

OPEN.



FOR
BUSINESS.



OCP
SUMMIT

OPEN.



FOR
BUSINESS.



OCP
SUMMIT



Distributed Compute

Hans-Juergen Schmidtke

Director of Network Hardware Engineering

FACEBOOK

OPEN. FOR BUSINESS.





100 Million

Hours of video watched
on Facebook daily



95+ Million

Photos and videos posted
to Instagram daily



400 Million

Voice and video chat
on Messenger monthly



Prineville, OR



Forest City, NC



Luleå, Sweden



Altoona, IA



Huntsville, AL



Fort Worth, TX



Clonee, Ireland



Los Lunas, NM



Odense, Denmark



Eagle Mountain, UT



Papillion, NE



New Albany, OH



Henrico, VA

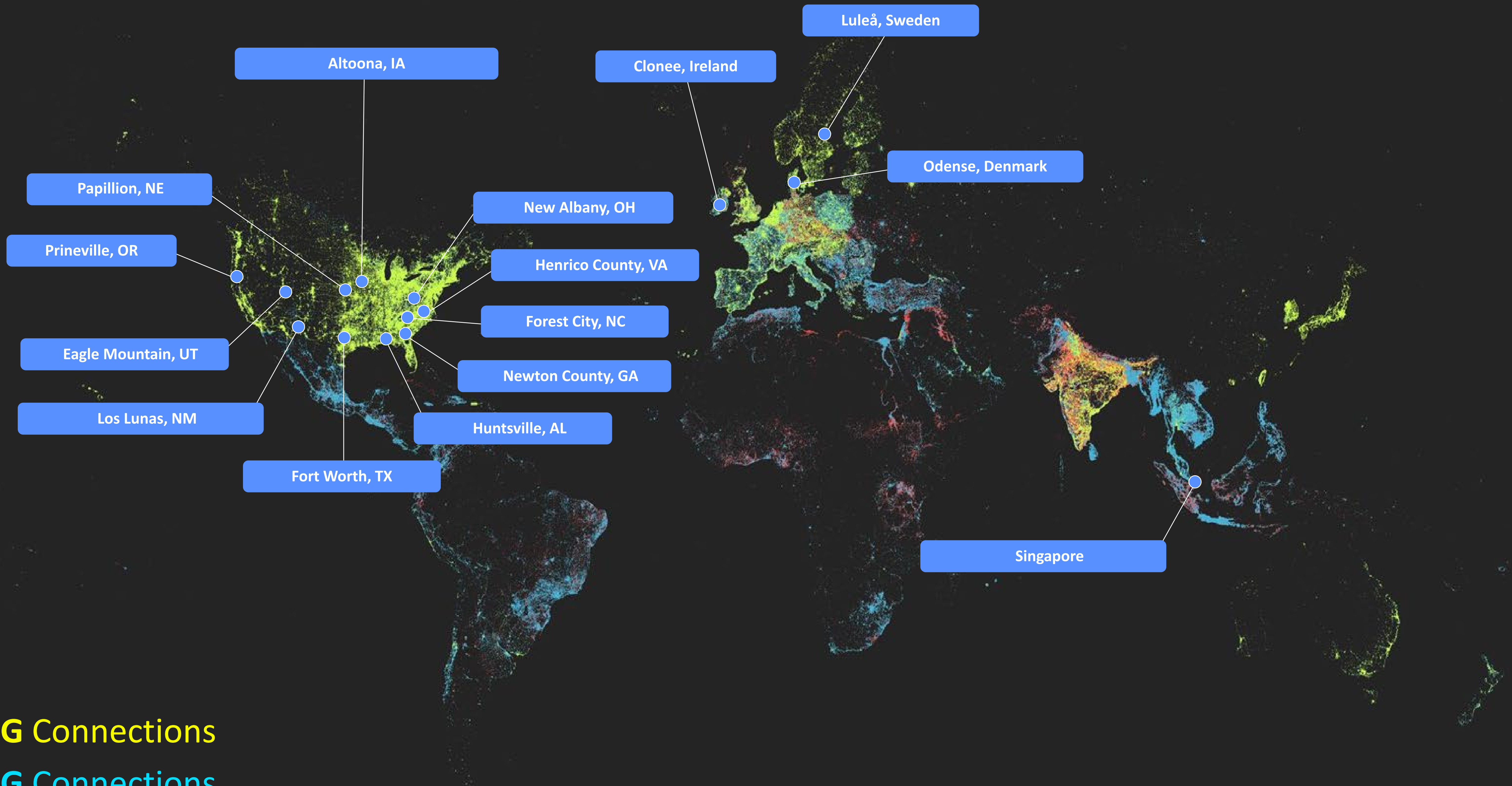


Newton County, GA



Singapore

4G Connections
3G Connections
2G Connections



“Wow, Italy is
so beautiful!”



