Performance Evaluation and Analysis of Tensor Processing Unit against CPU and GPU

Dhileepan Thangamanimaran[1] *, M. Sharat Chandar[2], S. Chandia[3]
[1] **III Year M.Sc. (Software Systems)**, Department of Computing**, Coimbatore Institute of Technology**, Coimbatore, India.
[2] **III Year M.Sc. (Software Systems)**, Department of Computing, **Coimbatore Institute of Technology**, Coimbatore, India.
[3]**Assistant Professor**, Department of Computing, **Coimbatore Institute of Technology**, Coimbatore, India.
*Corresponding author E-mail: dhileepan123@gmail.com.

***Abstract*—**Tensor Flow is an Open Source Software Library used for Machine Learning Applications. Machine Learning demands profound computation power due to its heavy workloads. Architecture of CPU and GPU restricts it from being an efficient platform for processing of Tensor Flow applications. Tensor Processing Unit abbreviation TPU are tailored specifically for Tensor Flow and Machine Learning Applications. TPUs are said to be 30 times faster than GPU and much more energy efficient. TPUs power Google Datacentres since 2015. This Paper provides a brief comparative study of CPU, GPU and TPU on deep learning models using Tensor Flow Library. We aim to access the performance of the above processing units and determine its scope in real life scenarios and the economic vs efficiency comparison.

Keywords—TPU, TensorFlow, Machine Learning.

I.INTRODUCTION

Unmatched Natural Intelligence is what separate us from all other inferior creatures. Humans developed machines to aid them in their chores. At the beginning machines were programmed step by step to perform their required functions. Now the machines are required to think and improvise on their own. Evolved from the study of pattern recognition and computational learning theory in artificial intelligence, machine learning explores the study and construction of algorithms that can learn from and make predictions on data. [1] Machines are required to analysis and processes complex datasets before recognizing the pattern and making predictions. The processing of complex data demand higher computing power than CPUs and GPUs to be efficient and optimal. That is where TPUs succeeds. The Objective of the paper is to perform a comparative study on the processors and its real-life implications.

II. TENSORFLOW

TensorFlow is an open source software library for high performance numerical computation. It is a symbolic math library and is also used for machine learning applications such as neural networks. TensorFlow is a cross platform library that runs on nearly everything. The TensorFlow distributed execution engine abstracts away the many supported devices and provides a high performance-core implemented in C++ for the TensorFlow platform. TensorFlow allows developers to create and describe how data moves through a graph, or a series of processing nodes. Each node in the graph represents a tensor. It uses python as a front-end API for the development of the models. However, the actual math operations are performed by High level C++ binaries. [2][3]

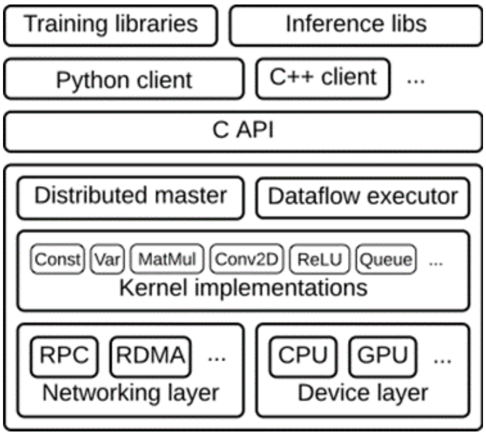


Figure 1: TensorFlow Architecture.

III. CPU

The three typical components of a CPU - Arithmetic Logic Unit (ALU), which performs arithmetic and logical operations. Control Unit (CU), which extracts instructions from memory and decodes and executes them, calling on the ALU when necessary. And Memory management unit, which is available in most high-end microprocessors to translating logical addresses into physical RAM addresses. The CPU comes with single and multi-core variant. Multi-core CPU can perform more than one process simultaneously. Running TensorFlow models on CPUs can consumed lot of resource and time. The Architecture of CPU does not allow for execution of huge number of process simultaneously.

