# GPU ACCELERATED COMPUTING & DEEP LEARNING

## A Performance and Power Analysis

Dipti Chaudhari
CSCS: 570 Term Paper Assignment
California State University,
Long Beach, CA
dipti.chaudhari.24@gmail.com

***Abstract*—Data Science is future. Application developed on deep learning are revolutionizing the perspective how the problems were solved. In recent years the Deep Learning has impressed world with applications like Baidu Speech 2, Alpha Go and Autonomous vehicle. These application needs very high computing power, GPUs serve best in such cases. Even the smallest GPU(Tegra X1) has shown significant improvement in performance as compared to the powerful CPUs.**

**In this study we look at the roadmap of development of GPUs and how it has boosted performance of Deep Learning. We will look at the the performance analysis using Caffe framework and cuDNN library using Alexnet environment.**

***Index Terms*—GPU, Deep learning, performance analysis**

### *[1] Introduction*

A lot, mean a lot of data is being generated every sec due to all kinds of activities carried out on internet. Let it be surfing on internet, internet shopping, online banking or booking flight tickets. With this data the Data Science has taken boom. It is growing more than ever and has surprised the computer scientist with significant discoveries in Bioinformatics, Healthcare, Cyber security and many more areas. With advancement in these application the Artificial intelligence is becoming more and more popular among computer scientist and flourishing Machine learning.

Machine learning is an application of artificial intelligence. Neural Network is similar branch of science where the programs are given ability to think like human brain. As a analogy for human brain's neuron a computational unit is developed and as a brin is complex network of billions and trillions of neurons, similarly a dense network of artificial neurons is developed. The Deep learning is a part of neural networks which supports development of such deep neural networks which gives applications which were beyond human imagination.

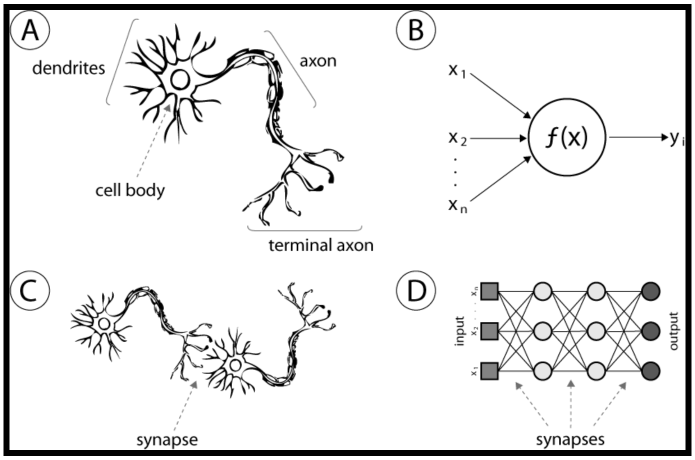


Fig 1: Human neuron and Artificial neuron.

As an example, Google's Deep mind's Alpha go made revolutionary impression by defeating the world's champion. The Deep Learning and Machine learning applications are build as network shown in fig 1, where we pass input and it processes and gives output. There is a development process has phases like training and testing. The empirical data we have is used to perform these phases. Large datasets needs a lot of computation capability. Where the GPU comes in picture. There are many libraries available to implement neural network using GPU computing like Google Tensorflow and frameworks like cuDNN. Here in this study we are going to look at these aspects and how growth of GPU has boosted the Deep learning performance.

We will see the performance comparison for Caffe deep learning framework (http://caffe.berkeleyvision.org/) and cuDNN-4 library (https://developer.nvidia.com/cudnn) on different GPU versions. We will also look at in which phase namely, inference and training phase it helps a lot.

## *[2] High performance computing paradigm*

High performance computing touches our life to a great extend. Most of the application we depend on these days involve very high performance computing. The field range from medical, education, stock and trading and what not. Need of high computation for such applications demands supporting hardware. Relying only on CPU's sequential execution becomes insufficient. Even with multithreading (https://computing.llnl.gov/tutorials/pthreads/ ) the scalibility is restricted to number of cores. Whereas the GPU comes with thousands of cores.

