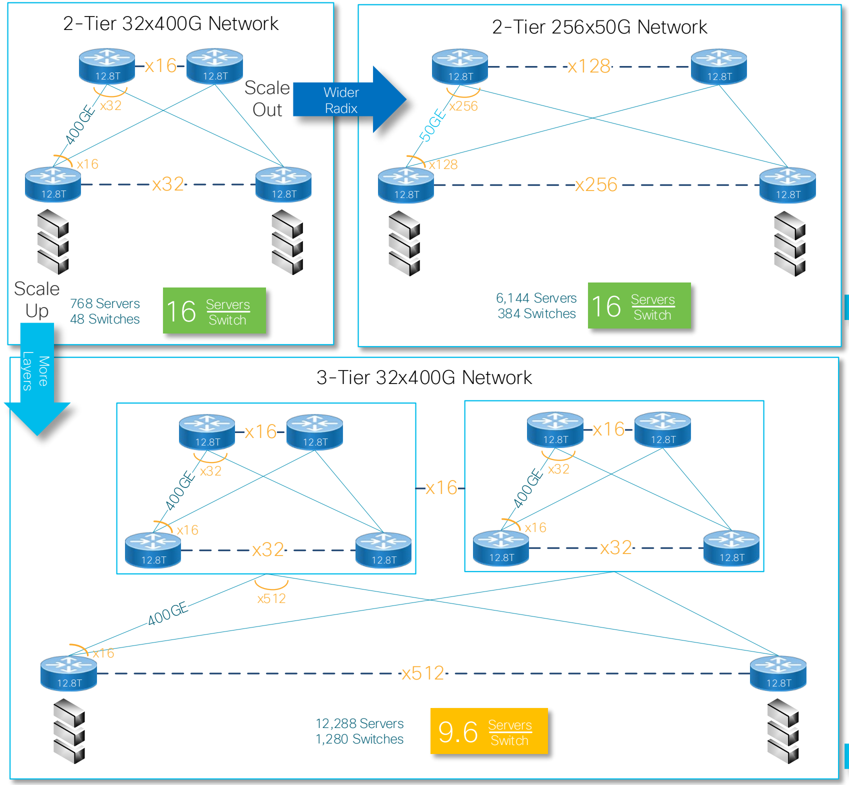
Case against adding Network Layers

Adding a layer adds 2x-256x more servers
depending on the switch radix



Increasing layers adds network cost and power

more switches and optics per server



2-Tier 32x400G Network

768 Servers
48 Switches

16 Servers/Switch

2-Tier 256x50G Network

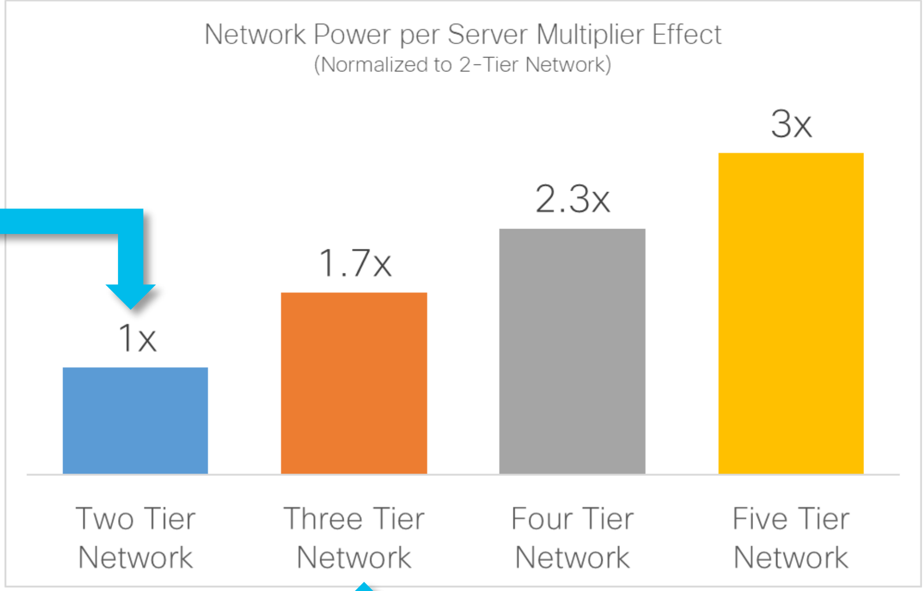
6,144 Servers
384 Switches

16 Servers/Switch

3-Tier 32x400G Network

12,288 Servers
1,280 Switches

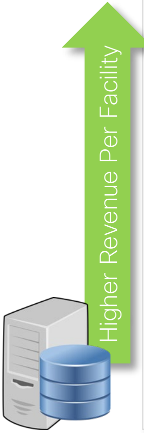
9.6 Servers/Switch

