KEYNOTE PRESENTATION

LOOKING BEYOND 400G - A SYSTEM VENDOR PERSPECTIVE

Presenter: Rakesh Chopra, *Cisco Systems*

TECHNOLOGY EXPLORATION FORUM

ethernet alliance

www.ethernetalliance.org

Looking Beyond 400G

A System Vendor Perspective

Rakesh Chopra
Cisco Fellow
January 25, 2020

Many thanks to Cisco engineers, insightful customers, and amazing partners …


rakchopr@cisco.com

www.linkedin.com/in/rakesh-chopra/

@Rakesh_Chopra1

System Architectures

Fixed*

Standalone Switch

Retimer

Port

Port

* - Can be interconnected to create a disaggregated chassis

Centralized

Standalone Switch

Optional Redundancy

Mux

Mux

Port

Port

Distributed

FE Switch

Retimer

LC Switch

Retimer

Port

Port

LR

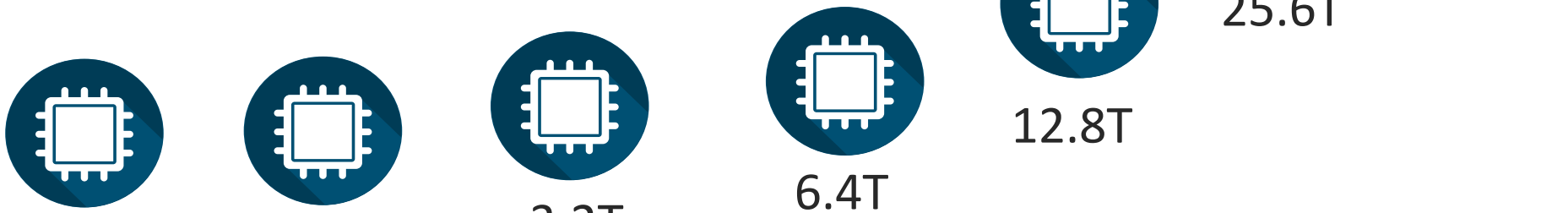
VSR

Relentless Advancement – Switch Silicon Bandwidth

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

102.4T?

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
# SerDes	x64	x128	x256			x512
SerDes Speed	10G	28G		56G		112G
Optics	QSFP+	QSFP28		QSFP-DD56		QSFP-DD800

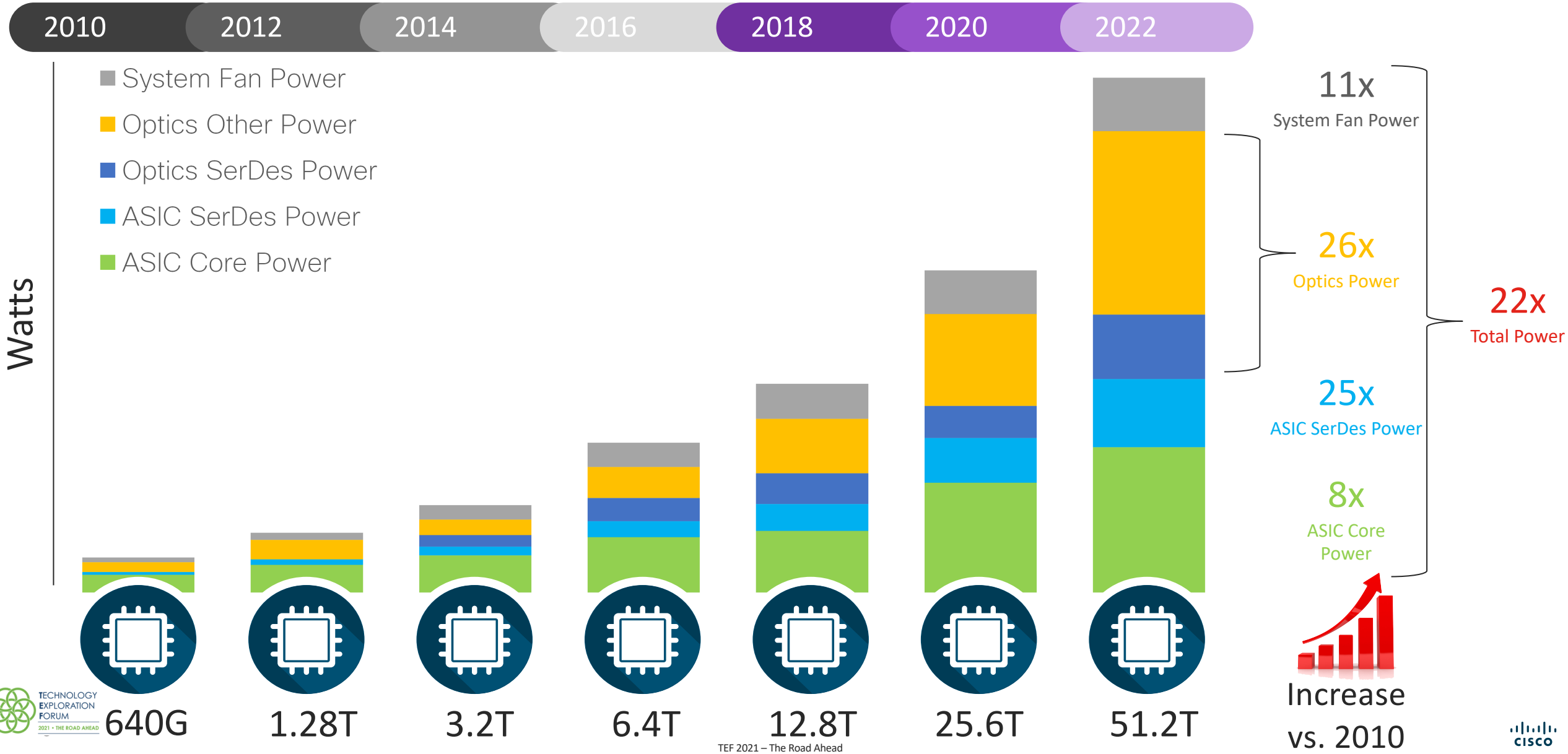
?

