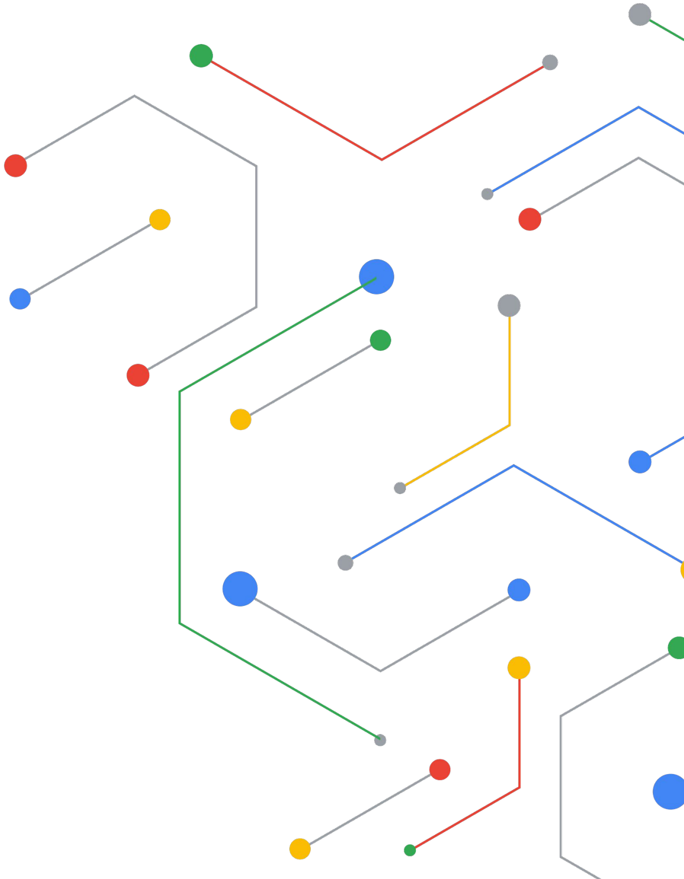
Google Cloud

“Without networking, there is no cloud.”

AI/ML Cloud Networking Innovation



Shaowen Ma, shaowen@google.com
Group Product Manager
Dec 2023

7DEOH RI &RQWHQWV

- Cloud Network Evolution 01
- AI Fabric Innovation 02
- AI Smart Offload Transport Innovation 03
- Summary 04

Cloud Network Evolution

01

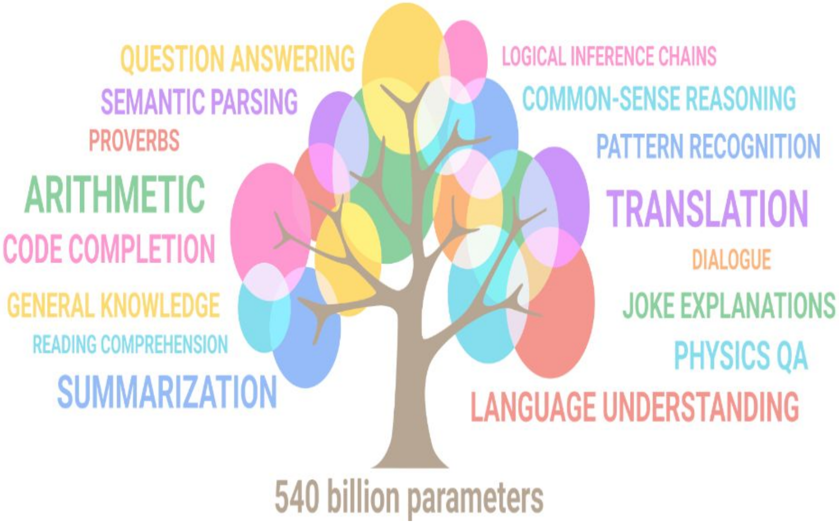
Google Cloud

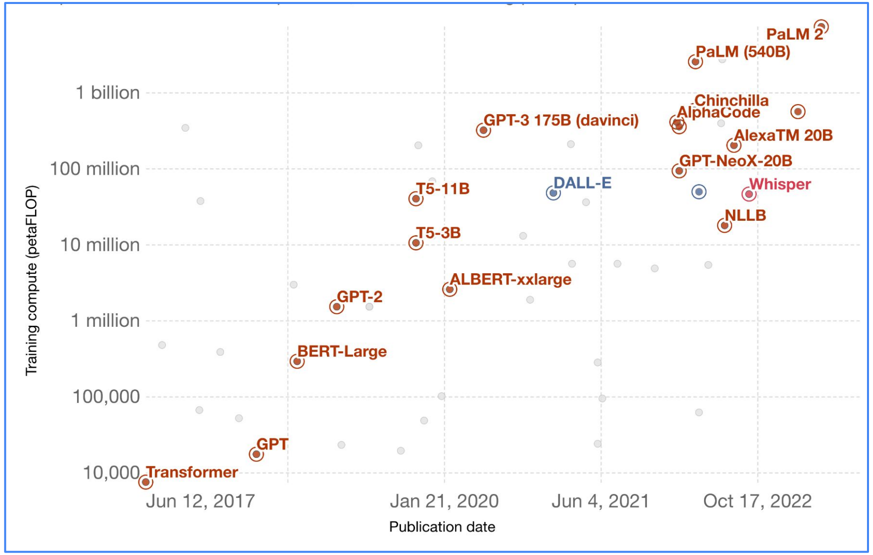
Exploding cost of powerful AI

Model & Dataset Size

Capabilities

Costs



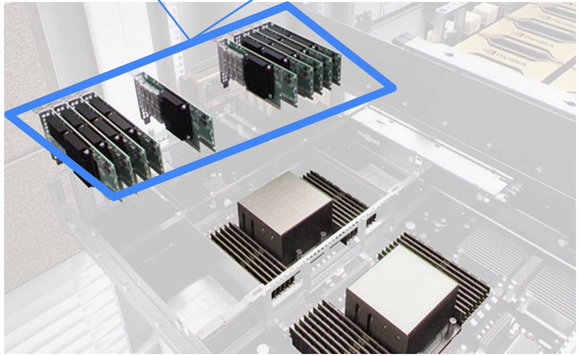


Source: Google's PaLM blogpost

Source: "Computation used to train notable intelligence systems", One World Data

)URP FOXVWHUV WR ZDUHKRXVH VFDOH FRPSXWHU

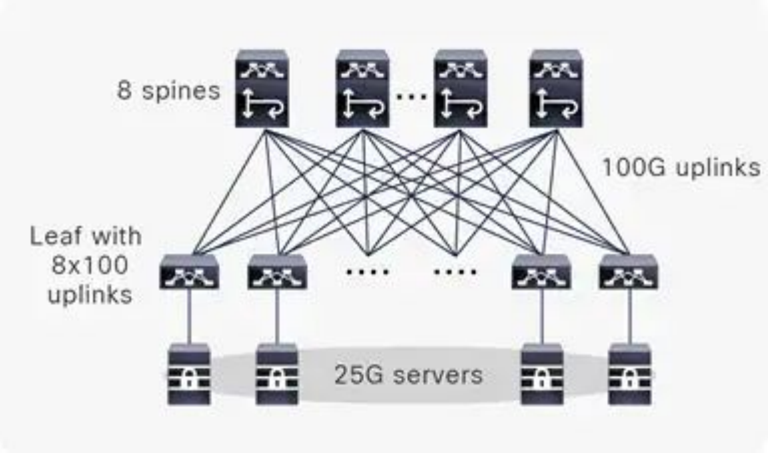
$, 1HWZRUN 'HVLJQ 'LIIHUHQFH 6HUYHU



8x400GE= 3.2Tbps+
vs CPU(25/100Gbps)