IV. GPU

GPU, the Graphics Processing Unit is a specialized electronic circuit designed to render 2D and 3D graphics together with a CPU. GPUs are great for Machine learning applications like TensorFlow because the type of calculations they were designed to process are the similar to those performed in deep learning. GPUs compared to central processing units (CPUs), are more specialized at performing matrix operations and several other types of advanced mathematical transformations. This makes deep learning algorithms run several times faster on a GPU compared to a CPU. Learning times can often be reduced from days to mere hours. The TensorFlow library allows algorithms to be described as a graph of connected operations that can be executed on various GPU-enabled platforms ranging from portable devices to desktops to high-end servers. GPUs can run TensorFlow up to 50% faster than CPUs. [4]

VI. TPU

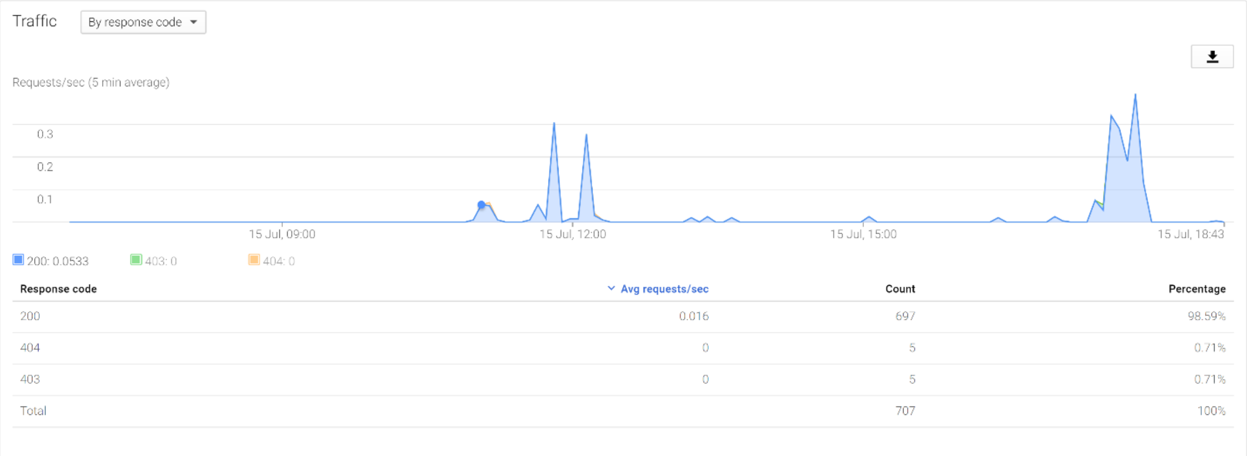
The Tensor Processing Unit (TPU) is a custom-made accelerator ASIC that has been specifically designed for Google's TensorFlow framework. TPU was announced in 2016 at Google I/O. It is customized to give high performance and power efficiency when running TensorFlow. TPU is unique in that it uses fewer computational bits. This allows more operations per second. Google's TPUs are proprietary and are not commercially available. A TPU is a coprocessor, it cannot execute code in its own right, all code execution takes place on the CPU which just feeds a stream of microoperations to the TPU. Compared to a graphics processing unit, it is designed for a high volume of low precision computation with higher IOPS per watt and lacks hardware for texture mapping. TPUs give much better performance/price and support all common neural network operations used in real world production networks. [5][6][7]

VII. EVALUATING TPU USING RETINANET MODEL

RetinaNet is a Machine Learning Model used extensively for object detection. The Backbone network computes a convolution map over an image. RetinaNet consists of two task specific subnetworks. The first subnet classifies the backbone output and the second subnet performs convolution bounding box regression. [9]. The RetinaNet model is trained on COCO dataset. COCO is a large-scale object detection, segmentation, and captioning dataset that has several in-build object detection features.

Training RetinaNet on Cloud TPU:

We Created a Google Cloud Project and accessed the Cloud TPU using Cloud TPU Provisioning Utility. A Cloud Storage bucket was created. We created an instance to access the compute engine. The RetinaNet model was cloned form GitHub. The Dataset needs pre-processing before it can be used for training. The RetinaNet model has been configured to train on the COCO dataset. The tpu/tools/datasets/download_and_preprocess_coco.sh script converts the COCO dataset into a set of TFRecords that the training application expects. This requires at least 100GB of disk space for the target directory and takes approximately 1 hour to complete. Then We ran the Model for 100000 steps using 100GB of training data on 8 Core TPU with 7GB of RAM. The Process took almost 6 hours. The following are the resultant graphs of the process.