TECHNOLOGY EXPLORATION FORUM 2021 • THE ROAD AHEAD

cisco

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



Power is THE Problem to Solve

Apollo 13 – Universal Pictures

"Power is Everything"*

John Aaron- Apollo 13 Flight Controller

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

Adopt a power first design and deployment methodology

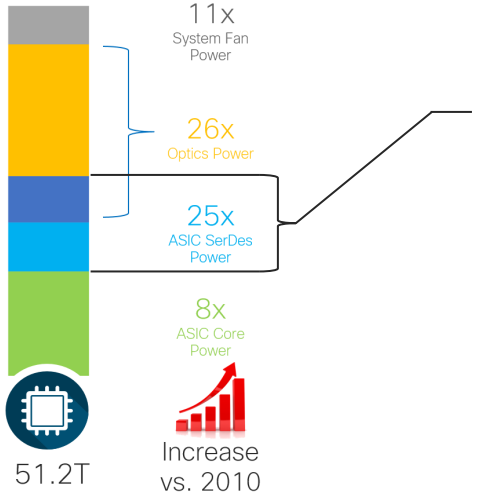
TECHNOLOGY EXPLORATION FORUM 2021 • THE ROAD AHEAD

* - Thanks to Kraig Owen for the reference

cisco

Co-packaged Optics Is Inevitable

Power savings drives requirement



Must minimize SerDes power

SerDes power increases with distance

Trends plot from premier CMOS wireline 2018 conference

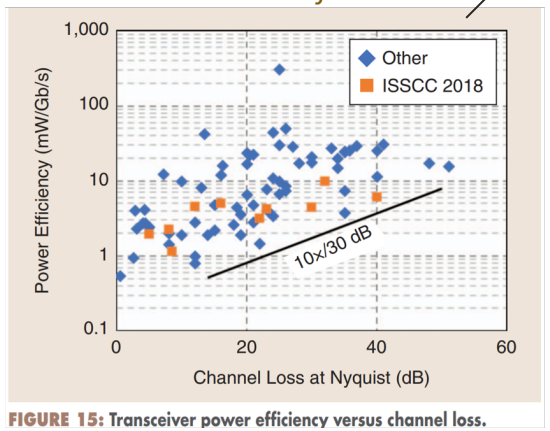
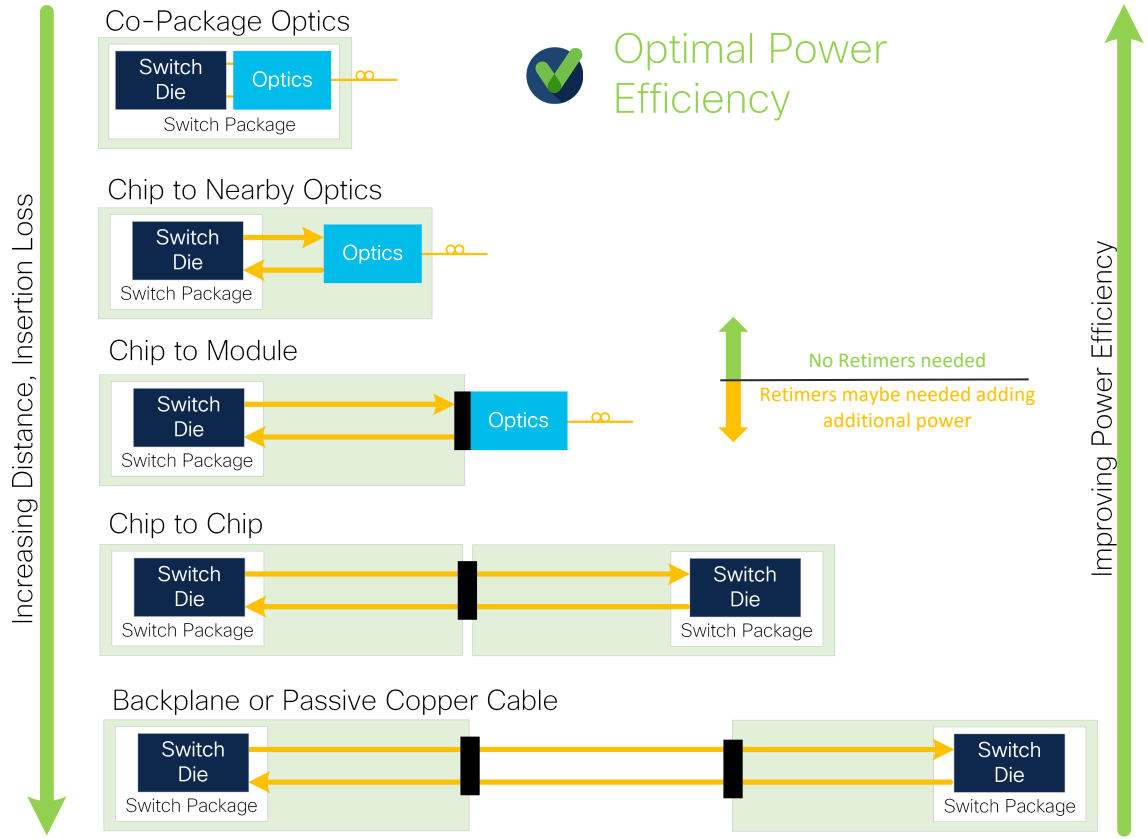


FIGURE 15: Transceiver power efficiency versus channel loss.