Huge bandwidth GPU/TPU
1.6Tbps/3.2Tbps

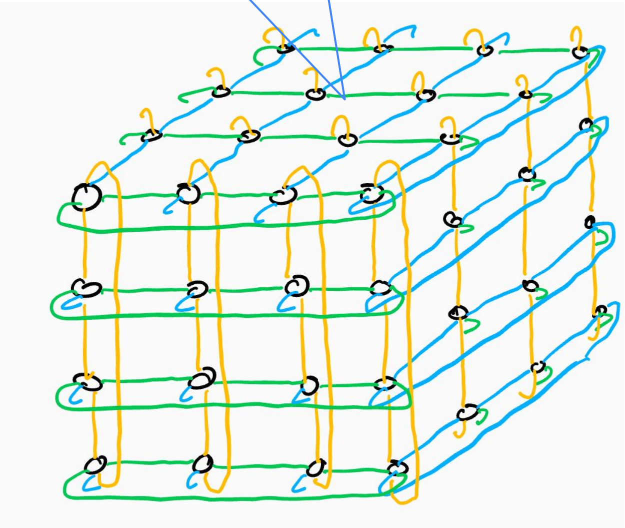
32~128 times
>>



CPU Server
still 25Gbps/100Gbps

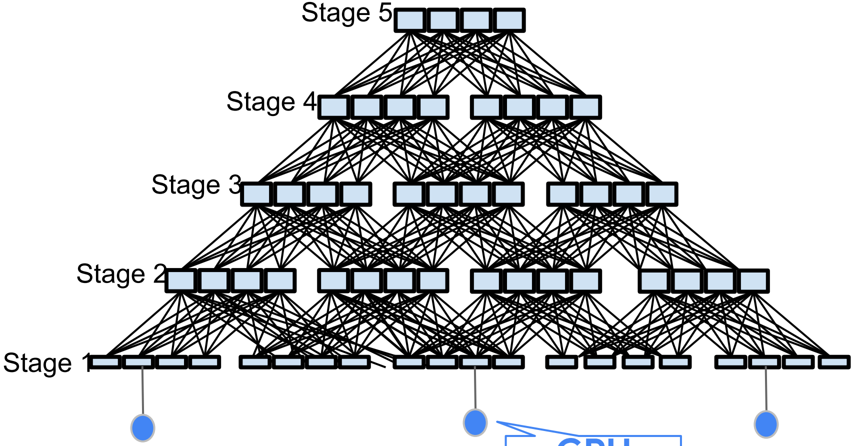
$, 1HWZRUN 'HVLJQ 'LIIHUHQFH 1HWZRUN IR

TPU Server



3D Torus
Each node Direct Connect to 6 others

Google Jupiter Rising

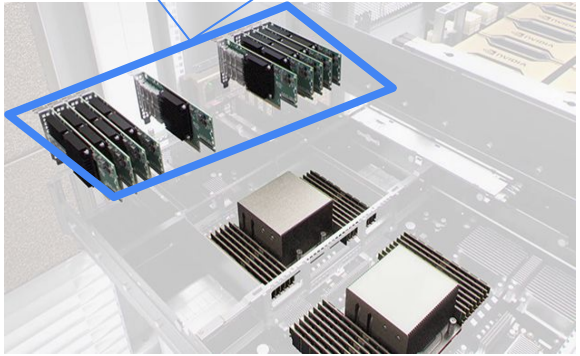


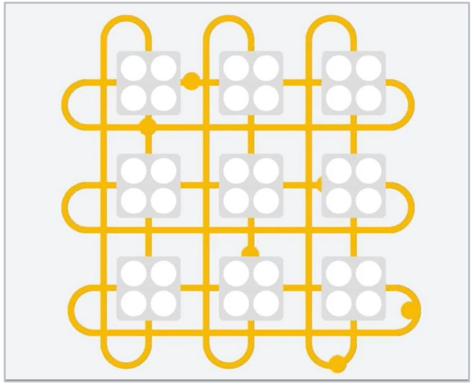
CPU Server

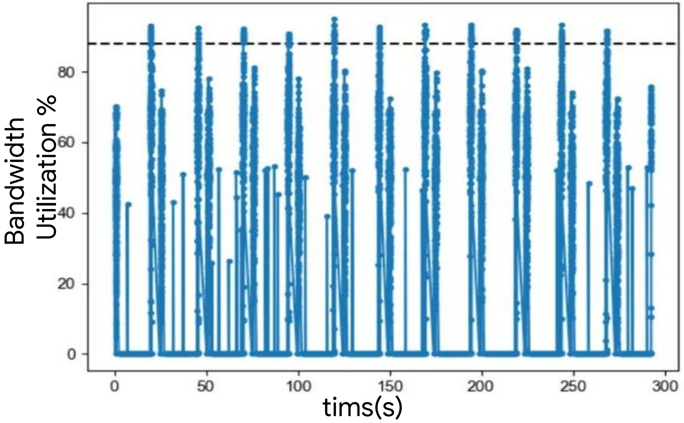
CLOS
Each node up to 9 hops(switches) to others

$, 1HWZRUN 'HVLJQ 'LIIHUHQFH

8x400GE= 3.2Tbps+
vs CPU(25/100Gbps)







Huge bandwidth GPU/TPU
(compute still 25G/100G Server)

AllReduce, All2All
(compute still CLOS)

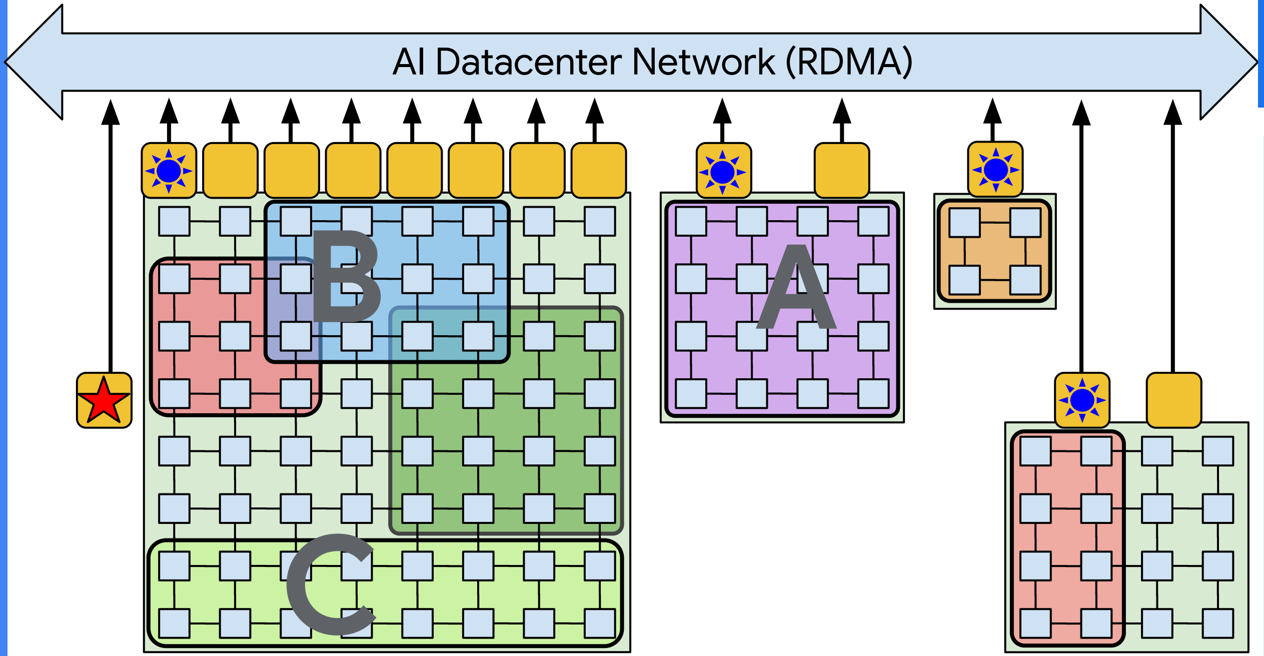
Burst to 3.2Tbps/100%

AI Fabric Innovations

02

Google Cloud

$, 1HWZRUN &OXVWHUV 7RSRORJ\



AI Accelerator Slices (with tightly coupled interconnects), Compute Servers, Storage

🟨 Host (many per island)

