Walchand College of Engineering, Sangli
Department of CSE

Seminar on
“The Performance Analysis & an In-depth Look at Google’s Tensor Processing Unit”

Roll No.: 2018MTECSCO005

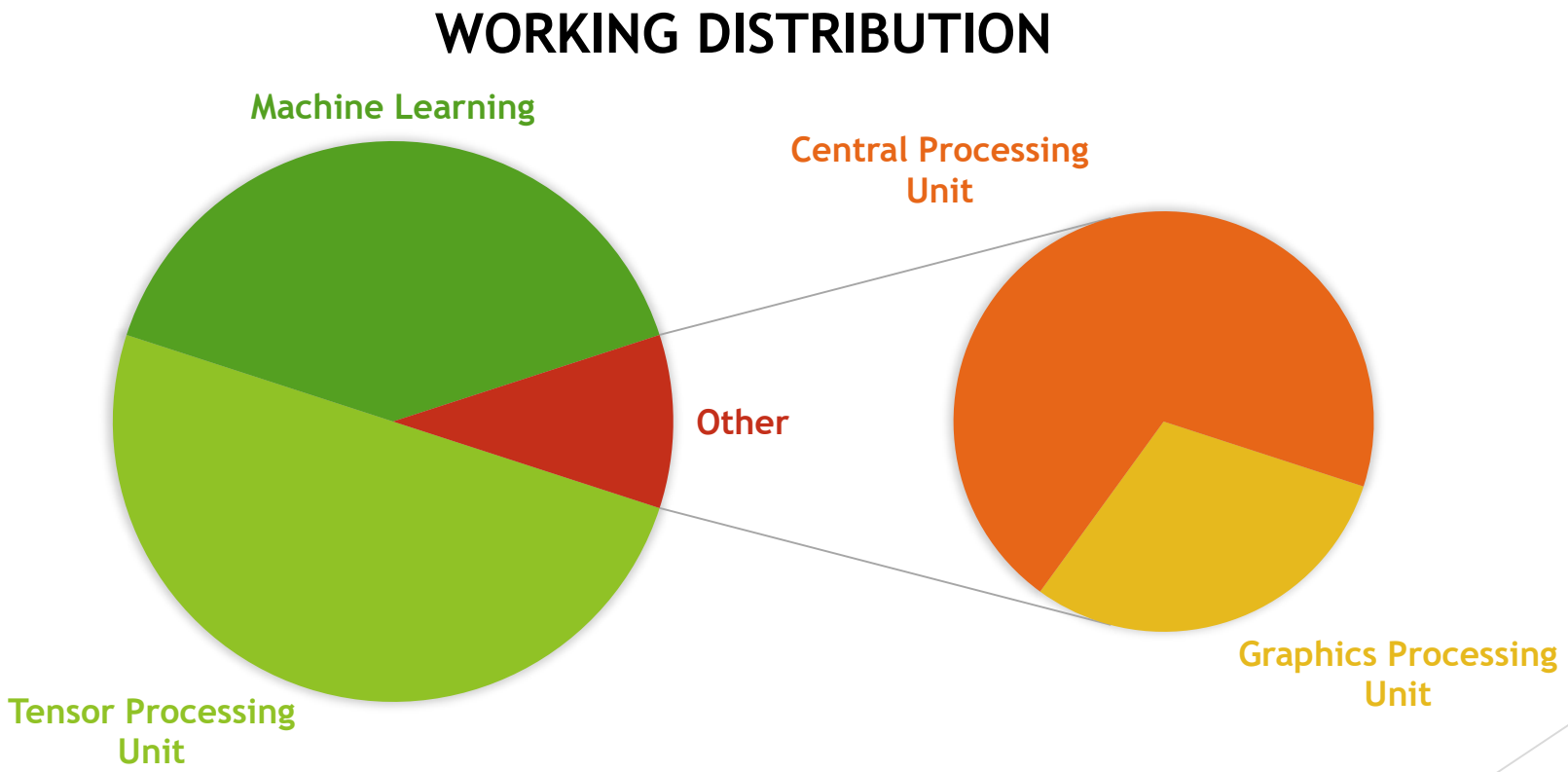
Name : Durgesh Sunil Naik

Contents

Research Area

- Research Area : Architecture & Working of Tensor Processing Unit
- Journal : IEEE Micro
- Paper Title : Motivation for and Evaluation of the First Tensor Processing Unit
- Authors : Norman P. Jouppi, Cliff Young, Nishant Patil, David Patterson
- Publication Year : 2018 (Published by Google)

Technology


WORKING DISTRIBUTION
Machine Learning
Central Processing Unit
Other
Tensor Processing Unit
Graphics Processing Unit

Why Tensor Processing Unit ?

- To study the detail architecture of Tensor Processing Unit(TPU).
- Recent upgrades in Hardware chips.
- Fast computations are necessary for the application deployed on DNN.
  - Deep Neural Network
- FPGA's were not power-efficient enough.
- ASIC design was selected for power and performance benefits.
- Device would execute CISC instructions on many networks.