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-packaged Optics Is Inevitable

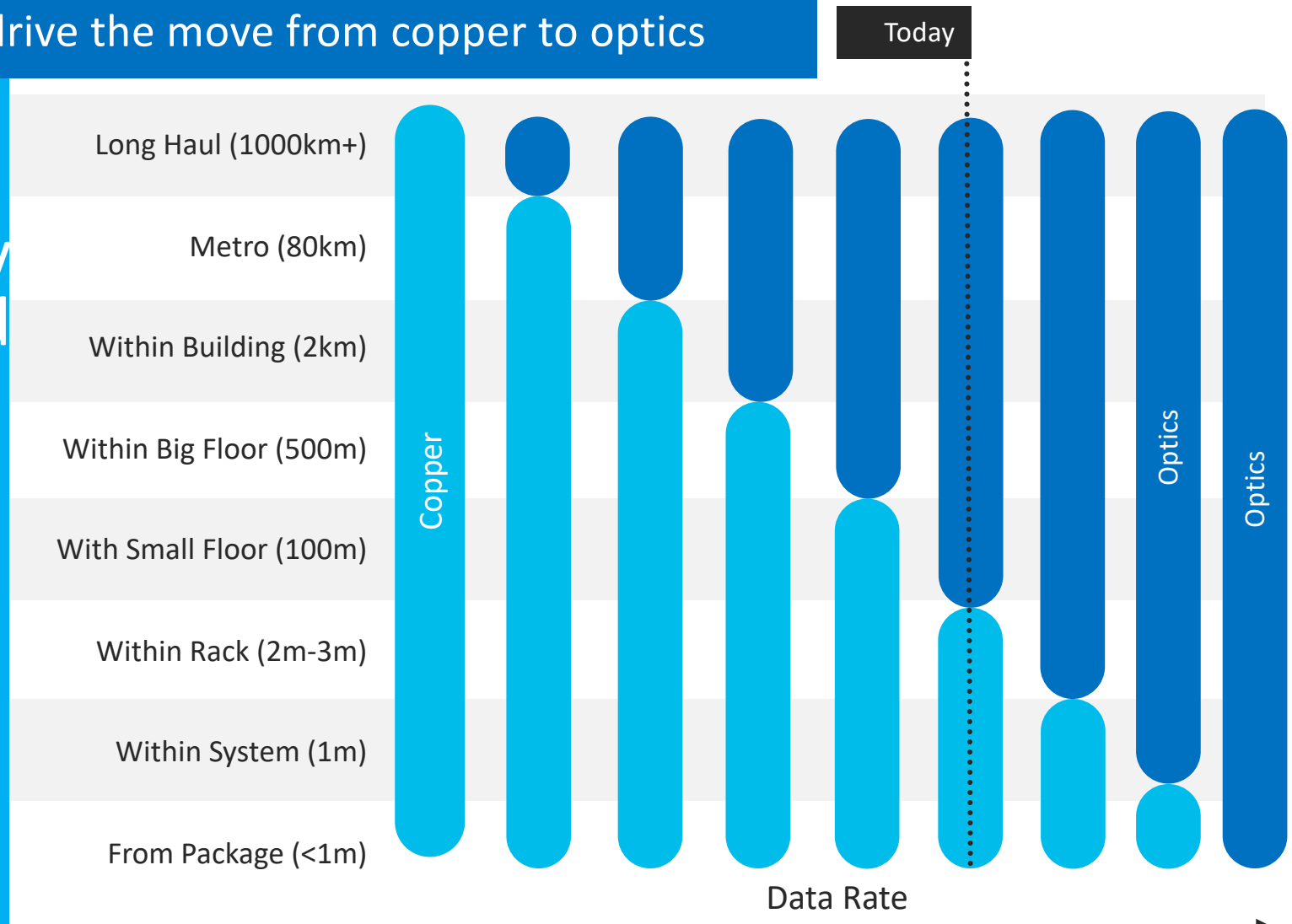
and viable in the 51.2T generation

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

Future innovations will only be possible with **silicon** and **optical** integration

