Google

New Opportunities and Challenges for Optics in Datacenters

Cedric F. Lam
On Behalf of Google Platforms Optics Team

ARPA-E Annual Meeting, July 19th, 2022, Long Beach California

Outline

- Introduction
  - Google’s datacenter infrastructure and growths
  - Google’s datacenter network architecture and needs
- Opportunities and challenges for in-rack connection
- Opportunities and challenges for campus networking
- Improving energy efficiency with SDM
  - Energy efficiency vs. spectral efficiency
- Network scaling challenges and optical switching opportunities
- Conclusion

Google Products

1997 — Search
Search ranking
Speech recognition

2004 — Gmail
Smart Reply
Spam classification

2005 — Maps
Street View image
Popular times

2006 — YouTube
Recommendations
Better thumbnails

2008 — Android
Keyboard input
Assistant

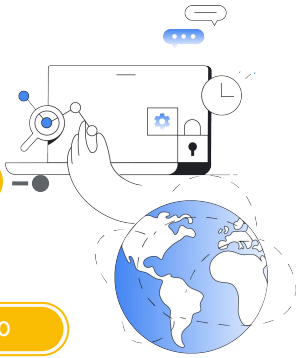
2008 — Chrome
Hack detection
Translation

2012 — Play
Keyboard input
Assistant

2012 — Drive
Quick Access
Data Loss Prevention

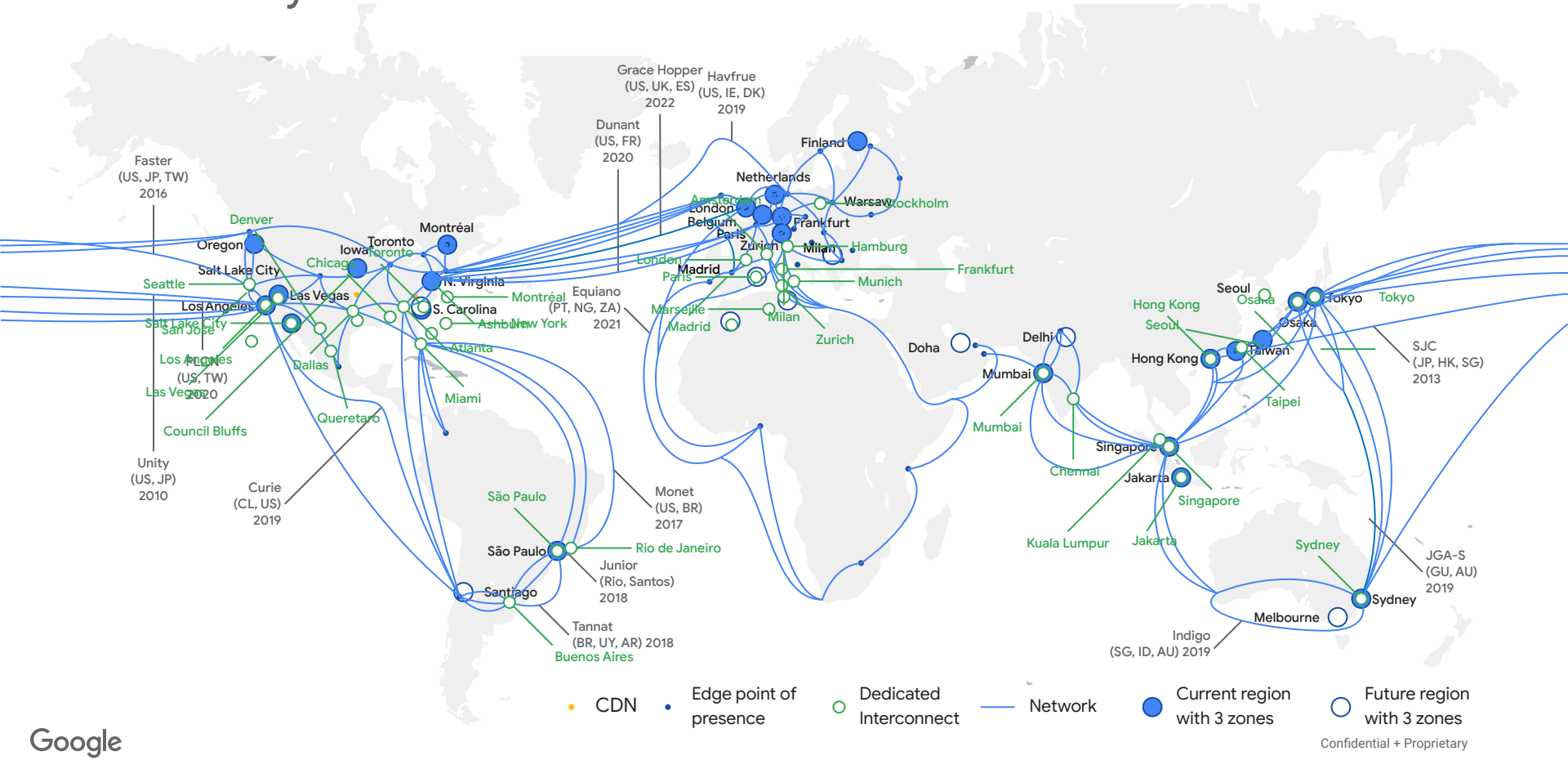
2015 — Photos
Photo search
Smart sharing

2020 — Translate
Translate text and/or
entire web pages



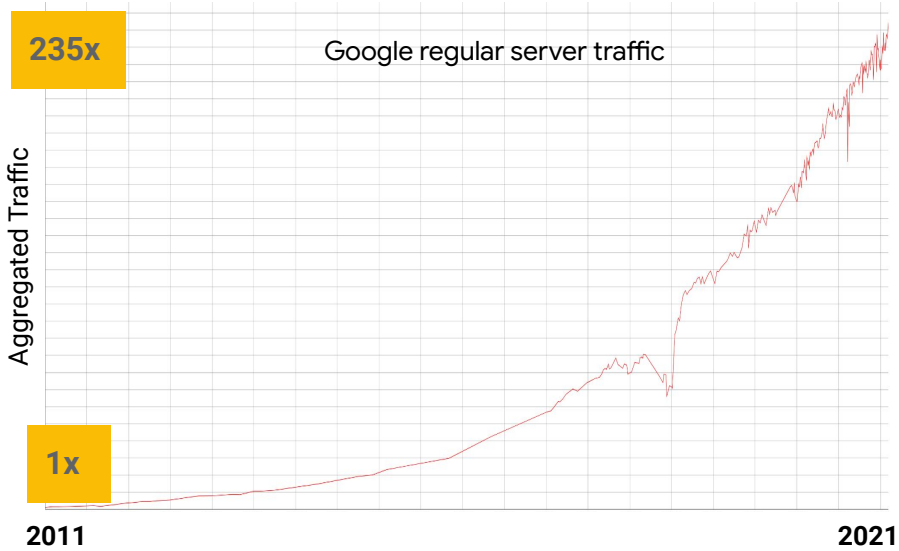
Powered by:
Google Cloud Platform

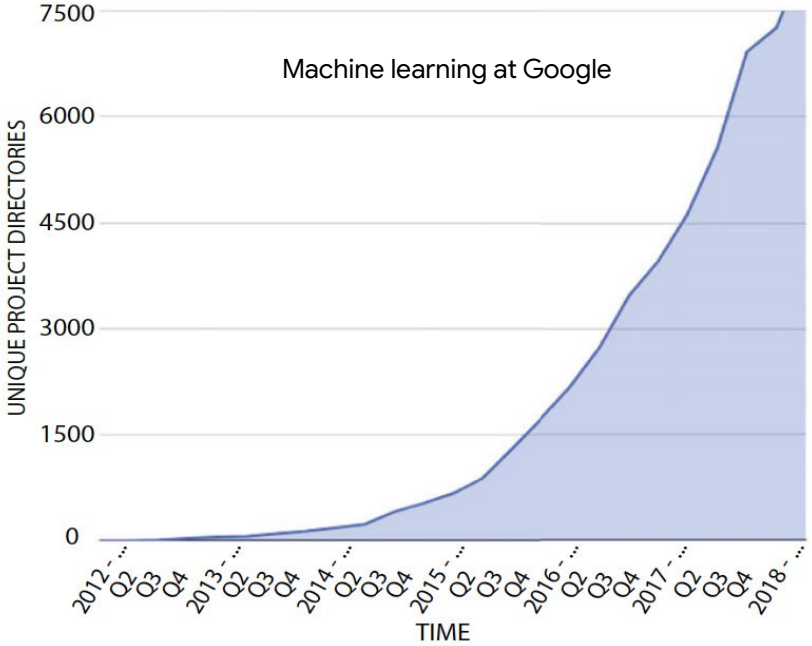
Powered by a Global Network of Datacenters



