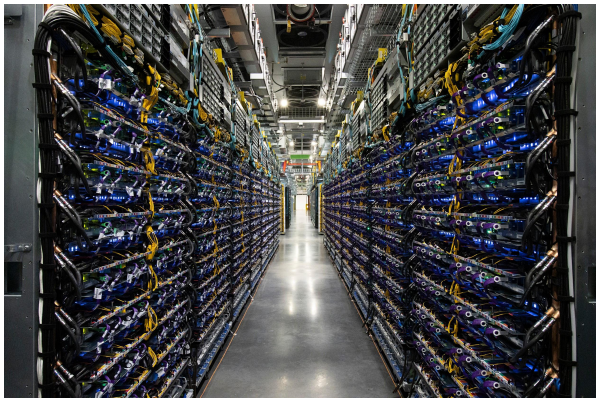
Google

A Machine Learning Supercomputer With An Optically Reconfigurable Interconnect and Embeddings Support

Norman P. Jouppi and Andy Swing
With contributions from many others
Hot Chips August 29, 2023

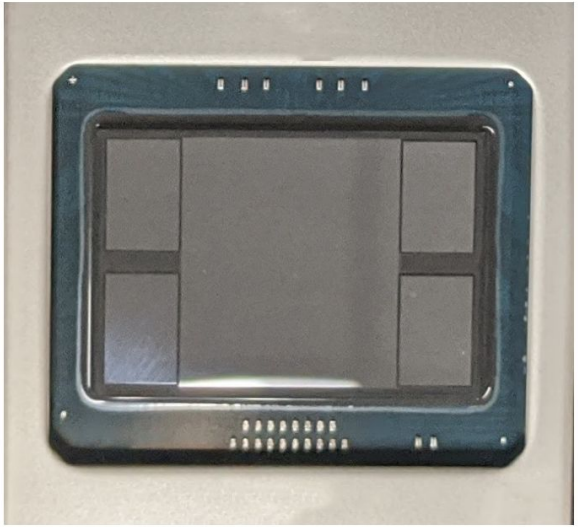


Breakthrough Innovations in Our 4th-Generation System

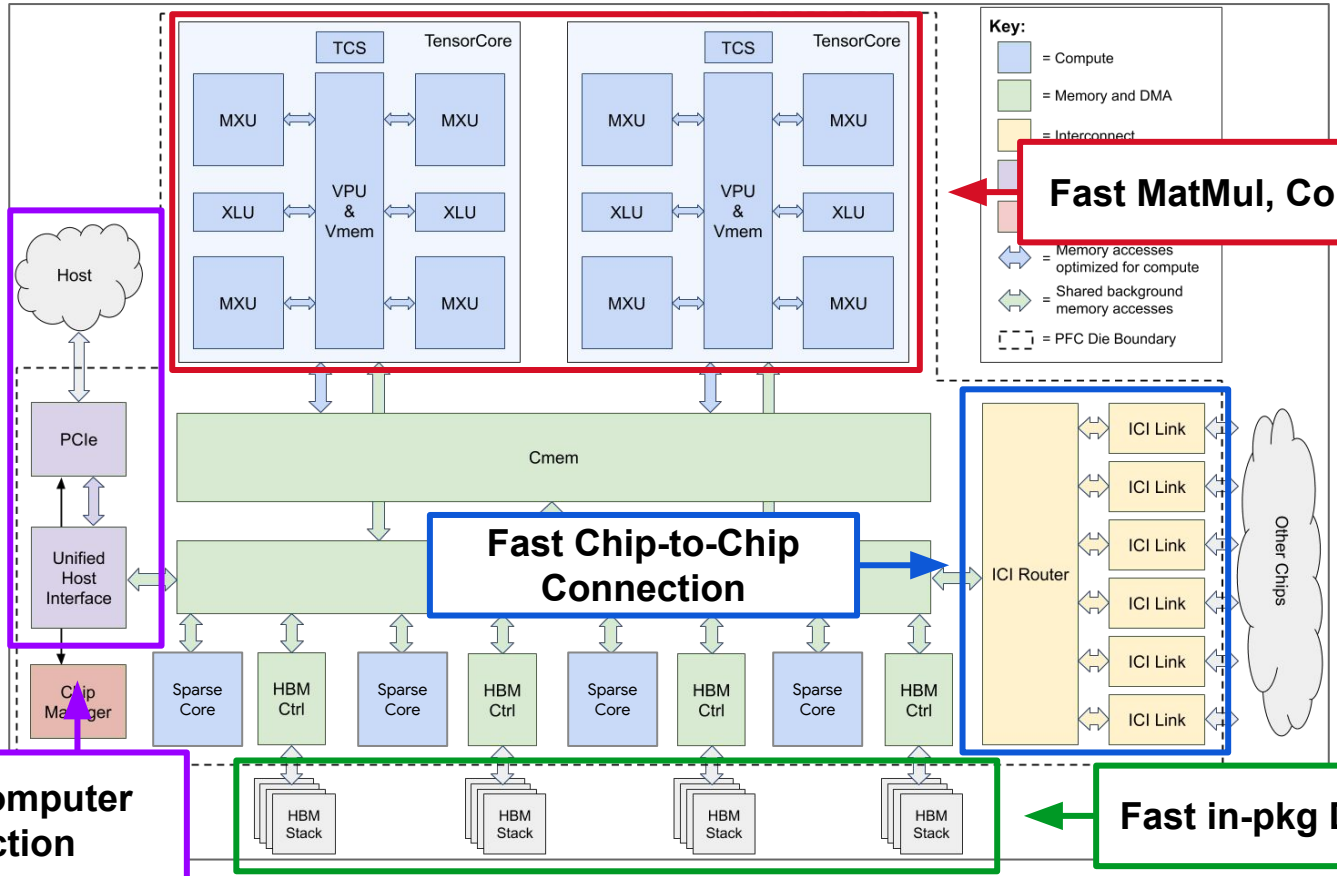
- Optical circuit switching (OCS)
- 3rd-generation embeddings coprocessor (Sparsecore)
- Industry-leading compute power efficiency
- 4096 chips all sharing 256 TiB of HBM memory
- In production at hyperscale from **2020**

The TPUv4 Chip

- 7nm process
- Optimized for training, superset of TPU v4i:
  - Two TPUv4i Tensorcores
  - 2X HBM of TPUv4i
  - 3D vs. 2D torus
- 275 peak TFLOPS
  - BF16 with FP32 accumulation
  - Also supports int8 like TPUv4i
- Typical power ~200W
  - TDP is higher to guarantee SLOs and prevent throttling
  - Peak power and water cooling are cheaper than SLO violations