- Device was made to be programmable, but operate on matrices instead of vector/scalar

Literature Survey

- N. P. Jouppi, C. Young, N. Patil, D. Patterson, Motivation for and Evaluation of the First Tensor Processing Unit, 2018
- N. P. Jouppi et al., “In-datacenter performance analysis of a tensor processing unit,” In Proceedings of the 44th Annual International Symposium on Computer Architecture (ISCA), 2917, pp. 1–12, 2017.
- S. Markidis, S. Chien, E. Laure, NVIDIA Tensor Core Programmability, Performance & Precision, 2018.
- White Paper by E. B. Olsen, Proposal for a High Precision Tensor Processing Unit, 2017.

Points to be discuss..

- Neural Network
- Architecture of TPU
- TPU Stack
- CPU & GPU
- Performance of TPU
- Matrix Multiply Unit(MXU)

Deep Neural Network(DNN)

- In deep-learning networks, each layer of nodes trains on a distinct set of features based on the previous layer's output.
- There are two phase of DNN:
  - Training (Learning)
  - Inference(Prediction)

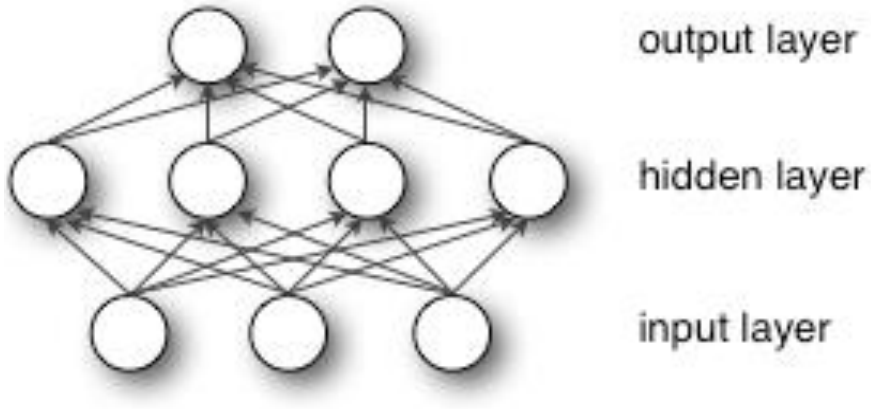


Figure 1 : DNN

Architecture of Tensor Processing Unit (TPU)