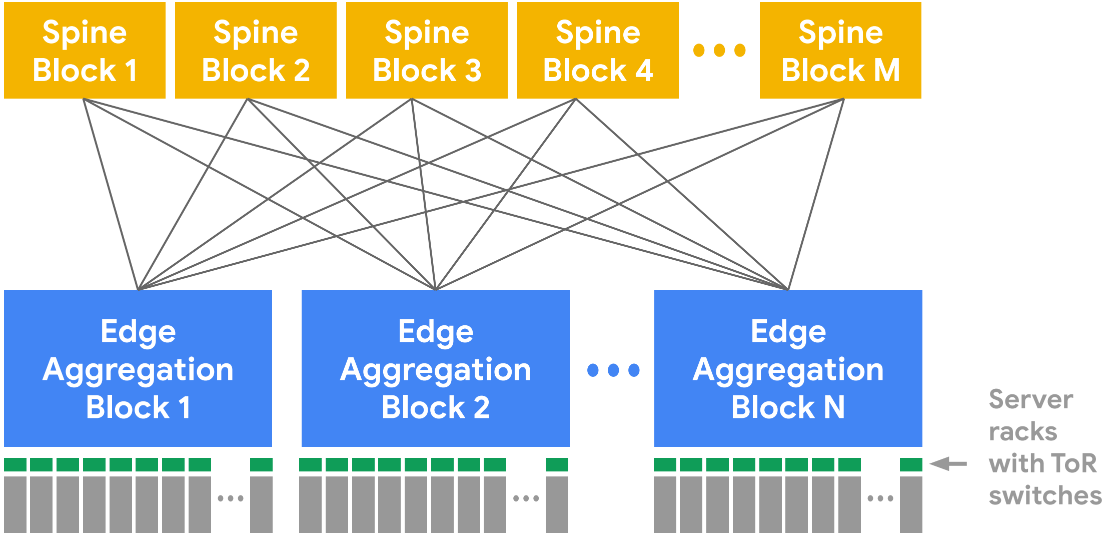
★ Resource Manager (global)

✹ Scheduler (per island)

WHAT WOULD IT TAKE...

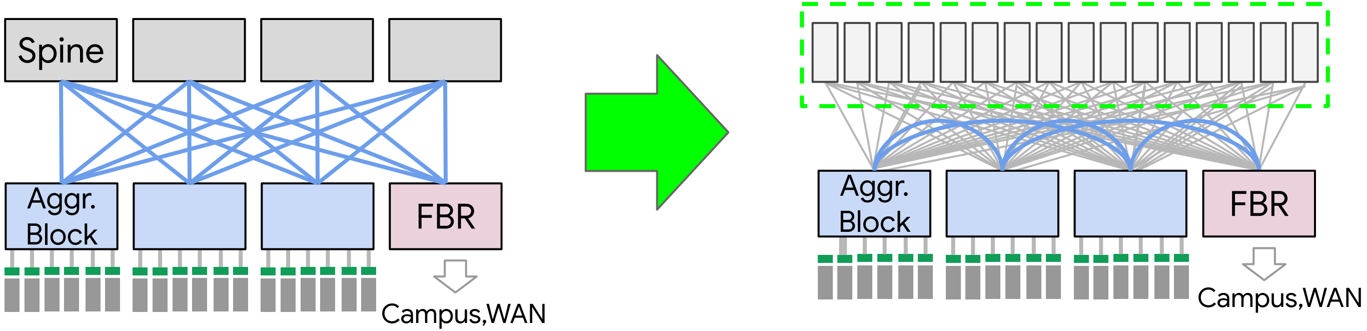
... to achieve **Performance**, **Isolation** and **Efficiency** at scale for High Bandwidth, Low Latency Workloads on **today's Datacenter Ethernet** Networks?

Five generations of clos topologies and software-defined networking



A scalable, commodity data center network architecture, SIGCOMM 2008
Jupiter Rising: A Decade of Clos Topologies and Centralized Control in Google's Datacenter Network, SIGCOMM 2015
Orion: Google's Software-Defined Networking Control Plane, USENIX 2021

/LJKWZDYH)DEULFV IRU 'DWDFHQWHU 1HWZR



Electrical Packet Switch-based DCN

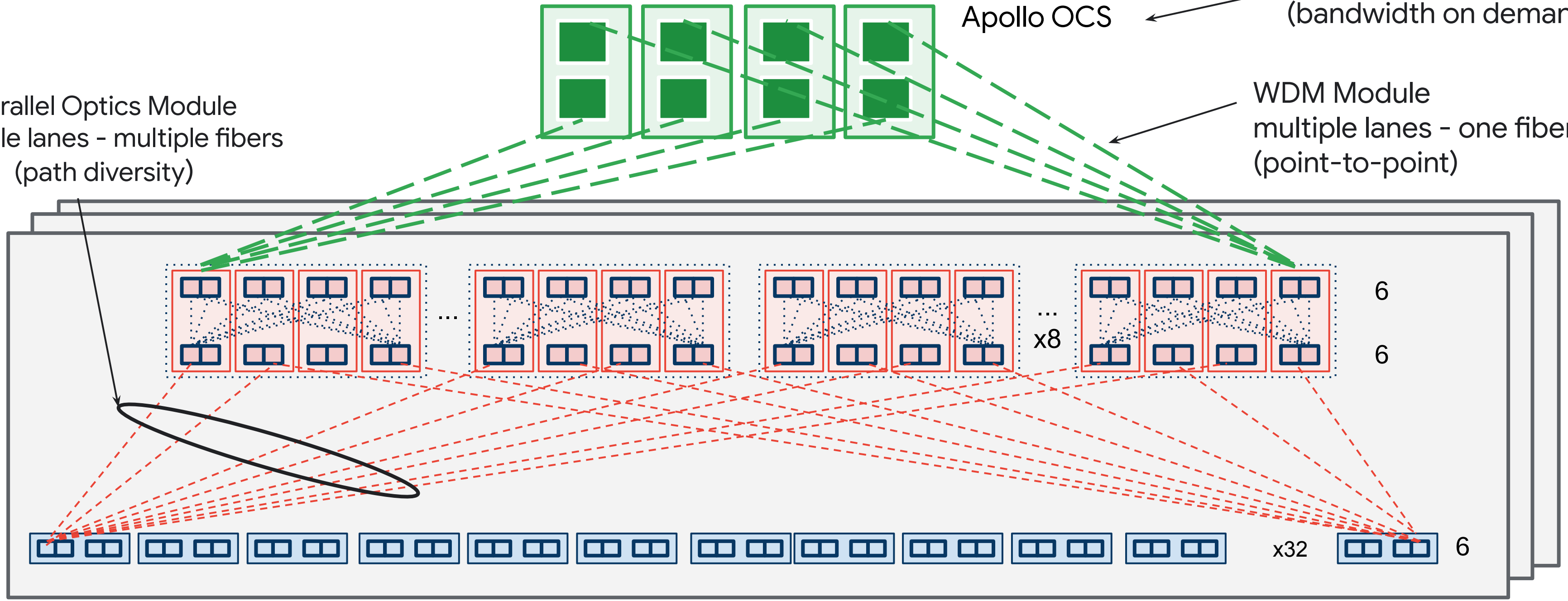
[Jupiter Rising, Sigcomm'15]

OCS-based LW Fabric

[Jupiter Evolving, Sigcomm'22]

- Lightwave Fabric with OCS enables Datacenter Networks with
  - 30% reduction in CapEx and 40% reduction in power consumption
    - No Fiber Change for multiple generations. 100G to 200G to 400G+
    - No Electronic switching in LW Fabric, all Optical Switching.
  - Expansion, topology engineering, heterogeneous networking

2SWLFDO FLUFXLW VZLWFKLQJ LQ WKH GDWD



Apollo OCS

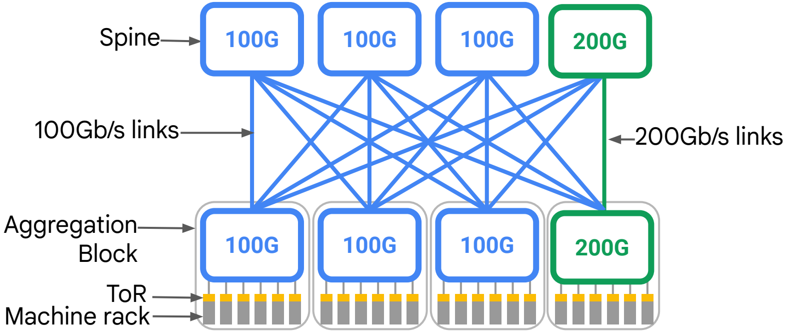
Dynamic functionality (bandwidth on demand)