### CPU vs GPU

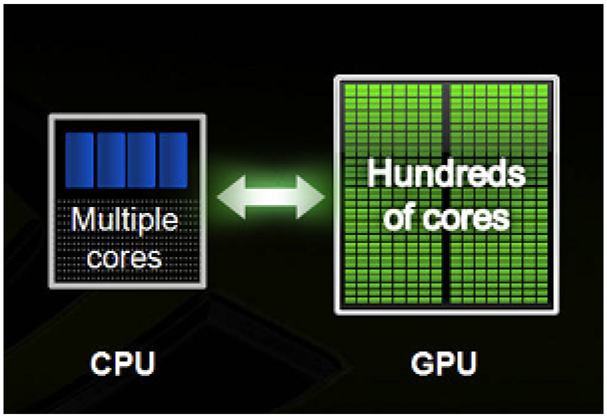


Fig2: CPU vs GPU

CPU is Central processing unit and it is considers as the brain of computer. But GPU is graphics processing unit which is soul of computer as now all the extended scalability and computational power is provided it. The chip rendered in every computer for graphical purpose is capable only of performing very basic image processing for applications like MS office's powerpoint. GPU are far beyond that. GPU along with graphics processing it gives computational capability to programer along with configurable environment. GPU is intensively used for performing intensive computational tasks.

### Accelerating Insights

Companies like Google, Facebook, Amazon have huge data centers. Let's take look at Google's and Stanford University's AI labs DCs info:

| | |
|---|---|
| Google Data Center<br>600 K Watts<br>$5,000,000 | 1000 CPU Servers<br>2000 CPUs * 16,000 Cores |
| Stanford AI laboratory<br>4 K Watts<br>$33,000 | 3 GPU-Accelerated Servers<br>12 GPUs * 18,432 cores |

Table 1: Google's data center vs Stanford AI lab infrastructure

An experiment conducted by Google of predicting cats in image used the resources mentioned in first row. Whereas the second row represents the same experiment conducted by Stanford Data Scientists leaded by Andrew Ng. These massive data centers now are possible to build with less cost. Wired (https://www.wired.com/2013/06/andrew_ng/ ) mentioned "Now you can built Google's $1M Artificial brain on the cheap". More and more companies are moving towards GPU computing.

### From HPC to Enterprise Data Centers

Accelerated computing is adopted by almost all companies from various domains. Let's take a look at examples:

- **Oil and Gas**: Schlumberger, BR, Petrobras, ENI, Statoli
- **Higher Ed**: Harward, Georgia Tech, University of Cambridge, Stanford

- **Government**: Raytheon, NASA, Naval Research Laboratory
- **SuperComputing**: Oak Ridge, Lawrence, Livemore National Laboratory
- **Finance**: JP Morgan, BArclays, Standard Life, Murex
- **Consumer Web**: Salesforce, BAidu, amazon.com, Yandex

Rapid adoption of accelerated computing in all the fields have remarkably observed in year of 2013. And 28700 apps were built using accelerated computing by 2014.

And out of all this adoption 85% of GPU were Nvidia GPUs.

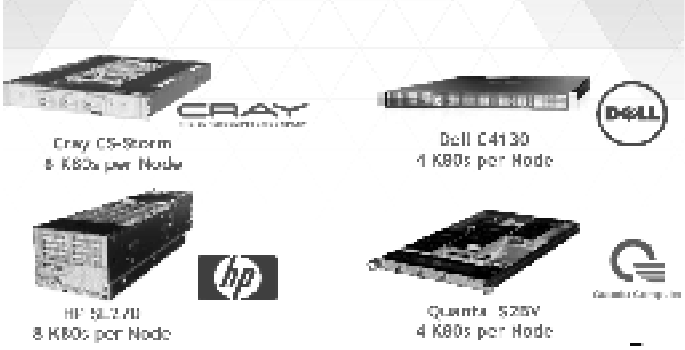


Fig 3: High GPU servers are widely used and could be found in every brands

### GPU roadmaps

Data center infrastructure and Development both are being accelerated. System Solutions, Communication and infrastructure Management are widely available for Data center infrastructure.

Many programming language is supported like Python,C/C++, OpenACC. Development tools like allinea DDT and software solutions like Matlab, Kitware provide wide scope for programmers to develop GPU based application. Let us view the development roadmap.

The GPU development is shown in Fig 4.

TESLA K80 world's fastest accelerator for data analytics and scientific computing with Dual-GPU Accelerator for Max Throughput. It is 2X faster, have double memory (24 GB) and gives dynamically maximum performance by boosting GPU.