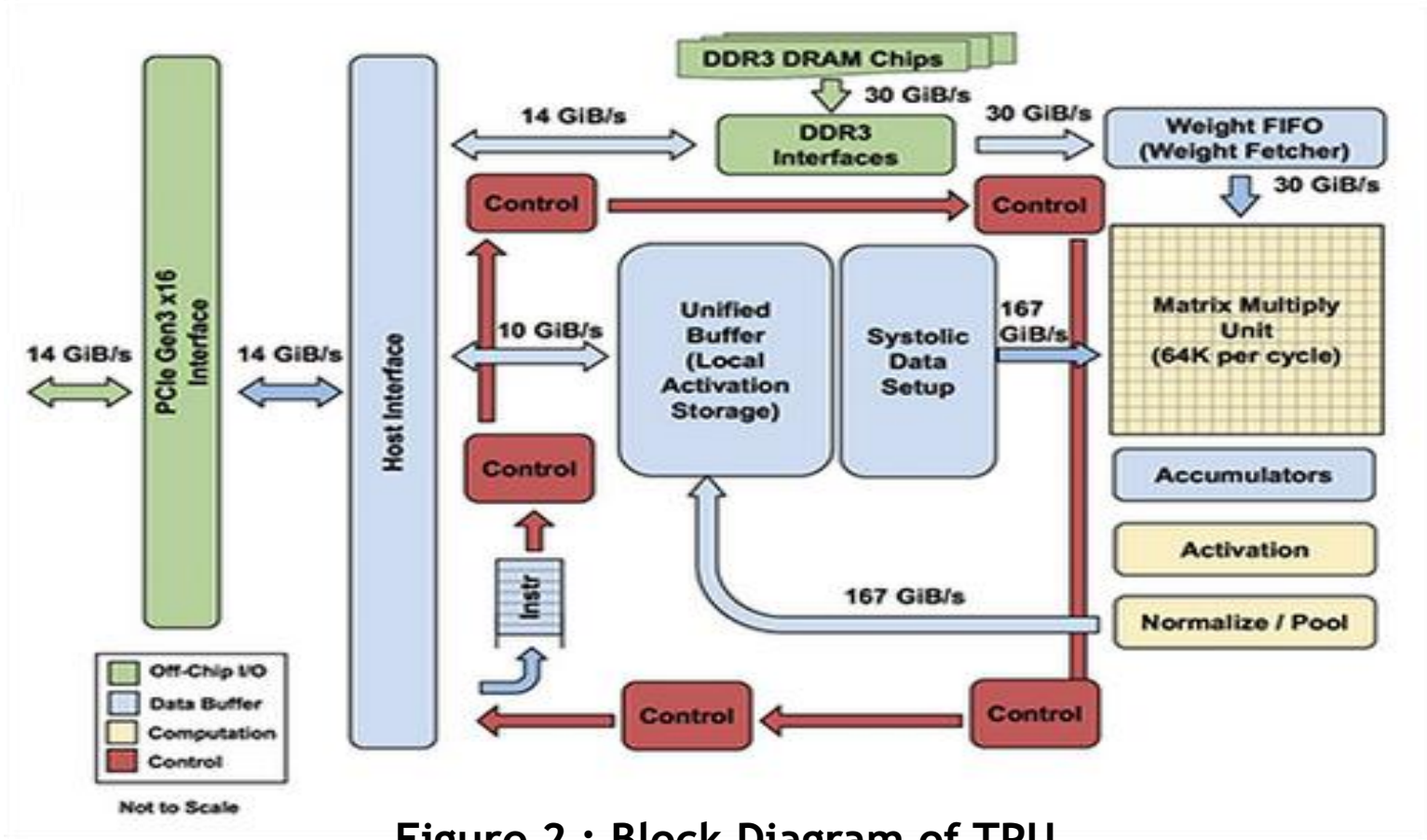


Figure 2 : Block Diagram of TPU

TPU Stack

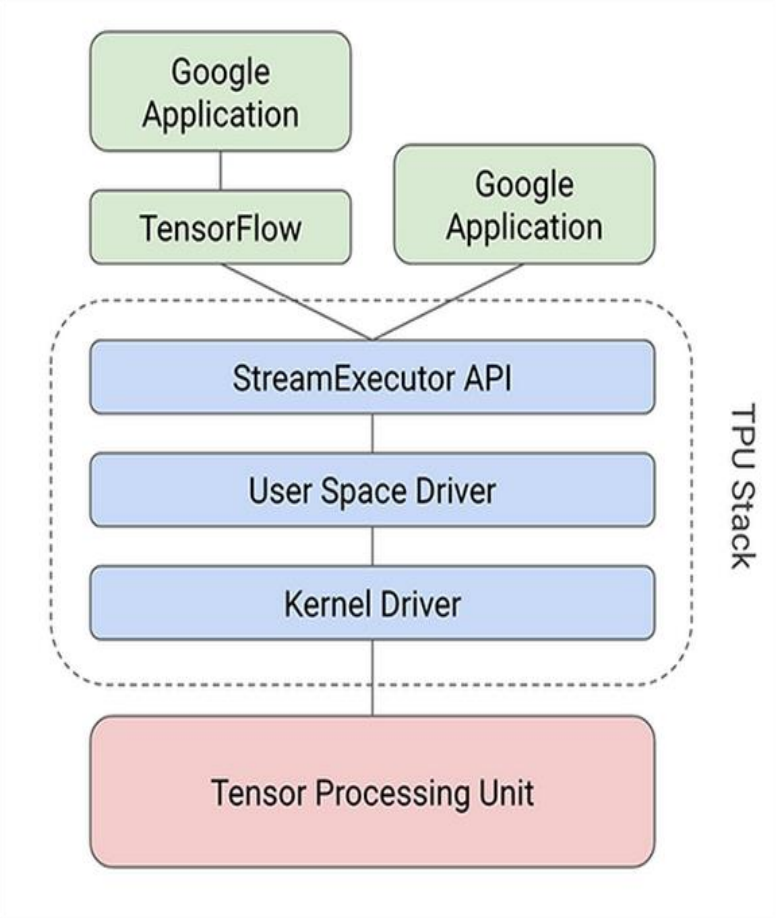


Figure 3: TPU Stack

- TPU stack translates the API calls into TPU instructions
- The Kernel Driver is lightweight and handles only memory management and interrupts.
- The User Space driver changes frequently. It sets up and controls TPU execution, reformats data into TPU order, translates API call into TPU instruction, and turns into them application binary.