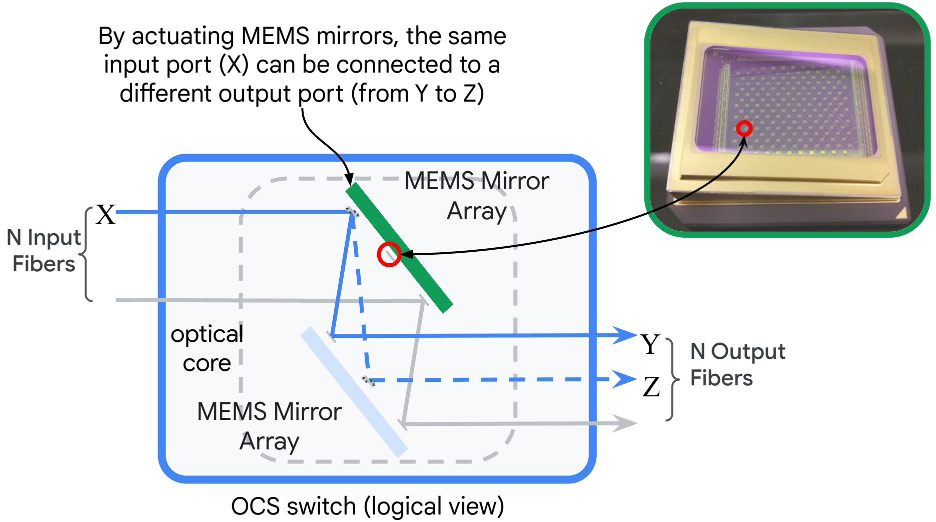
WDM Module
multiple lanes - one fiber
(point-to-point)

Parallel Optics Module
multiple lanes - multiple fibers
(path diversity)

2SWLFDO FLUFXLW VZLWFKLQJ LQ WKH GDWD

30% lower cost
40% lower power
New capability: application-specific topology!



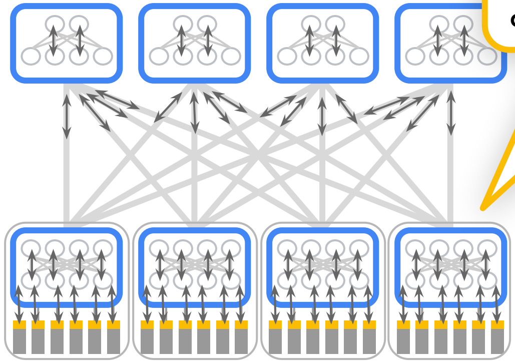

By actuating MEMS mirrors, the same input port (X) can be connected to a different output port (from Y to Z)

MEMS mirror package die with an array of individually controllable micro mirrors

OCS switch (logical view)

7UDGLWLRQDO YV VRIWZDUH GHILQHG QHWZ

Traditional network



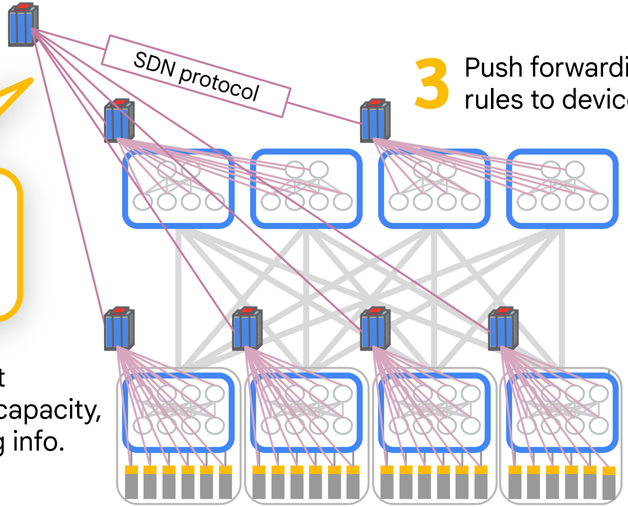
- O(log N) rounds to converge
- O(N log N) messages
- Global consistency and correctness near impossible

Each router exchanges routing info, independently builds reachability graph, populates switching table, and forwards traffic.

Software-defined network (SDN)

2 (Hierarchical) Central SDN controller
Compute best paths across network.

1000x
compute
power



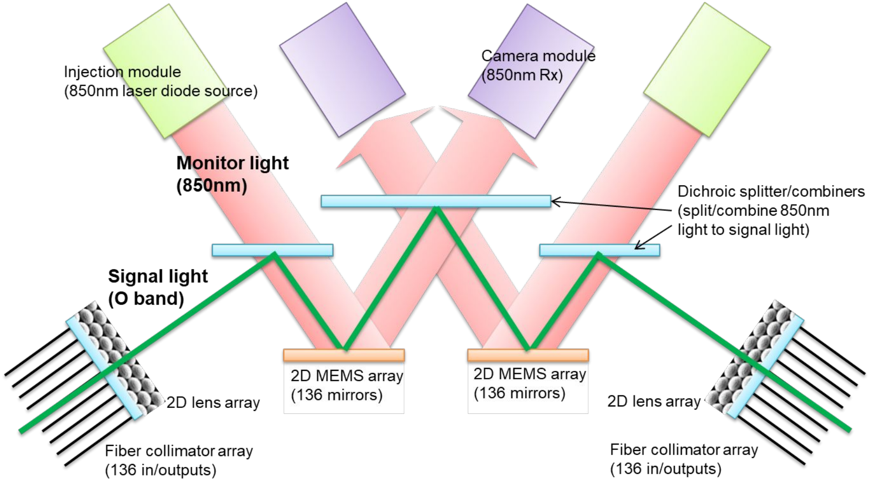
3 Push forwarding rules to devices.

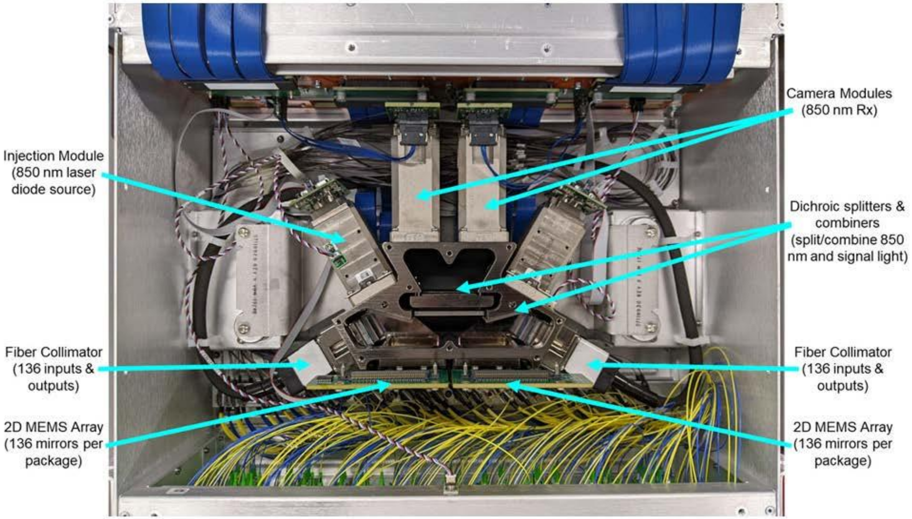
- O(1) rounds to converge
- O(N) messages
- Transition n/w from one globally consistent state to another