TPUv4 Chip Architecture



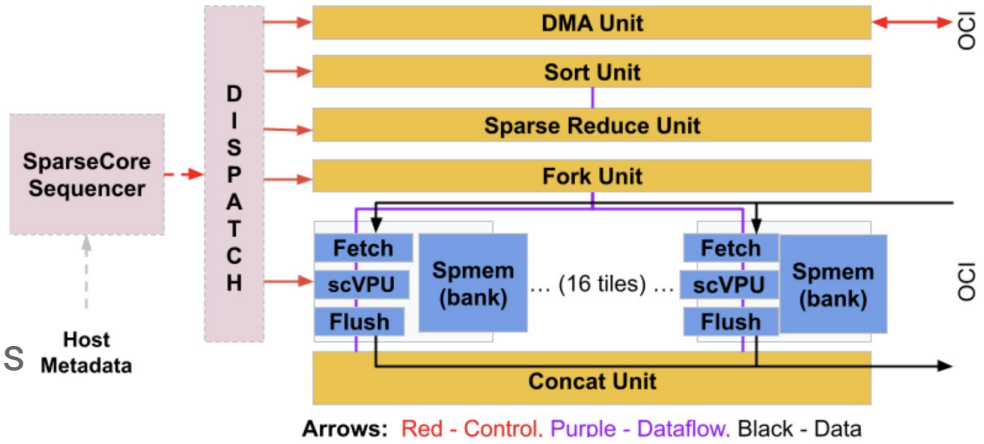
TPUv4 vs TPUv3: >2X Performance at Lower Power!

- Key stats:

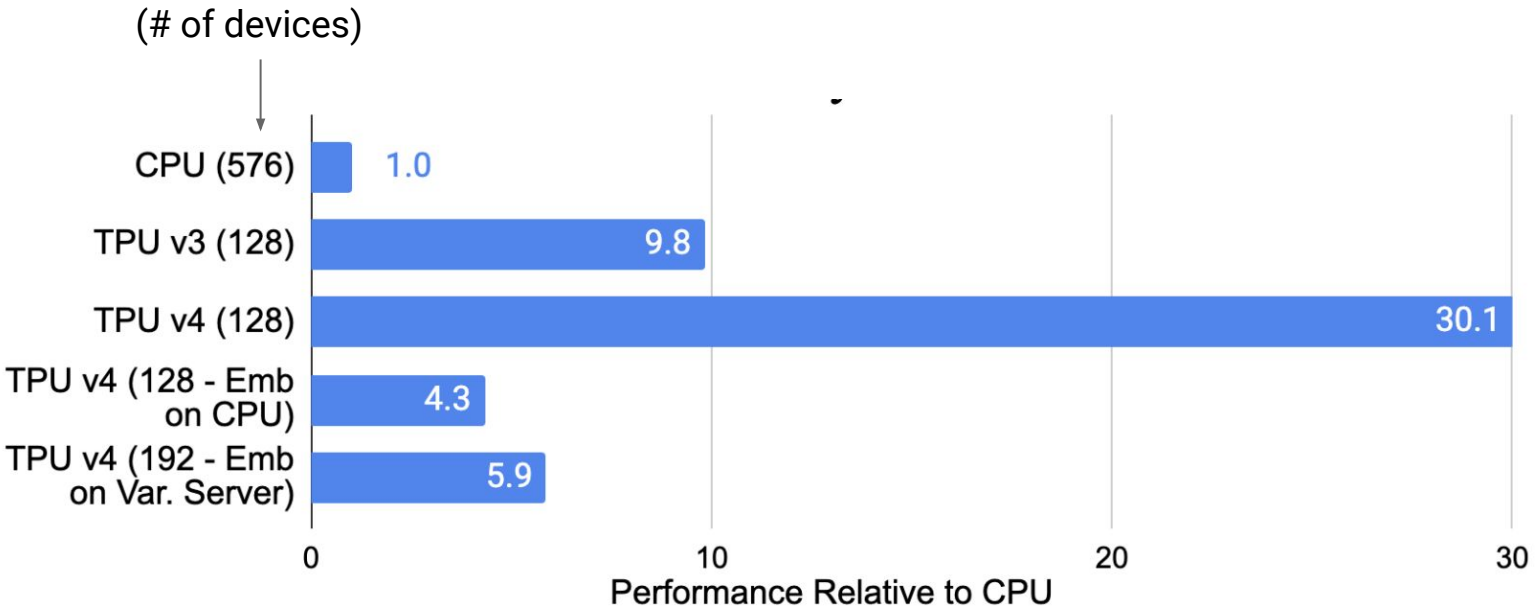
	Google TPUv4	TPUv3
Production deployment	2020	2018
Peak TFLOPS	275 (bf16 or int8)	123 (bf16)
Clock Rate	1050 MHz	940 MHz
Tech. node, Die size	7 nm, <600 mm2	16 nm, < 700 mm2
Transistor count	22 billion	10 billion
Chips per CPU host	4	8
TDP	N.A.	N.A.
Idle, min/mean/max power	90, 121/170/192 W	123, 175/220/262 W
Inter Chip Interconnect	6 links @ 50 GB/s	4 links @ 70 GB/s
Largest scale configuration	4096 chips	1024 chips
Processor Style	Single Instruction 2D Data	Single Instruction 2D Data
Processors / Chip	2	2
Threads / Core	1	1
SparseCores / Chip	4	2
On Chip Memory	128 (CMEM) + 32 MiB (VMEM) + 10 MiB (spMEM)	32 MiB (VMEM) + 5 MiB (spMEM)
Register File Size	0.25 MiB	0.25 MiB
HBM2 capacity, BW	32 GiB, 1200 GB/s	32 GiB, 900 GB/s

SparseCore: Embeddings Accelerator inside the TPU

- Programmable accelerator mainly for embedding computations (used in recommendation models)
- 3rd generation SparseCore
- SparseCores leverage non-coherent shared memory across a pod
- Massive memory parallelism (millions of outstanding references accessing any node in the pod) is exploited with multithreading