- The User Space driver compiles a model the first time it is evaluated, caching the program image and writing the weight image into the TPU's weight memory; the second and following evaluations run at full speed.

CPU & GPU

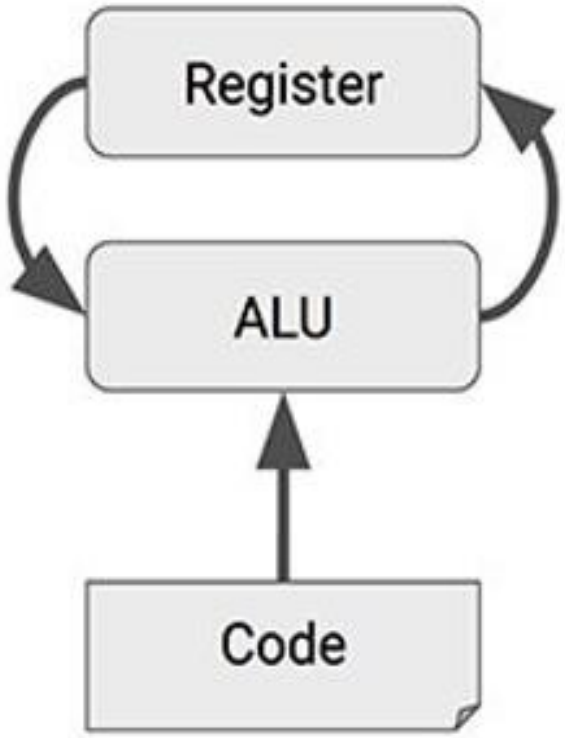


Figure 4: Work Flow of CPU & GPU

- CPUs and GPUs store values in registers.
- A program tracks the read/operate/write operations.
- A program tells ALUs :
  - Which Register to read from
  - What operation to perform
  - Which Register to write to