1 Extract state, capacity, routing info.

4 Forward traffic along computed paths.

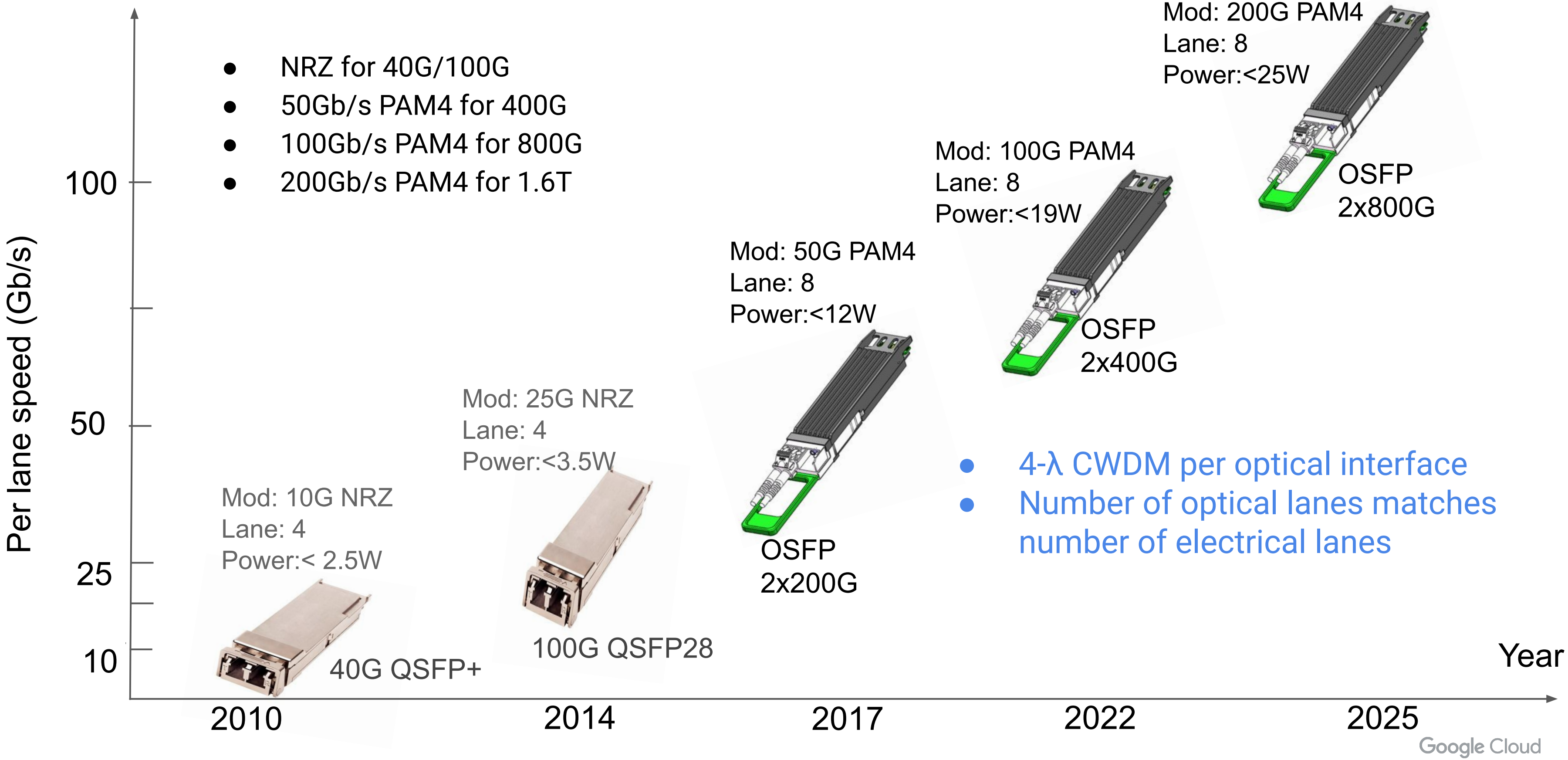
3DORPDU 2SWLFDO &LUFXLW 6ZLWFK





- 136x136 input/output ports
- Camera-based mirror control scheme simplifies design/manufacturing

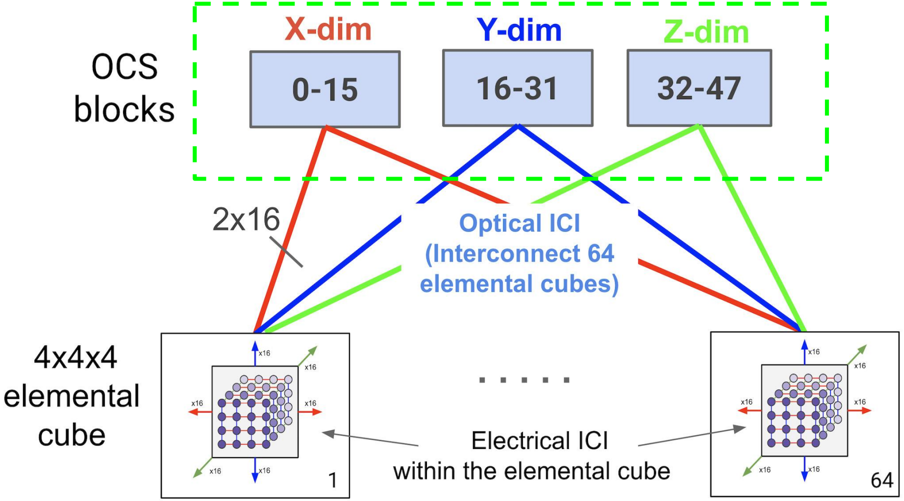
:'0 2SWLFDO 7UDQVFHLYHU IRU '&1



- NRZ for 40G/100G
- 50Gb/s PAM4 for 400G
- 100Gb/s PAM4 for 800G
- 200Gb/s PAM4 for 1.6T

- 4-λ CWDM per optical interface
- Number of optical lanes matches number of electrical lanes

3DORPDU 2SWLFDO &LUFXLW 6ZLWFK



OCS blocks
X-dim
0-15
Y-dim
16-31
Z-dim
32-47
2x16
Optical ICI (Interconnect 64 elemental cubes)
4x4x4 elemental cube
x16
1
. . . . . .
Electrical ICI within the elemental cube
64

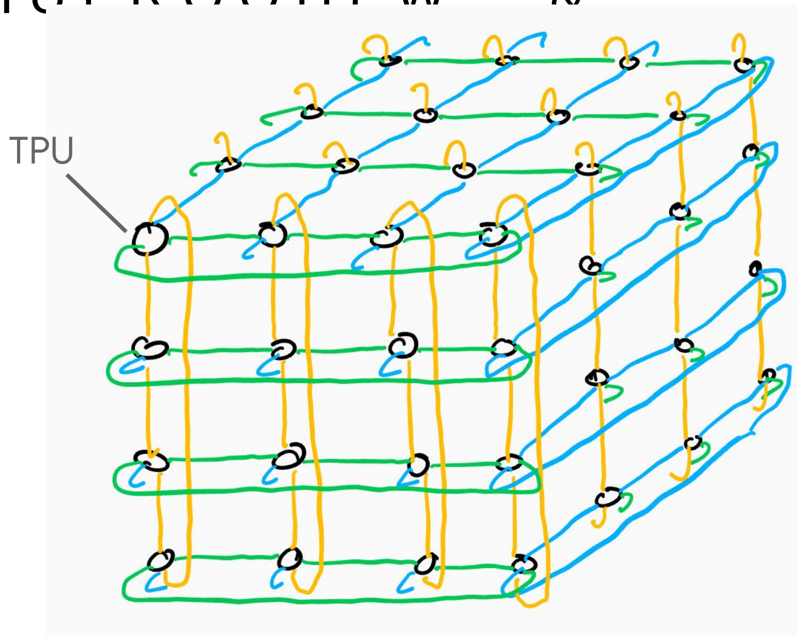
[TPUv4, ISCA’23]

- 4x4x4 multi-TPU cubes tied together by LW Fabric
- LW Fabric enables reconfigurable interconnection between elemental cubes
  - ML Systems with improved scale, availability, utilization, modularity, deployment, security, power, and performance