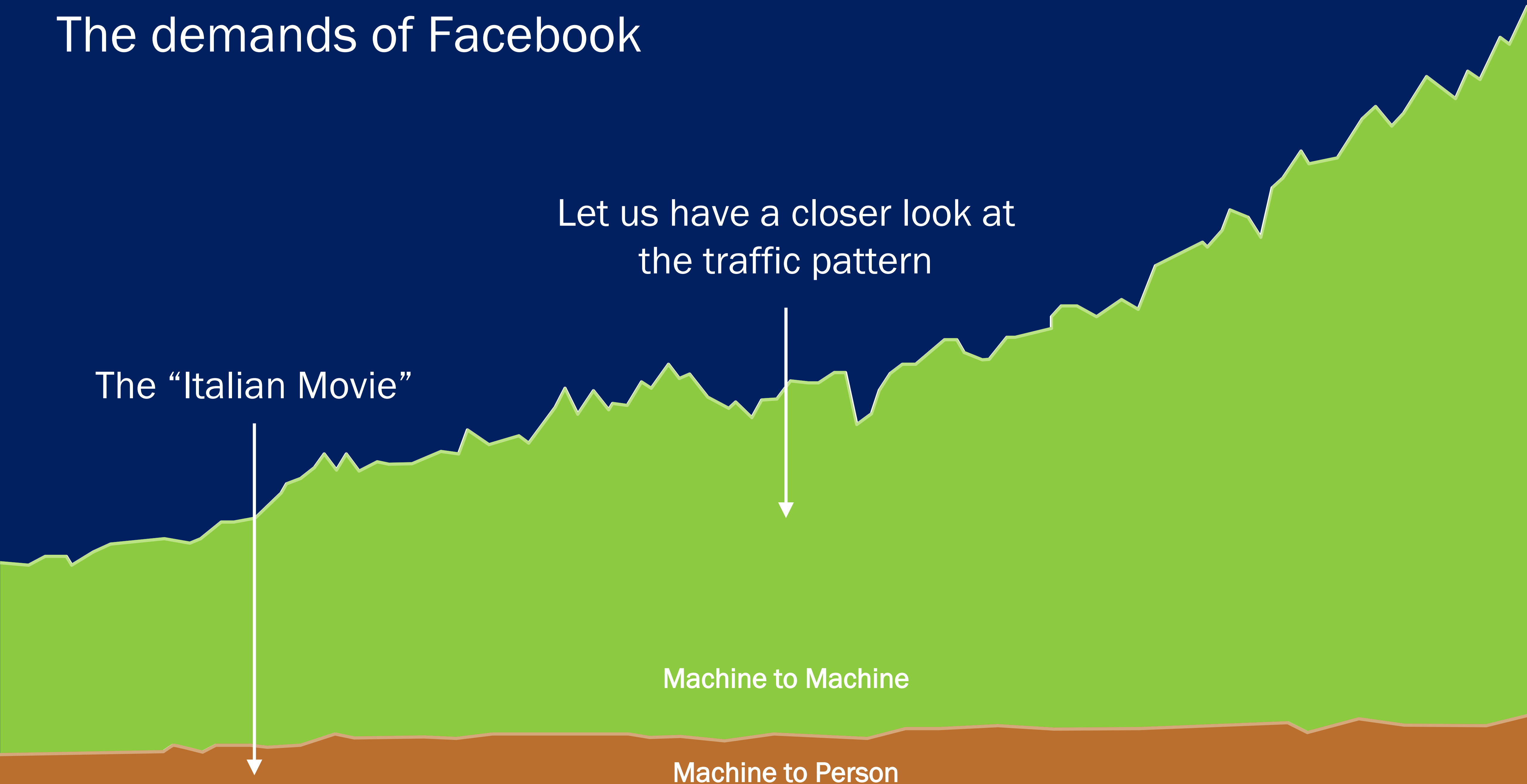
The demands of Facebook

Let us have a closer look at
the traffic pattern

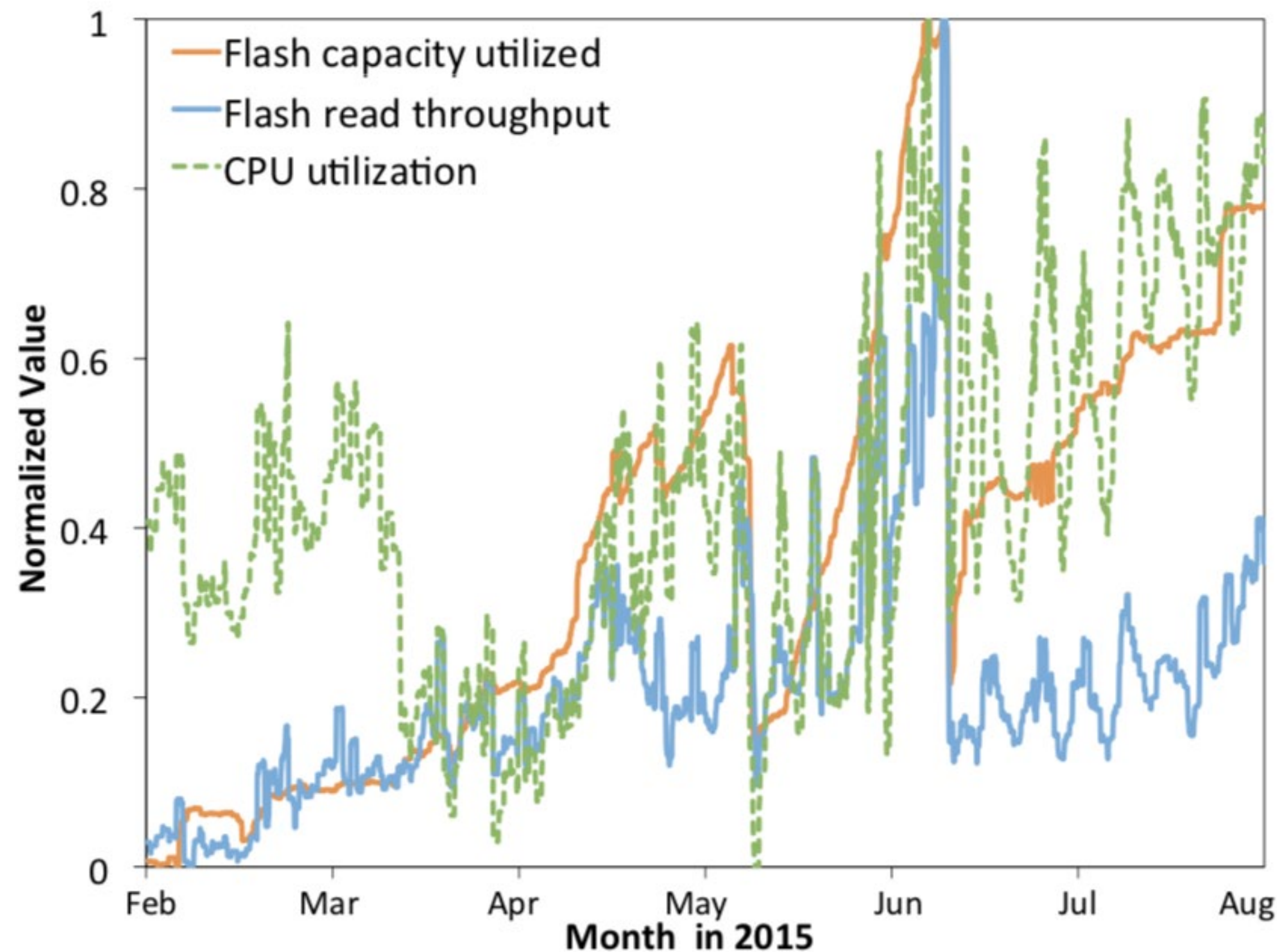
The “Italian Movie”

Machine to Machine

Machine to Person



Workload: Example of server traffic patterns

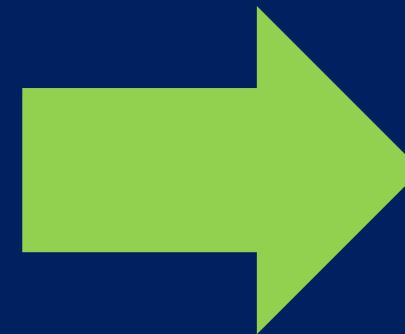


- Prediction of resources for CPU, Flash and I/O Flash difficult
- Flash utilization at times fairly low
- Requires scaling of Compute, Storage and Network resources independently
- Dilemma: Utilization of dense flash devices require more compute clients than what fits in one rack

Approaches for DC Optimization

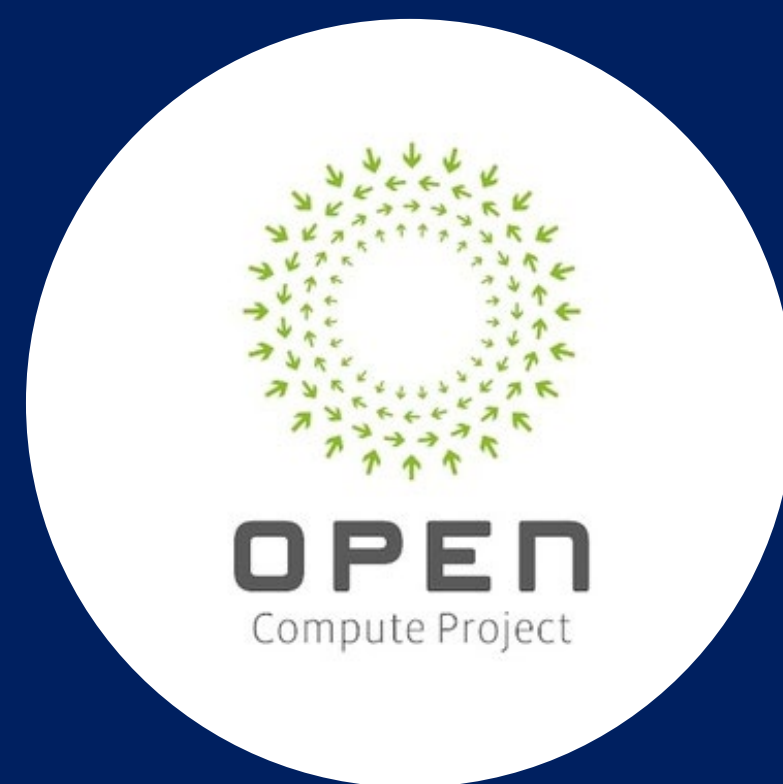
Resources to be optimized for:

- People
- Scale
- Power
- Thermal / Airflow
- Space
- Economical deployment
- Time



Decisions to be balanced:

- Operations over Feature
- Scale out vs. Scale up
- SW and HW
- Commodity over specialization, use or enable economy of scale



OCP 2011

Hardware Disaggregation

Open Interfaces /
Implementations



TIP 2016

End-to-end Solutions

Community Collaboration

Storage (HDD / Flash)