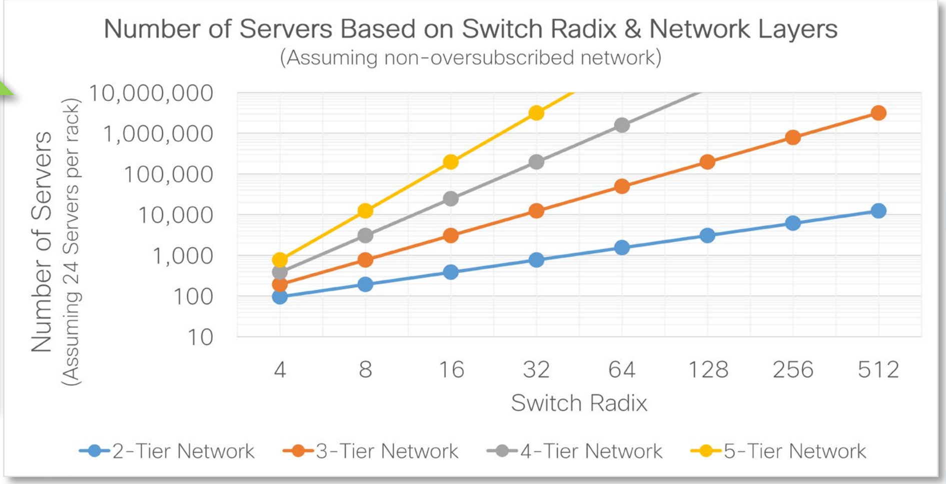

Network Power per Server Multiplier Effect
(Normalized to 2-Tier Network)

Assuming no extra components needed to scale out (reverse gearboxes, etc....)
Ignoring ECMP hash efficiency impact for "goodput" of the network

Building Your Data Center

Impact of Switch Radix

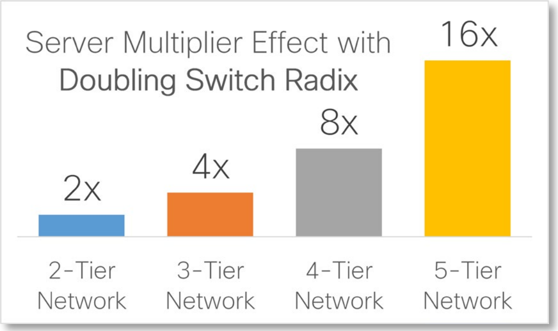




Graph concept leveraged from R. Nagarajan, Ilya Lyubomirsky, "Next-Gen Data Center Interconnects: The Race to 800G"
Adjusted to hold servers per rack constant

