TPU v4: An Optically Reconfigurable Supercomputer for Machine Learning with Hardware Support for Embeddings

Norman P. Jouppi, George Kurian, Sheng Li, Peter Ma, Rahul Nagarajan, Lifeng Nai, Nishant Patil, Suvinay

Subramanian, Andy Swing, Brian Towles, Cliff Young, Xiang Zhou, Zongwei Zhou, and David Patterson

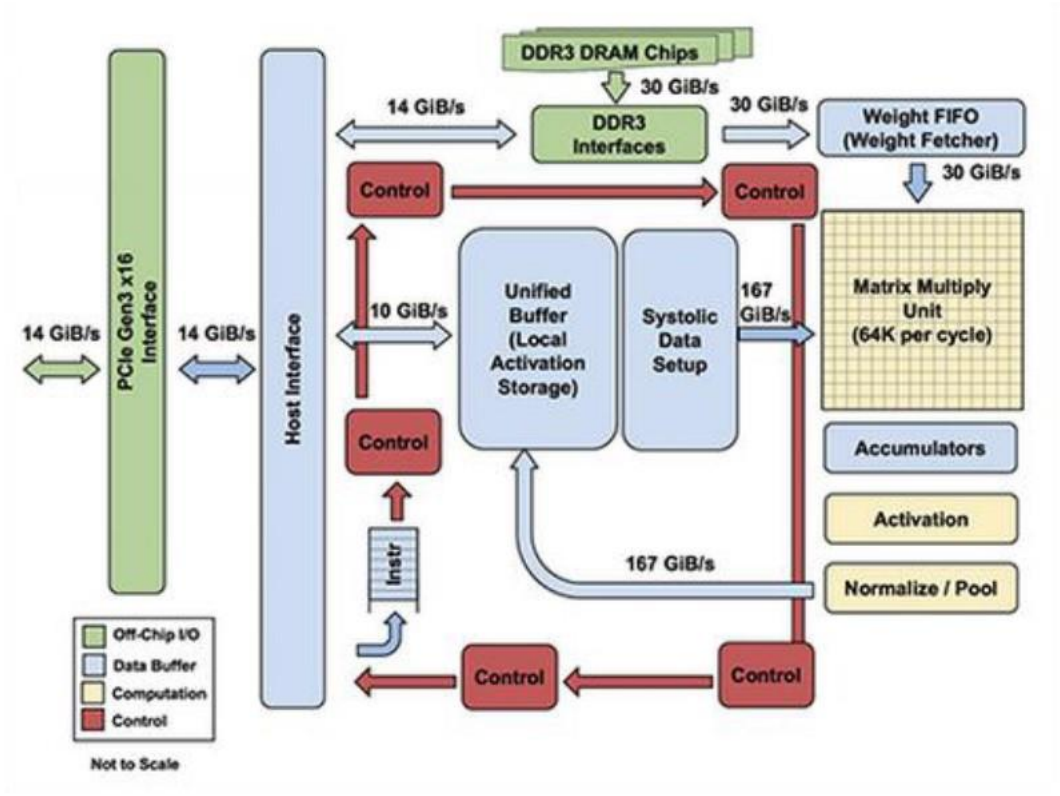
Google, Mountain View, CA

Presenter R12631055 林東甫

Background of TPUs

- TPUv1 was announced in May 2016 at Google I/O
- TPUs are designed for a high volume of low precision computation (e.g. as little as 8-bit precision) with more input/output operations per joule.
- Comparison to CPUs and GPUs : TPUs are well suited for CNNs, while GPUs have benefits for some fully-connected neural networks, and CPUs can have advantages for RNNs.

TPU



Google Cloud TPU Pod (Hot Chips 2017)

Matrix Multiplier Unit (MXU): 65,536 8-bit multiply-and-add units for matrix operations
Unified Buffer (UB): 24MB of SRAM that work as registers
Activation Unit (AU): Hardwired activation functions

Source: Google

Comparison of TPUv1 CPU and GPU

CPU
GPU
TPU

Model	Die										Benchmarked Servers				
	mm^2	nm	MHz	TDP	Measured Idle	Measured Busy	TOPS/s 8b	TOPS/s FP	GB/s	On-Chip Memory	Dies	DRAM Size	TDP	Measured Idle	Measured Busy
Haswell E5-2699 v3	662	22	2300	145W	41W	145W	2.6	1.3	51	51 MiB	2	256 GiB	504W	159W	455W
NVIDIA K80 (2 dies/card)	561	28	560	150W	25W	98W	--	2.8	160	8 MiB	8	256 GiB (host) + 12 GiB x 8	1838W	357W	991W
TPU	NA*	28	700	75W	28W	40W	92	--	34	28 MiB	4	256 GiB (host) + 8 GiB x 4	861W	290W	384W

Table 2. Benchmarked servers use Haswell CPUs, K80 GPUs, and TPUs. Haswell has 18 cores, and the K80 has 13 SMX processors. Figure 10 has measured power. The low-power TPU allows for better rack-level density than the high-power GPU. The 8 GiB DRAM per TPU is Weight Memory. GPU Boost mode is not used (Sec. 8). SECDEC and no Boost mode reduce K80 bandwidth from 240 to 160. No Boost mode and single die vs. dual die performance reduces K80 peak TOPS from 8.7 to 2.8. (*The TPU die is ≤ half the Haswell die size.)

Figure 3. TPU Printed Circuit Board. It can be inserted in the slot for an SATA disk in a server, but the card uses PCIe Gen3 x16.