Google

Growth Drivers of Datacenter Connectivity





- Aggregate regular server traffic increased 235x from 2011 to 2021
- New ML use case drives more efficient networking with exponential bandwidth growth

Google

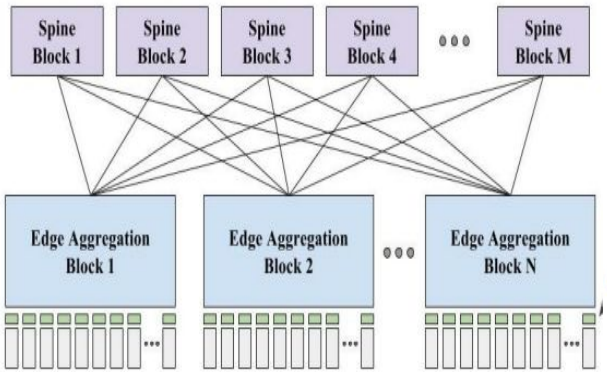
Ref: "The Datacenter as a Computer", 3nd edition, Luiz Barroso, et al

A Hyperscale Datacenter

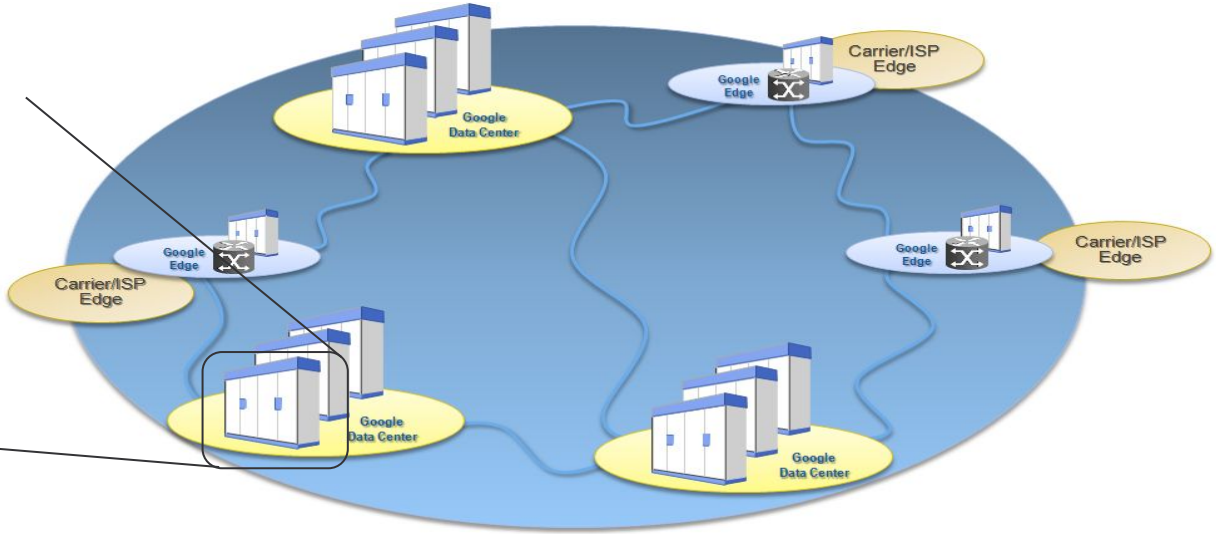
Google Datacenter Networks

Intra-DC (Clusters)	Intra-Campus (Campus)	Intra Metro POP - POP (Metro)	Inter-metro DC - DC; DC - POP (Backbone, LH / Subsea)
< 1km	6-10km	~40-80 km	1000s km

Distance

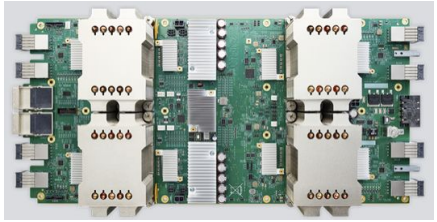


Cluster

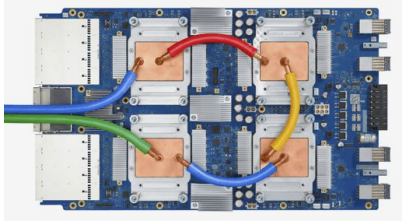


Google

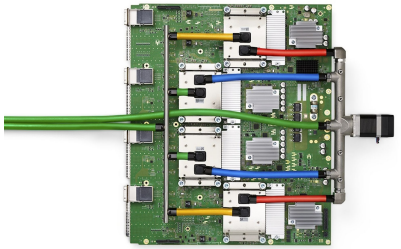
Four Generations of TPU at Google



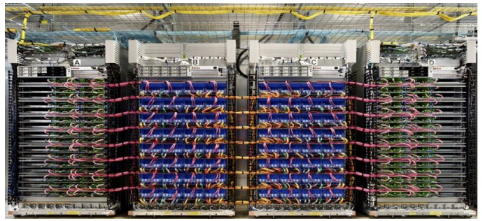
TPU v2
180 teraflops, 64 GB HBM



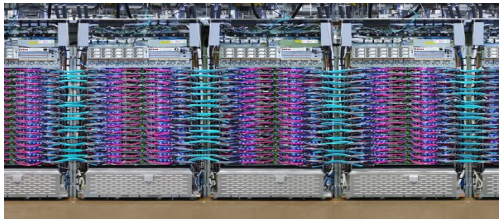
TPU v3
420 teraflops, 128 GB HBM



TPU v4
1.1 petaflops, 128 GB HBM



TPU v2 Pod - 2017
11.5 petaflops, 4 TB HBM
2D Torus network (256 chips)