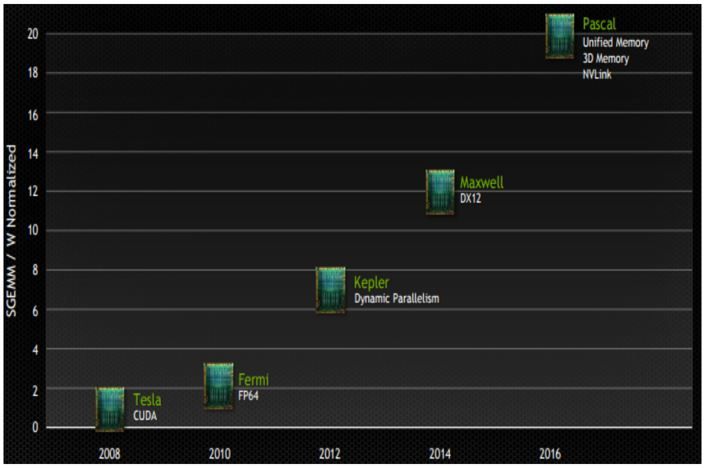


Fig 4: Road- map of GPU development

Pascal GPU Features NVLINK and Stacked Memory.

NVLINK GPU high speed interconnect 80-200 GB/s 3D whereas Stacked Memory 4x Higher Bandwidth (~1 TB/s). It boosts 3x Larger Capacity and 4x More Energy Efficient per bit.

### Accelerating Computing 5X to 10X

By year 2013, accelerated computing was 5X higher than before and just within 2 years there was 10X growth in GPU computing.

The GPU computing made the high intense computation very easy. Field ranging from material science, bioinformatics, atmospheric models to modern robotics and video gaming. These areas require very high computational efficient environment.

The table below we will see that in short period how the popularity for these applications increased and how cuda was adopted by developers.

Also universities started including gpu programming and supporting assignments in curriculum.

Table below mentions the growth:

|  | 2008 year | 2015 year |
|---|---|---|
| Downloads | 150,000 | 3 Millions |
| Cuda Apps | 27 | 319 |
| Universities Teaching | 60 | 800 |
| Academic Papers | 4000 | 60000 |
| Tesla GPUs | 6000 | 450000 |
| Supercomputing Teraflops | 77 | 54000 |

Table 2: 5X to 10X GPU computing adoption statistics

### *[3] Deep Learning*

Deep Learning(DL) is a movement. Everyone is talking about it and the jobs in this branch if AI is attracting engineers. So what is Deep Learning? Deep Learning is one of the hardest thing happening in areas with most efficient and skilled resources. DL helps us to imagine life beyond expectation from the machines. For example an average person takes at least around 1000 photographs in an year. When he wants to take look at one specific photo he has to go and look manually scrolling in his phone but DL enables him to just mentions some key objects in may be he can just say them and the phone automatically pull out the desired images. This may involve, voice recognition, image processing, predictive analysis and many machine learning application. Isn't it wonderful?

### Deep Learning overview

Deep learning is solution from branch of neural networks. It is a multi layers neural network and the layers signifies the word '*deep*'. It is because the analysis it does. Billions and millions input data is passed through this network and it slowly recognizes edges, then it recognizes what formation is termed as nose, eyes. The network learns to extract these features from image and slowly trains itself to recognize the object present in image. This phase is called as training phase. And partial data is used to test the accuracy of result given by trained network. It is termed as testing phase or inference phase.

### Traditional Machine Learning Approach vs Deep Learning Approach

Traditional machine learning had perception to extract hand-crafted features. The problem solutions were restricted due to lack of availability of resources. The Deep learning is capable of solving more complex and complicated tasks. More amount of Data is available to rely on, the data is just exploding every now and then. More than 500 million of images are uploaded to facebook, this is big data. This amount of data needs to be crunched a lot. Either you can buy 1000 CPU servers or you can buy 3 CPU servers with GPU. However this is just in terms of images. We will discuss most significant applications of DL which has surprised the computer science world.