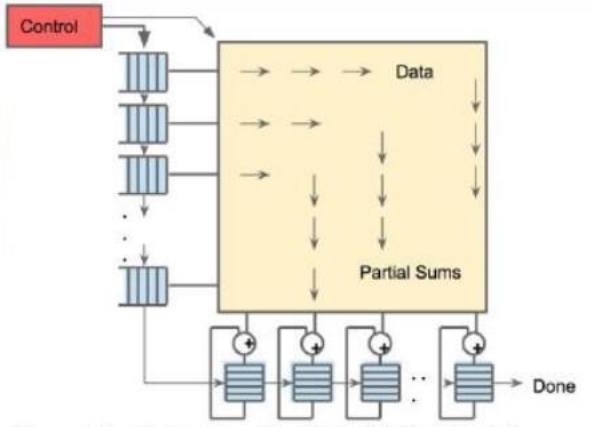
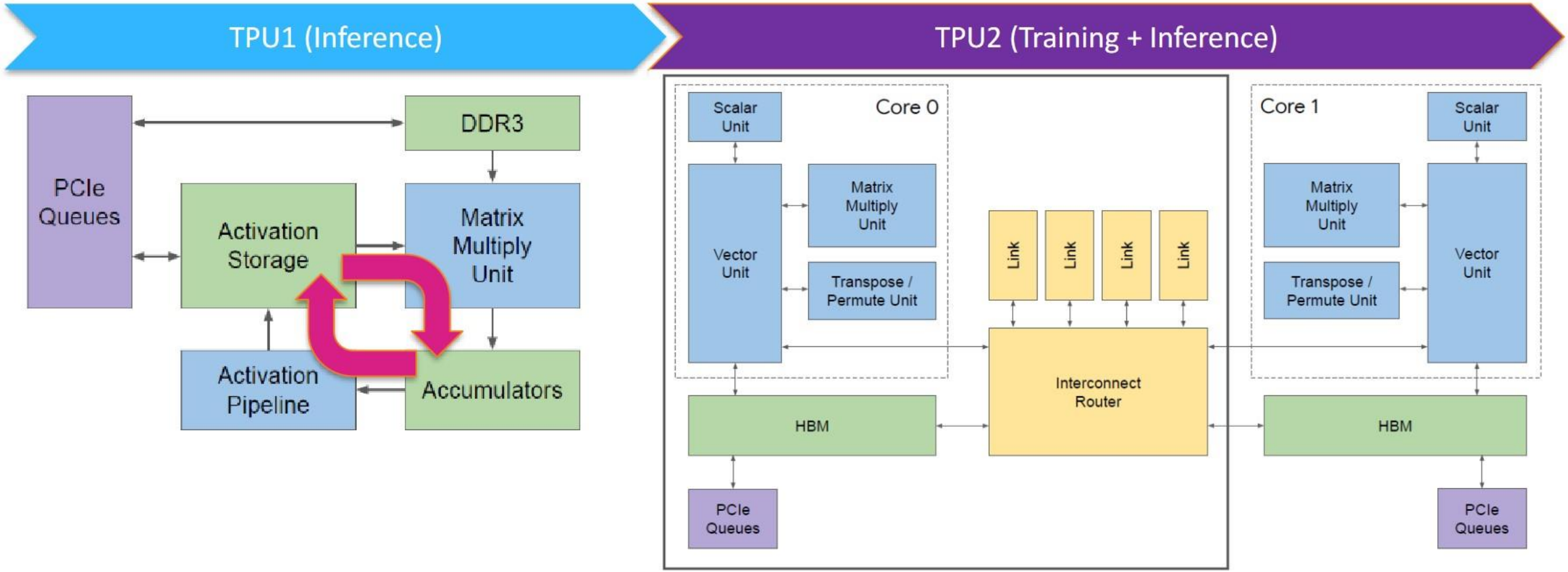


Figure 4. Systolic data flow of the Matrix Multiply Unit. Software has the illusion that each 256B input is read at once, and they instantly update one location of each of 256 accumulator RAMs.