Si Silicon + **Op** Optics

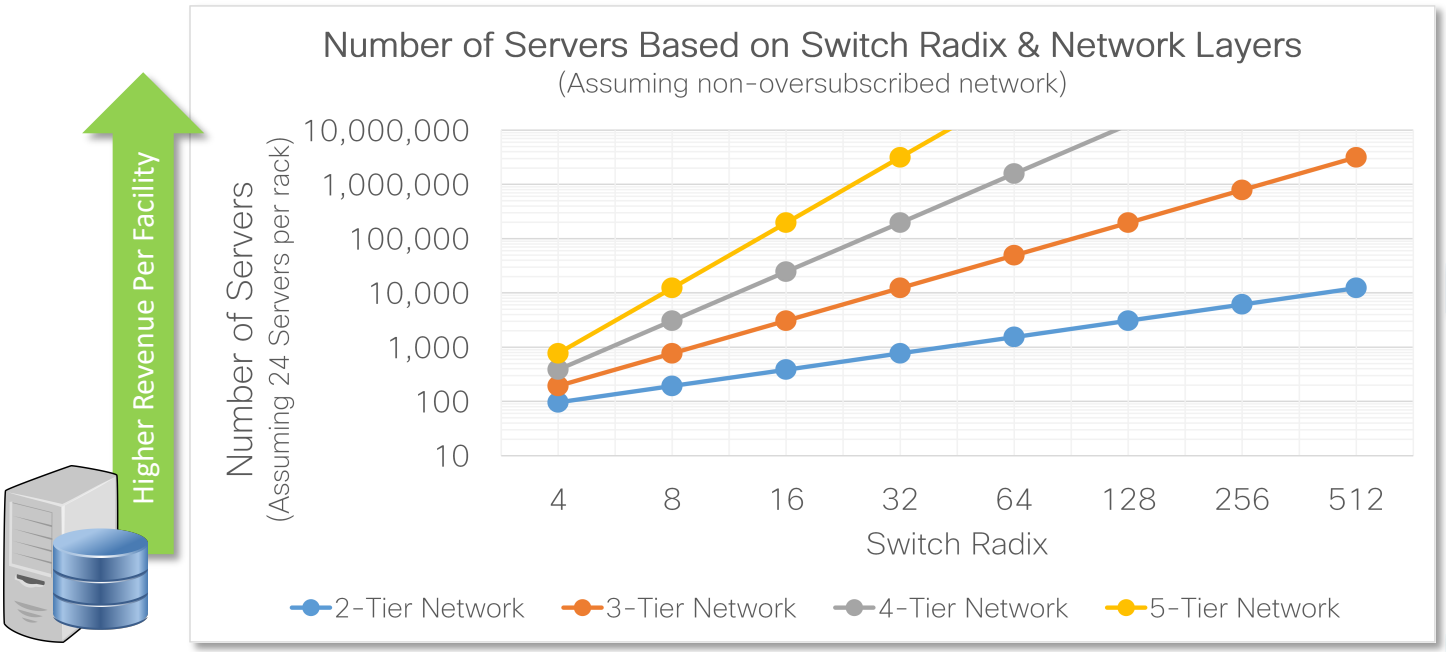
51.2T



TECHNOLOGY EXPLORATION FORUM 2021 • THE ROAD AHEAD

Building Your Data Center

Impact of Switch Radix

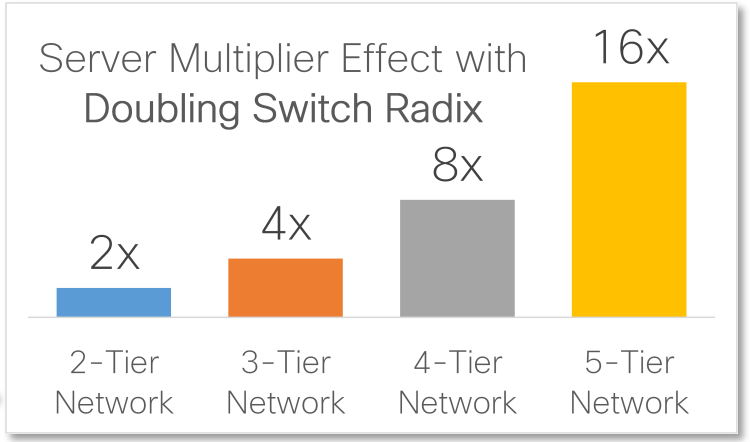


Graph concept leveraged from R. Nagarajan, Ilya Lyubomirsky, "Next-Gen Data Center Interconnects: The Race to 800G"
Adjusted to hold servers per rack constant