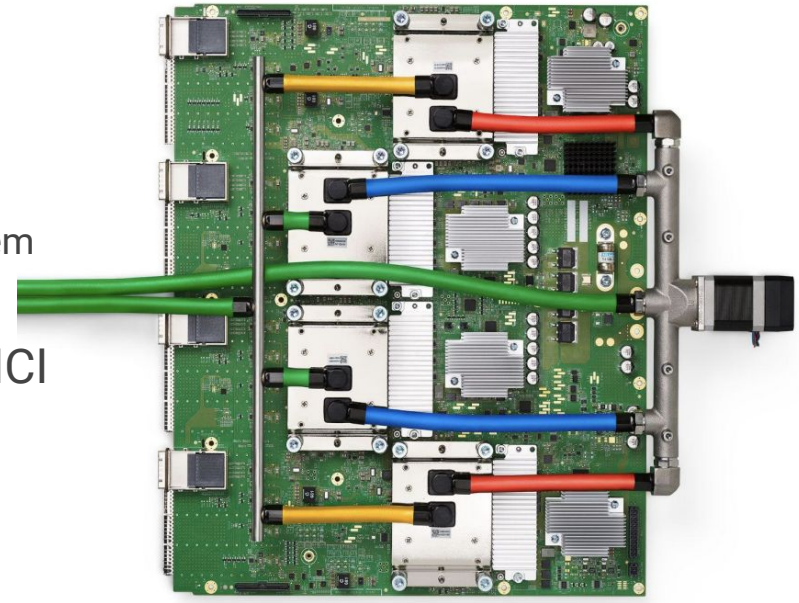


SparseCore Performance



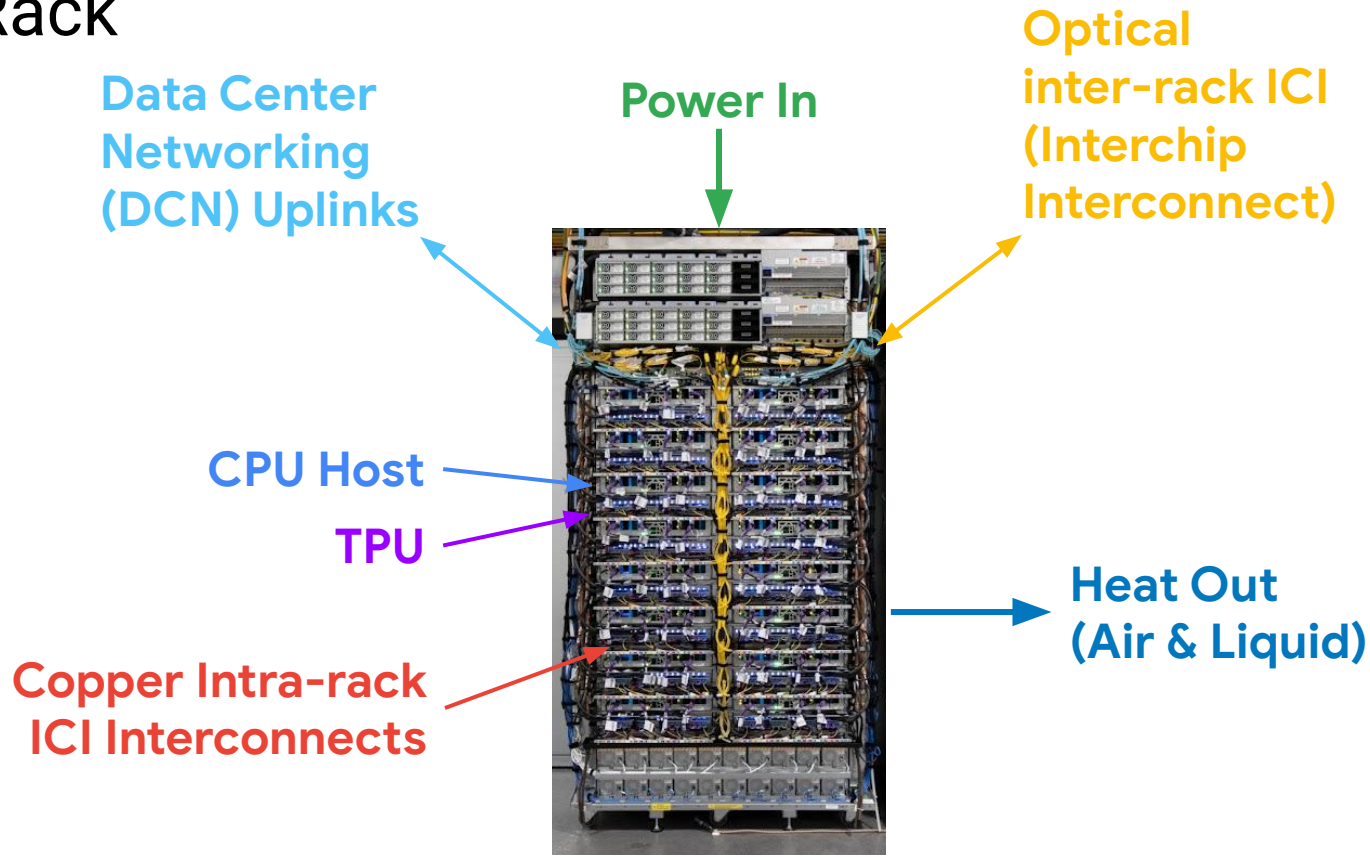
The Board

- 4 TPUs per board
- Liquid cooled
  - 4 chips with parallel water flow
  - Flow rate controlled by valve
  - Similar to fan speed control in an air-cooled system
- PCIe Gen3x16 per TPU for host I/O
- 4 OSFP[1] connectors per TPU for off-board ICI
  - Each OSFP supports 400Gbs each direction
  - 2 more links per chip on-board for interconnect



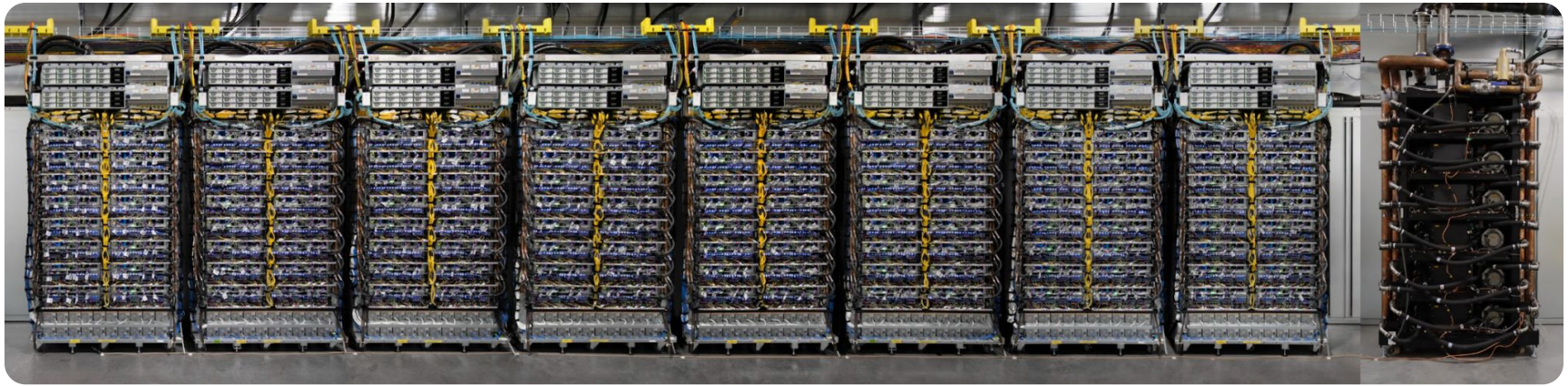
[1]OSFP = “Octal Small Form Factor Pluggable”

The Rack


Data Center Networking (DCN) Uplinks
Power In
Optical inter-rack ICI (Interchip Interconnect)
CPU Host
TPU
Heat Out (Air & Liquid)
Copper Intra-rack ICI Interconnects