Google Cloud

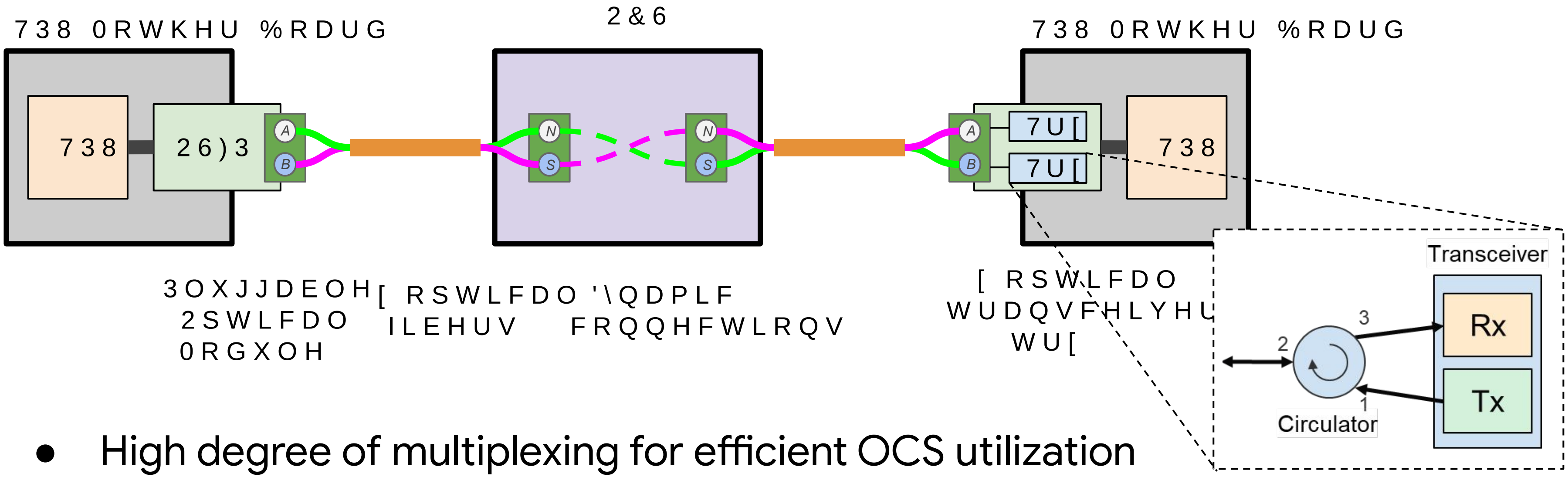
738 LQWHU FKLS LQWHUFRQQHFW &

- TPUs are interconnected using a torus network
  - Each dimension is a ring, has three dimensions
- Well matched to requirements
  - High neighbor bandwidth: “allreduce”
  - Low radix
    - Workload is latency tolerant
    - Simpler router, integrated with TPU
  - Low cost ($-per-BW)
    - Mostly in-rack, passive electrical links
- Torus have a long, but niche history
  - Notable examples: Cray T3x, IBM BlueGene, Fujitsu K computer



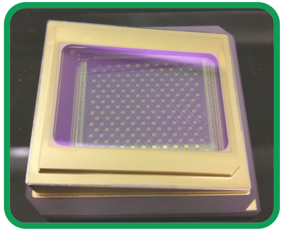
64 TPUs arranged as a 4×4×4 torus network. Each TPU addressed by X,Y,Z coordinates and connected to six neighbors along X+, X-, Y+, Y-, Z+ and Z-.

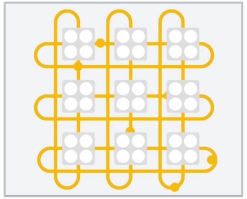
/:)DEULF %XLOGLQJ %ORFNV 2&6 :'0 WU[

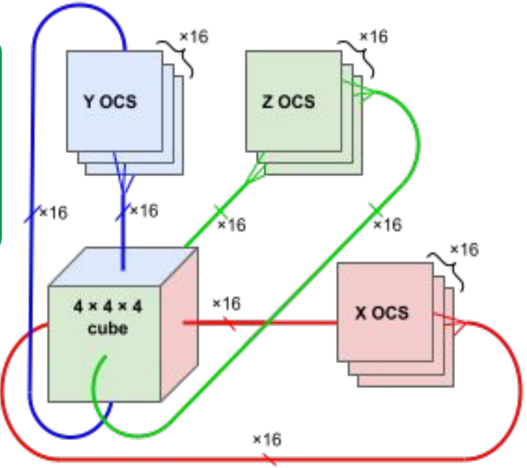


- High degree of multiplexing for efficient OCS utilization
  - WDM transceivers
  - Circulator-based bidirectional links
- Benefits: OCS/fiber is data rate agnostic for extensibility, cost amortization
- Challenges: Higher losses, Multi-Path Interference (MPI) effects

OCS vs IB/NVLink and Ethernet

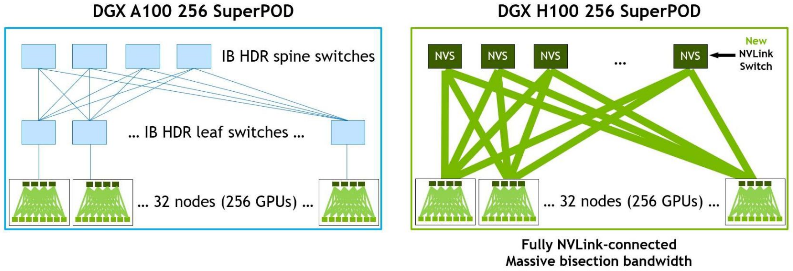


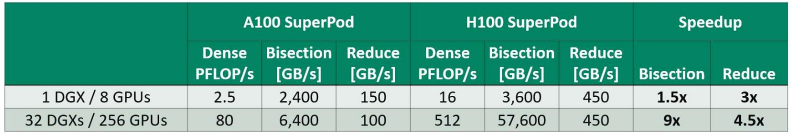




GCP Innovation with OCS

Great for upgrade





	A100 SuperPod			H100 SuperPod			Speedup	
	Dense PFLOP/s	Bisection [GB/s]	Reduce [GB/s]	Dense PFLOP/s	Bisection [GB/s]	Reduce [GB/s]	Bisection	Reduce
1 DGX / 8 GPUs	2.5	2,400	150	16	3,600	450	**1.5x**	**3x**
32 DGXs / 256 GPUs	80	6,400	100	512	57,600	450	**9x**	**4.5x**

Industry with

Infiniband/ Nvlink Switch

2SWLFDO &LUFXLW VZLWFKLQJ YV (WKHUQHW

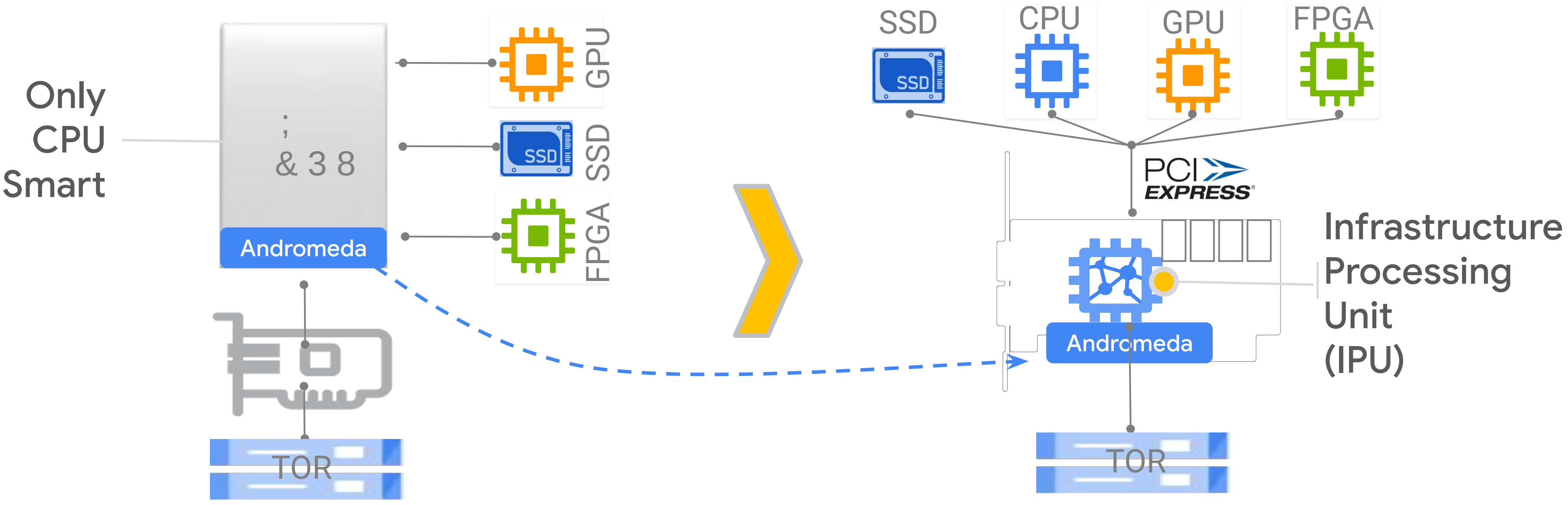
	Optical Circuit switching	Ethernet switching
Speeds	100G/200G/400G/800G	100G/200G/400G/800G
Upgrade	No Change	New Box
Upgrade cable	No Change	New Cables
Fanout	136	128
Latency	Low	High
Flow Control	Per Port/Channel	Per MAC/IP flow
Software Define/Error Isolation	Yes	yes
Power	40%	100%
Cost	30%	100%