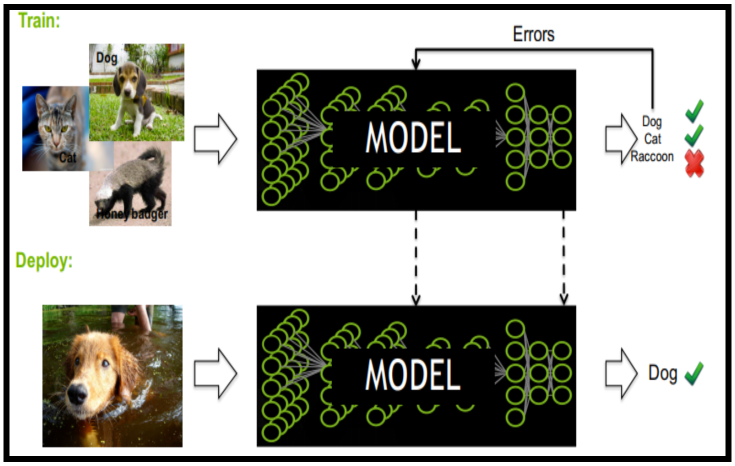


Fig 5: Deep Learning training

Deep learning training compared to inference. In training phases a large data set is used to train the model. In inference phase, the trained network is used to predict the results on similar data but the amount of data used is small.

### Most Significant Applications of Machine Learning

In these section we will look at some applications which has made remarkable contribution to the growth of deep learning field. Many big companies like Google

Microsoft use ML and DL for famous products like Siri, Google Now, Face recognition etc.

### Baidu Deep Speech 2

Fig: Baidu Deep Speech 2

English and Mandarin are most widely used languages. Baidu speech 2 is termed as Google of China. This application is end-to-end deep learning approach speech recognition for these language. The translation is also possible and it is made a lot simpler with help of DL. It requires no feature engineering or Mandarin specifics is required.

### Alpha Go

Fig: DeepMind AlphaGo

First time in history a computer software defeated the world's champion in games. Deep Mind is Google's part working on DL. Japanese game Go was implemented using DNN and was trained for 3 weeks. 340 million training steps on 5 GPUs were used.

Alpha Go is very great achievement in field of artificial intelligence.

### DL Autonomous Vehicle

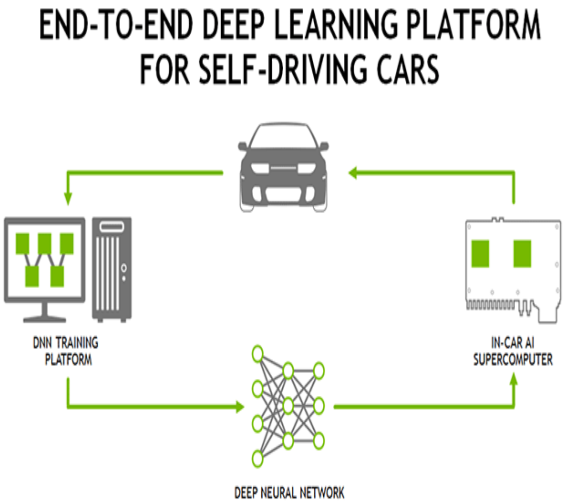


Fig: Self Driving car

Autonomous vehicles are being developed by all top brands like Audi, Mercedes, Ford. These are self learning cars who are capable for driving in traffic, in scenarios of unclear vision and few also successfully find parking spots.

### *[4] Power and Performance analysis*

So far we have seen how the GPU computing has become popular in past few year. Also influence of Deep Learning is our walks of life ranging from medical to drug development, from stocks and trading to finance and marketing. For development of DL applications very high computing infrastructure is needed. When we discussed things supported growth of DL one was infrastructural support. High computational ability of GPU is one of the most imp factor. We are going to compare the performance of DL application on different hardware. Namely, an NVIDIA Jetson™ TX1 developer kit, an NVIDIA GeForce GTX™ Titan X, an Intel Core™ i7 6700K, and an Intel Xeon™ E5- 2698 v3.