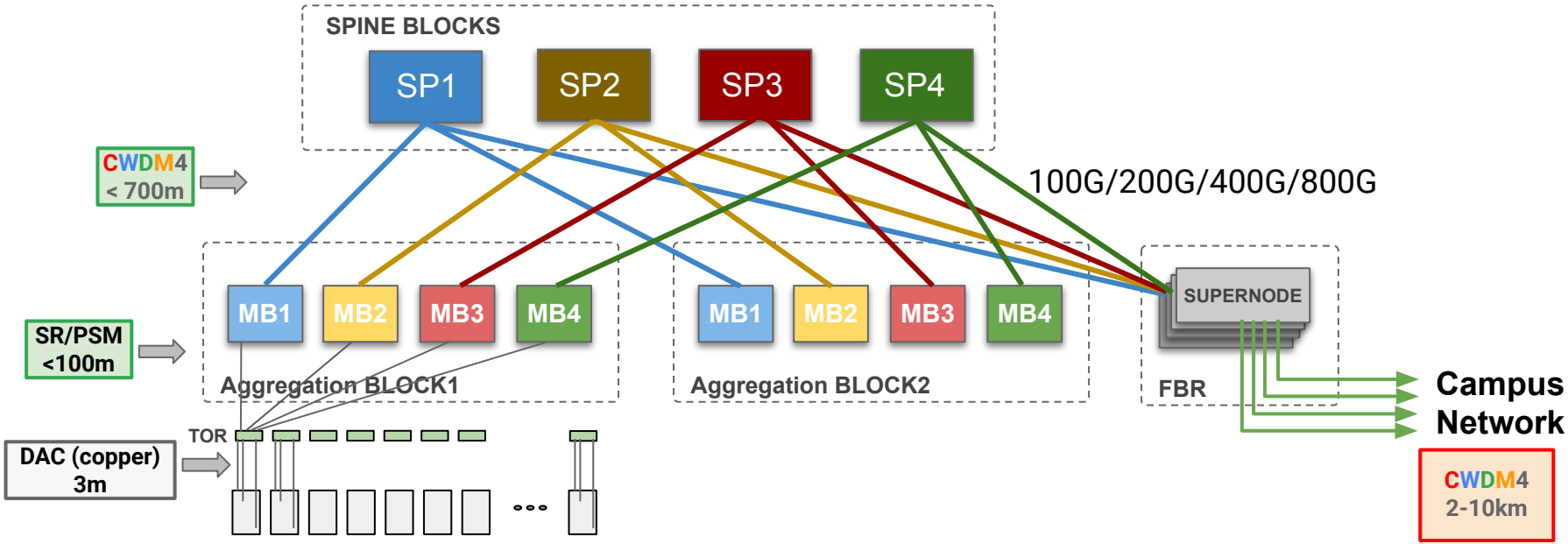


TPU v3 Pod - 2018
100 petaflops, 32 TB HBM
2D Torus network (1024 chips)

TPU v4 Pod - 2021
1 Exaflops, 132 TB HBM
3D Torus network (4096 chips)

- TPUs provide more computing power and require more efficient networking infrastructure.

Heterogeneous Interfaces in a Cluster Fabric



- Pluggable modules provides flexibility, maintainability and ease of operation.
- Copper links are the highest number of connections
- Backward compatibility is important for fabric interconnects
- Campus connections are growing fast and campus networks are growing beyond 2km to 10km

Google Interconnect Evolution (Copper)

- Pluggables with Optics and Copper Cables
- NRZ for up to 100Gb/s
- PAM4 for 400Gb/s and 800Gb/s
- TBD for 1.6Tb/s (PAM4 or PAM6)

Per lane speed (Gb/s)

200
100
50
25
10

Year

2007
2010
2014
2017
2020
2023

Mod: 10G NRZ
Lane: 1

56.5 x13.4 x 8.5

10G SFP+

Mod: 10G NRZ
Lane: 4
Optics Power< 2.5W
Cu Reach <=7m

40G QSFP+

Mod: 25G NRZ
Lane: 4
Optics Power< 3.5W
Cu Reach <=5m

100G QSFP28

Mod: 50G PAM4
Lane: 8
Optics Power< 15W
Cu Reach <=3m

400G OSFP

Mod: 100G PAM4
Lane: 8
Optics Power< 20W
Cu Reach <=2m

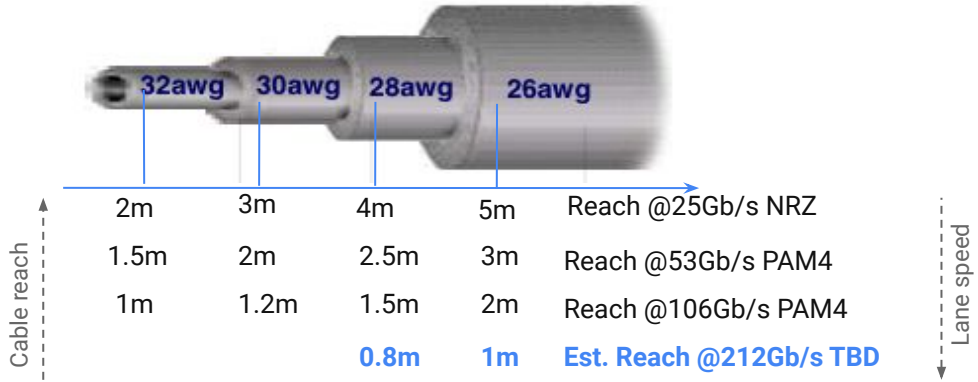
800G OSFP112

OSFP200

Mod: 1.6T PAMn?
Lane: 8
Optics Power< 23W
Cu Reach <=1m

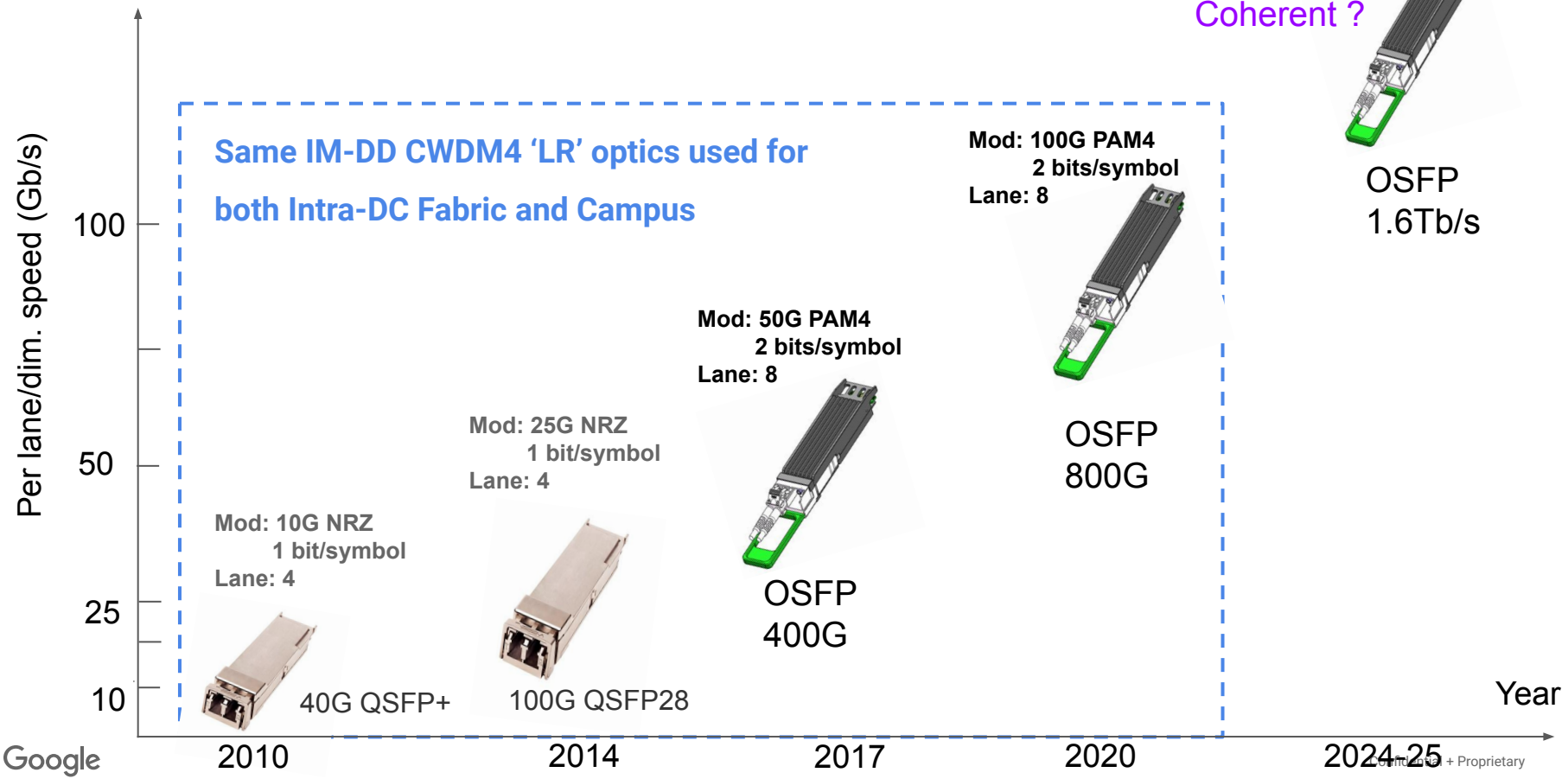
Intra-Rack Connections Opportunity and Challenges