Bryce Canyon

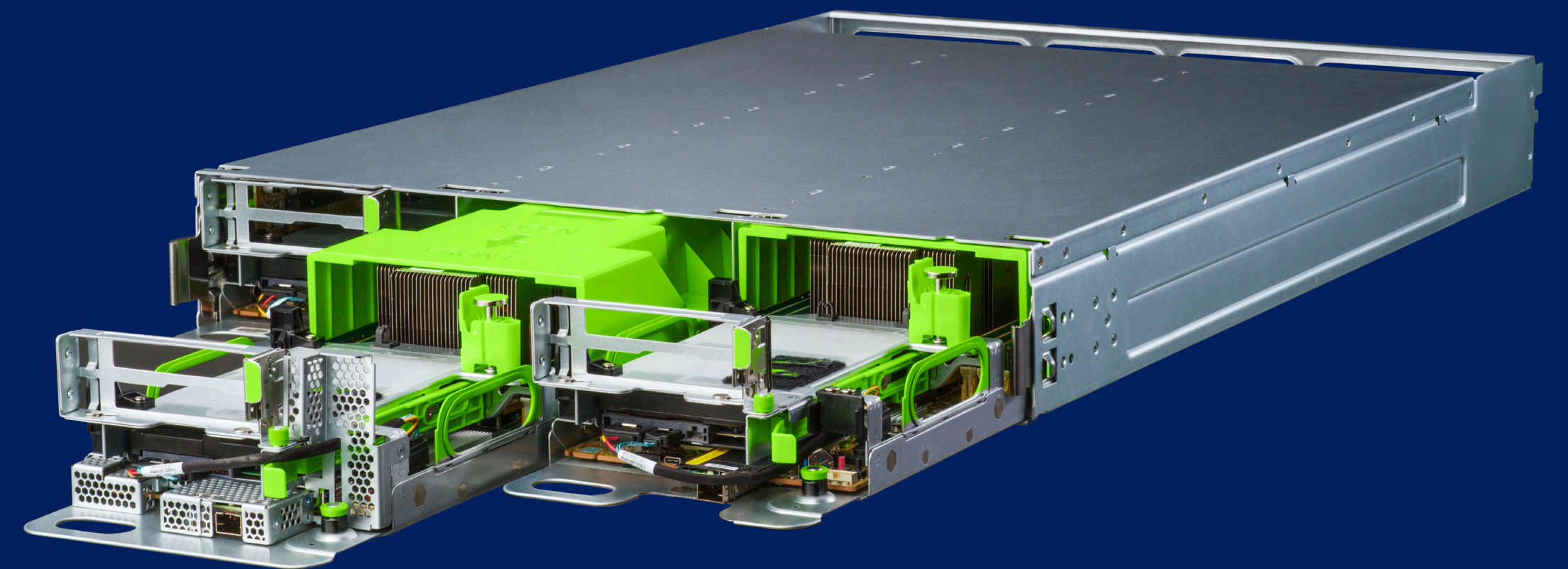


Lightning

Memory

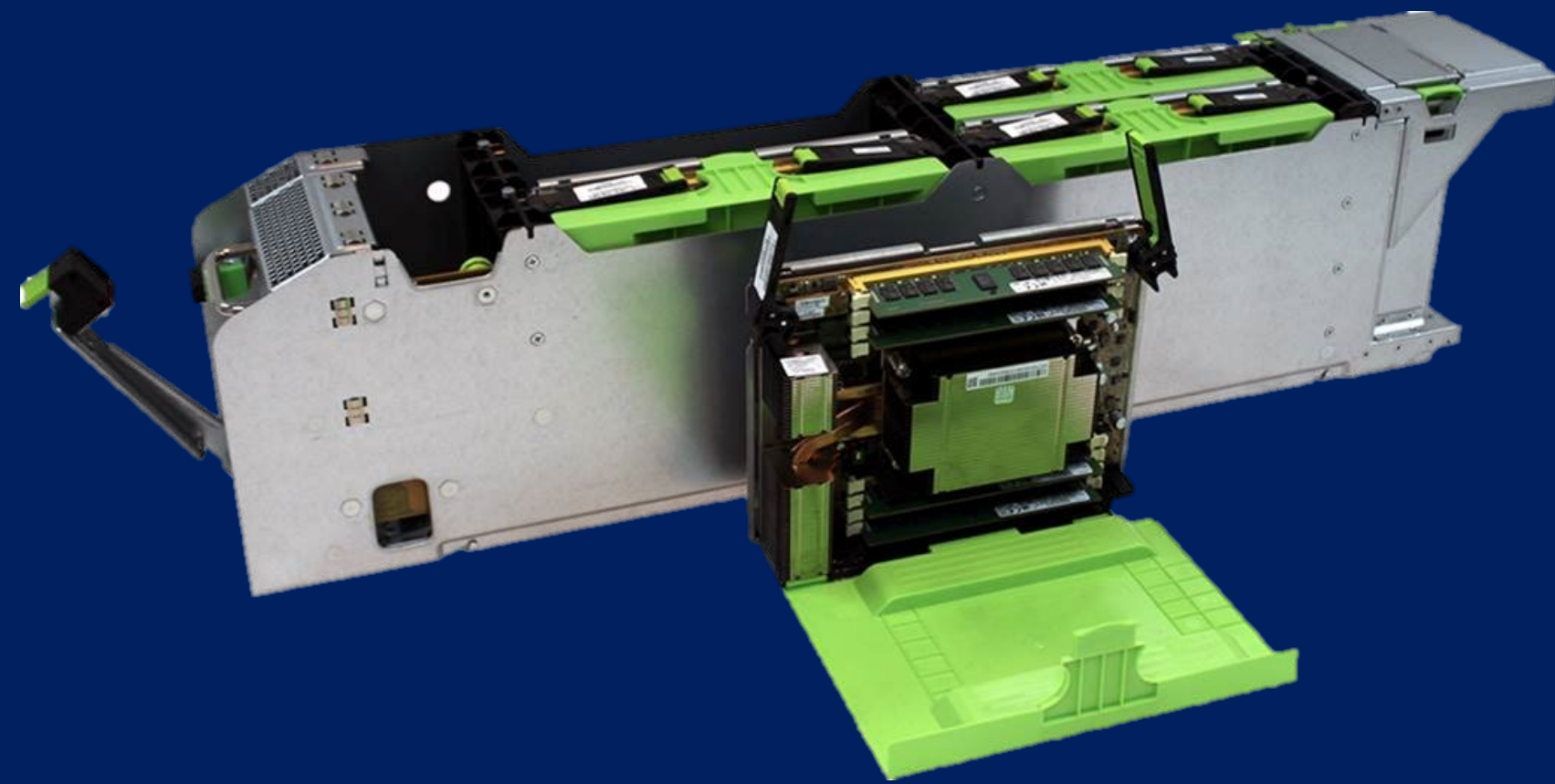


Tioga Pass Dual-Socket CPU Sled



Tioga Pass Chassis

Compute

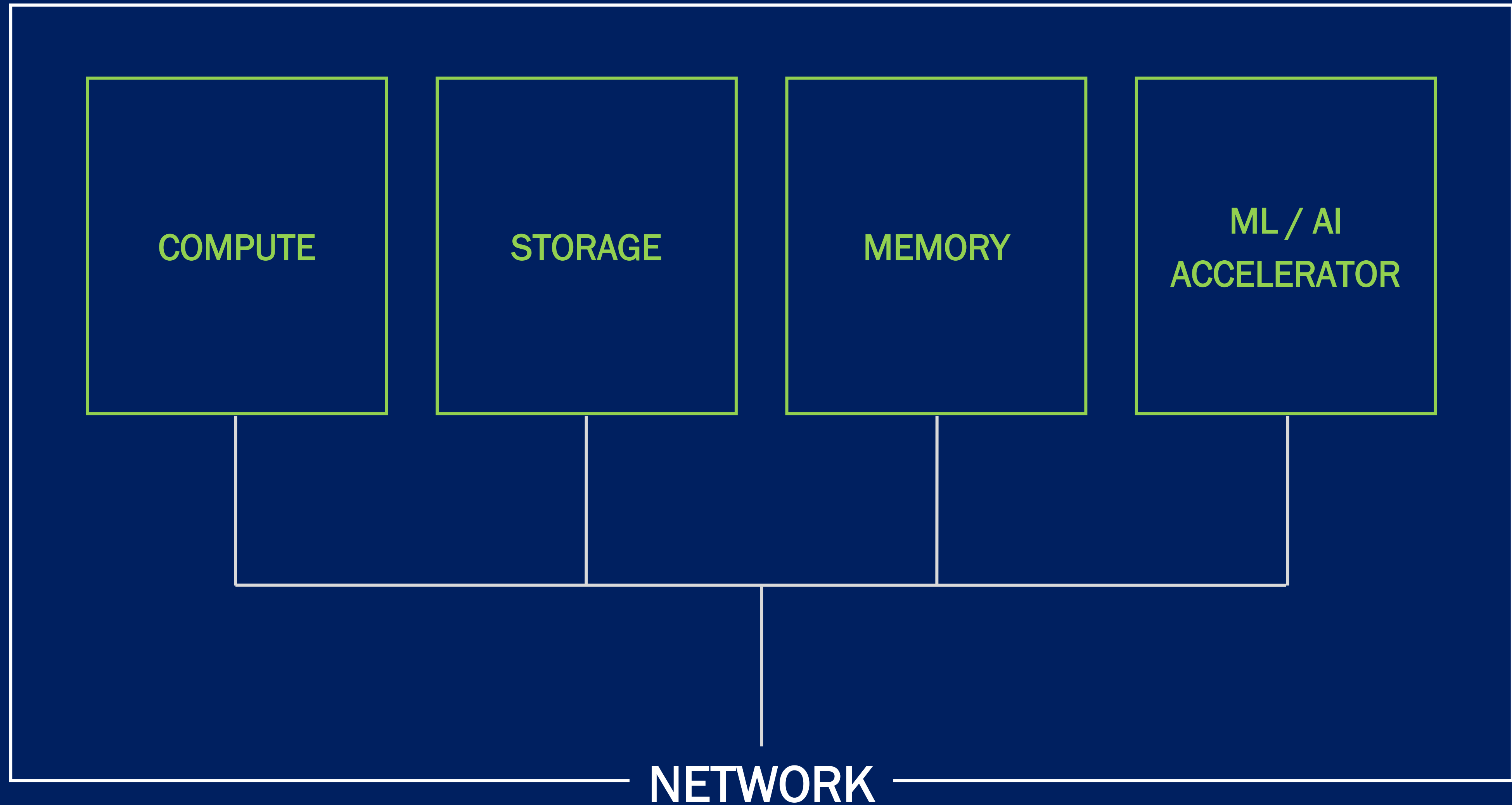


Twin Lakes Server Card



Yosemite V2 Chassis

DATA CENTER



CPU Performance: Demand and Supply. Expectation and Experience

Relative Performance Increase

2012

2018

CPU Generation (Baseline 2012)