Doubling Radix adds 2x-16x more servers

Scale Out

Wider Radix

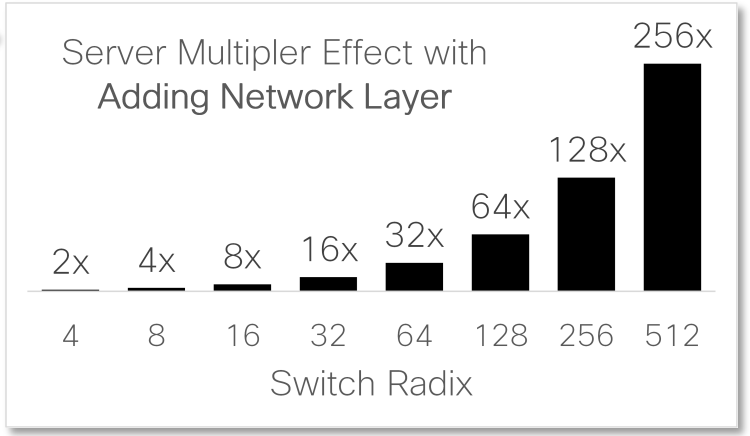


Complex Cabling
Worse ECMP Hashing
Lower Link Utilization

Adding a layer adds 2x-256x more servers

Scale Up

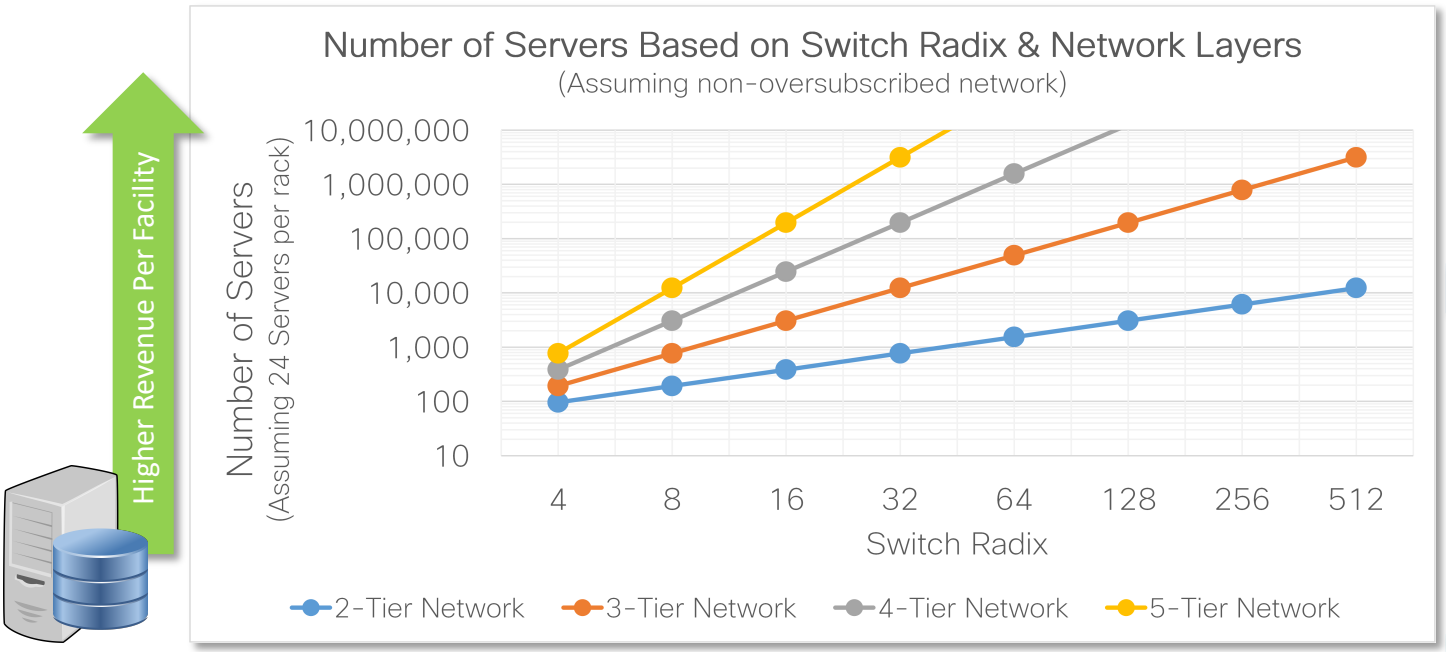
More Layers



Higher Networking Power Per Server

Building Your Data Center

Impact of Switch Radix

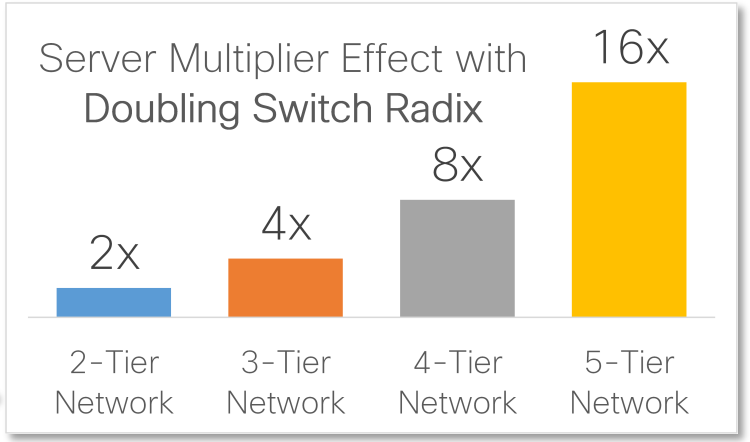


Graph concept leveraged from R. Nagarajan, Ilya Lyubomirsky, "Next-Gen Data Center Interconnects: The Race to 800G"
Adjusted to hold servers per rack constant