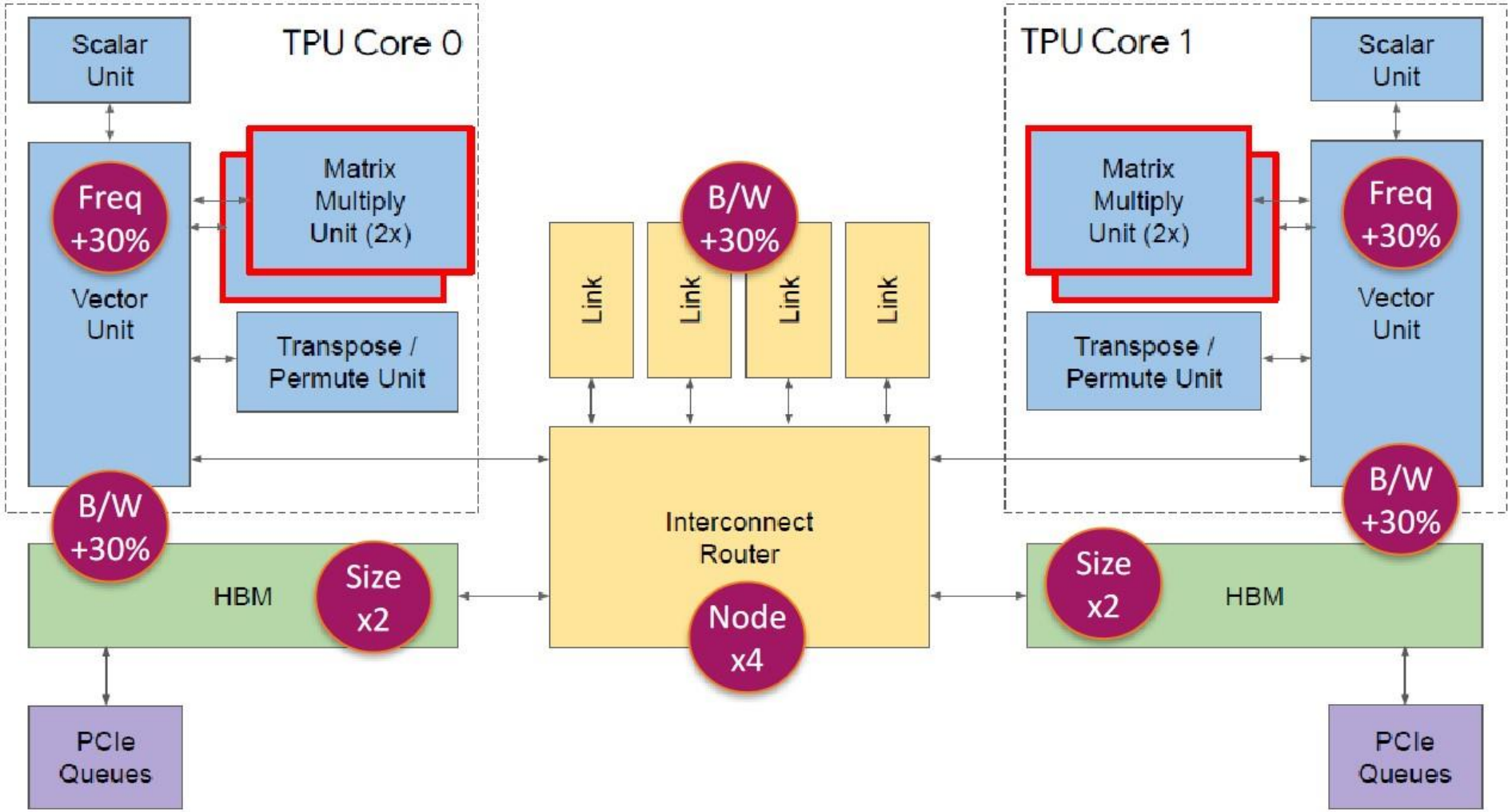
Source: Google

TPU2 (Connect-Oriented)



- VLIW Architecture → Leverage known compiler, 322bit VLIW
- Linear Algebra ISA → Scalar, vector, and matrix
- Built for generality

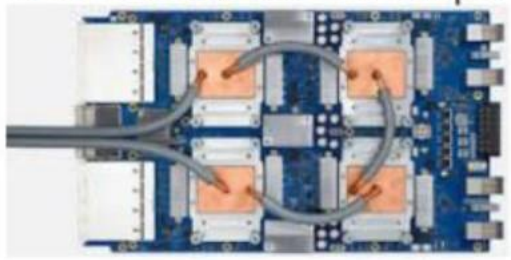
TPU3 (upgraded version of TPU2)



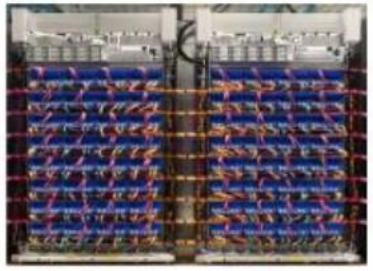
From TPUv2 to TPUv3

TPUv2 boards = 4 chips

TPUv3 boards = 4 chips

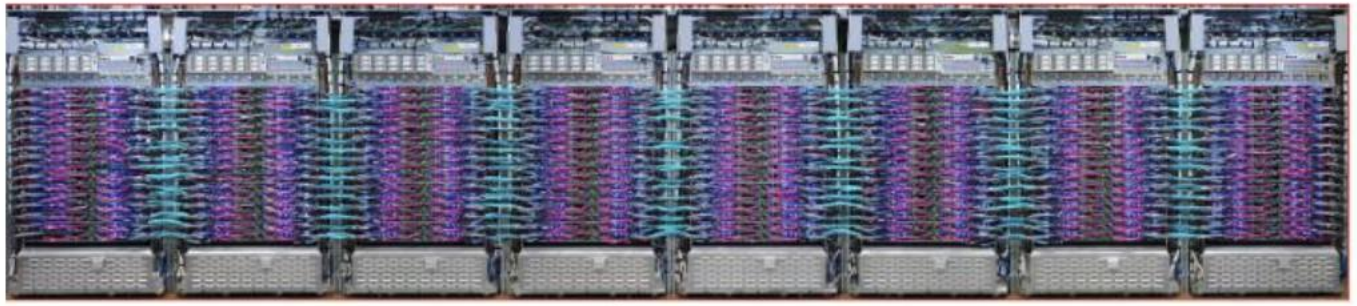


TPUv2 supercomputer (256 chips)



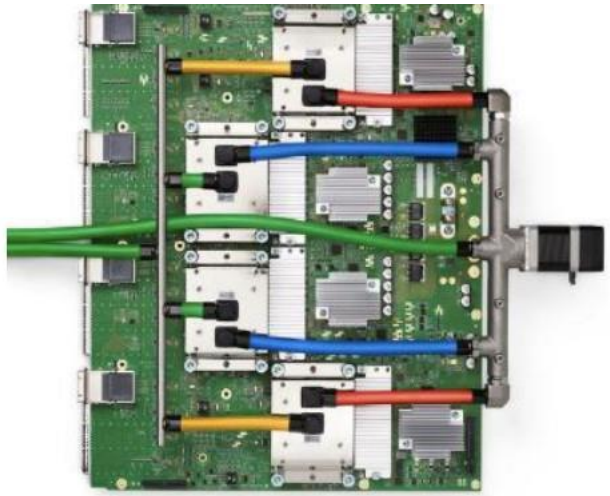
11.5 petaflops
4 TB HBM
2-D torus
256 chips

TPUv3 supercomputer (1024 chips)



> 100 petaflops
32 TB HBM
Liquid cooled
New chip + larger-scale system
1024 chips