Scale Out
Wider Radix

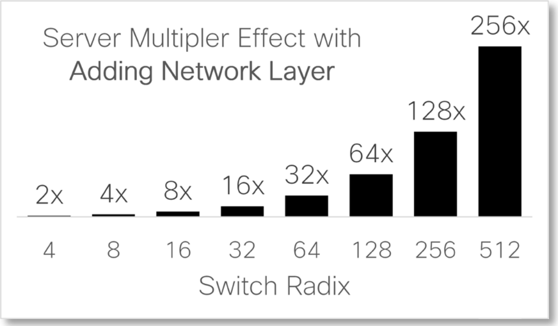
Doubling Radix adds 2x-16x more servers



✓ Power Efficiency

Scale Up
More Layers

Adding a layer adds 2x-256x more servers



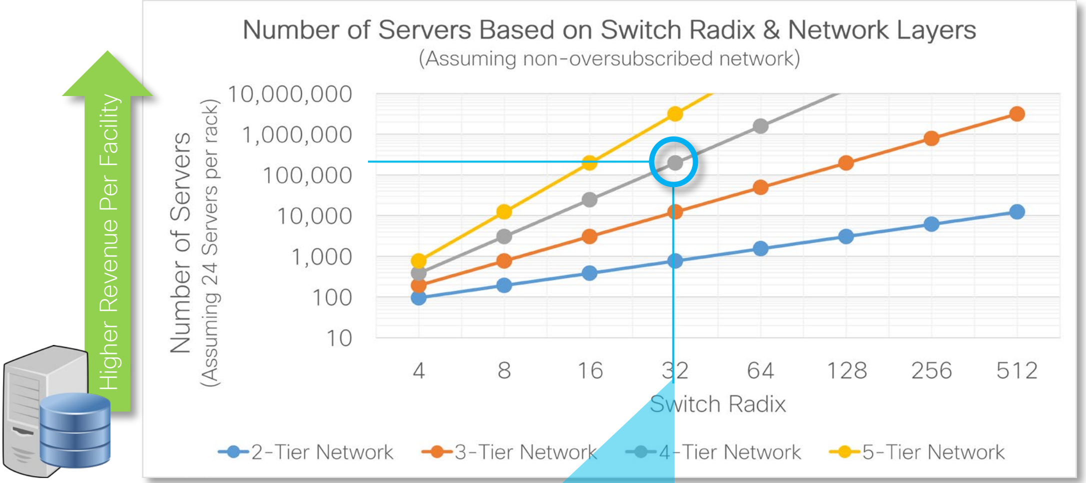
✓ Link Efficiency

There is no free lunch, every engineering choice has trade-offs

Balancing act between radix, MAC speed, and layers in the network…

Building Your Data Center

Scale-Out vs. Scale-Up- A Balancing Act



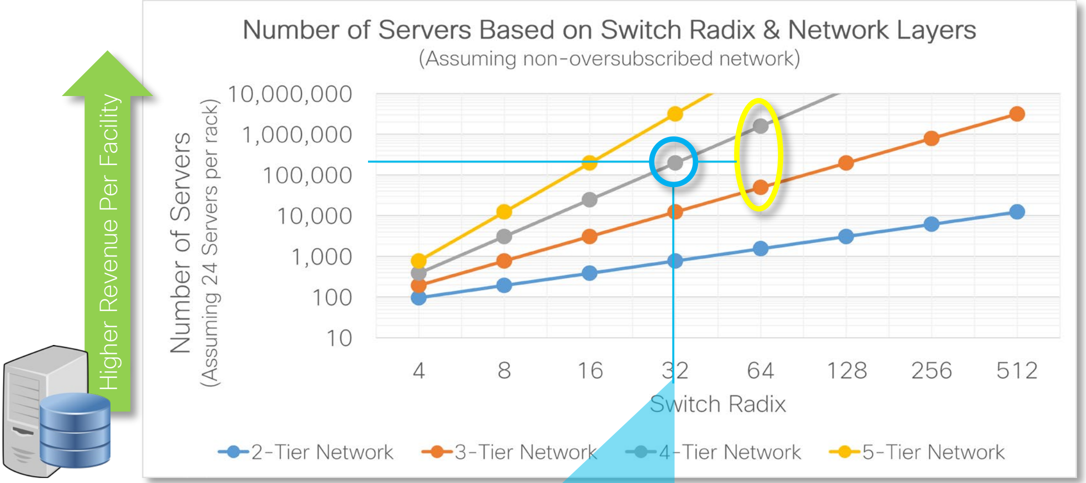
Graph concept leveraged from R. Nagarajan, Ilya Lyubomirsky, "Next-Gen Data Center Interconnects: The Race to 800G"
Adjusted to hold servers per rack constant