Performance of TPU

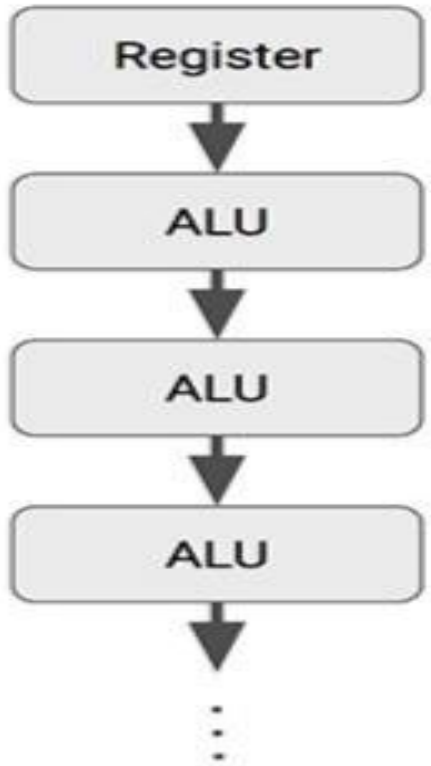


Figure 5: Work Flow of TPU

- TPU consists of Matrix Multiplier Unit (MXU).
- MXU performs hundreds of thousands of operations per clock cycle.
- Reads an input value only once.
- Inputs are used many times without storing back to register.
- Wires connect adjacent ALUs.
- Multiplication and addition are performed in specific order.
- Design is known as systolic array.

Matrix Multiply Unit(MXU)

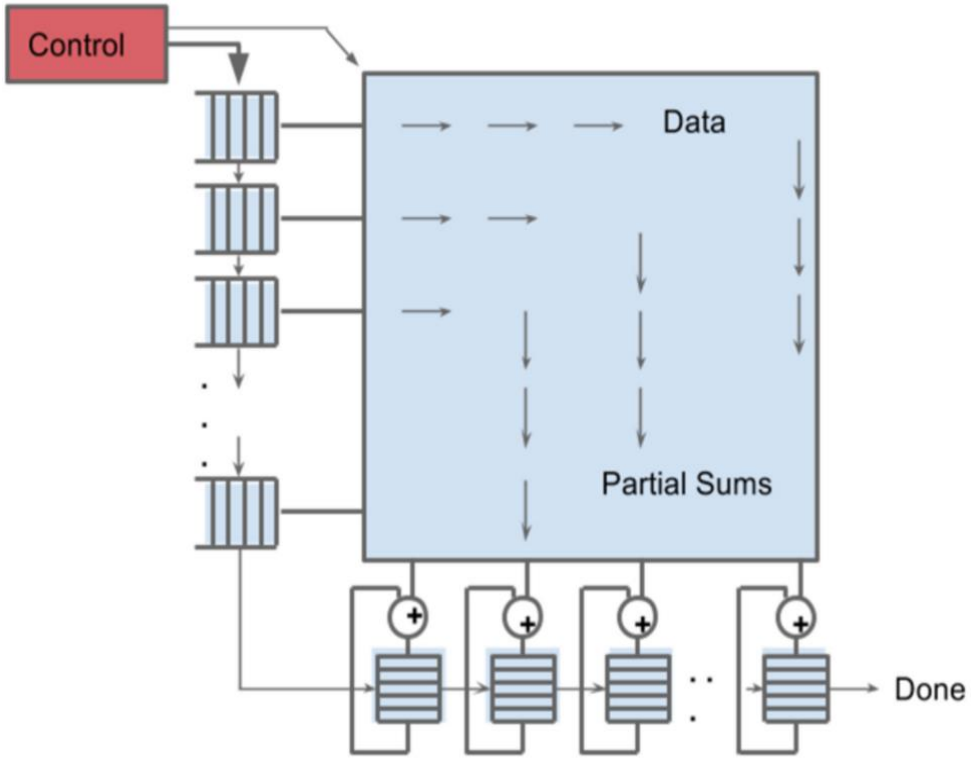


Figure 6: Work Flow of MXU

- TPU Matrix Multiply Unit has a systolic array mechanism that contains 256x256 = 65,536 ALU's.
- That means TPU can process 65,536 multiply-and-add for 8 bit integers every cycle.
- TPU runs at 700 MHz, it can compute 92 Teraops per second in Matrix Unit.