TPU v4



- 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

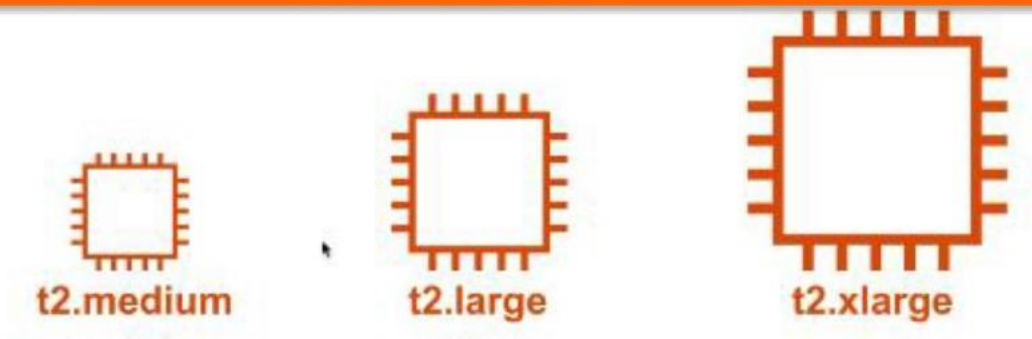
Scale-Up vs Scale-Out

When utilization increases and we are reaching capacity we can:

Scale up (Vertical Scaling)

Increasing the size of instances

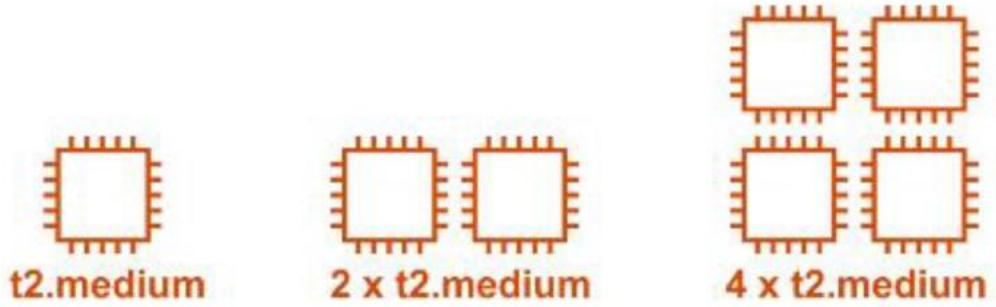
- Simpler to manage.
- Lower availability (if a single instance fails service becomes unavailable)



Scale out (Horizontal Scaling)

Adding more of the same

- More complexity to manage.
- Higher availability (if a single instance fail it doesn't matter)



Source: ExamPro, "Scale Up and Scale Out", AWS Solutions Architect Associate 2019, https://www.youtube.com/watch?v=ukhSRu_Uymw

Techniques improve of TPUv4

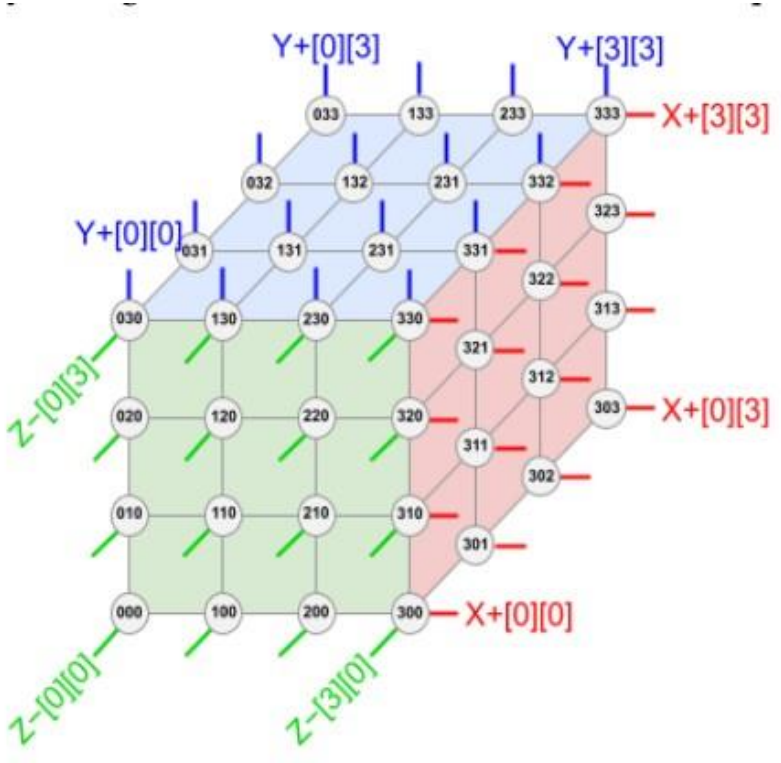
- Optical circuit switches (OCSes) dynamically reconfigure its interconnect topology to improve scale, availability, utilization, modularity, deployment, security, power, and performance.

- SparseCores: dataflow processors that accelerate models that rely on embeddings by 5x–7x yet use only 5% of die area and power.