Switch BW	SerDes	Radix x32
12.8T	56G	400GE x8
25.6T	112G	800GE x8
51.2T	112G	1.6TE x16
102.4T?	212G?	3.2TE x16

- x32 and x128 radix are prominent today
  - Ethernet rates are lagging for x32 radix
  - Will x32 networks migrate to x64?

Building Your Data Center

Scale-Out vs. Scale-Up- A Balancing Act



Graph concept leveraged from R. Nagarajan, Ilya Lyubomirsky, "Next-Gen Data Center Interconnects: The Race to 800G"
Adjusted to hold servers per rack constant

Switch BW	SerDes	Radix x32	Radix x64
12.8T	56G	400GE x8	200GE x4
25.6T	112G	800GE x8	400GE x4
51.2T	112G	1.6TE x16	800GE x8
102.4T?	212G?	3.2TE x16	1.6TE x8

Wider Radix – Scale Out

More Layers – Scale Up

Improved Power Efficiency →

← Improved Link Utilization

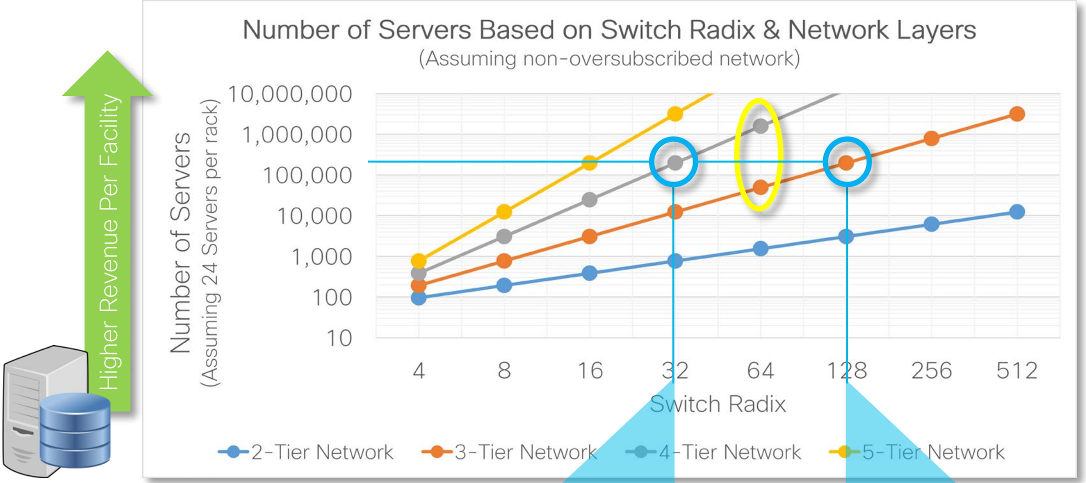
- x32 and x128 radix are prominent today
  - Ethernet rates are lagging for x32 radix
  - Will x32 networks migrate to x64?

Radix 64

- **Potential** need for 800GE with 8x112G Lanes
  - 51.2T
  - 64 x QSFP-DD800 (carrying 1x800GE) – 2RU
- **Potential** need for 1.6TE with 8x224G Lanes
  - 102.4T
  - 64 x QSFP-DD1600 (Carrying 1x1.6TE) – 2RU

Building Your Data Center

Scale-Out vs. Scale-Up- A Balancing Act



Graph concept leveraged from R. Nagarajan, Ilya Lyubomirsky, "Next-Gen Data Center Interconnects: The Race to 800G"
Adjusted to hold servers per rack constant

Switch BW	SerDes	Radix x32	Radix x64	Radix x128
12.8T	56G	400GE x8	200GE x4	100GE x2
25.6T	112G	800GE x8	400GE x4	200GE x2
51.2T	112G	1.6TE x16	800GE x8	400GE x4
102.4T?	212G?	3.2TE x16	1.6TE x8	800GE x4

Wider Radix – Scale Out

More Layers – Scale Up

Improved Power Efficiency

Improved Link Utilization

- x32 and x128 radix are prominent today
  - Ethernet rates are lagging for x32 radix
  - Will x32 networks migrate to x64?