TPU Instructions & Functions

Table 1.TPU Instruction & Function.

TPU Instruction	Function
Read_Host_Memory :	it reads data from the CPU host memory into the Unified Buffer (UB).
Read_Weights :	it reads weights from Weight Memory into the Weight FIFO as input to the Matrix Unit.
MatrixMultiply/Convolve :	it multiply & convolve with data and weights accumulate the result.
Activate :	it performs the nonlinear function of the artificial neuron, with options for Sigmoid, and so on.
Write_Host_Memory :	it writes data from the Unified Buffer into the CPU host memory.

Implementation : Environment

Table 2.Environmental Setup.

CPU	GPU	TPU
Intel Xeon CPU	Nvidia Tesla K80	Google Edge TPU
CPU Core → 1 (Hyperthreaded)	CUDA Cores → 2496	TPU Cores → 2
CPU Clock → 2.20GHz	GPU base clock → 875MHz	TPU Clock → 700MHz
L3 Cache → 45 MB	Memory → 12,288 MB GDDR5	Weight Memory → 8,192 MB DRAM

Implementation

- **Fashion MNIST with Keras and CPU**
- **Fashion MNIST with Keras and GPU**
- **Fashion MNIST with Keras and TPU**

Implementation

- The programs which are used for demonstration purpose, contains training set of 60,000 labeled images which having 0-9 labels as shown in above Table No.3. The testing set contains 10,000 of unlabeled images which are to be classified.

Table 3. Labels of Test set images.

Label	Description
0	T-shirt/top
1	Trouser
2	Pullover
3	Dress
4	Coat
5	Sandal
6	Shirt
7	Sneaker
8	Bag
9	Ankle boot