- Passive copper provides low cost, low power, dense interconnect within a rack
  - Limited reach (target 1m) at 200Gbps
  - Thicker & bulkier as data rate increases
- Opportunities
  - Active copper cable
  - Active optical cable (enable new architectures and applications)
- Challenges: dirt cheap cost



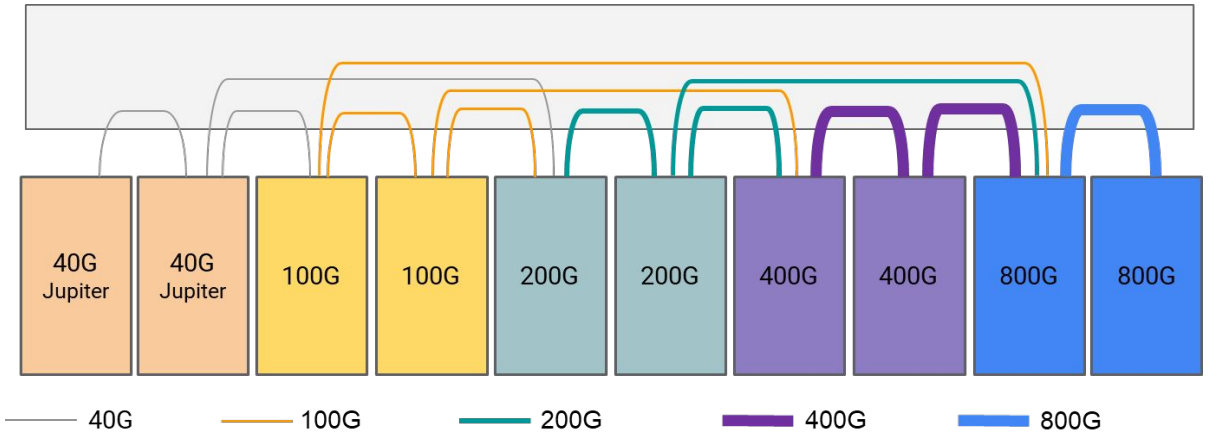
800G OSFP* Cu cable

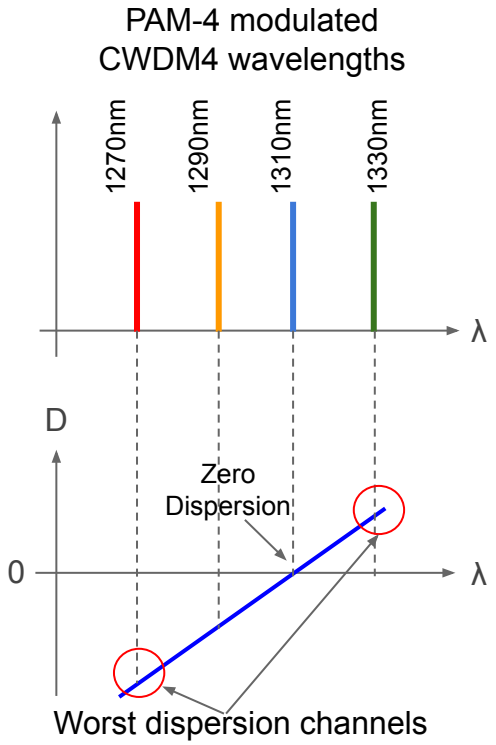
Google DC Optics Trend



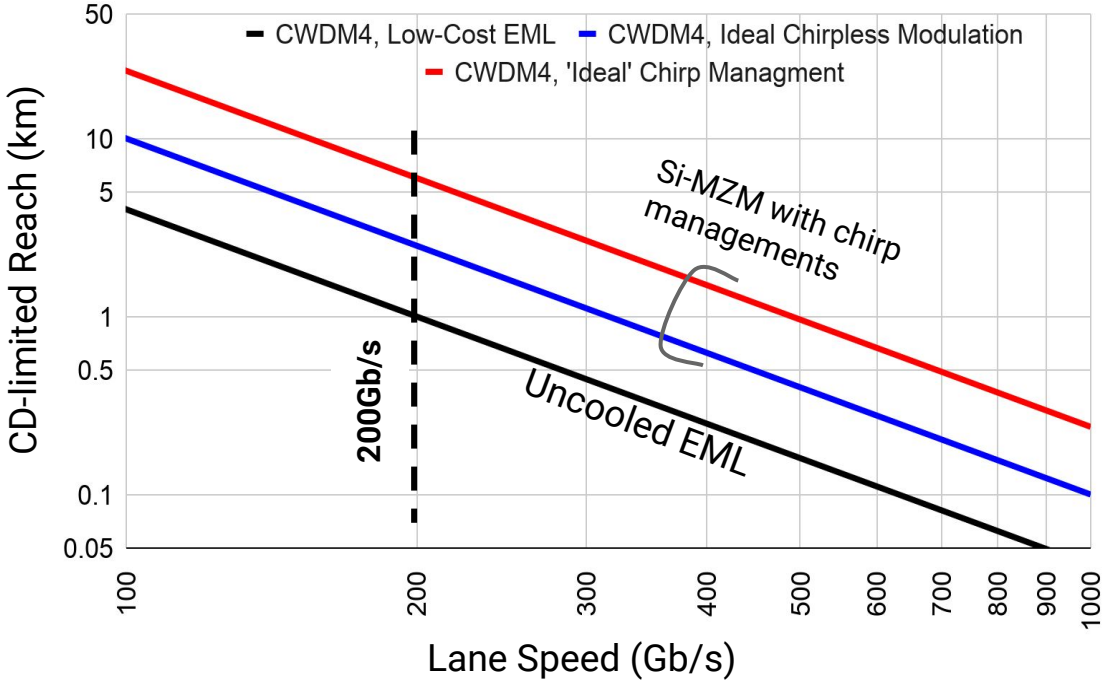
Campus Connectivity

- Traditionally served by the same 20nm spaced CWDM-4 optics used for the spine fabrics
- Backward compatibility in the optical layer





Challenge of Chromatic Dispersion

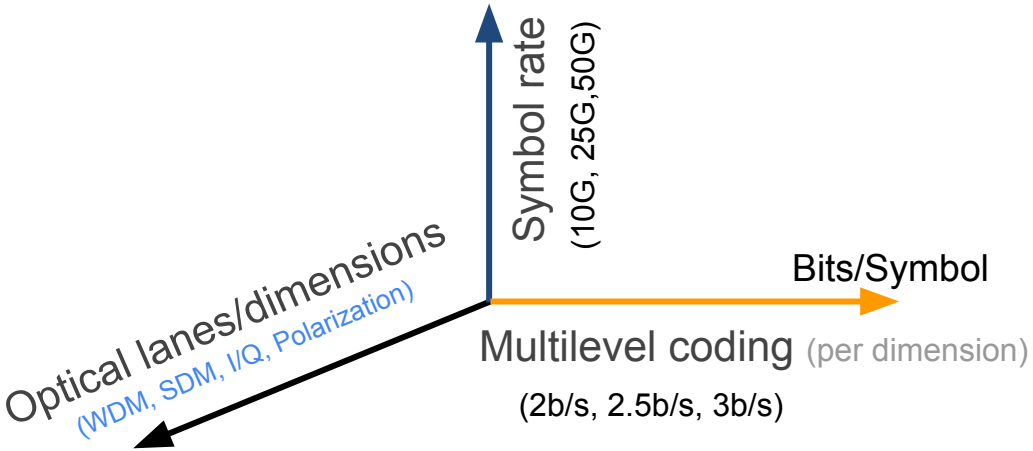


- **CD-limited reach (with CWDM4-EML)**
  - 100Gb/s PAM4: ~4km
  - 200Gb/s PAM4: ~1km
  - 400Gb/s PAM4: ~0.25km
- Sophisticated chirp management techniques could make incremental improvement in dispersion-limited reach , but face loss-budget challenge.