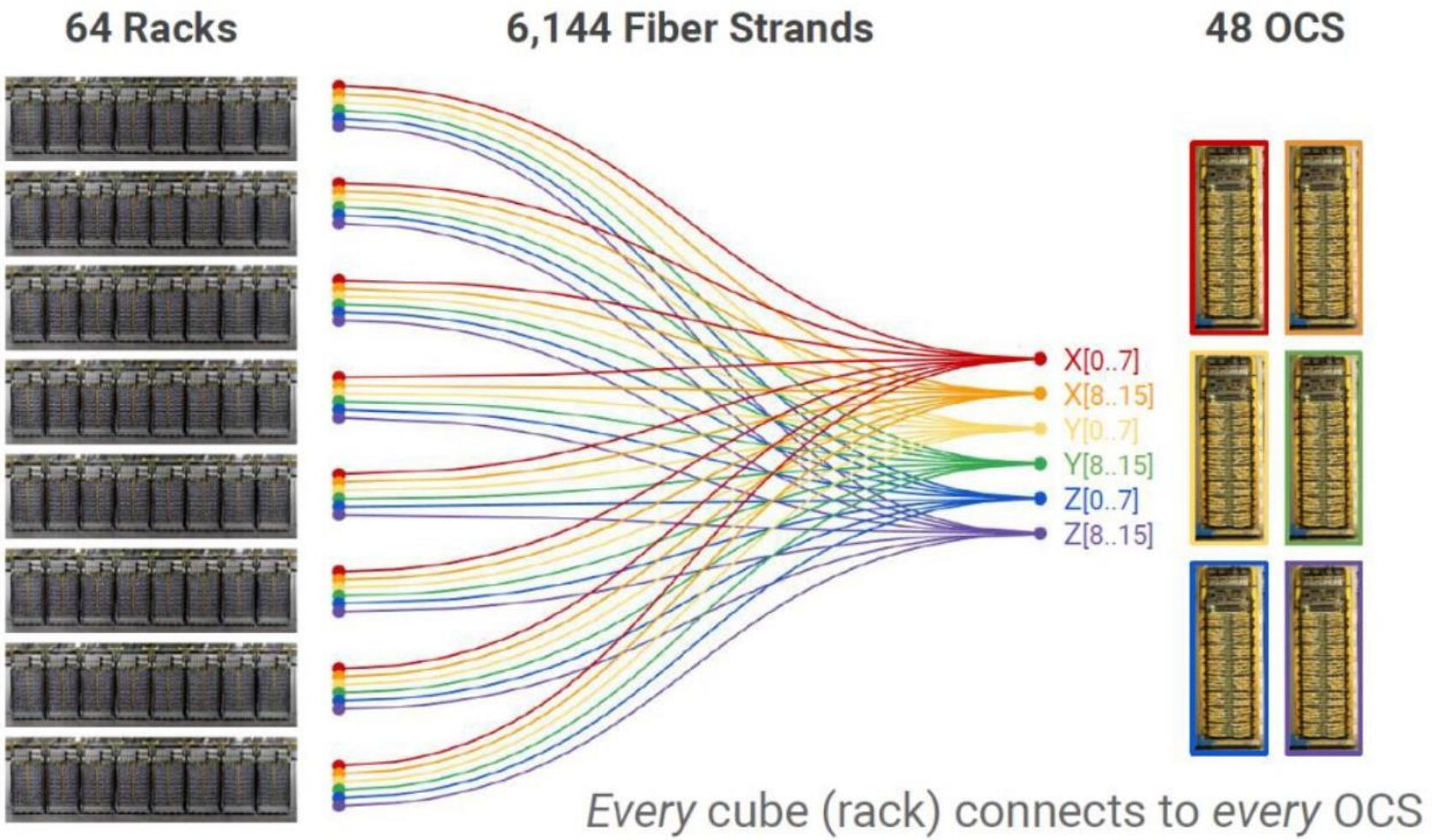
Construction of the TPU v4 Supercomputer

There are 16 links per face, totaling 96 optical links per block that connect to OCSes. To provide the wraparound links of a 3D torus, the links on the opposing sides must connect to the same OCS. Thus, each 43 block connects to $6 \times 16 \div 2 = 48$ OCSes.

The Palomar OCS is 136×136 (128 ports plus 8 spares for link testing and repairs), so 48 OCSes connect the 48 pairs of cables from 64 43 blocks (each 64 chips), yielding the desired total of 4096 TPU v4 chips.



OCS connection



Optical Circuit Switching

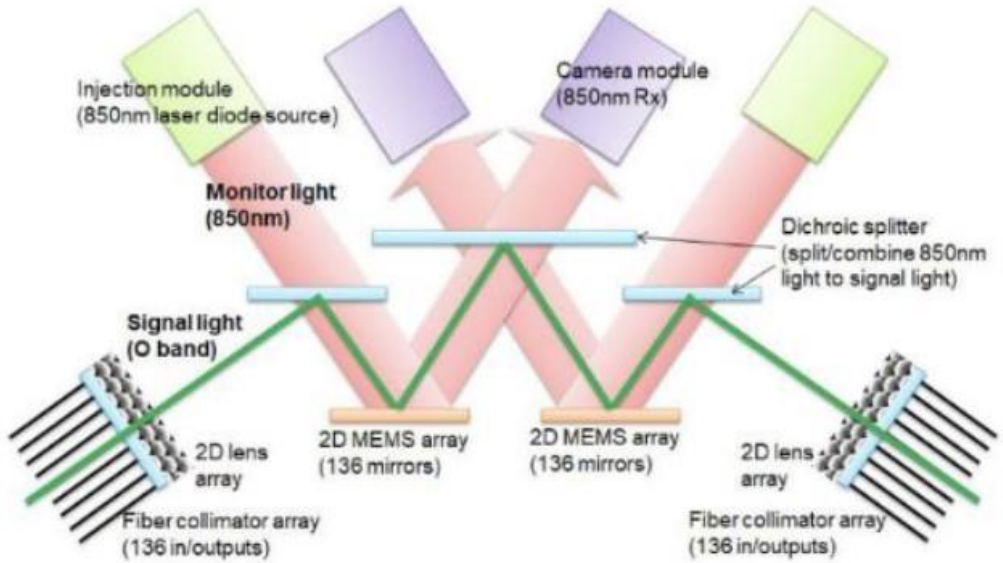
- Given the distance between TPU v3 racks, some wrap-around links of its 2D torus topology were so long that they had to be optical due to the reach limitation of electrical interconnects.