— Experience

- - Expectation

What was

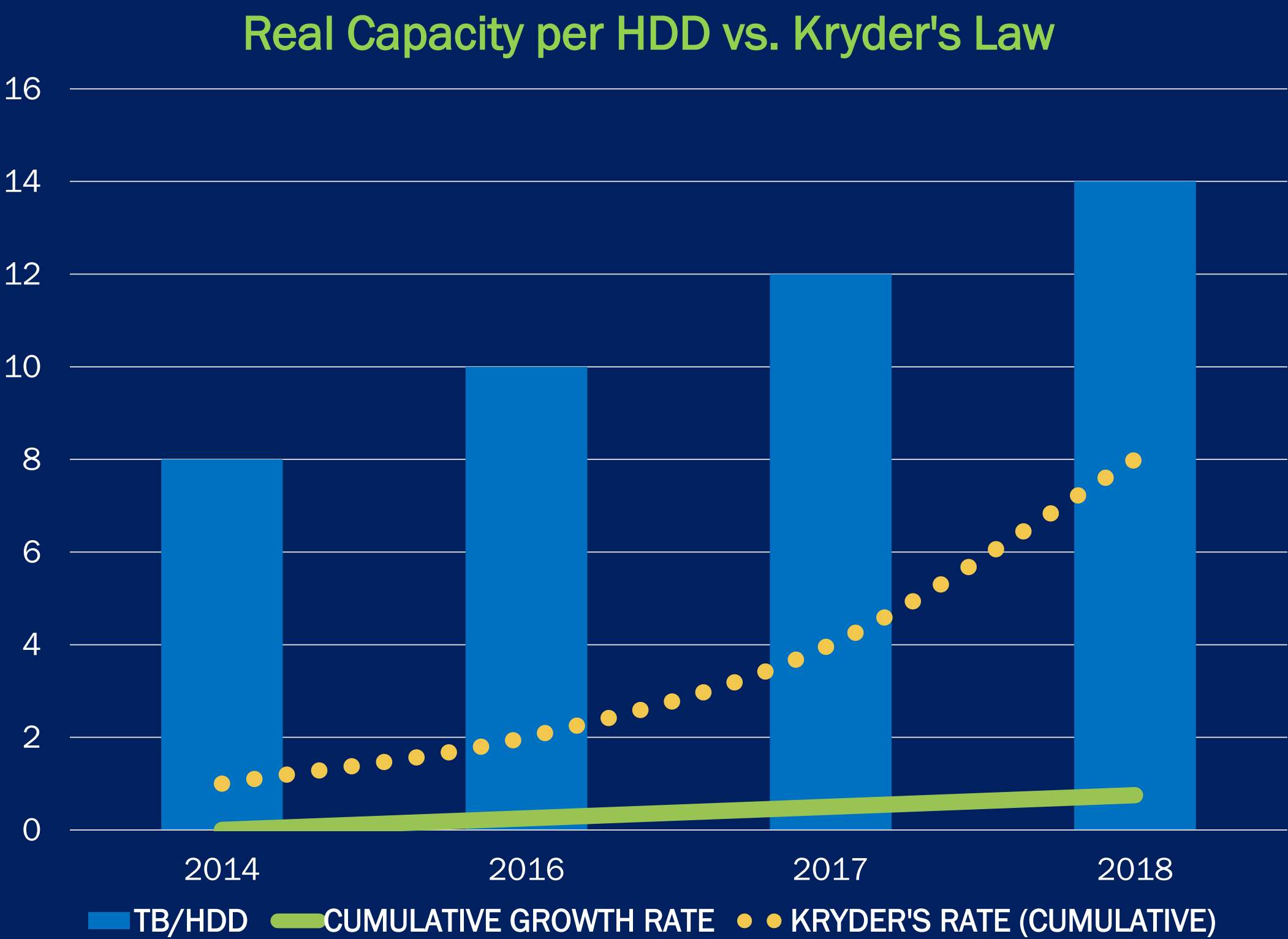
What is

The Gap

Sources for: Yellow Line: Wikipedia
https://en.wikipedia.org/wiki/Moore%27s_law
https://en.wikipedia.org/wiki/Dennard_scaling
https://en.wikipedia.org/wiki/Pollack%27s_rule

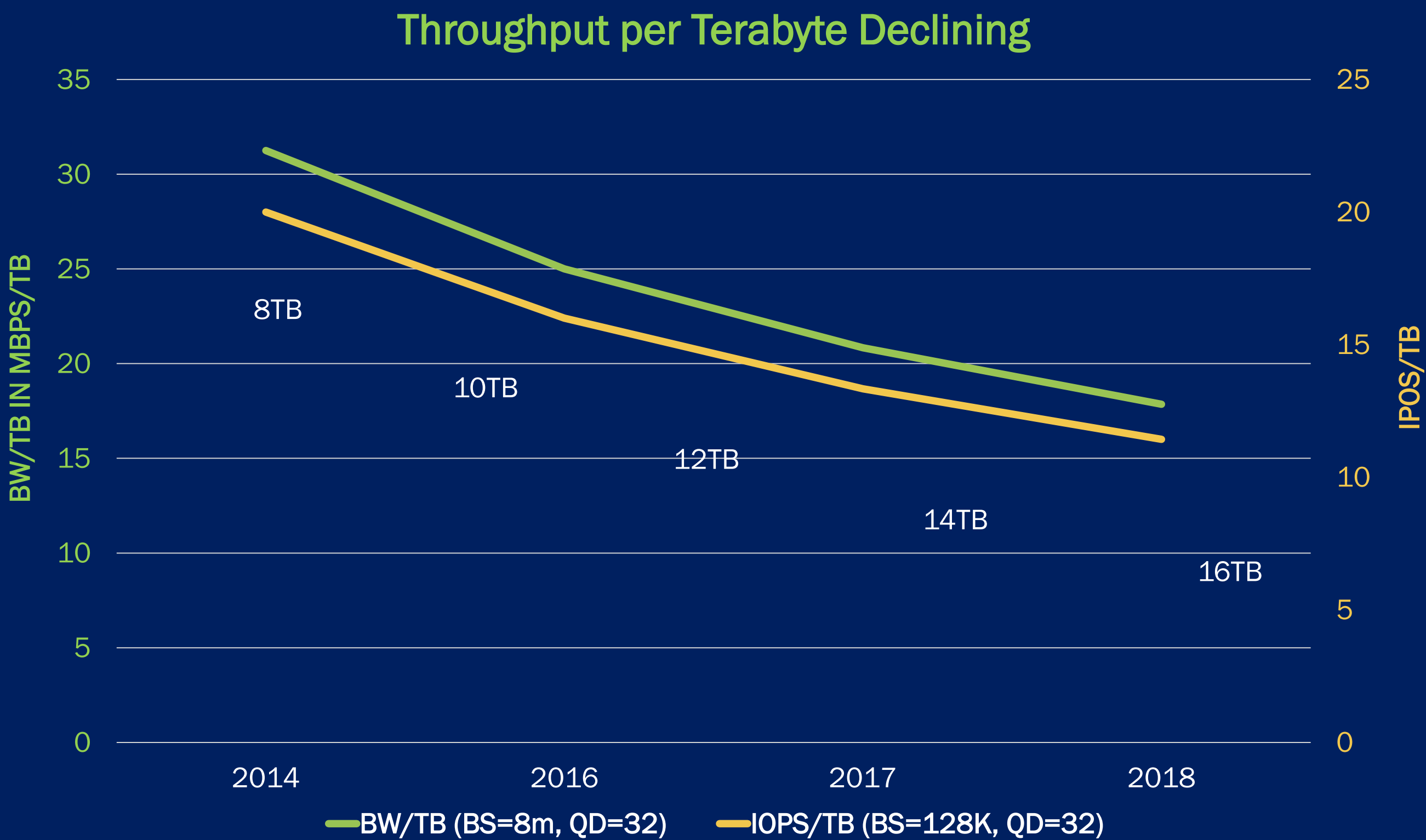
Source for Blue Line:
Composite of SpecInt Rate Scores mapped against TDP /
W from a variety of mainstream processors since 2012.
Intended to represent an industry norm, not any particular
product or supplier.

Supply: HDD



- HDD storage not increasing much

Sources for: Kryder
Kryder, Mark H.; Chang Soo Kim (October 2009). "After Hard Drives - What Comes Next?" (PDF). *IEEE Transactions on Magnetics*. **45** (10). doi:10.1109/TMAG.2009.2024163.



- I/O challenge becoming apparent

Source for Throughput per Terabyte:
Facebook composite of random (IOPS/TB) and sequential (BW/TB) IO scores, across a variety of mainstream products originating since 2014, and forward-projected into 2020. Intended to represent an industry norm, not any particular product or supplier.

Dilemma of scaling up and wide

Prediction ↓	Year	Throughput	SERDES	Ideal optical speed	Implemented @scale
	2016	3.2T	25G	100G	100G
	2018	6.4T	25G	200G	100G
	Estimates				
	2019	12.8T	50G	400G	100G
	2021	25T	50G	800G	200G
	2023	50T	100G	1.6T	400G
	2025	100T	100G/200G	3.2T	400G?
				Optics Running 1-2 Generations Behind	Power Limit

- Challenge of optical manufacturing
- Challenge of SERDES scale
- Challenge of Chip edge I/O
- Challenge of Power

Demand vs. Supply

Exponential demand meets reduced technology supply

- Technology work-arounds not trivial e.g. making dies larger
- Manifests itself in power challenges
- Breaking: Optical supply

Options how to deal with demands e.g.

- Topology and protocol optimization
- Balance of “high” vs. “wide”
- ASIC – dedicated HW for specific applications

Examples for approaches

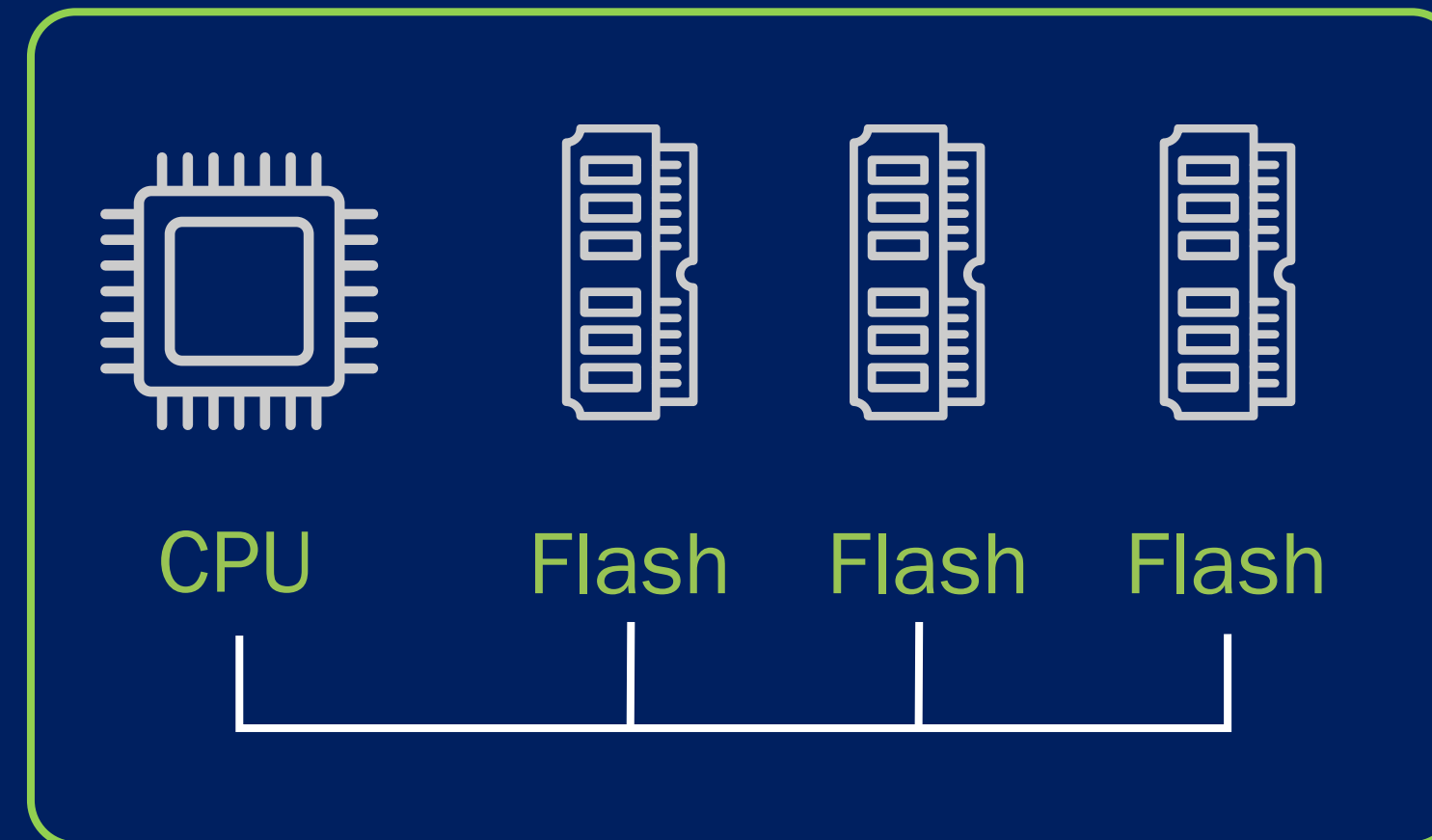
Topology optimization, resource distribution, disaggregation