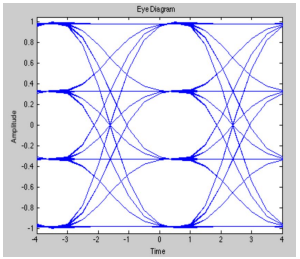
- **200Gb/s per lane PAM4 can support ~ 1.5km reach with low-cost CWDM4**
  - **Good enough for 800G Intra-DC use**
  - But challenging to support extended campus reach (<10km)

Opportunity for Coherent-Lite Transmission in Campus

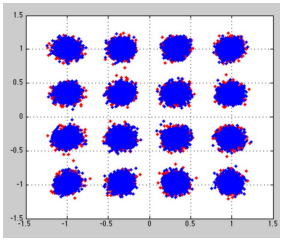
- Significant dispersion and link budget challenges as campus links grows beyond 5km



2b/s PAM4



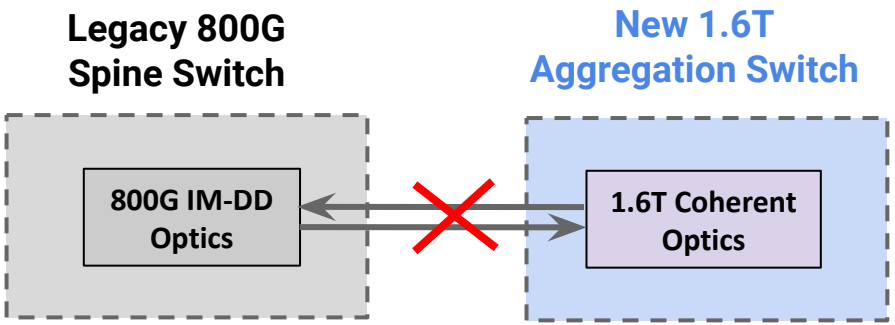
16QAM (2D-PAM4)

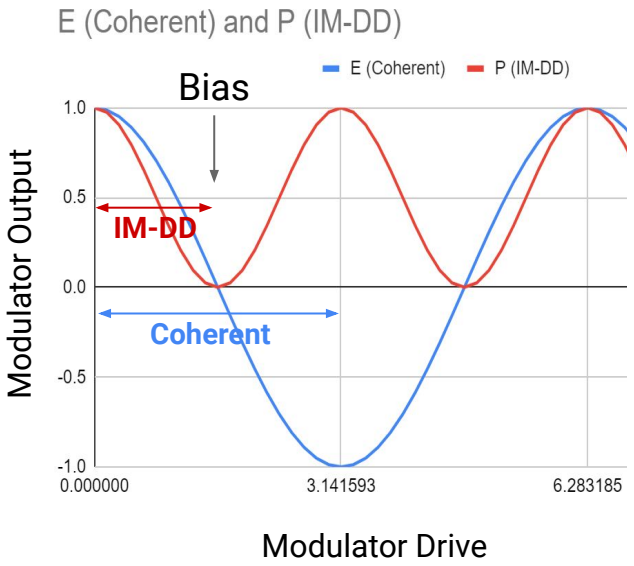


- More bandwidth-efficient modulation/detection technology
- Dispersion compensation with electronic DSP

Challenges for Coherent-Lite Transmission

- Power consumption for coherent-lite optical modules
  - Need optimized DSP without burden for long haul processing
  - Low V-pi & low loss optical modulator is key
- Backward compatibility with legacy systems





Coherent for Telecom and Datacom

Datacenter	Telecom
Cost, density and power efficiency is key ● High volume usage, lots of short distance fiber ● Large numbers of transceivers in a densely populated chassis	Spectral efficiency is key ● Long haul fiber is scarce and expensive ● Line system and amplifier huts are costly.
Link budget limited ● Unamplified. External amplifiers add cost, power consumption, operational complexity and another active element to fail.	OSNR limited ● Cascaded amplifiers is common

Power Constrained Systems - Rethink SDM

Spatial Dimension

$$C = m \times W \times log_2\left(\frac{P}{N} + 1\right)$$

Information Capacity (bits/s)

Bandwidth

Signal-to-noise Ratio

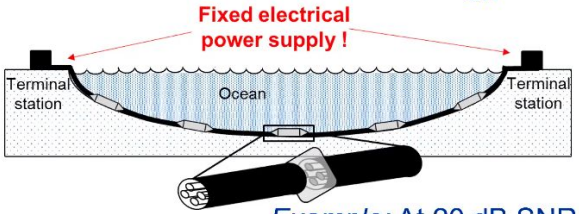
- Information capacity is logarithmic w.r.t. Signal Power under the same spatial and spectral constraints.
- Exponentially higher power is necessary to achieve higher spectral efficiency
  - For better energy efficiency, one should exploit the spectral and spatial dimensions.

Capacity Scaling with Power Constraints Using SDM

System Capacity = Bit rate per (sub)-channel x Number of (sub-)channels

Modulation x Multiplexing

$C = B_{ch} \times \log_2(1 + SNR) \times 2 \times 2 \times M \times N$



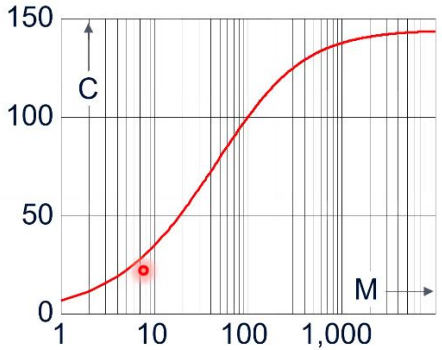
$$SNR = \frac{P}{2MNB_{ch}N_0}$$

Example: At 20 dB SNR, can we do better by going parallel?