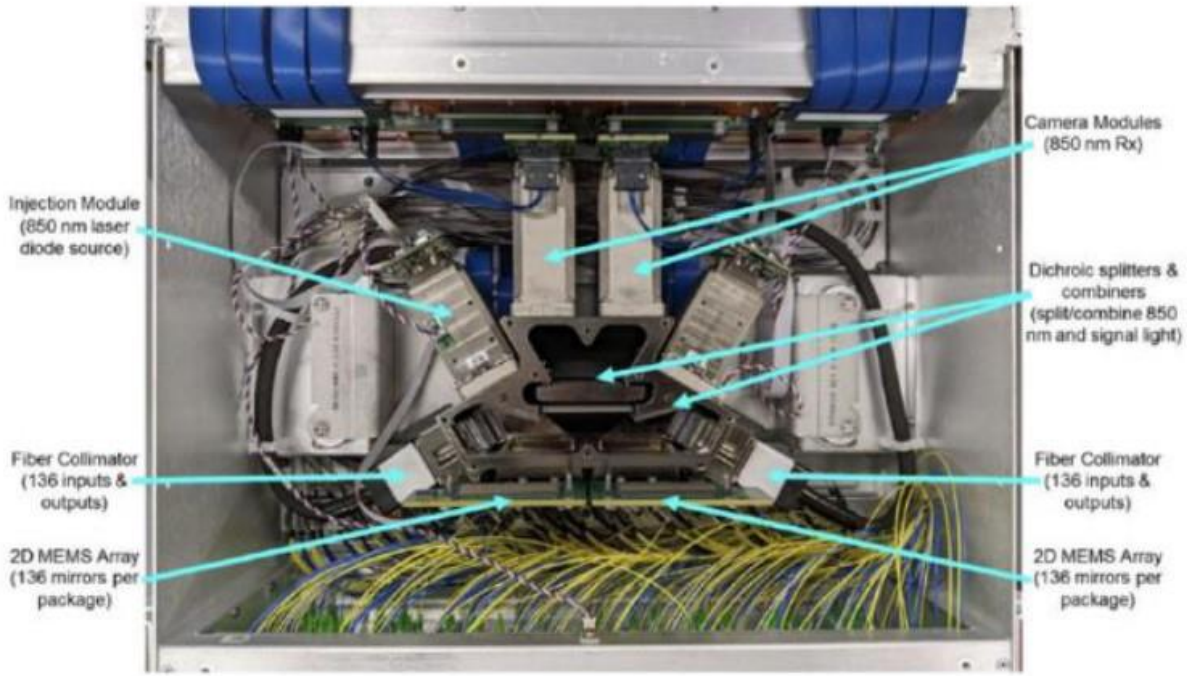
- To improve data center networking, Google advanced the state-of-the-art in reliability and cost of optical transceivers and OCSes

- The resulting Google Palomar OCS is based on 3D Micro-Electro-Mechanical Systems (MEMS) mirrors that switch in milliseconds. They employ circulators to send light both ways in a fiber, halving the number of required ports and cables.

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





SPARSECORE: EMBEDDINGS SUPPORT

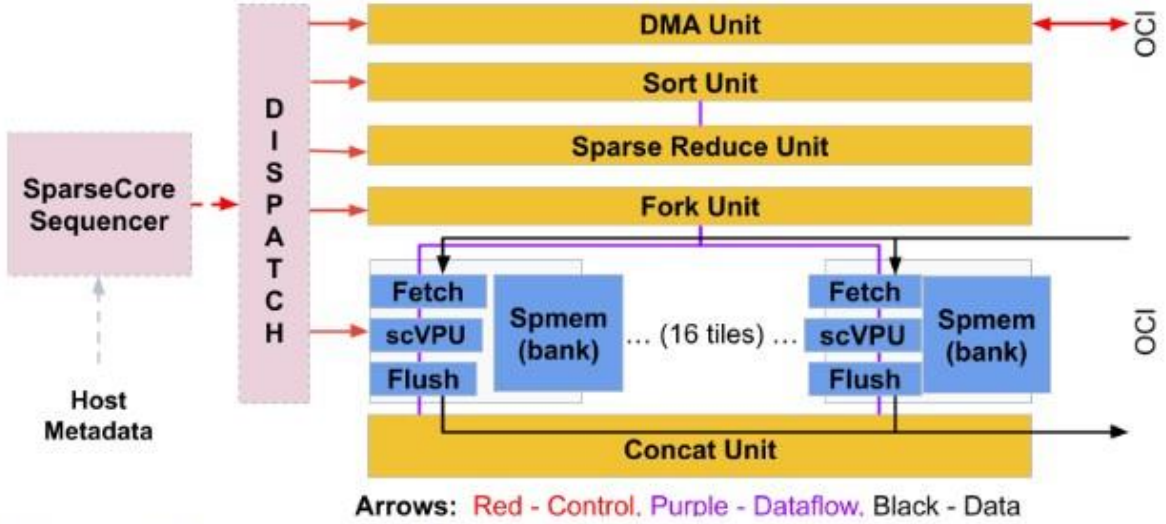


Figure 7: SparseCore (SC) Hardware Architecture.

- The SC is a domain-specific architecture for embedding training starting with TPU v2, with later improvements in TPU v3 and TPU v4.
- SCs are relatively inexpensive, at a total of only ~5% of the die area and ~5% of the power.