Radix 64

- **Potential** need for 800GE with 8x112G Lanes
  - 51.2T
  - 64 x QSFP-DD800 (carrying 1x800GE) – 2RU
- **Potential** need for 1.6TE with 8x224G Lanes
  - 102.4T
  - 64 x QSFP-DD1600 (Carrying 1x1.6TE) – 2RU

Radix 128

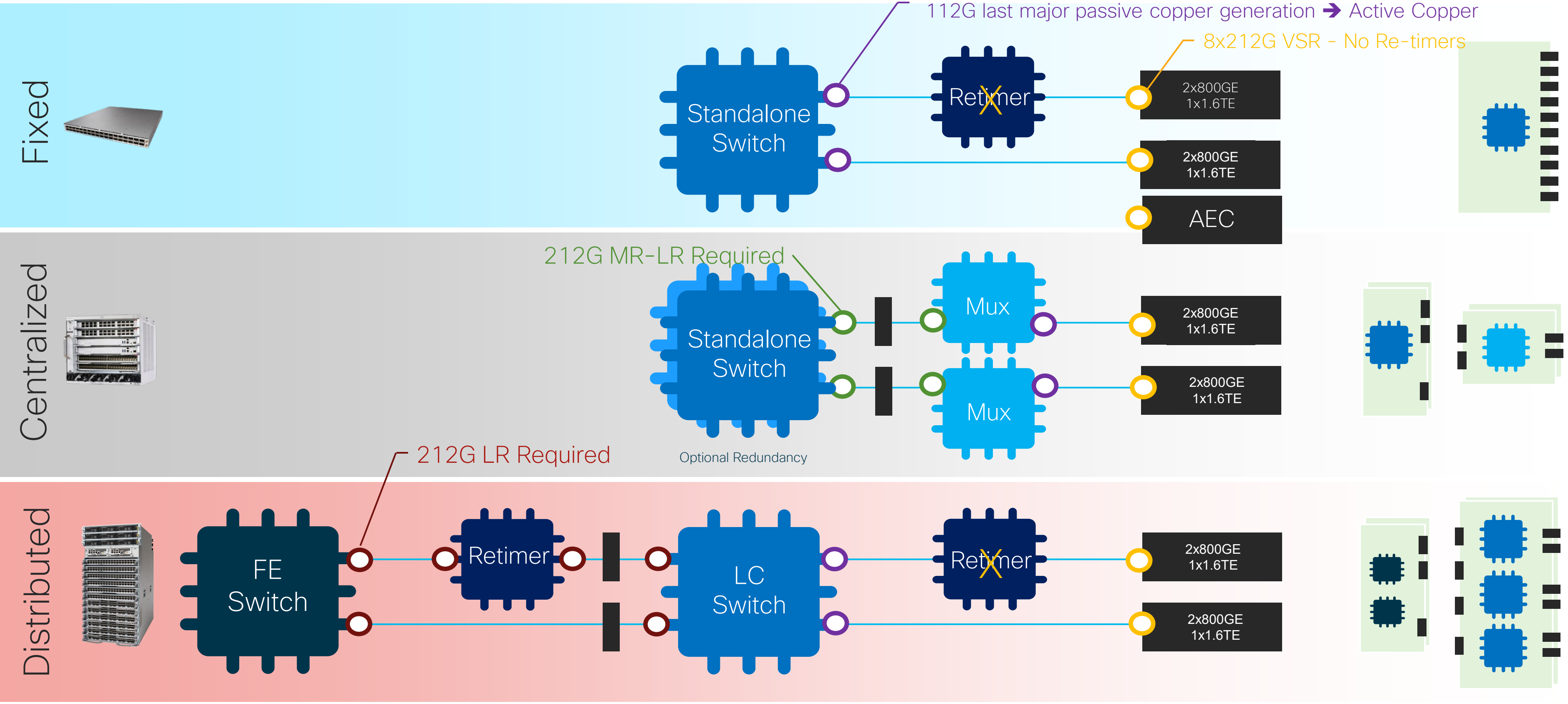
- **Clear** need for 800GE with 4x224G Lanes
  - 102.4T with 128-Radix
  - 128 x QSFP-800 (carrying 1x800GE) – 4RU