1 spatial path: $\log_2(1+100)$ … 6.6 b/s/Hz

2 spatial paths: $2\log_2(1+50)$ … 11.3 b/s/Hz

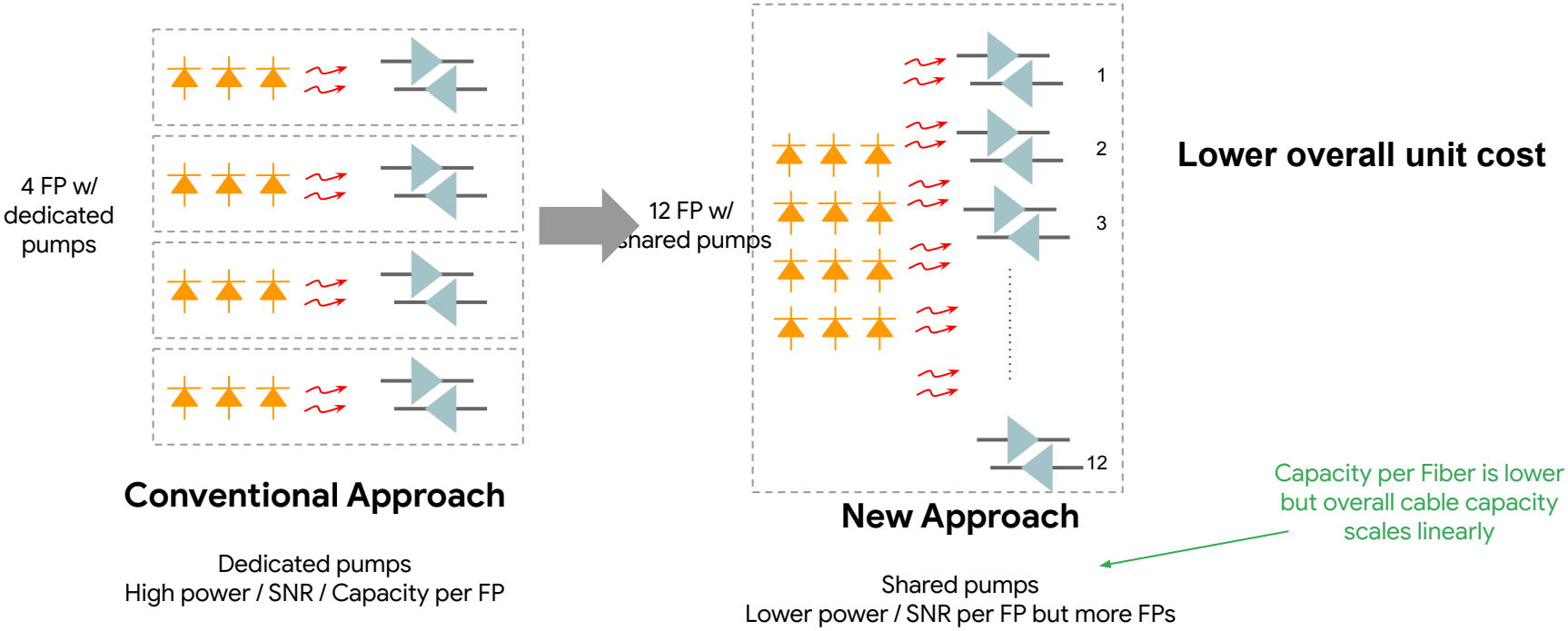
4 spatial paths: $4\log_2(1+25)$ … 18.8 b/s/Hz



[Dar et al., Proc. ECOC, Tu.1.E (2017) and JLT (2018)]

Optimizing Capacity of the Whole Cable vs. Individual Fiber Pairs

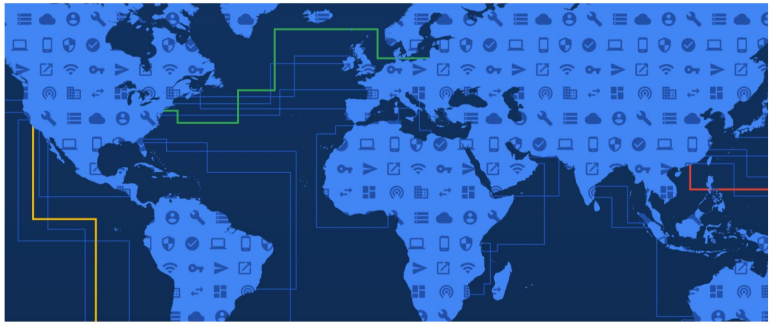
US Patent 9,755,734



Google Trans-Atlantic Dunant Cable w/ Record 240Tb/s Capacity

INFRASTRUCTURE

A quick hop across the pond: Supercharging the Dunant subsea cable with SDM technology



Vijay Vusirikala
Director of Network Architecture and Optical Engineering

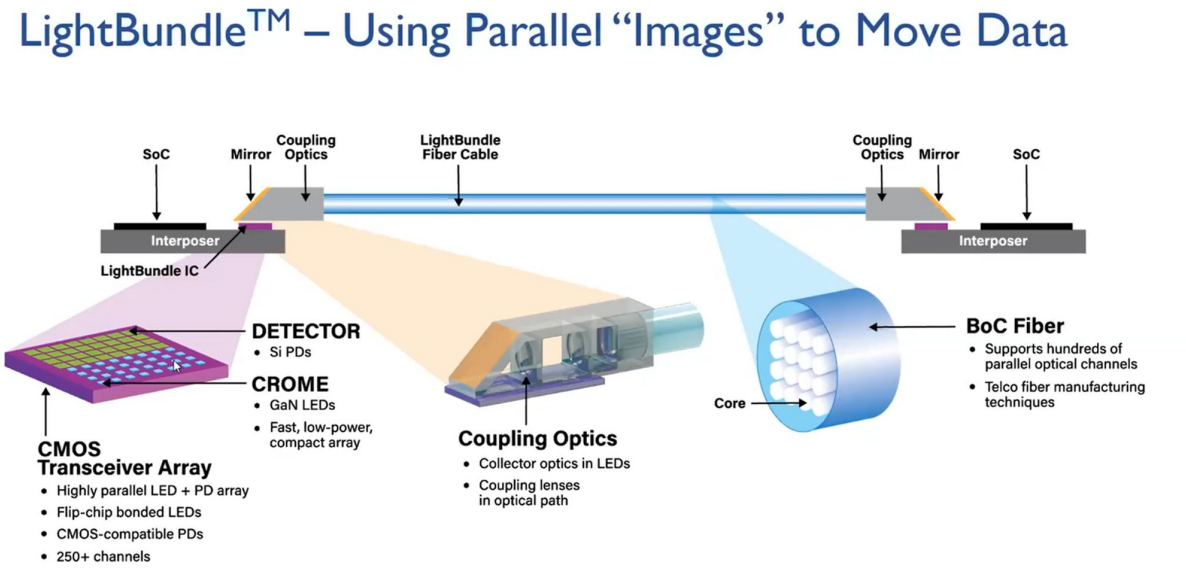
April 5, 2019

In 1858, Queen Victoria sent the first transatlantic telegram to U.S. President James Buchanan, sending a message in Morse Code at a rate of one-word per minute. In Q3 of 2020, when we turn on our private Dunant undersea cable that connects the U.S.A. and France, it will transmit 250 Terabits of data per second—enough to transmit the entire digitized Library of Congress three times every second.

To achieve this record-breaking capacity, Dunant will be the first cable in the water to use space-division multiplexing (SDM) technology. SDM increases cable capacity in a cost-effective manner with additional fiber pairs (twelve, rather than six or eight in traditional subsea cables) and power-optimized repeater designs. These advancements were created in partnership with SubCom, a global partner for undersea data transport, which will engineer, manufacture and install the Dunant system utilizing their SDM technology and equipment.