The System

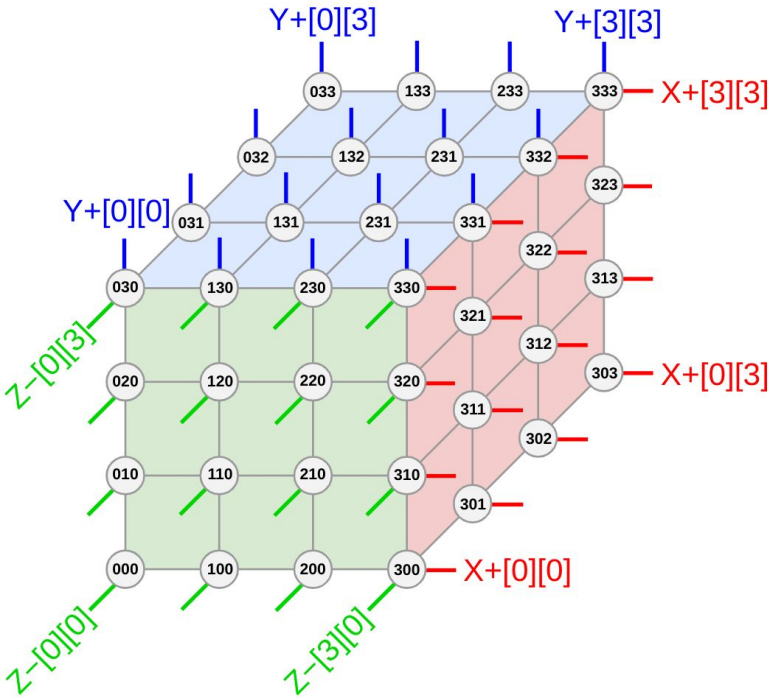
- Each system consists of 64 Google racks, deployed in 8 groups of 8
  - 4096 interconnected chips sharing 256TiB of HBM memory
  - Total compute >1 ExaFLOP
  - Each group of 8 racks gets a CDU (Coolant Distribution Unit)
- Dozens of systems deployed [Sundar, Google I/O]
  - Up to 8 superpod systems in a single cluster!

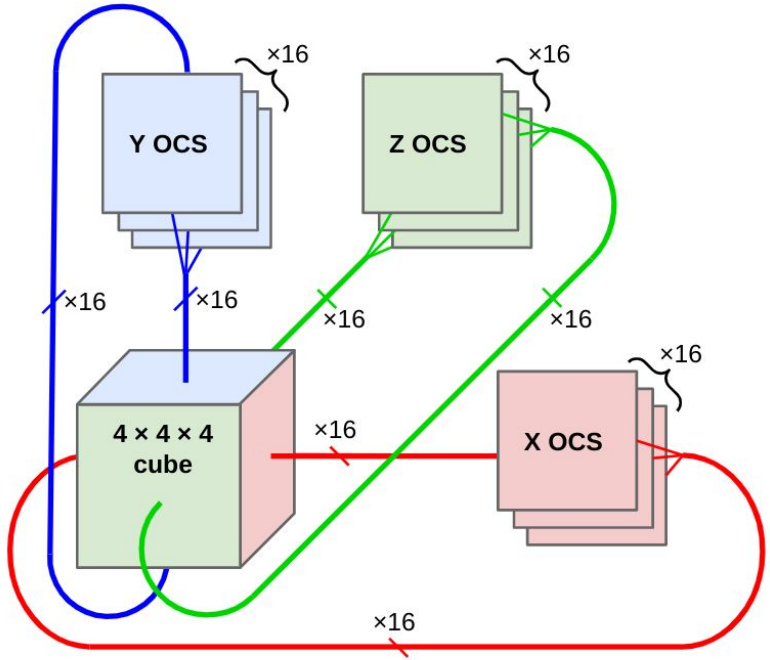


8x TPU Racks | 1x CDU

Each Rack is a 64 4x4x4 Cube, Connected With 48 OCS

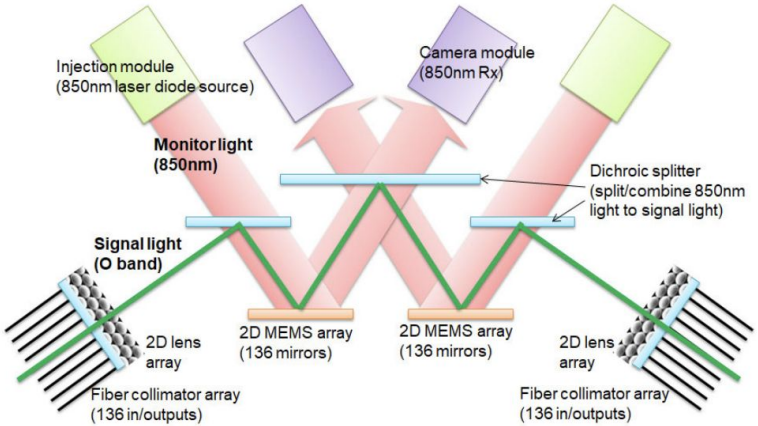
- Different ranks of OCS switch different dimensions and indices

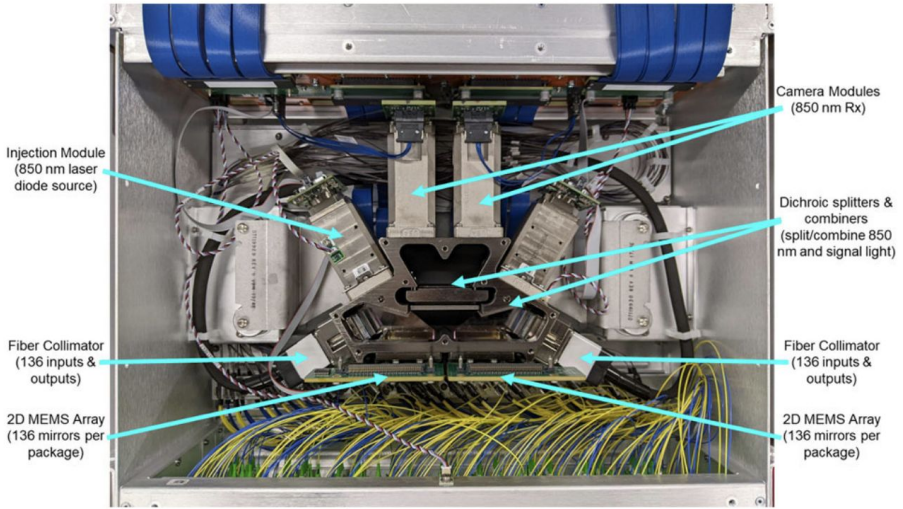




The Optical Circuit Switch (OCS)