Response code	Avg requests/sec	Count	Percentage
200	0.016	697	98.59%
404	0	5	0.71%
403	0	5	0.71%
Total		707	100%

Figure 2: Google Compute Engine API Traffic Graph

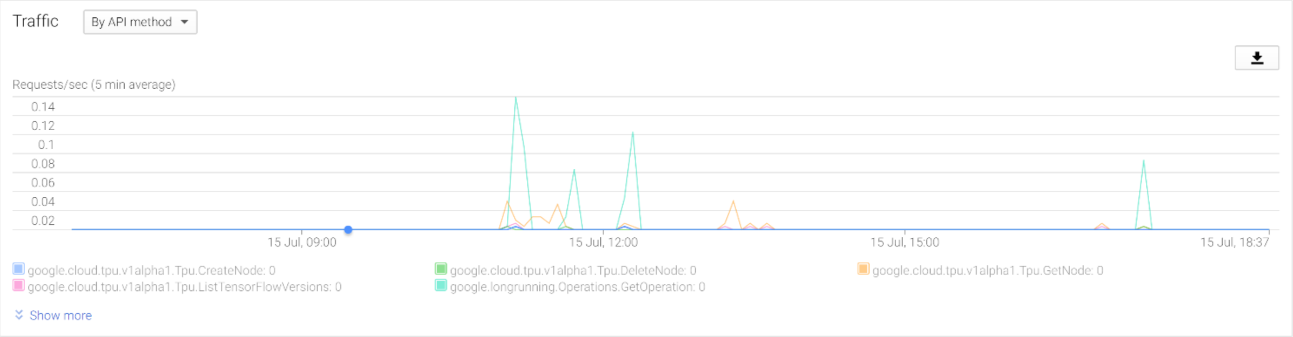

Traffic
By API method
Requests/sec (5 min average)
0.14
0.12
0.1
0.08
0.06
0.04
0.02
15 Jul, 09:00
15 Jul, 12:00
15 Jul, 15:00
15 Jul, 18:37
google.cloud.tpu.v1alpha1.Tpu.CreateNode: 0
google.cloud.tpu.v1alpha1.Tpu.ListTensorFlowVersions: 0
google.cloud.tpu.v1alpha1.Tpu.DeleteNode: 0
google.longrunning.Operations.GetOperation: 0
google.cloud.tpu.v1alpha1.Tpu.GetNode: 0
Show more