Component development – co-packaged optics

Scale out of Fabric Aggregator

Topology optimization, resource distribution, disaggregation

Defined
resources

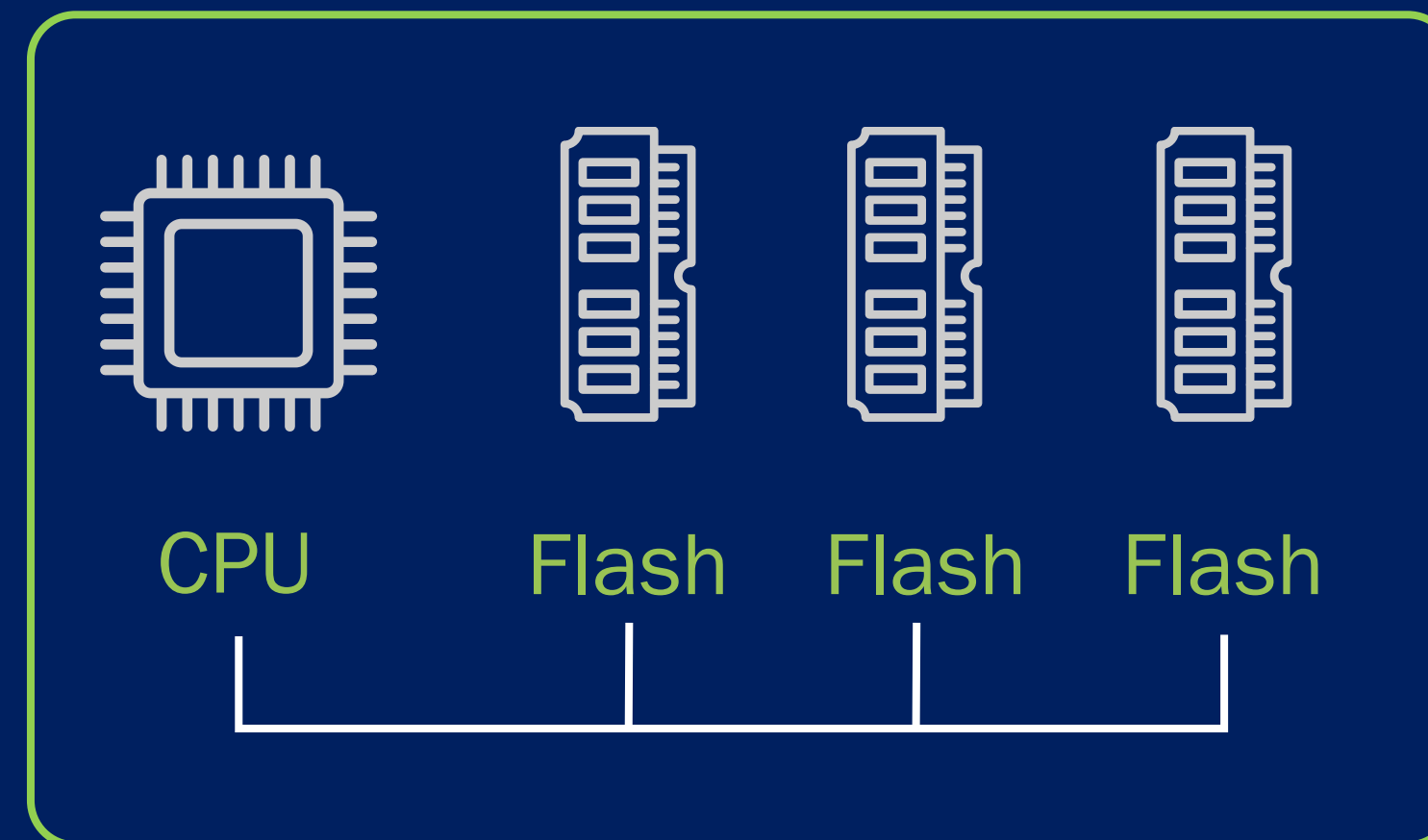


Disaggregated

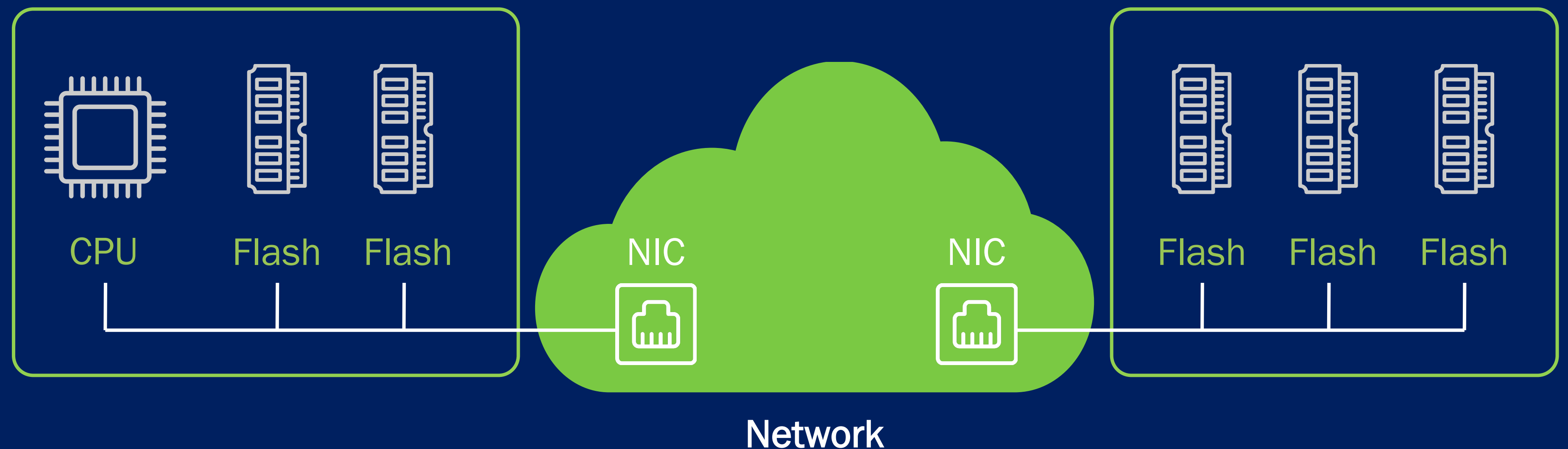


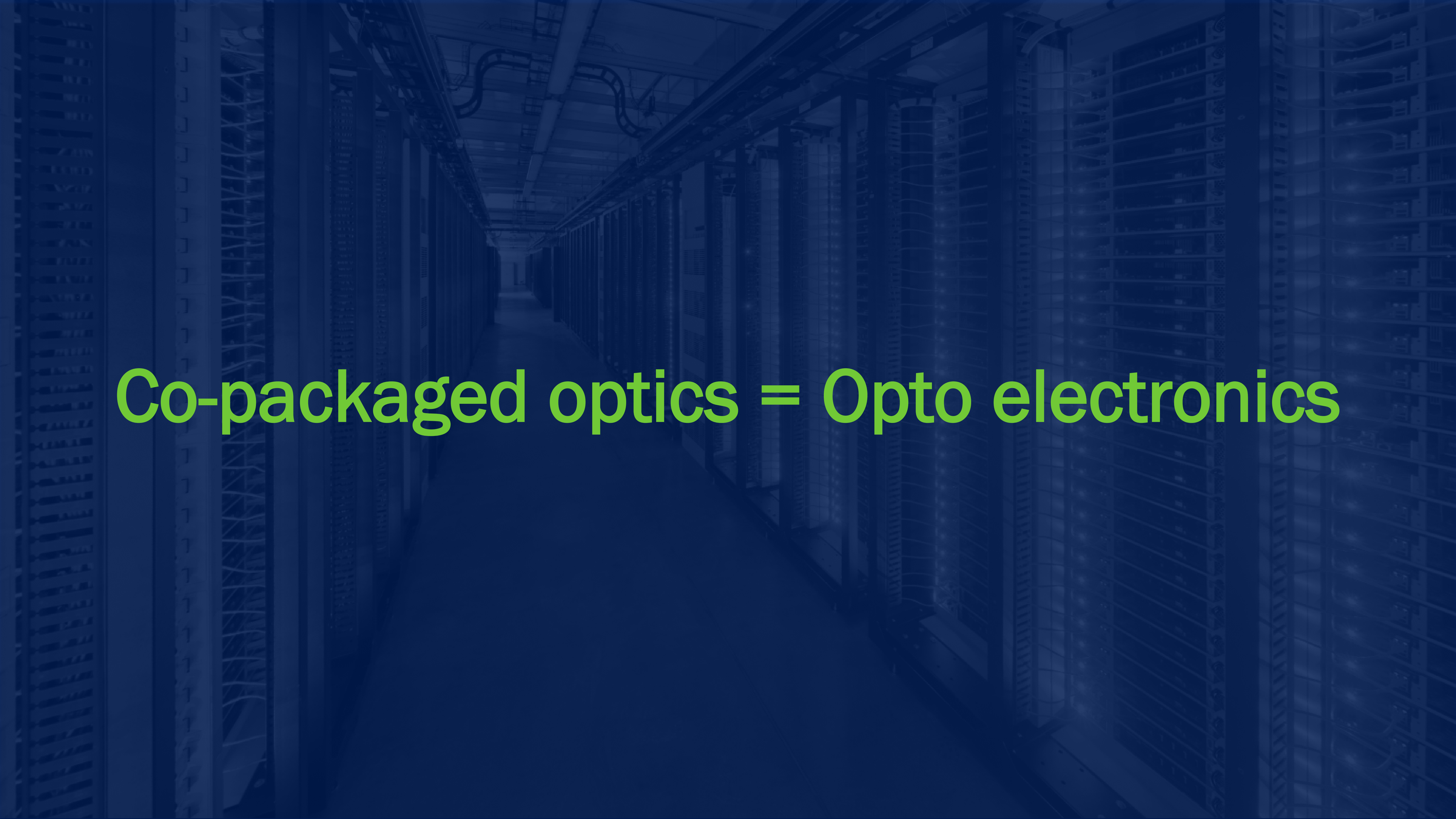
Topology optimization, resource distribution, disaggregation

Non-aggregated



Disaggregated





Co-packaged optics = Opto electronics

Datacenter Power