Doubling Radix adds 2x-16x more servers

Scale Out

Wider Radix

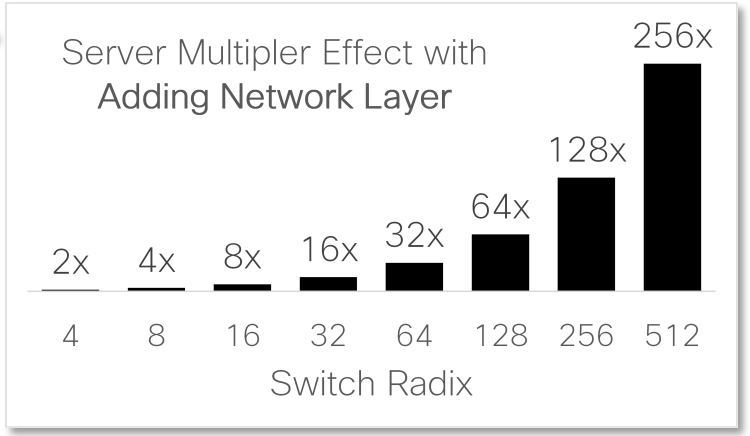


Power Efficiency

Adding a layer adds 2x-256x more servers

Scale Up

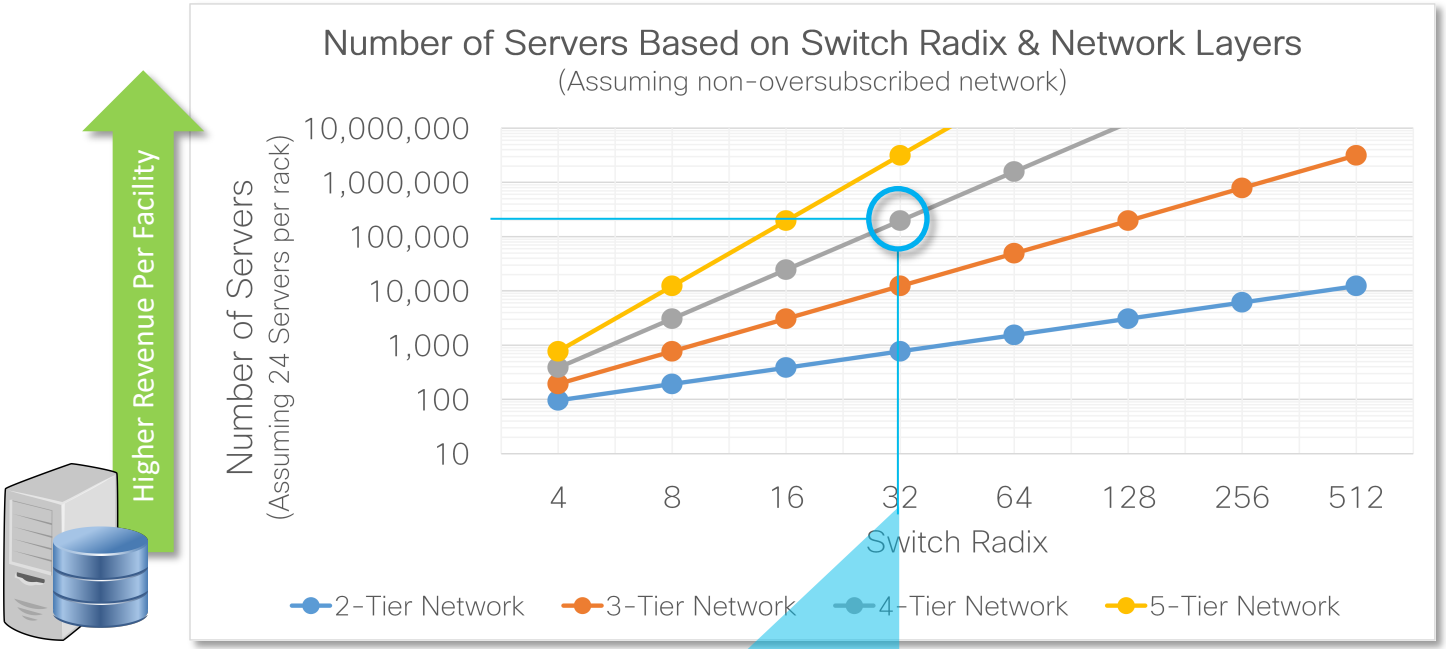
More Layers



Link Efficiency

Building Your Data Center

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



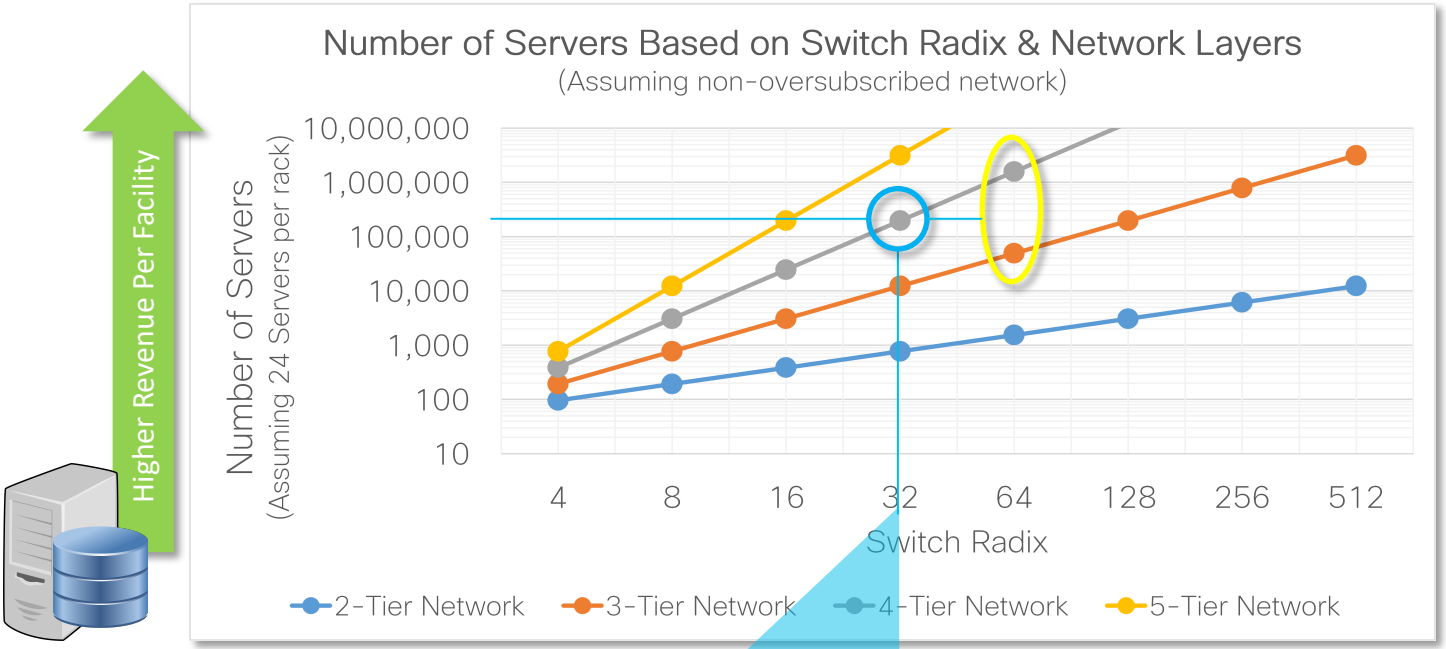
Graph concept leveraged from R. Nagarajan, Ilya Lyubomirsky, "Next-Gen Data Center Interconnects: The Race to 800G"
Adjusted to hold servers per rack constant

Switch BW	SerDes	Radix x32
12.8T	56G	400GE x8
25.6T	112G	800GE x8
51.2T	112G	1.6TE x16
102.4T?	224G	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?	224G	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

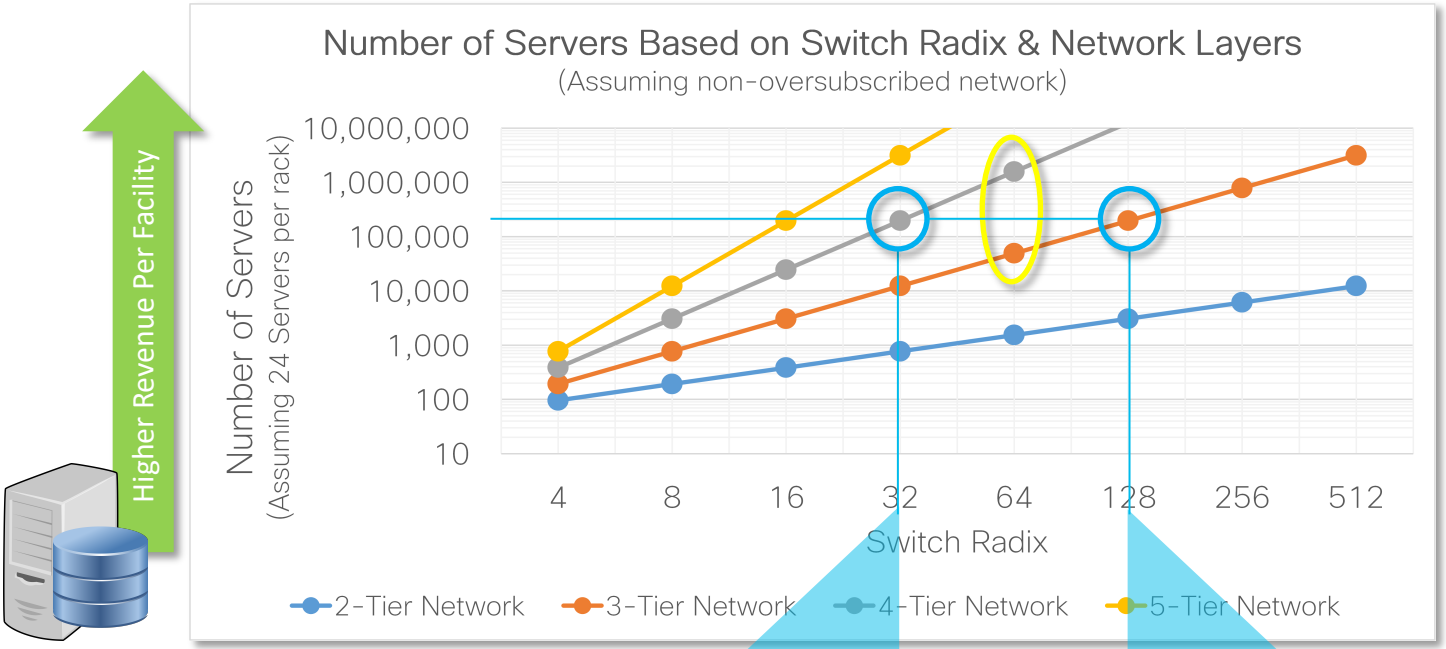
- **Potential** need for 800GE with 8x112G Lanes
  - 51.2T
  - 64 x QSFP-DD800 (carrying 1x800GE) – 2RU
- **Potential** need for 1.6TE with 8x224G Lanes
  - 102.4T
  - 64 x QSFP-DD1600 (Carrying 1x1.6TE) – 2RU