Figure 3: Google TPU API Graph

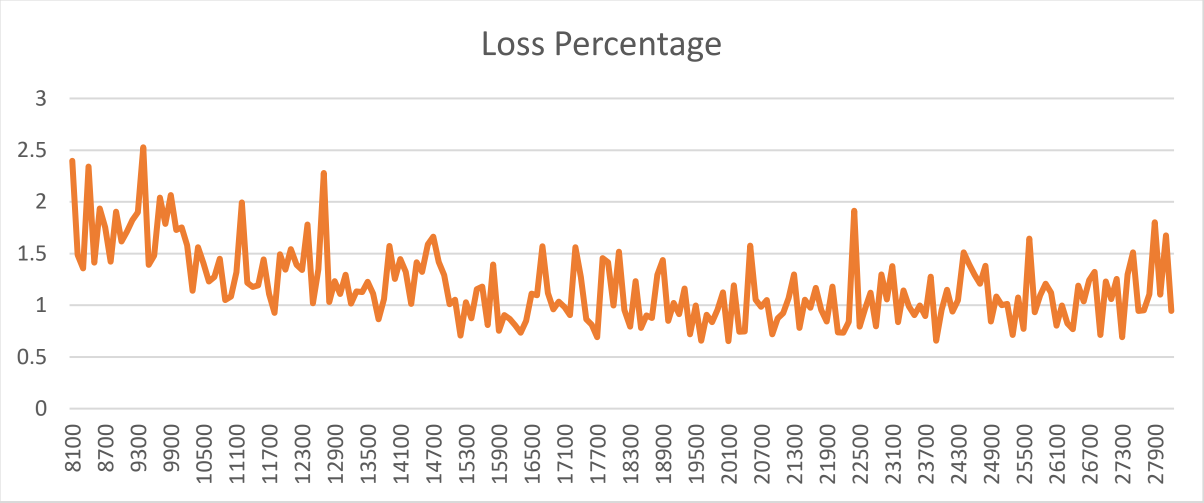


Figure 4: Loss/ Training Data Graph

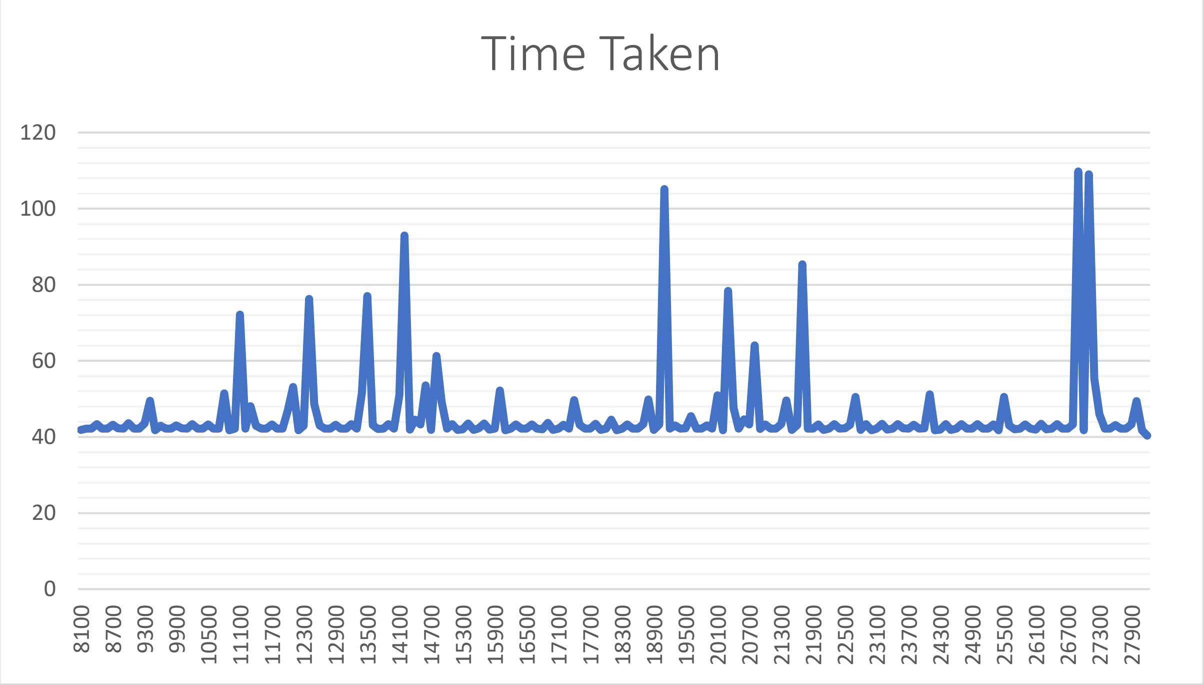


Figure 5: Time Taken/Training Data Graph

VIII. PERFOMANCE COMPARISION

Referencing the Benchmark Reports by Elmar Haußmann at Rise ML, below are the report on experiments with ResNet and Inception. [8]

Experiments for TPUs and P100 were run on Google Cloud Platform on n1-standard-16. For the V100 GPU, p3.2xlarge instances on AWS were used. All systems were running Ubuntu 16.04. On ResNet-50, a single Cloud TPU is ~7.3 faster than a single P100 and ~2.8 times faster than a V100. For InceptionV3, the speedup is almost the same. With higher precision, the V100 loses a lot of speed. Based on the above referenced experiment we can have a definite conclusion that a TPU performs significantly better than CPUs and GPUs in running a TensorFlow Application.