    or

  - 64 x QSFP-DD1600 (carrying 2x800GE)-2RU

212G Generation Traditional System Architectures

Viable with Traditional System Designs



VSR – Optimize for Optics

112G last major passive copper generation ➔ Active Copper

8x212G VSR – No Re-timers

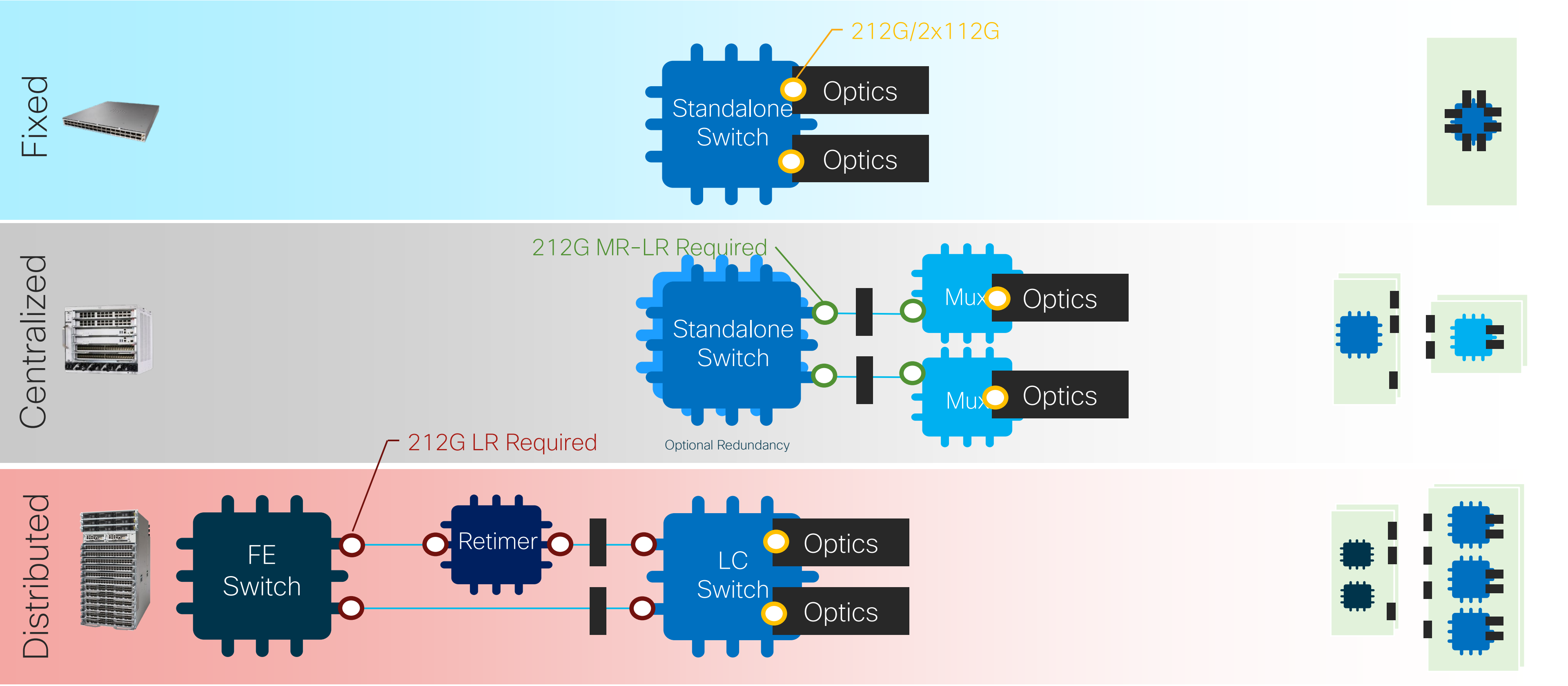
212G MR-LR Required

Optional Redundancy

212G LR Required

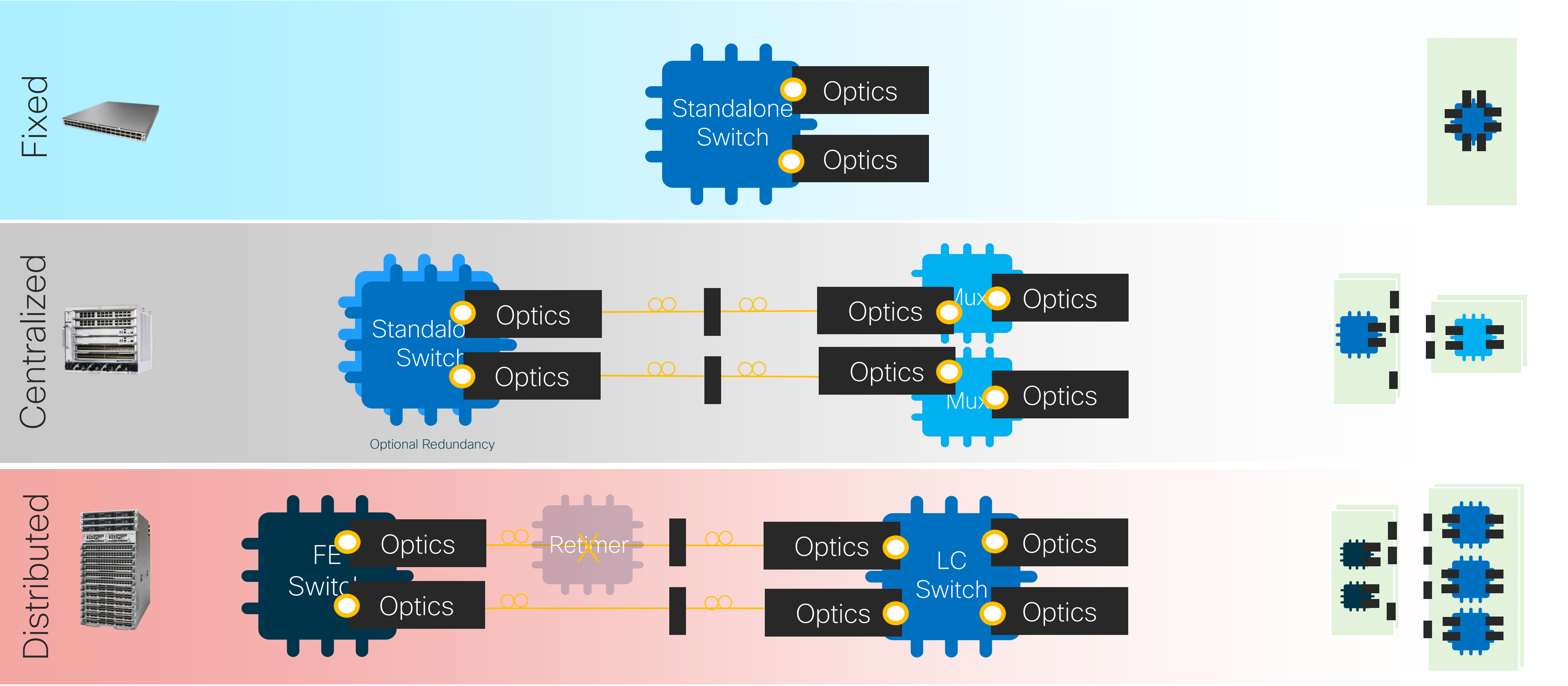
212G Generation CPO System Architectures