TECHNOLOGY EXPLORATION FORUM 2021 • THE ROAD AHEAD

cisco

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?	224G	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

224G Generation Traditional System Architectures

Viable with Traditional System Designs

VSR - Optimize for Optics

112G last major passive copper generation ➔ Active Copper

8x224G VSR - No Re-timers

Fixed

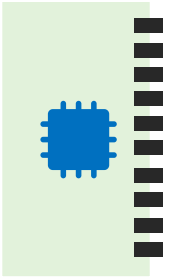
Standalone Switch

Retimer

2x800GE 1x1.6TE

2x800GE 1x1.6TE

AEC



224G MR-LR Required

Centralized

Standalone Switch

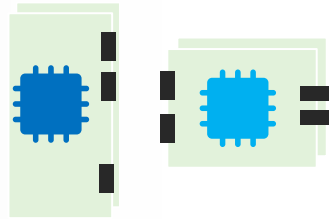
Optional Redundancy

Mux

Mux

2x800GE 1x1.6TE

2x800GE 1x1.6TE



224G LR Required

Distributed

FE Switch

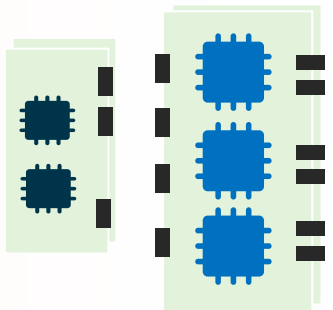
Retimer

LC Switch

Retimer

2x800GE 1x1.6TE

2x800GE 1x1.6TE



224G Generation CPO System Architectures

Power Optimized ; Introduced first on Client-Side Optics

Fixed

224G/2x112G

Standalone Switch

Optics

Optics

Centralized

224G MR-LR Required

Standalone Switch

Mux

Optics

Mux

Optics

Optional Redundancy

Distributed

224G LR Required

FE Switch

Retimer

LC Switch

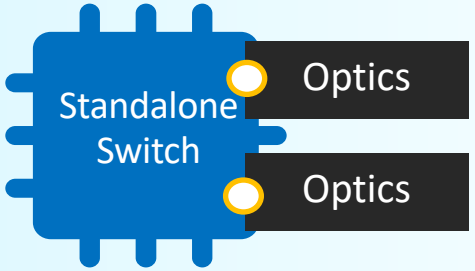
Optics

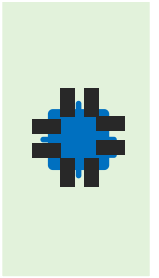
Optics

Future CPO Architectures

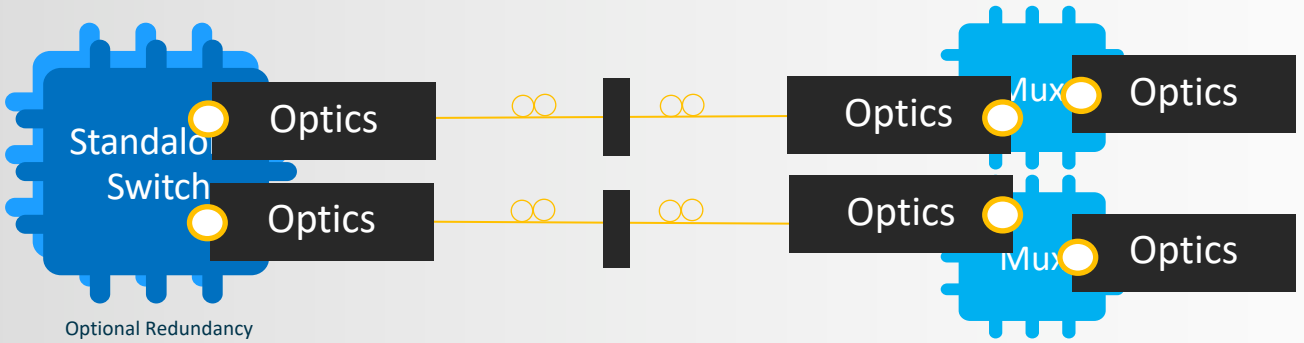
Eventually Optics replace high speed data interconnect