Workload Mix Changes

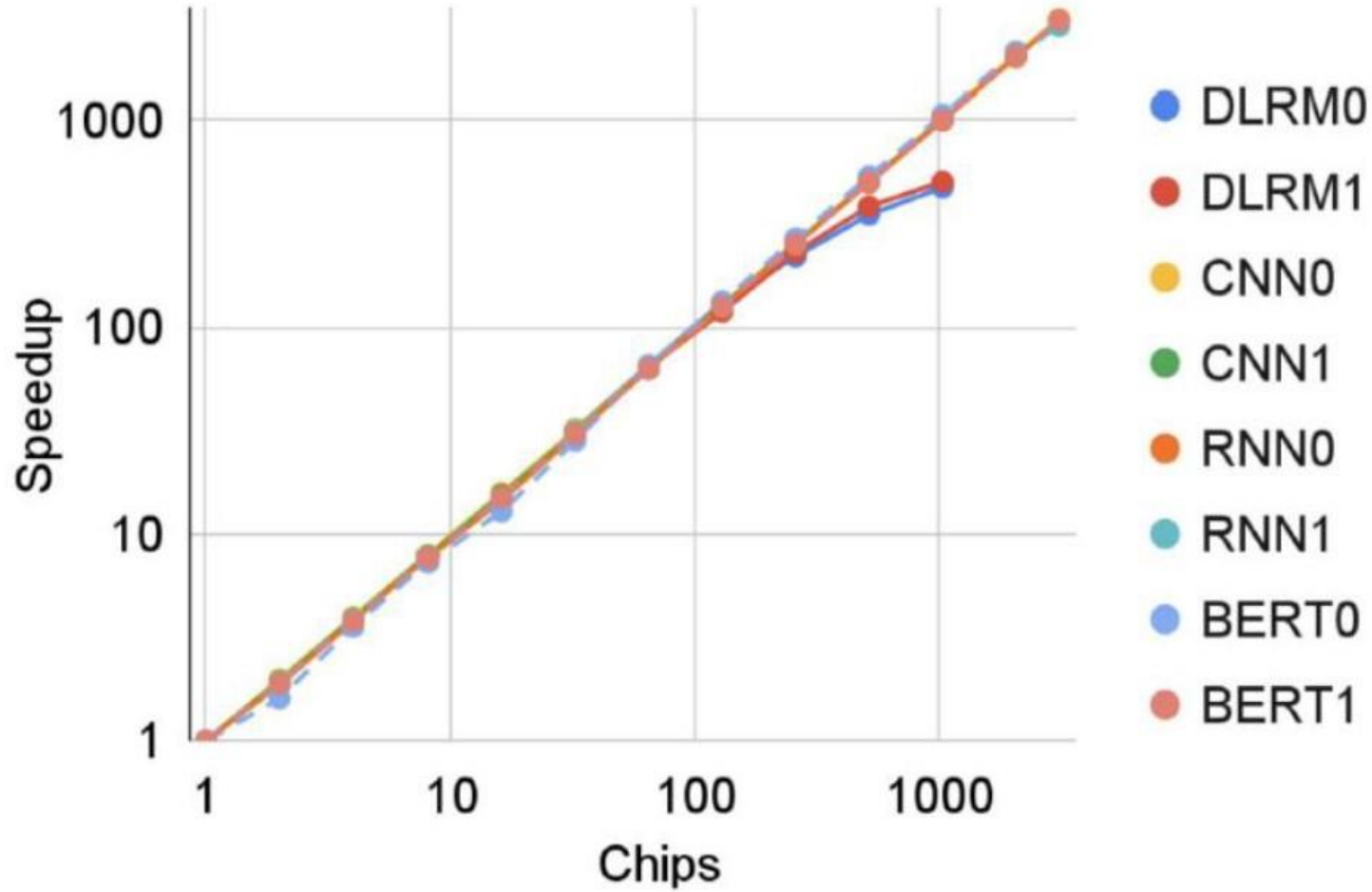
Significant changes in workload mix over the last 6 years