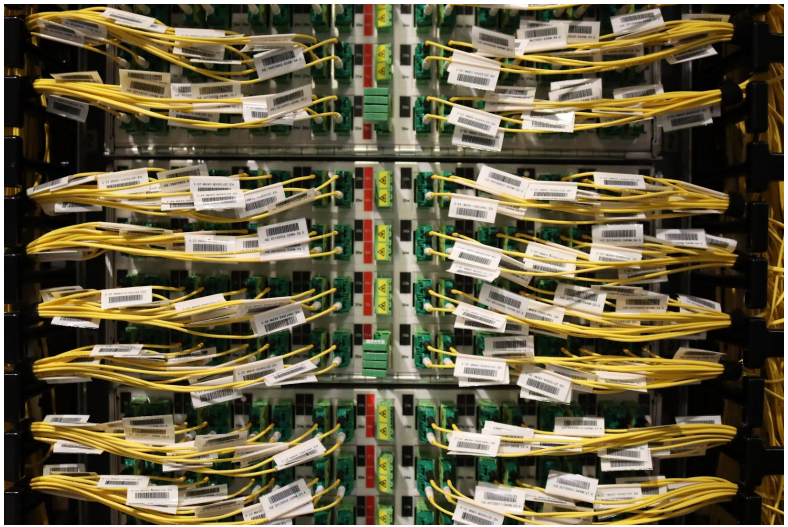
- Builds a direct light connection between two optical fibers using mirrors
  - Set up at the beginning of each job's slice allocation
- No switching of packets and multiple protocol levels like an electrical switch
  - A direct fiber connection requires less power and incurs less latency, no congestion, etc.
  - Enables efficient distributed shared memory across up to 8K Tensorcores and 16K Sparsecores

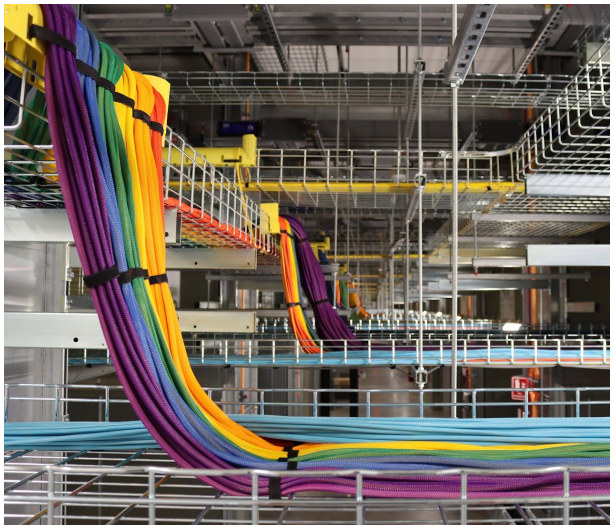




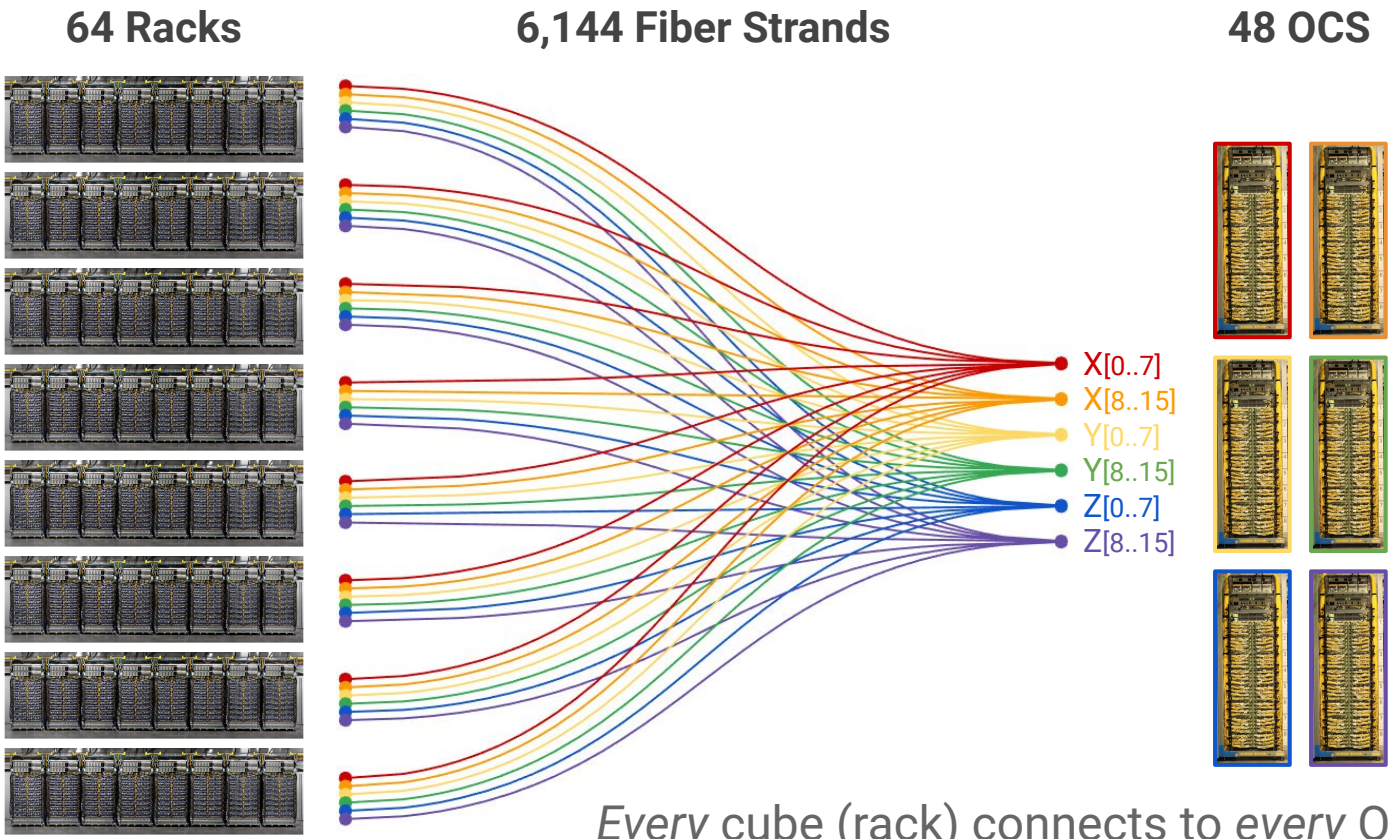
The Fiber

- Each Superpod has enough fiber to encircle the state of Rhode Island!
- Over 16,000 individual connections
- Major focus on deployability and serviceability