Implementation : Running Time vs. Platform

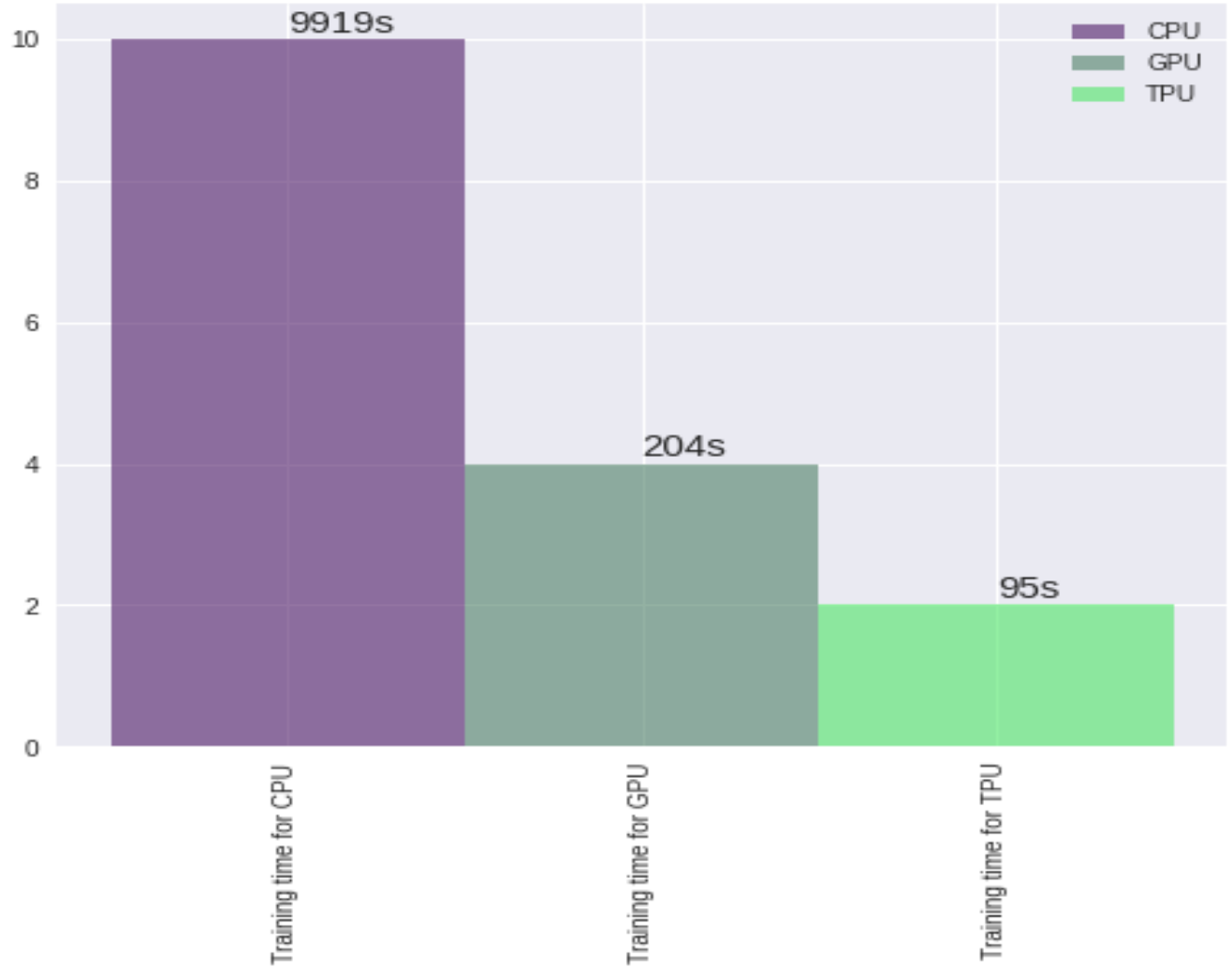


Figure 7 : Performance Plot of CPU, GPU, TPU

Result & Analysis

Table 3. 99th-percentile response time versus throughput (inferences per second) for MLP0.

Type	Batch Size	99th % Response	Inferences/s (IPS)	Speedup Ratio
CPU	16	7.2ms	5,482	42%
	64	21.3ms	13,194	100%
GPU	16	6.7ms	13,461	37%
	64	8.3ms	36,465	100%
TPU	200	7.0ms	225,000	80%
	250	10.0ms	280,000	100%

Uses

- RankBrain algorithm used by Google search.
- Google Photos
- Google Translate
- Google Cloud Platform

Future Development

- Allows to build machine learning supercomputers called “TPU Pods”.
- Improvement in training times.
- Allows mixing and matching with other hardware which includes Skylake CPUs and NVIDIA GPUs.