AI Smart Offload Transport Innovation

03

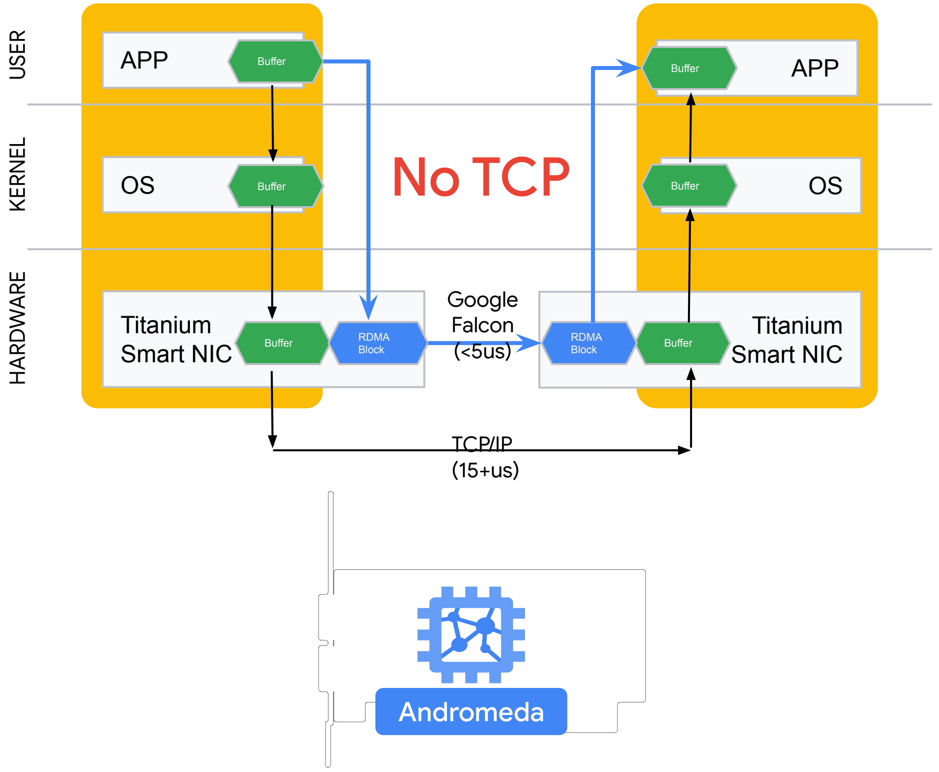
CPU or SmartNic?

Cloud Offload from X86



AI Transport Smart Offload From Google

Falcon RDMA



Predictable Efficiency performance @ warehouse-scale:

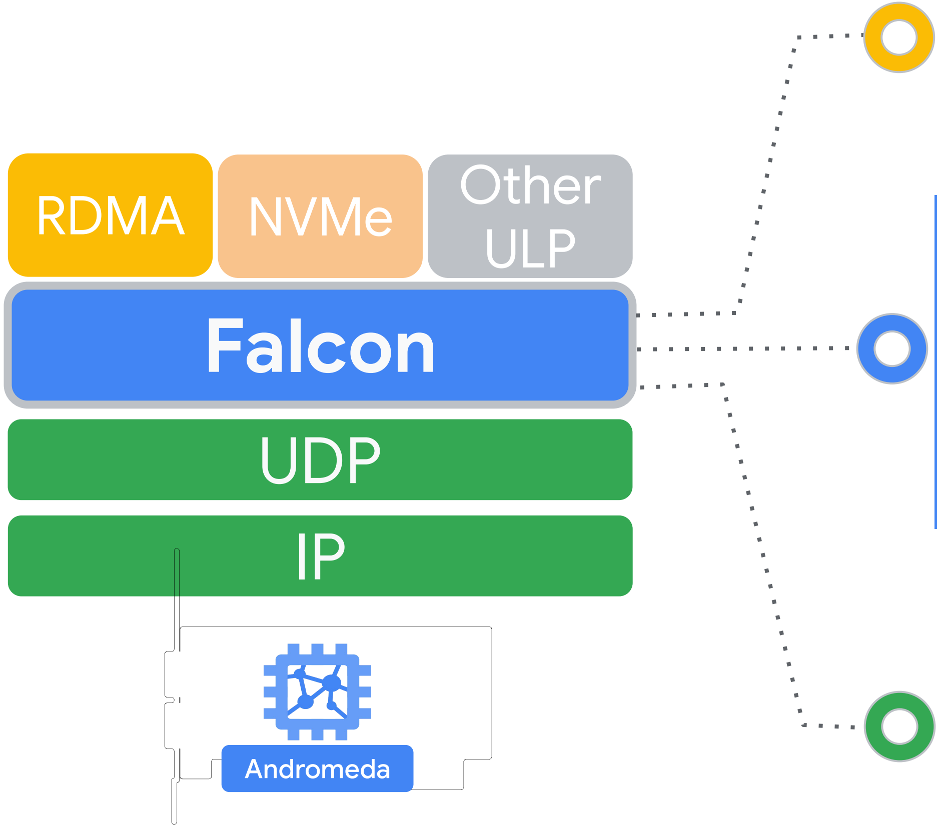
hardware acceleration, offloads CPU from data movement, Low-latency with OS-bypass, massive application bandwidth, mitigating congestion and efficient network utilization.

Google Falcon

introduces usability and scalability improvements via relaxed ordering and robust error handling.

Need of the day:

meets requirements of critical workloads, HPC and AI; also good for offloading Storage and RPC.



Tail Latency in Ethernet networks

→ HW assisted **delay-based Congestion Control**

→ **Selective ACKs** for fast loss recovery

→ **Multipath** capable connections

Isolation and Visibility at scale

→ µs-granularity per-flow **Traffic Shaping**

→ Fine-grained Stats for **Debuggability, Software Defined Network control**

Efficiency and Security

→ **Implemented in HW** for Low Latency, High Op Rate using Industry-standard Interfaces, and PSP encryption