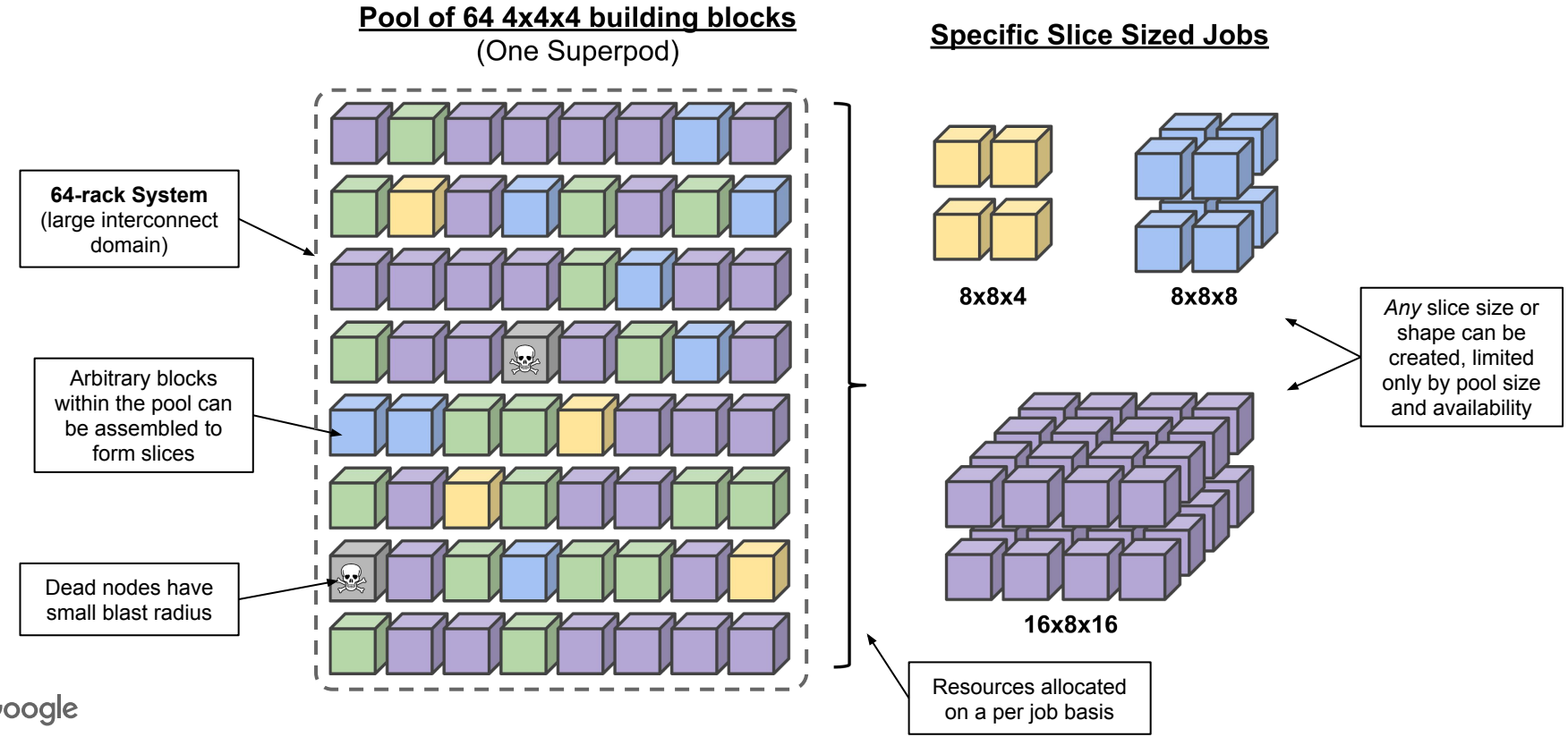
The Physical System



Every cube (rack) connects to *every* OCS

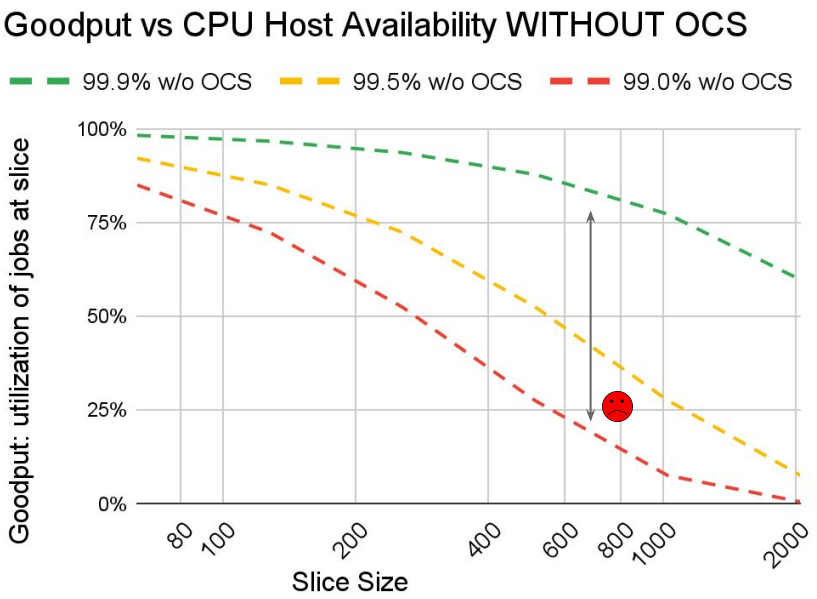
The Logical System

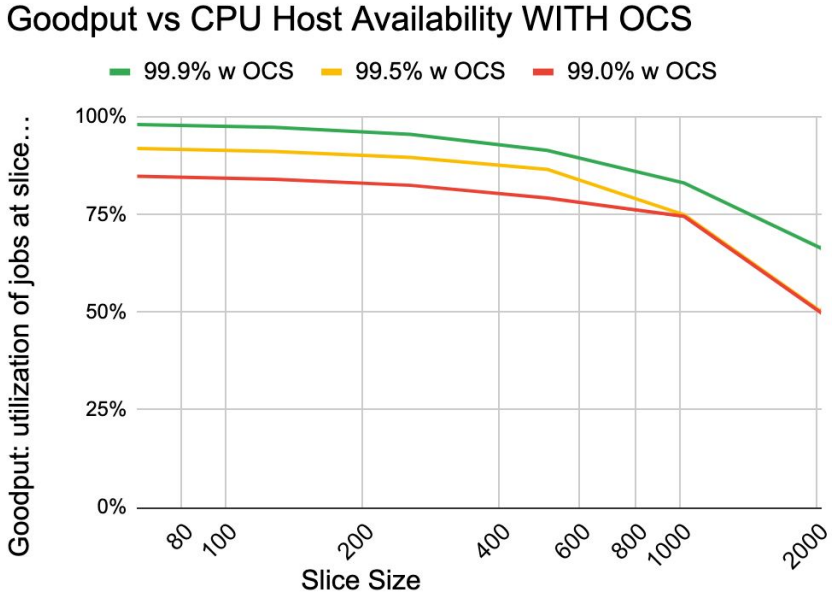
A large pool of building blocks that can be connected on a per-job basis to form larger slices.



Slice Availability Without and With Optical Circuit Switch[1]

Enables reliable provisioning of larger slices vs. previous hardwired designs

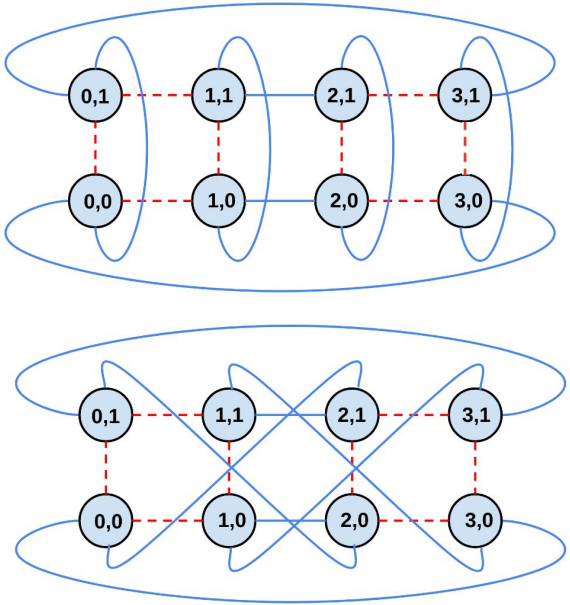




[1]Results are modeled based on different node availabilities.