### Inference versus Training

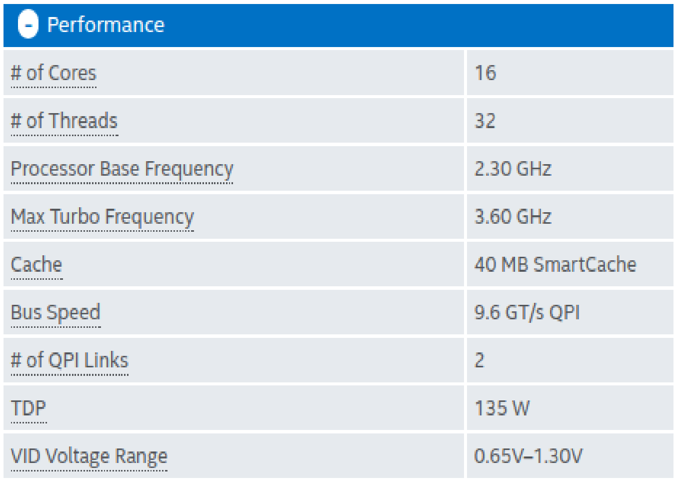
In section IV we saw how a neural network is trained. The phases Inference and training both start with forward propagation. Forward propagation means the data is being transferred from input layer to output layer in forward direction. Result from previous computational unit is forwarded to the next unit. But in training, the result obtained from forward propagation is compared with the actual result and for incorrect result the result is propagated backwards for rectification. Hence in training the per image workload is high. Hence to maximize throughput training phase needs to be attended.

### Experiment Detailing

Ebay's neural network processing implementation of maxDNN has shown how the performance can be boosted with use of GPU(for more than 96%). On other hand it also depends the consistency of the environment uses. Hence for this study we are considering Caffe deep learning framework (http://caffe.berkeleyvision.org/). Caffe is developed by Berkeley Vision and Learning center(BVLC). Along with Caffe framework for this study we are using cuDNN library (https://developer.nvidia.com/cudnn ) developed by nvidia. Its Deep Neural Network Library of GPU accelerated computing developed by Nvidia. In cuDNN 4 these were significant changes over last version like cuDNN 3 computed convolutions using algorithm called precomputed implicit GEMM(generalized matrix-matrix product). For powerful GPU's like GeoForce GTX Titan X with 24 SMs for 3072 CUDA cores. Using cuDNN 3 restricted to two thread blocks or eight wraps in total. Each thread block is assigned to only one GPU's Streaming Multiprocessor. Out of 24 only two were used. This was overcome by cuDNN 4 along with other changes which optimized the utilization of GPU's capability.

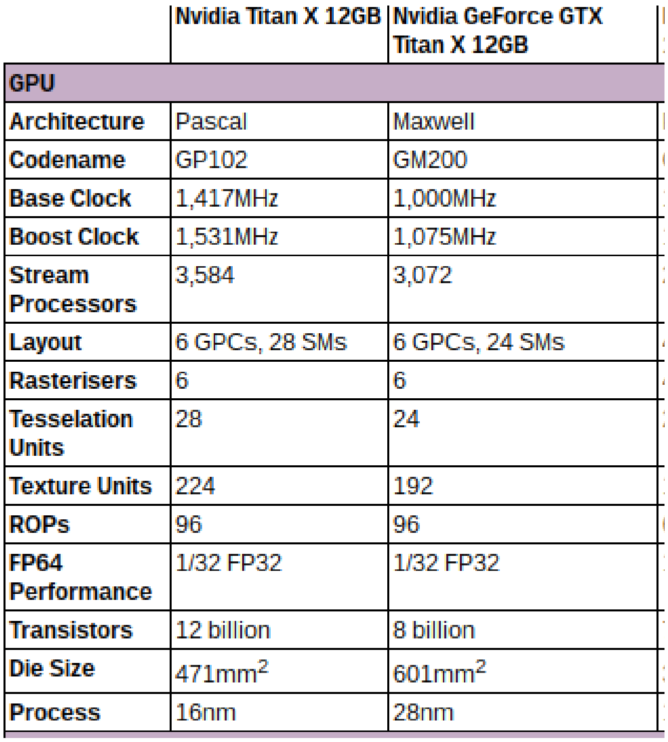
### Setup and Testing

The NVIDIA Tegra X1 and the Intel Core i7 6700K used as client-side processors for this experiment, and the NVIDIA GeForce GTX Titan X and a 16-core Intel Xeon E5-2698 v3 as high-end processors. Stable and consistent versions of cuDNN 4 and Caffe were used. Alexnet is image processing neural network.