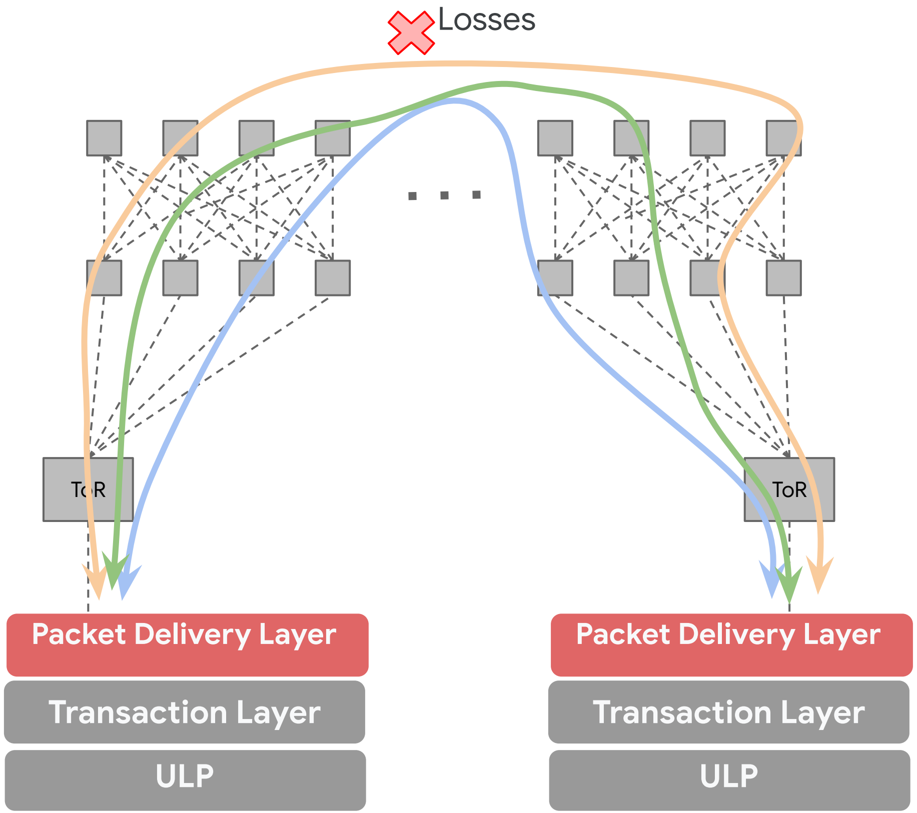
Falcon Packet Delivery Layer

Delay-based Congestion Control for low latency and high utilization.

Leverages multiple paths in the network fabric transparently to applications.

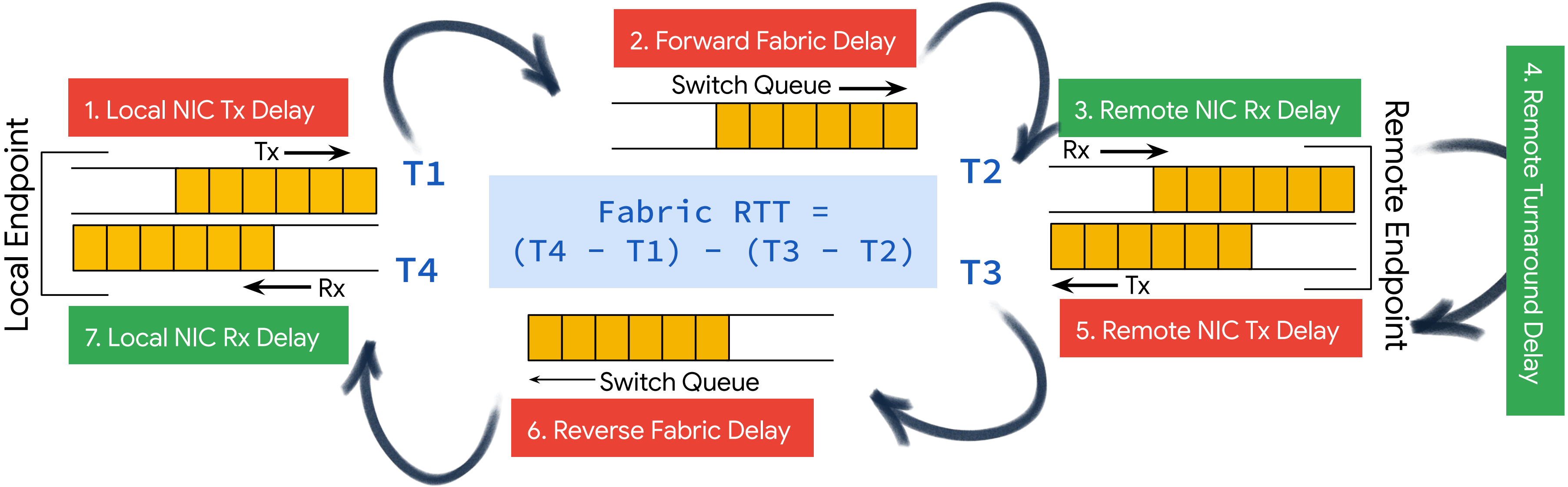
End-to-end reliable delivery

- Timely retransmission of lost packets.
- Hardware based retransmission.
- Ack coalescing/piggybacking for high Op rate.



Google

Swift Congestion Control as Baseline



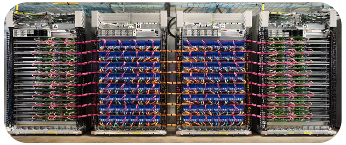
Swift* is a delay based congestion-control for Datacenters that achieves low-latency, high-utilization, near-zero loss implemented completely at end-hosts and NICs supporting diverse workloads like large-scale incast across latency-sensitive, bursty and IOPS-intensive applications working seamlessly in heterogeneous datacenters.

*Swift: Delay is simple and effective for congestion control in the Datacenter, SIGCOMM 2020.

Summary

04

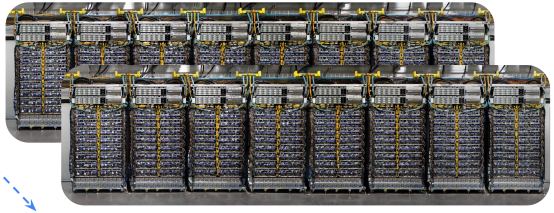
Rapid Innovation with Cloud TPUs



Cloud TPU v2

- Domain-specific AI supercomputing
- 256 chips distributed shared memory

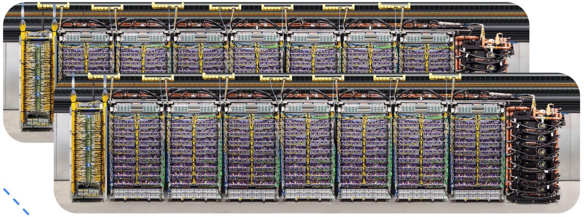
8x



Cloud TPU v4

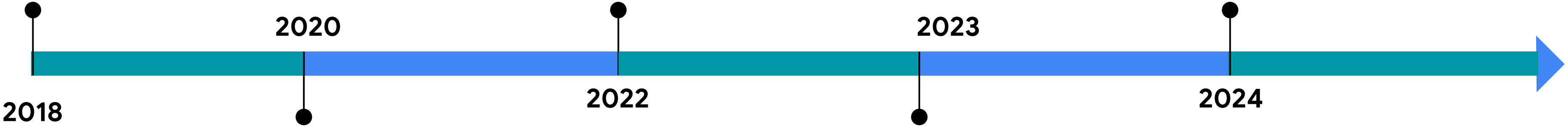
- Optically reconfigurable 3D Torus
- 4k chips with distributed shared memory

20x



Cloud TPU v5p

- Programmable Sparsecores for embeddings
- 9k chips with distributed shared memory



2018

2020

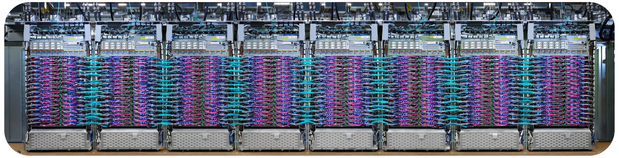
2022

2023

2024

Cloud TPU v3

- Liquid cooling
- 1k chips distributed shared memory



Cloud TPU v5e

- Efficient and scalable training and serving
- 256 chips, horizontally scalable to 10s of k

6XPPDU\

- AI Fabric Innovation: What are the Benefits of Lightwave Fabrics for DCN and ML?
  - Provides direct optical connections (circuits) between network endpoints
  - Comprised of an optical circuit switch (OCS), WDM optical transceivers (trx) co-designed with OCS, circulators, and the hardware/software control plane
  - Reconfigurable, extensible fabric for both datacenter networks (DCN) and ML
  - Enables performant, cost & energy efficient DCNs and & ML supercomputers
    - DCN: 30%/40% reduction of CapEx/OpEx
    - ML: Ability to run large/multi-k node systems; Up to 3.3x speed up in model training

- AI Software Transport Innovation

  - Design new Protocol for Remote DMA for AI Network(GPU/CPU and TPU)

Google

Thank you