- Turned on in Q3 2020
- Space Division Multiplexing
  - Innovation against common thought of maximizing spectral efficiency per fiber
- 12 Fiber Pairs
  - 20Tb/s per pair
- Power efficiency

SDM in Short Reach Connection Example

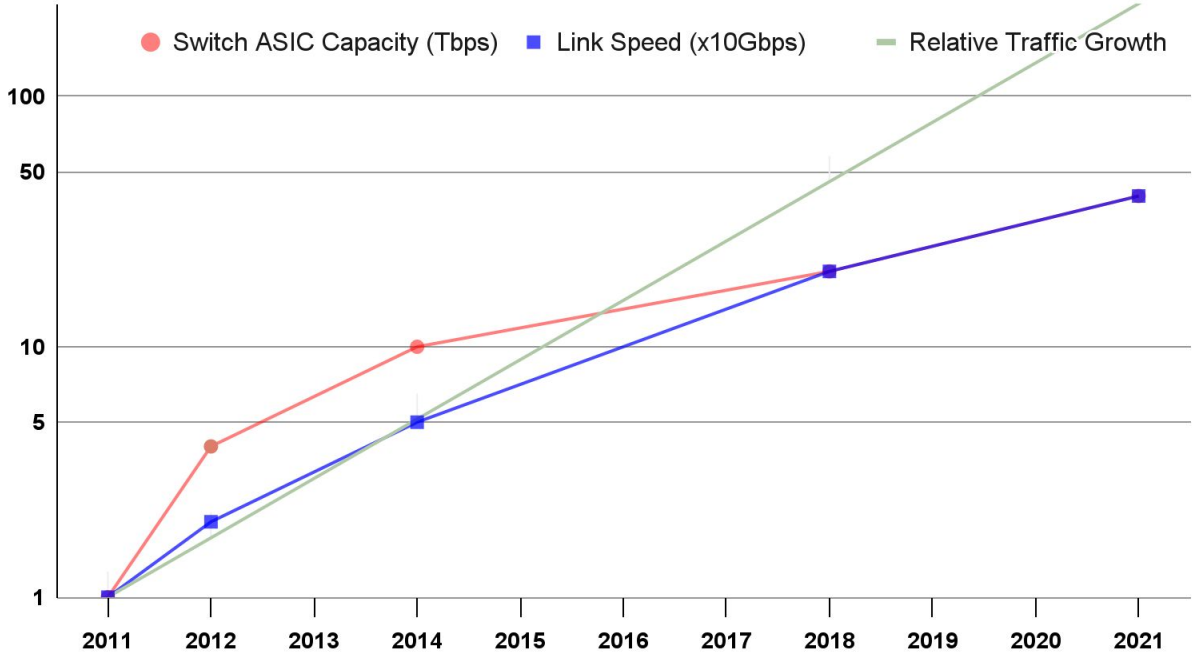


- Parallel Connection
  - No serdes
- Reach 10m
- < 0.5pJ/bit
  - Compare with 3pJ / bit for VSR serdes
- > 1 Tb/s per mm
- New Application:
  - Memory disaggregation for HPC and ML accelerators

avicena

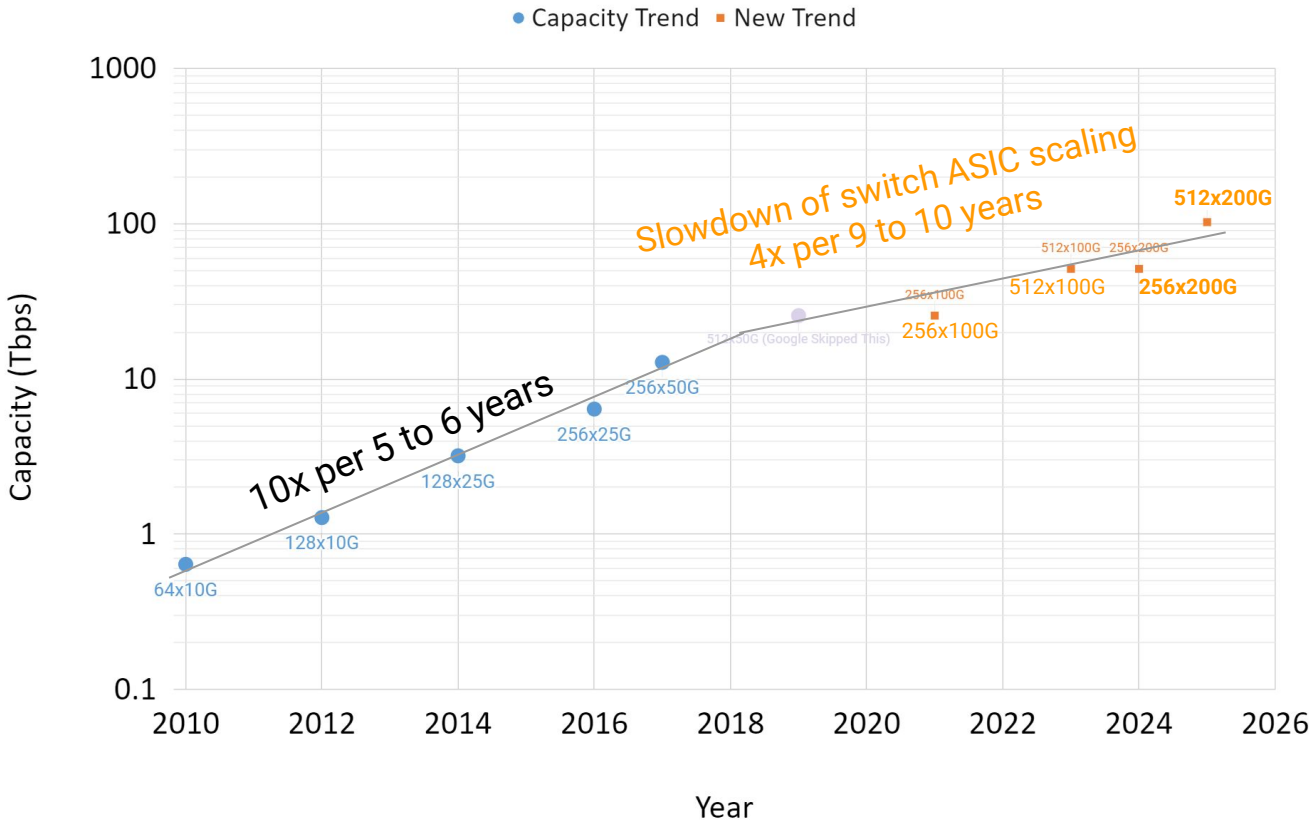
Ref: Bardia Pezeshki, Microled Array-Based Optical Links Using Imaging Fiber for Chip-to-Chip Communications - OFC 2022 - W1E.1

System Opportunity: Traffic Growth Outstrips Technology Growth



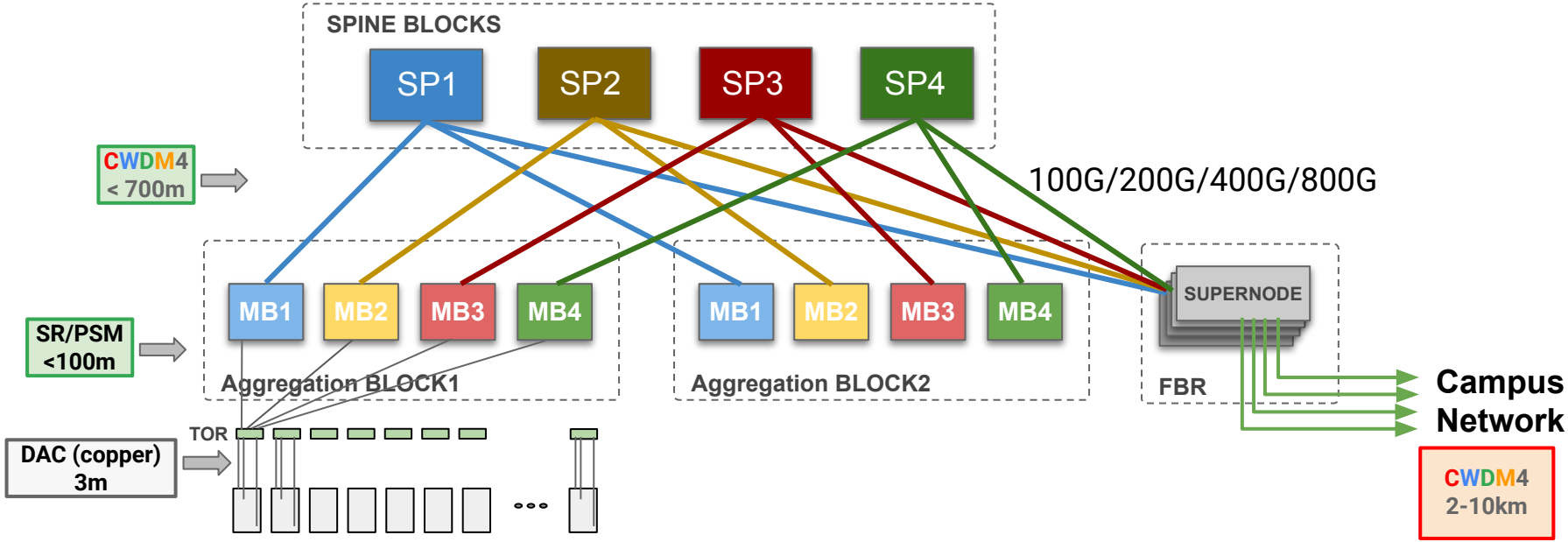
- New Ethernet standard overdue
- IEEE 802.3df task force working on 800G and 1.6T Ethernets
- Expected timeline for 800G/1.6T Ethernet is 2025