Optical Circuit Switch Enables Diverse Topologies

Supports regular and twisted tori

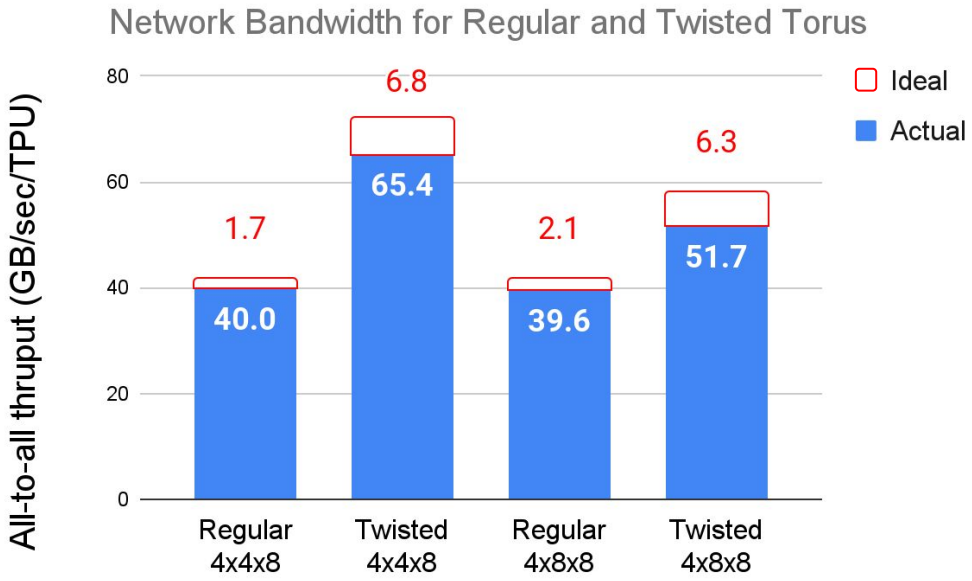


- Non-regular topologies support different parallelisms: model, data, pipelined, etc.
- No rewiring; just reload routing table

Case	Versions	Through-put (seqs/sec)	Hyper-Parameters (topology, partition spec [pipeline, data, model1, model2], 1D/2D activation/weight partitioning)
LLM	Novice's pick	17.9 (1.0x)	4×8×16, [1, 1, 16, 32], 2D/2D
	Best perf.	41.3 (2.3x)	8×8×8, [1, 1, 64, 8], 1D/2D
GPT-3 Pre-training	Expert's pick	21.0 (1.0x)	8×8×8, [8, 1, 8, 8], 2D/2D
	Best perf.	25.0 (1.2x)	4×8×16, [16, 4, 1, 8], 1D/1D

Benefits of Twisted Tori

Example for all-to-all communication:



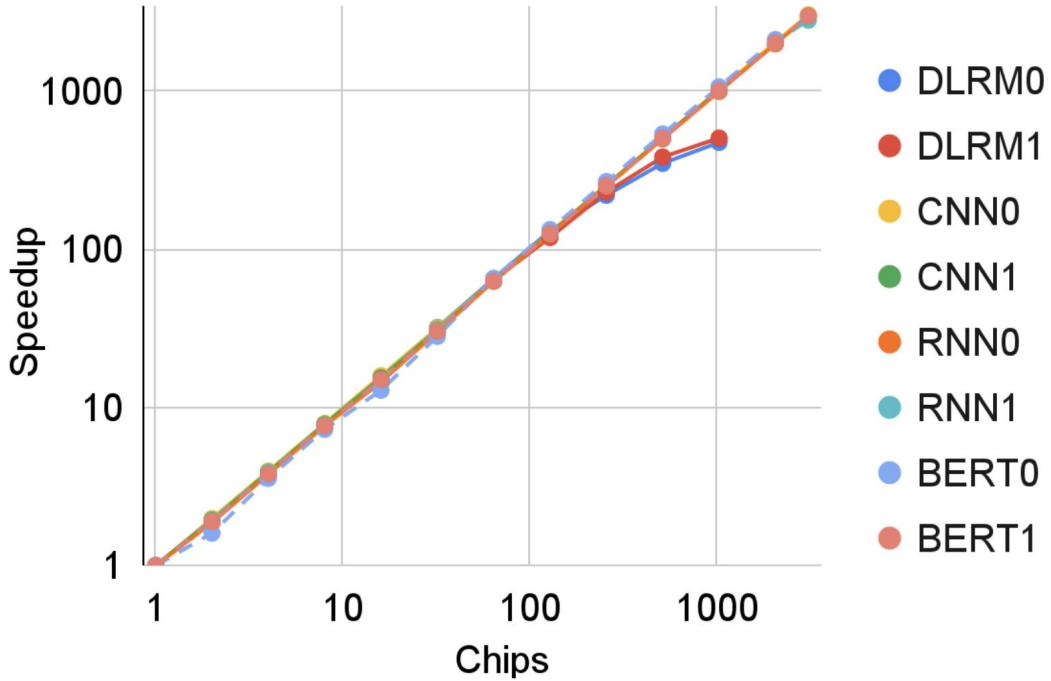
Putting It All Together

- Significant changes in workload mix over the last 6 years
- Requires significant system flexibility to be performant over lifetime of system

DNN Model	*TPU v1 7/2016 (Inference)*	*TPU v3 4/2019 (Training & Inference)*	*TPU v4 Lite 2/2020 (Inference)*	*TPU v4 10/2022 (Training)*
MLP/DLRM	61%	27%	25%	24%
RNN	29%	21%	29%	2%
CNN	5%	24%	18%	12%
Transformer	--	21%	28%	57%
(BERT)	--	--	*(28%)*	*(26%)*
(LLM)	--	--	--	*(31%)*