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

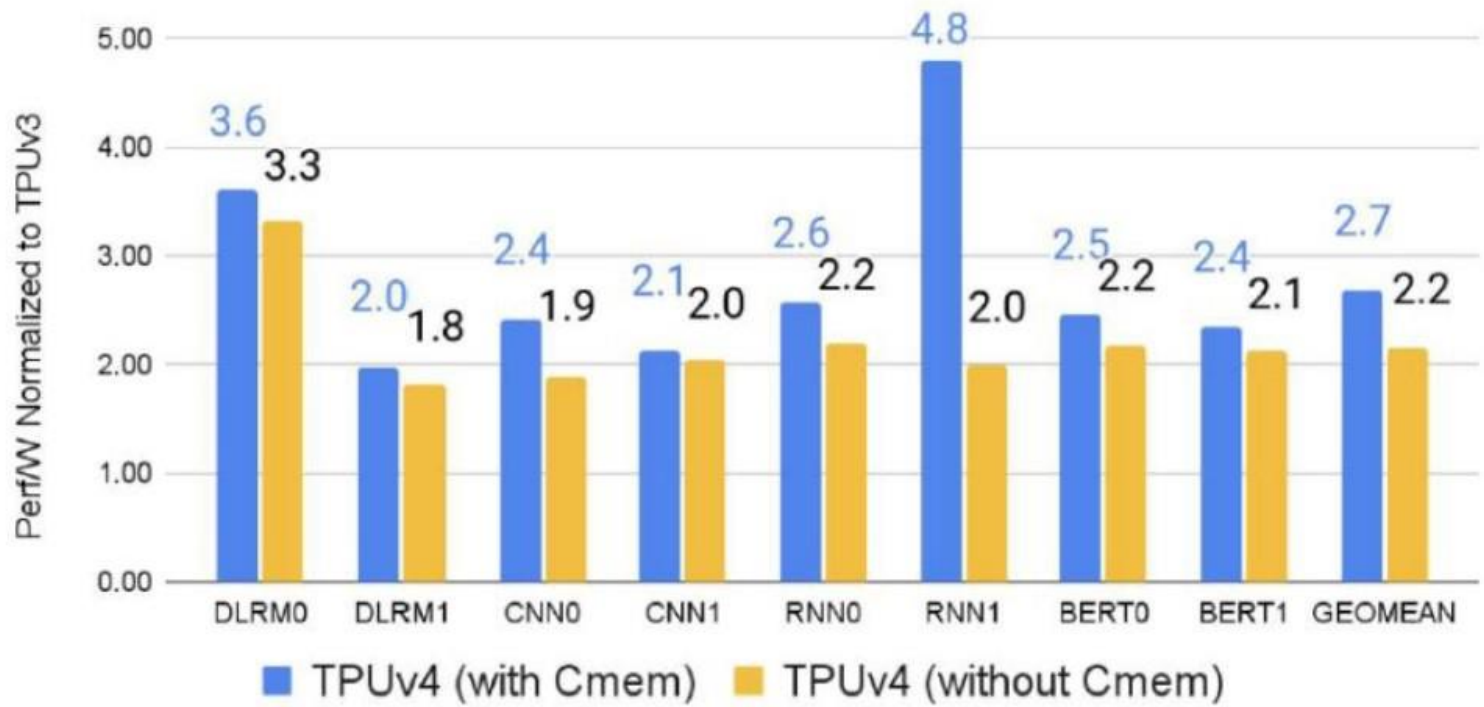


Perf/W Improvements from Increased On-chip Memory

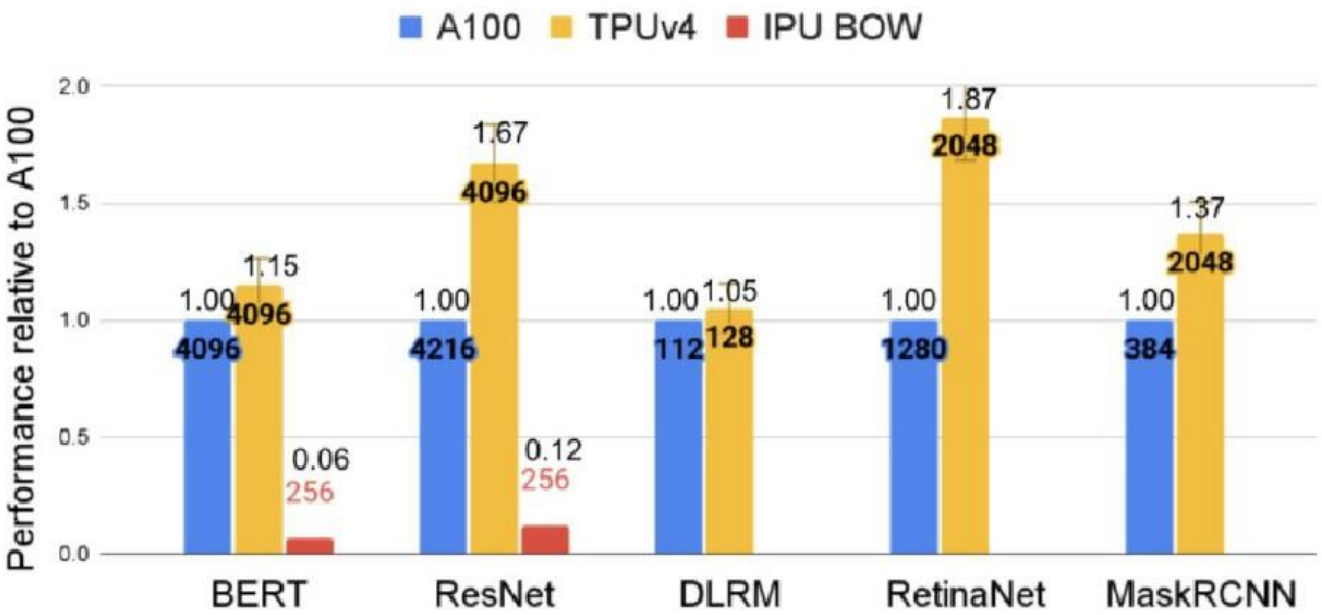
TPUv4 provides significant increases in perf/W over TPUv3: **2.7X geomean**

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

			Google				nVidia		Tesla	
	Feature		TPUv1	TPUv2	TPUv3	TPUv4i	NVIDIA T4 (Turing)	nVidia A100 SXM (Ampere)	Tesla D1	Tesla D1 Tile (25xD1)
	First deployed (GA date)		Q2 2015	Q3 2017	Q4 2018	**Q1 2020**	Q4 2018	**Q2 2020**	**Q2 2022**	**Q2 2022**
	Chip Technology		28 nm	16 nm	16 nm	**7 nm**	12 nm	**7nm**	**7nm**	**7nm**
AI	Peak TFLOPS	8bit	92 (8b int)			**138 (8b int)**	130 (8b int)			
AI		16bit (BF16)		46	123	**138**	65 (ieee fp16)	**312**	**362**	**9050**
AI		32bit (FP32)						**20**	**23**	**575**
AI	Inference or Training		Inf. only	Train. & Inf.	Train. & Inf.	Inf. only	Train. & Inf.	Train. & Inf.	Train. & Inf.	Train. & Inf.
Mem	Memory size (on-chip)		**28MB**	**32MB**	**32MB**	**144MB**	**18MB**	**40MB**	**442MB**	**11,050MB**
Mem	Memory size (off-chip)		**8GB**	**16GB**	**32GB**	**8GB**	**16GB**	**80GB (HBM)**	**No off-chip**	**No off-chip**
Mem	Memory GB/s / Chip		**34**	**700**	**900**	**614**	**320**	**2039 (HBM)**	-	-
Link	Network links (Gbits/s)		--	**4 x 496**	**4 x 656**	**2 x 400**	**?**	**600**	**16,000**	**36,000**
Link	Max chips / supercomputer		--	256	1024	--	?	?	3000	120 Tiles
	Chip Clock Rate (MHz)		700	700	940	1050	585 / (Turbo 1590)	?		
	Idle Power (Watts) Chip		28	53	84	55	36	?		
	TDP (Watts) Chip / System		75 / 220	280 / 460	450 / 660	175 / 275	70 / 175	400	400	15000
	Die Size (mm2)		< 330	< 625	< 700	< 400	545	826	645	15 x 645
	Transistors (B)		3	9	10	16	14	54		
	MXU Size / Core		1 256x256	1 128x128	2 128x128	4 128x128	8 8x8			
	Cores / Chip		1	2	2	1	40			