Servers and Network

Power is a shared resource



+

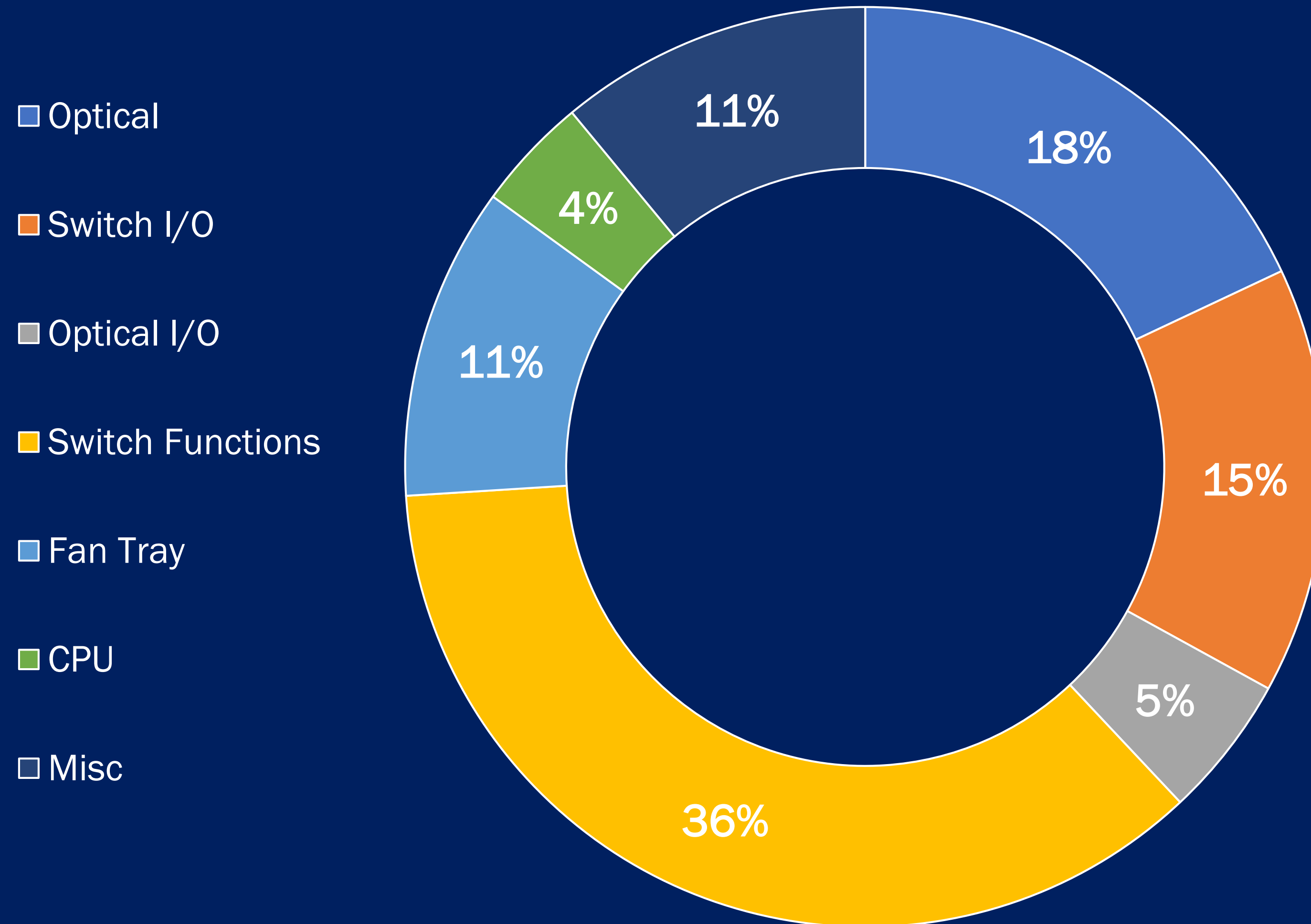


=



Power efficiency for today's switch and optical generation

Percentage of Network Power Per Function



Power for Optical Input/Output and Switch Input/Output is around 20% of the total power

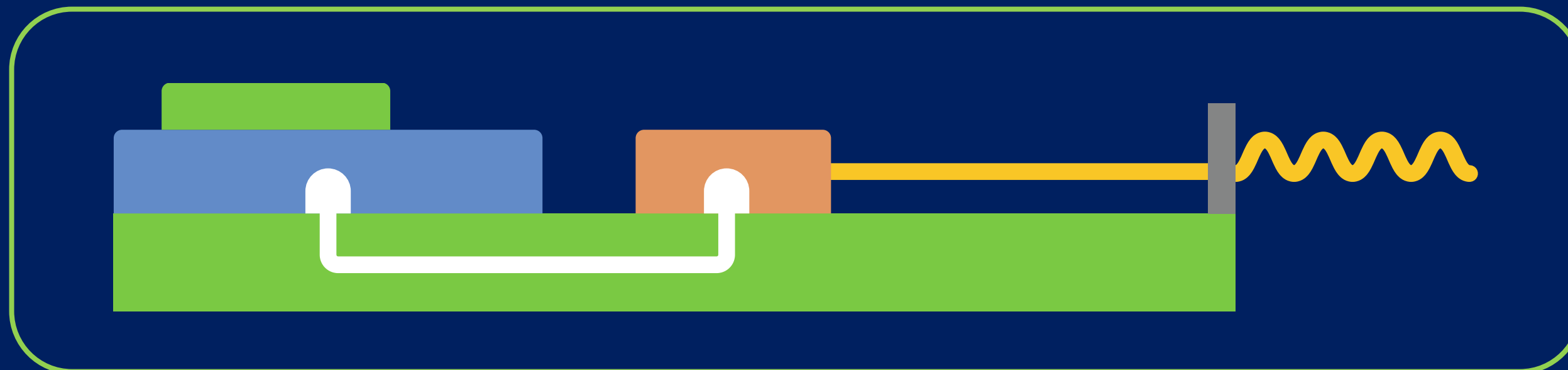
Design Goals:

- *Reduced total link power consumption*
- *Overcome SI limits of electrical links to edge of card*