Slowdown of Switching ASIC Capacity Growth



- Optical switching may be an opportunity to augment electronic switching and provide energy efficiency.

Where Should We Place Optical Switches?



Challenges for Optical Switching in DC

- Optical switch challenges
  - High switching radix
  - Low insertion loss
    - Recall DC links are power budget limited and unamplified
- SDMs using parallel fiber would incur parallel optical switches
  - Cost and management complexity
- Transceivers with higher link budget and higher spectral efficiency works better with OCS.
- How should we balance the tradeoff between transceiver complexity and optical switching complexity, electronic switching vs. optical switching to achieve the best overall system performance, energy efficiency and total cost of ownership?

SIGCOMM 2022

Amsterdam

Jupiter Evolving: Transforming Google's Datacenter Network via Optical Circuit Switches and Software-Define Networking

Leon Poutievski, Arjun Singh, Joon Ong, Rui Wang, Omit Mashayekhi, Mukarram Tariq, Jianan Zhang, Karthik Nagaraj, Rishi Kapoor, Hong Liu, Ryohei Urata, Virginia Beauregard, Lorenzo Vicisano, Jason Ornstein, Samir Sawhney, Stephen Kratzer, Nanfang Li, Junlan Zhou, Shidong Zhang, Patrick Conner, Steve Gribble, Amin Vahdat (Google)

Tuesday, August 23, 2022 CEST

11:00am - 12:30pm

https://conferences.sigcomm.org/sigcomm/2022/program.html

Jupiter Evolving: Transforming Google's Datacenter Network via Optical Circuit Switches and Software-Defined Networking

Paper #715. Experience-track submission. 12 pages body, 18 pages total.

ABSTRACT

We present a decade of evolution and production experience with Jupiter datacenter fabrics. In this period, we have evolved the Jupiter fabrics to deliver 5x higher speed and capacity, 30% reduction in capex, 41% reduction in power, as well as incremental deployment and technology refresh all while serving live production traffic. A key aspect of this journey has been *evolving Jupiter from a Clos to a direct-connect topology among machine aggregation blocks*. This paper describes the critical elements powering this change: A datacenter interconnection layer that employs Micro-Electro-Mechanical Systems (MEMS) based Optical Circuit Switches (OCSes) to enable dynamic topology reconfiguration, centralized Software-Defined Networking (SDN) control for traffic engineering, and automated network operations for incremental capacity delivery and topology engineering. Based on extensive evaluation from production data, we show that using a combination of traffic and topology engineering on direct-connect fabrics, we achieve similar throughput as Clos fabrics for our production traffic patterns. We also optimize for path lengths, wherein 60% of the traffic takes direct path from source to destination aggregation blocks, while the remaining transits one additional block, achieving an average block-level path length of 1.4 in our fleet of direct-connect fabrics today. The use of OCS also achieves 3x faster fabric reconfiguration compared to Clos fabrics using patch panel based datacenter interconnect in our earlier work.

1 INTRODUCTION

Advances in Software-Defined Networking [13] and Clos topologies [2, 3, 15, 25, 35] built with merchant silicon have enabled cost effective, reliable, and building-scale data center networks as the basis for Cloud infrastructure. A range of network services, Machine Learning (ML) training and inference, and storage infrastructure leverage uniform, high bandwidth connectivity among tens of thousands of servers to great effect.

While the progress has been tremendous, managing the heterogeneity and incremental evolution of a building-scale network has received comparatively little attention. Cloud infrastructure grows incrementally, often one rack or even one server at a time. Hence, filling an initially empty building takes months to years. Once initially full, the infrastructure evolves incrementally, again often one rack at a time with the latest generation of server hardware. There is no pre-existing blueprint for the types of servers, storage, accelerators, or services that will move in or out over the lifetime of the network. We have found the alternative of co-designing the network, storage, compute, and accelerator infrastructure for the lifetime of a service deployment to be unacceptably expensive both in absolute costs but also in human planning overhead. The realities of exponential growth and changing business requirements mean that the best laid plans quickly become outdated and inefficient.

Incremental refresh of compute and storage infrastructure is relatively straightforward: drain the capacity associated with perhaps one rack's worth of capacity among hundreds or thousands in a data center and replace that rack with a newer generation of hardware or one that matches the requirements of the services that have migrated into a data center over time. Such incremental refresh of the network infrastructure is more challenging as modern Clos-based topologies require pre-building at least the spine layer for the entire network as shown in Fig. 1. Doing so unfortunately restricts the data center bandwidth available to individual groups of machine capacity (*aggregation blocks*) and ultimately individual servers to the speed of the network technology available at the time of spine deployment.

Consider the simple example of a pre-built 40Gbps spine at the maximum-scale of 64 aggregation blocks each supporting 20Tbps burst bandwidth across the data center network [35]. Within a few years, the next generation of 100Gbps becomes available supporting, in theory, aggregation blocks with 51.2Tbps of burst bandwidth capability. Unfortunately, these new blocks would be limited to the 40Gbps link speed of the pre-existing spine blocks, reducing capacity to 20Tbps per aggregation block. Ultimately, individual server and storage capacity would be derated because of insufficient data center network bandwidth. A doubling of server core count without a doubling of network speeds leads to system imbalance and stranding of expensive server capacity. Unfortunately, the nature of Clos topologies are such that incremental refresh of the spine results in only incremental improvement in the capacity of new-generation aggregation blocks. One approach would be to refresh the entire spine layer as soon as new hardware technology becomes available. However, doing such a wholesale building-wide upgrade would be expensive, time consuming, and likely operationally disruptive given the need for fabric-wide rewiring.

Conclusions

- Continual scaling of datacenter presents new challenges and opportunities for optics.
- Heterogeneous interconnect technologies are required to address the different transmission challenges in an actual network
  - Pluggable modules presents flexibility, optionality, and ease of operation and maintenance.
- Coherent technologies help to overcome future challenges of link budget and fiber dispersion limits but requires
  - Optimized DSP and efficient optical modulators
- SDM is a way to trade spectral efficiency for energy efficiency
- Appropriate implementation of optical switching could augment electronic switching with more energy and cost efficiency, and bring better scalability to datacenter network fabrics.

Q&A

The Need for Pluggable Modules

- We need different types of PMDs to serve the network function requirements and achieve cost targets
  - Copper
  - SRn, DRn, PSMn
  - FR4, CWDM-4
- Operation considerations
  - Reliability
    - Yield for large-scale optoelectronic integration is still challenging
  - Serviceability
  - Flexibility
  - Mature technology ecosystem