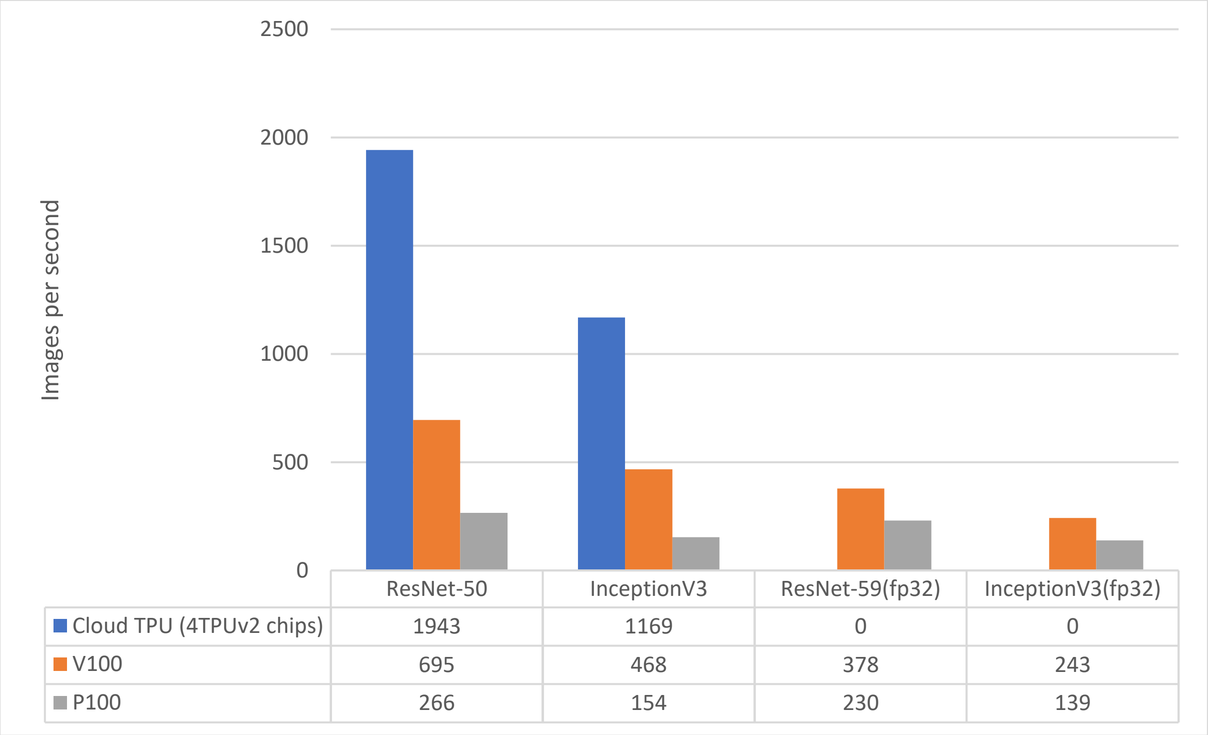


Figure 7: Performance Graph

ResNet-50 Performance				
	TPUv2	V100 fp16	P100	V100
Cloud	Google	AWS	Google	AWS
Price $ per hour	7.26	3.06	1.58	3.06
Images/second	1943	695	230	378
Performance images/s per $	268	227	146	124

Figure 6: Performance Table

XI. ENERGY COMPARISION

TPUs are also energy efficient when compared to CPUs and GPUs. According to Google, on production AI workloads that utilize neural network inference, the TPU is much more energy efficient, delivering 30 times to 80 times improvement in TOPS/Watt measure. The Energy Efficiency can be attributed to the Architecture of the TPUs.

X. CONCLUSION

Based on our study and the TPU evaluation using RetinaNet we conclude that TPU are better suited to TensorFlow Applications than CPUs and GPUs. TPUs adapt better to TensorFlow owing to their ASIC Architecture. TPUs also reduces the performance to cost ratio. TPUs are energy efficient and can be easily adapted in an existing GPU environment. Since TPUs are proprietary of Google their utilization is restricted. once TPUs are commercially available to a larger audience, they could become a real alternative to Nvidia GPUs. TPUs can propel Machine Learning few decades ahead.

XII. REFERENCES

[1].
Samuel, Arthur (1959). "Some Studies in Machine Learning Using the Game of Checkers". IBM Journal of Research and Development.

[2].
https://opensource.com/article/17/11/intro-tensorflow

[3].
https://www.tensorflow.org/extend/architecture

[4].
https://www.nvidia.com/en-us/data-center/gpu-accelerated-applications/tensorflow/

[5].
https://cloud.google.com/blog/big-data/2017/05/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu

[6].
https://cloudplatform.googleblog.com/2017/04/quantifying-the-performance-of-the-TPU-our-first-machine-learning-chip.html

[7].
https://www.nytimes.com/2018/02/12/technology/google-artificial-intelligence-chips.html

[8].
https://blog.riseml.com/benchmarking-googles-new-tpuv2-121c03b71384

[9].
https://medium.com/@14prakash/the-intuition-behind-retinanet-eb636755607d

[10].
In-Datacenter Performance Analysis of a Tensor Processing Unit TM Norman P. Jouppi, Cliff Young, Nishant Patil et al
The 44th International Symposium on Computer Architecture (ISCA) 2017.

[11].
TensorFlow: A system for large-scale machine learning
Martin Abadi , Paul Barham et al
12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)