Co-packaged optical solutions



- Conventional pluggable optics – not scalable



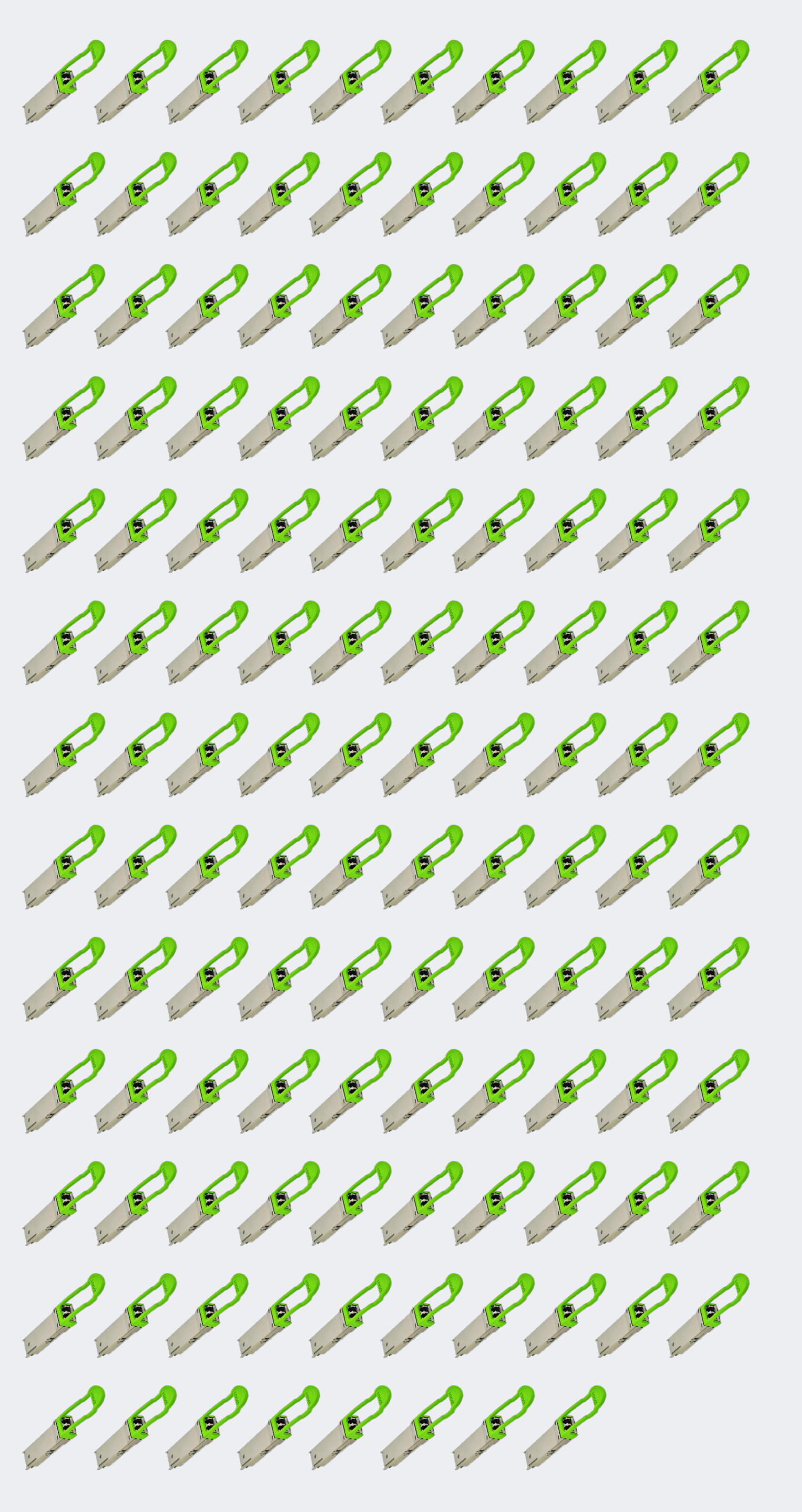
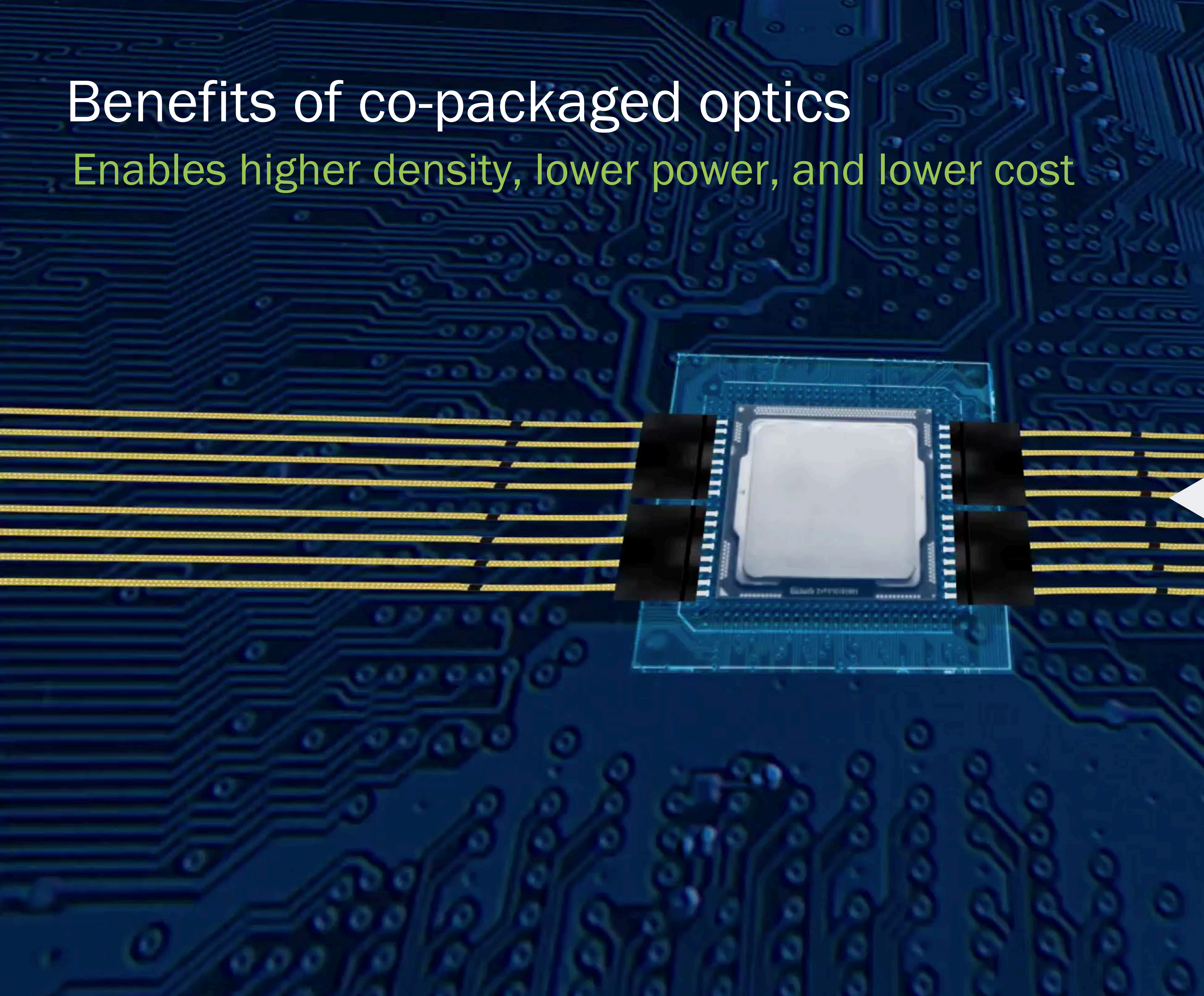
- On-board optics



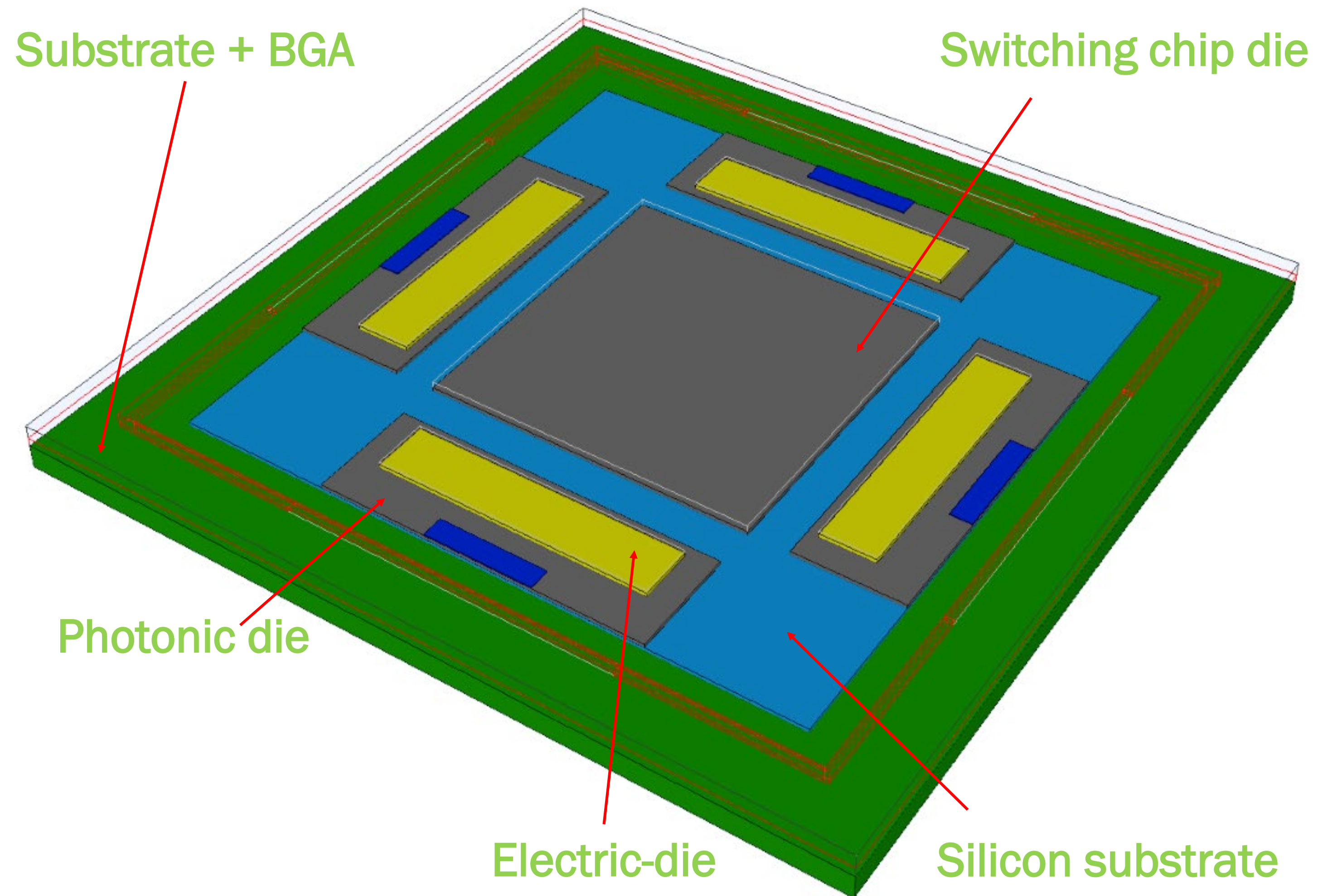
- Co-packaged optics
 - Wafer level integration/test
 - Packaging challenge to wafer
 - Better process control