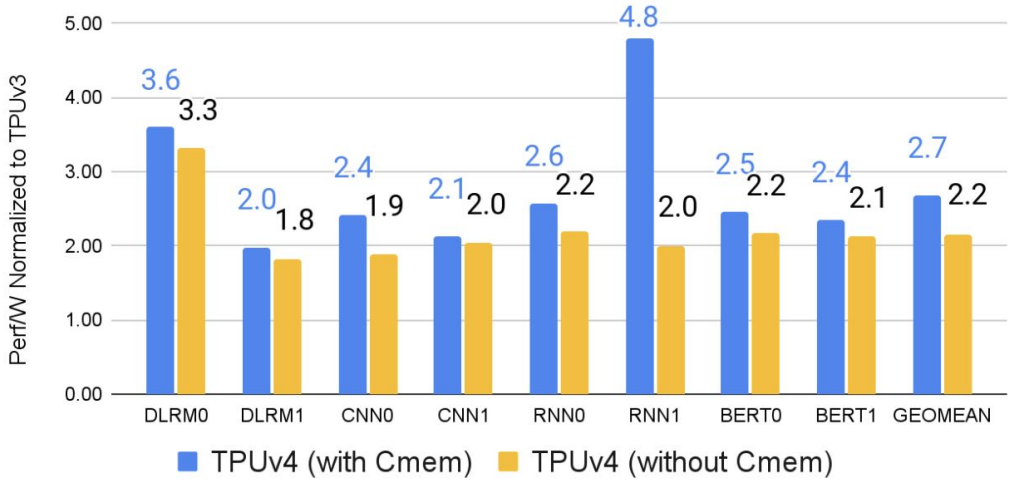
Scalability

- Goal was to create a highly scalable balanced system
- Hence TPUs connected by high BW to distributed shared memory
- We have ~linear speedups up to 3072 chips on internal workloads except for DLRMs

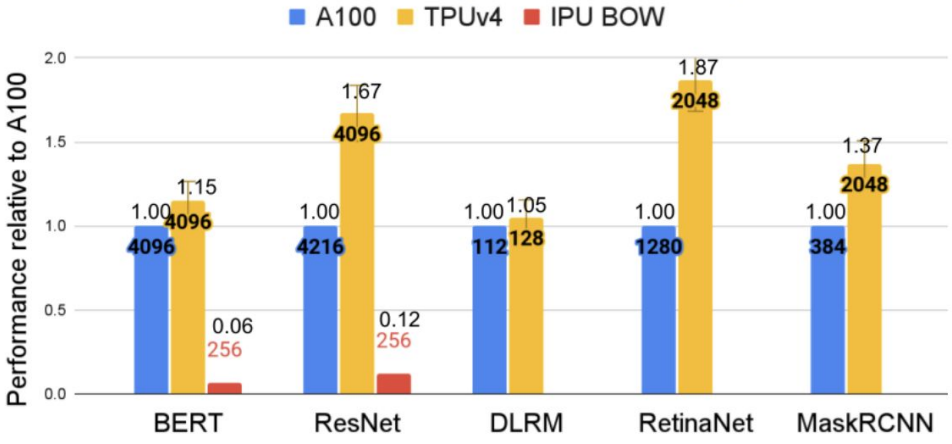


Perf/W Improvements from Increased On-chip Memory

- TPUv4 provides significant increases in perf/W over TPUv3: **2.7X geomean**
- CMem provides 128MB large-on-chip memory shared between TPUv4 Tensorcores
  - Helps reduce geomean power by 22%



Comparisons With Other Contemporary Platforms



Reported MLPerf Training 2.0 highest performance relative to A100

MLPerf Benchmark	A100	TPU v4	Ratio
BERT	380 W	197 W	1.93
ResNet	273 W	206 W	1.33

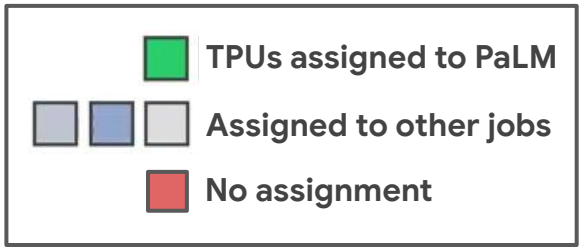
Mean power for DSA plus HBM for 64-chip systems running MLPerf

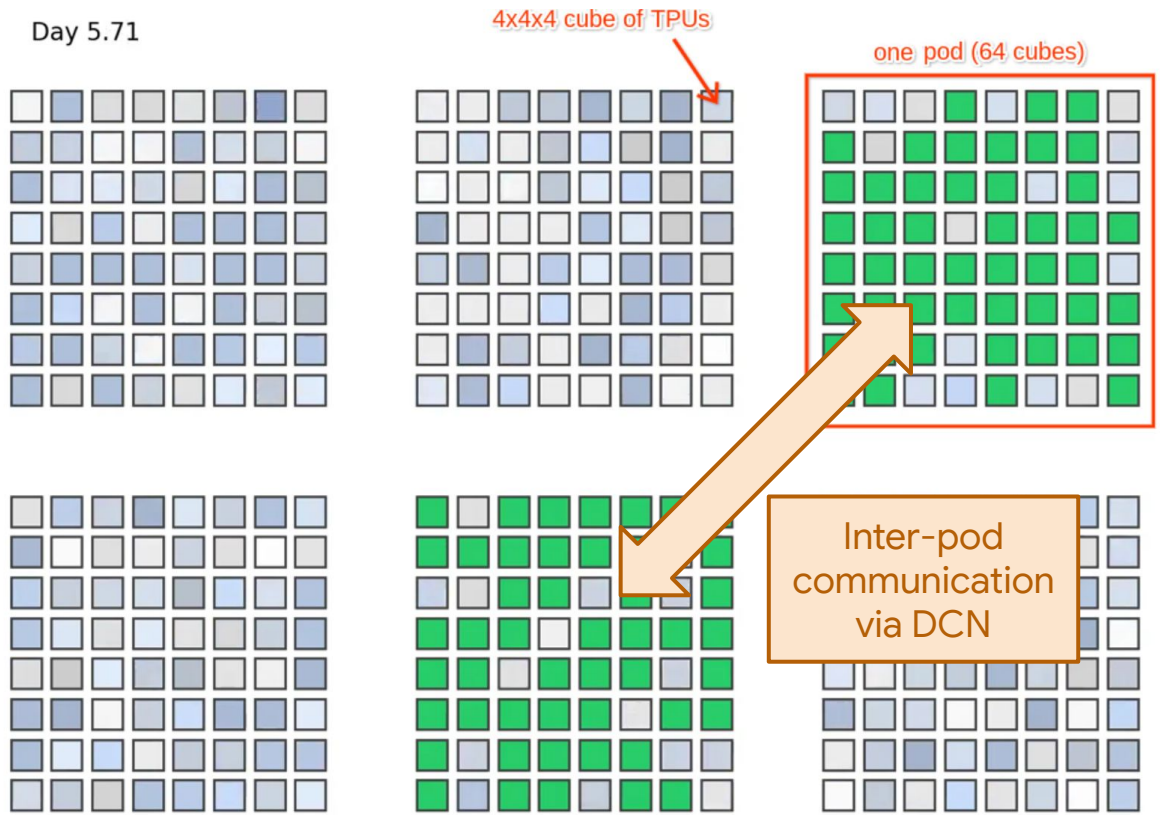
ICI and DCN working together: PaLM training

PaLM model training

- 500B parameters
- 6144 TPUv4s = 2 “pods”
- 56 days to train

Image shows the PaLM scheduling at a moment of time across 6 pods.





Video