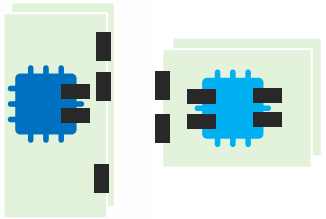
Fixed



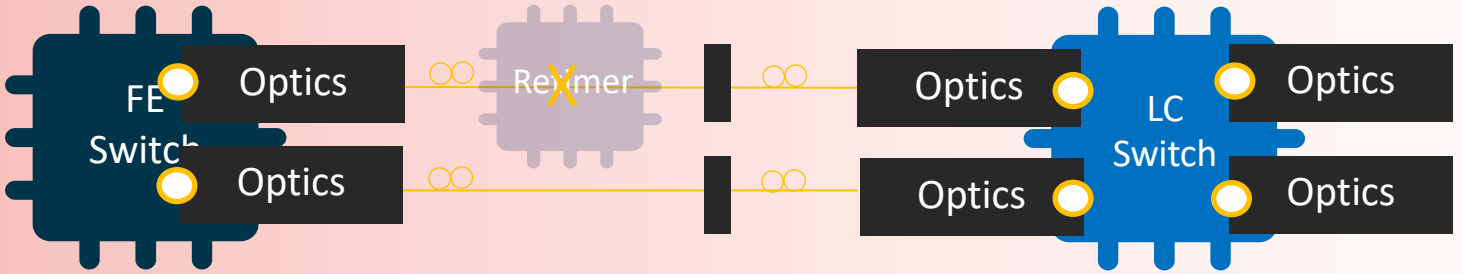


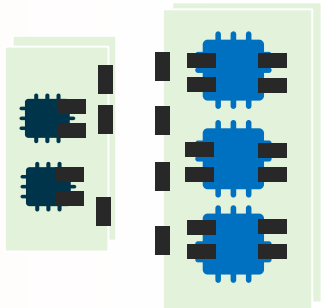
Centralized





Distributed

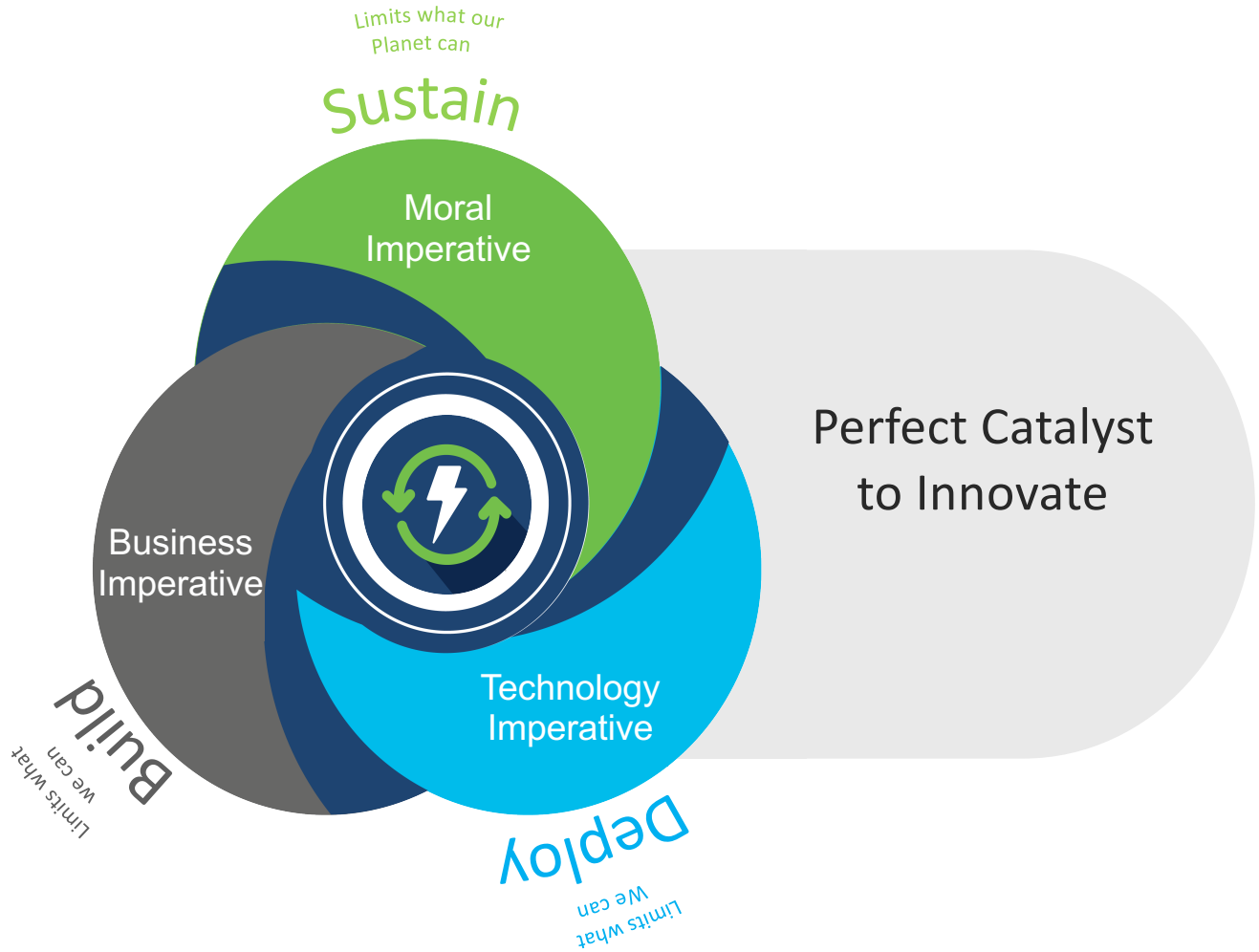




Call to Action

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
- **Must start on 224G … Yesterday**
  But let's do it right
- **800GE / 1.6TE Coming**
  800GE vs 1.6TE is just timing
  Importance is 224G development




CISCO

For our TEF 2021 on-demand content go to bit.ly/EATEF2021-OD

If you have any questions or comments, please email **admin@ethernetalliance.org**

@ethernetalliance

@EthernetAllianc

Ethernet Alliance