Benefits of co-packaged optics

Enables higher density, lower power, and lower cost



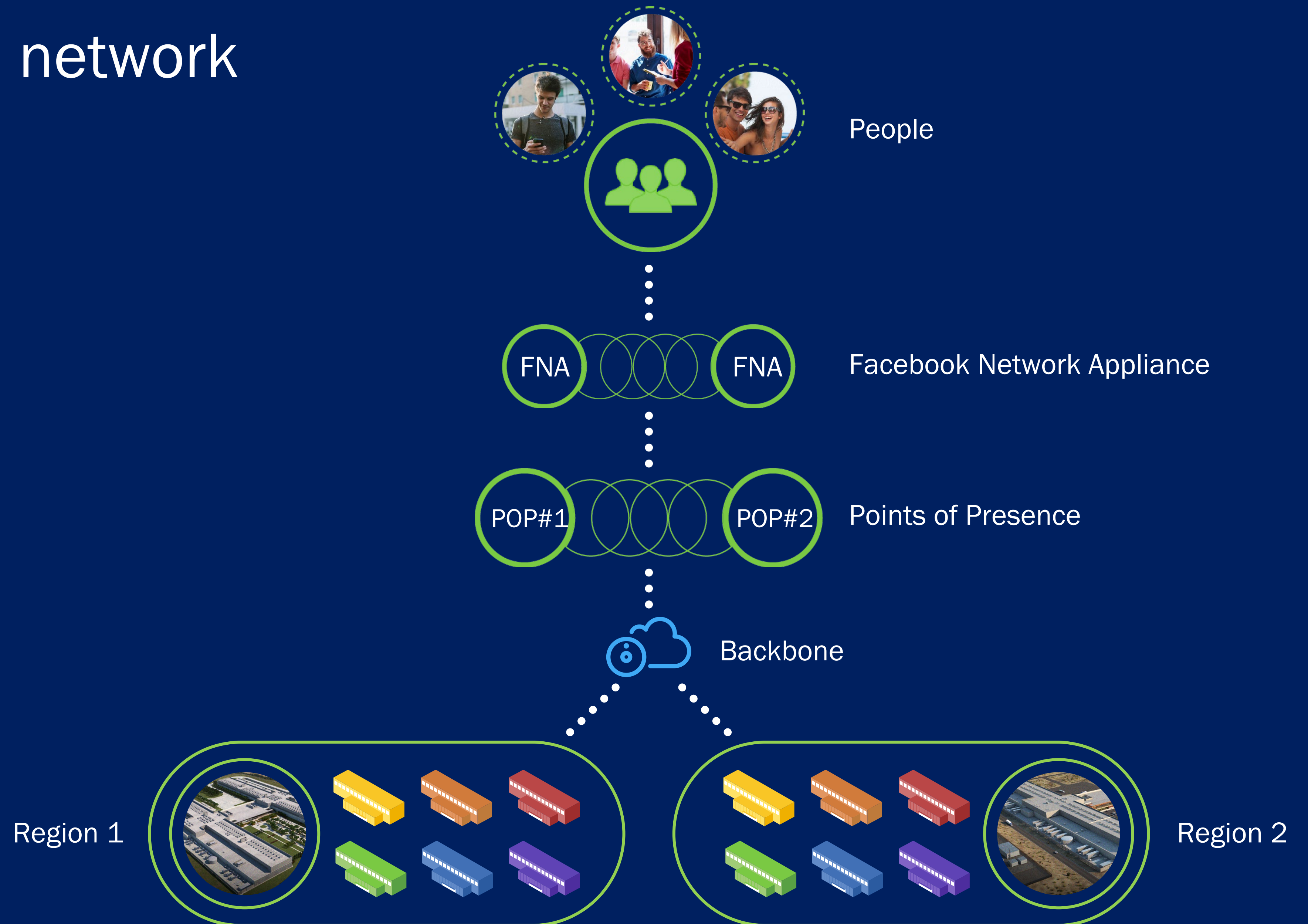
Co-packaged optics with ASIC



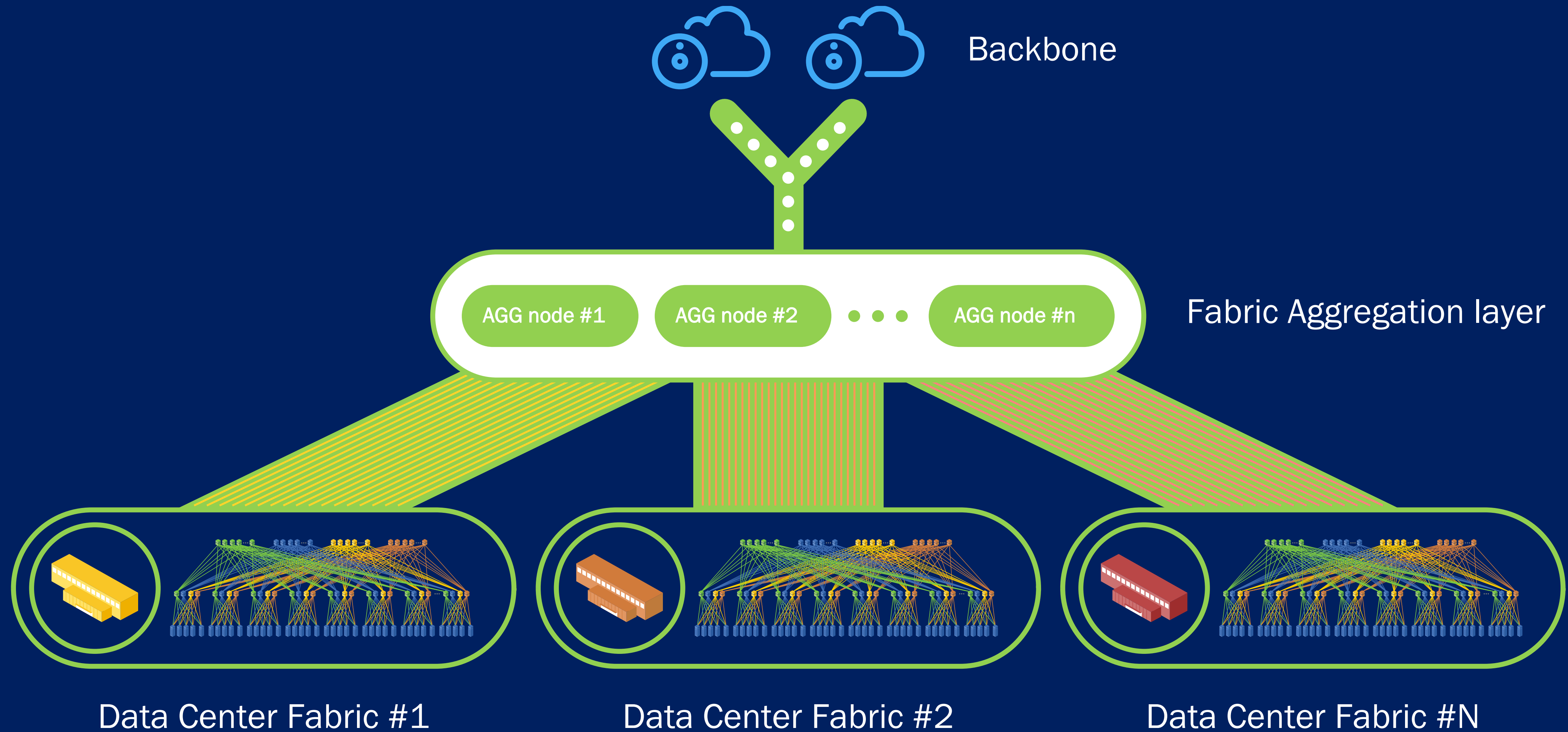


Scale out Fabric Aggregation

Traffic flow in to network



Role of fabric aggregation layer



The solution:
Fabric Aggregator



Integrating disaggregated building blocks



Wedge 100S

Integrating disaggregated building blocks



Wedge 100S

+



=



Cable backplane options



PIG-TAIL AOC



CWDM4



PSM4



DAC

Industry wide implications



Telco



Datacenter



Hyperscale

Summary

- Supply does not meet workload demand
- Solutions being worked on:
 - Topology optimization, resource distribution including compute disaggregation, flash disaggregation
 - Component development
 - co-packaged optics
 - Scale out of networks
 - FA & next generation fabrics
- Industry-wide implications
 - Any e2e cloud architecture can make use of tech