Power Optimized ; Introduced first on Client-Side Optics



Future CPO Architectures

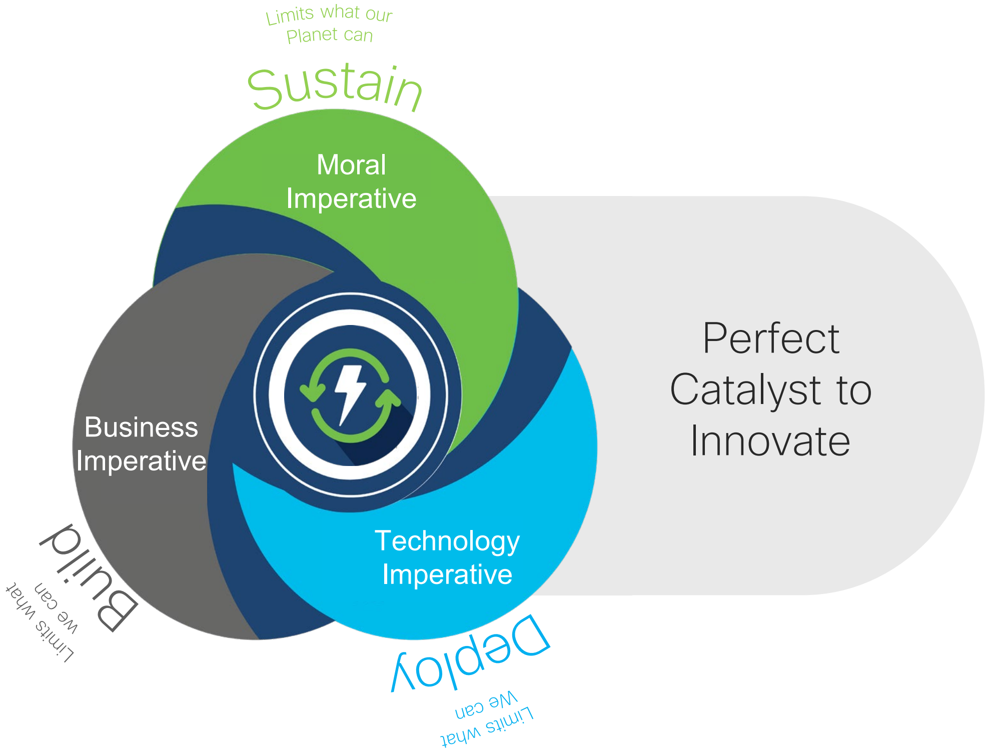
Eventually Optics replace high speed data interconnect



Call to Action

Power Driven Architecture

- ✓ 3 Main System Architectures
  Fixed, Centralized, Modular
- ✓ BW Doubling every 2 Years
  Not Slowing Down, Power Too High
- ✓ Co-package Optics are Coming
  51.2T Generation



Next Steps

Looking for study group to define a cost and power effective solution to these problems

Increasing Priority

1 Define 212G Electrical

- XSR, VSR as first priority to optimize power efficiency
  - Define VSR standard to ensure retimer-less designs
- Define MR, LR as second priority
- Focus on 212G* instead of 224G to optimize for Ethernet rates

2 Define 800GE MAC

- Over 212G to enable 102.4T with radix 128 (128x800GE)
- Over 112G to enable 51.2T with radix 64 (64x800GE)

3 Define 1.6TE MAC

- Only if there is a cost effective PMD solution
- Over 212G to enable 102.4T with radix 64 (64x1.6TE)

* - Final rate depends on future work

Thank You!