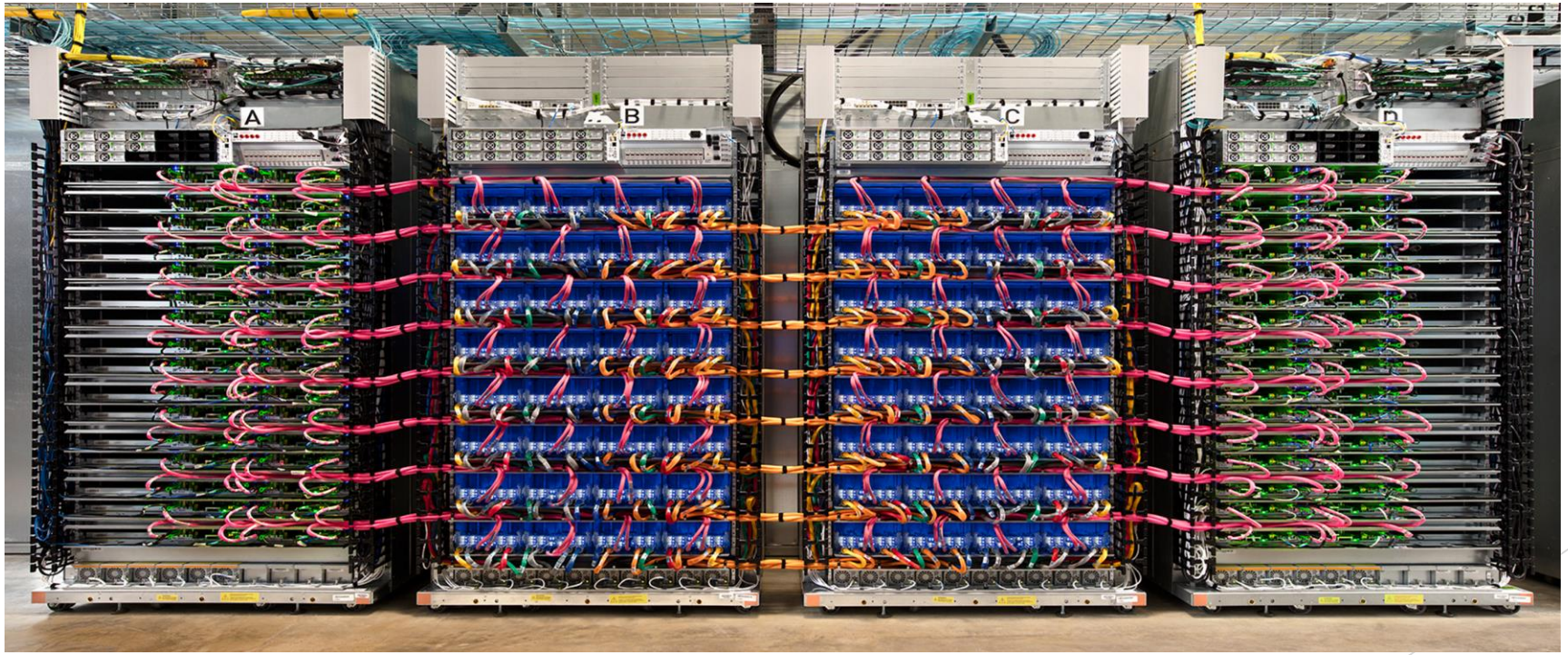


Figure 8: TPU Pods

Conclusion

- The TPU's DNN applications use 8-bit integers rather than 32-bit floating point to improve efficiency of computation, memory bandwidth, and memory capacity.
- The 2D organization enables systolic arrays, which reduce register accesses and energy
- Because the TPU is a DSA, it can drop features required by CPUs and GPUs that DNNs don't use. Such omissions make the TPU cheaper, save energy, and allow transistors to be repurposed for domain-specific optimizations.
- The TPU has one thread, while the K80 has 13 and the CPU has 18. A single thread makes it easier to stay within a fixed latency limit of our inference applications, as well as save energy.

References

[1] N. P. Jouppi, C. Young, N. Patil, D. Patterson, Motivation for and Evaluation of the First Tensor Processing Unit, 2018

[2] N. P. Jouppi et al., “In-datacenter performance analysis of a tensor processing unit,” In Proceedings of the 44th Annual International Symposium on Computer Architecture (ISCA), 2917, pp. 1–12, 2017.

[3] S. Markidis, S. Chien, E. Laure, NVIDIA Tensor Core Programmability, Performance & Precision, 2018.

[4] White Paper by E. B. Olsen, Proposal for a High Precision Tensor Processing Unit, 2017.

[5] https://cloud.google.com/tpu/docs/tpus

[6] https://keras.io/datasets/

Thank You…!