| Performance | |
|---|---|
| # of Cores | 16 |
| # of Threads | 32 |
| Processor Base Frequency | 2.30 GHz |
| Max Turbo Frequency | 3.60 GHz |
| Cache | 40 MB SmartCache |
| Bus Speed | 9.6 GT/s QPI |
| # of QPI Links | 2 |
| TDP | 135 W |
| VID Voltage Range | 0.65V–1.30V |

Fig 6: Configuration of Intel® Xeon® Processor E5-2698 v3



| | Nvidia Titan X 12GB | Nvidia GeForce GTX Titan X 12GB |
|---|---|---|
| **GPU** | | |
| **Architecture** | Pascal | Maxwell |
| **Codename** | GP102 | GM200 |
| **Base Clock** | 1,417MHz | 1,000MHz |
| **Boost Clock** | 1,531MHz | 1,075MHz |
| **Stream Processors** | 3,584 | 3,072 |
| **Layout** | 6 GPCs, 28 SMs | 6 GPCs, 24 SMs |
| **Rasterisers** | 6 | 6 |
| **Tesselation Units** | 28 | 24 |
| **Texture Units** | 224 | 192 |
| **ROPs** | 96 | 96 |
| **FP64 Performance** | 1/32 FP32 | 1/32 FP32 |
| **Transistors** | 12 billion | 8 billion |
| **Die Size** | 471mm$^2$ | 601mm$^2$ |
| **Process** | 16nm | 28nm |

Fig 7: Configuration of NVIDIA GeForce GTX Titan X

Two cases were considered. First where large batch of images is accepted as input. And second is latency-focused hence the smaller batch is feasible but for testing purpose no batching is considered(worst case). We use the default Caffe version provided by BVLC for FP32 benchmarking on Tegra X1, and

NVIDIA’s Caffe branch providing FP16 support and enabling more efficient convolutions [16] when benchmarking FP16 Tegra X1 and Titan X, respectively. The Intel CPUs run the most optimized CPU inference code available, the recently release Intel Deep Learning Framework (IDLF).

### Analysis on Small and Large GPUs

Results for batching and without batching are discussed in this section.

The results clearly shows that the comparison between the performance of Tegra X1 and Core i7 6700K. The results are taken considering batching and no- batching in training. The units used are performance -> (images/second), power -> Watts and energy efficiency -> images/second/Watt:

**For low-scale infrastructure**

| Network: AlexNet | Batch Size | Tegra X1 (FP32) | Tegra X1 (FP16) |
|---|---|---|---|
| Inference Performance | 1 | 47 img/sec | 67 img/sec |
| Power | 1 | 5.5W | 5.1W |
| Performance/Watt | 1 | 8.6 img/sec | 13.1 img/sec/ W |

Table 3: Performance of Tegra X1 for batch size 1

Table 3 describes the performance of Tegra X1 when the experiment is done using batch size 1. The performance is measured in terms of performance in inference phase.

Also the 32 bit and 16 bit versions are used to do this analysis.

| Network: AlexNet | Batch Size | Core i7 6700K |
|---|---|---|
| Inference Performance | 1 | 62 img/sec |
| Power | 1 | 49.7 W |
| Performance/Watt | 1 | 1.3 img/sec/W |

Table 4: Performance of Core i7 for batch size 1

Table 4 describes the performance of Core i7 when the experiment is done using batch size 1. The performance is measured in terms of performance in inference phase.

Core i 7 made use of intel deep learning framework.

| Network: AlexNet | Batch Size | Tegra X1(FP16) | Tegra X1 (FP16) |
|---|---|---|---|
| Inference Performance | 128 | 155 img/sec | 258 img/sec |
| Power | 128 | 6.0W | 5.7W |
| Performance/Watt | 128 | 25.8 img/sec/ W | 45.0 img/sec/ W |

Table 5: Performance of Tegra X1for batch size 128

Table 5 describes the performance of Tegra X1 when the experiment is done using batch size 48. The performance is measured in terms of performance in inference phase.

Also the 32 bit and 16 bit versions are used to do this analysis.

Table 6 describes the performance of Core i7 when the experiment is done using batch size 1.

| Network: AlexNet | Batch Size | Core i7 6700K |
|---|---|---|
| Inference Performance | 48 | 242 img/sec |
| Power | 48 | 62.5 W |
| Performance/ Watt | 48 | 3.9 img/sec/W |

Table 6: Performance of Core i7 for batch size 48

The performance is measured in terms of performance in inference phase.
Core i 7 made use of intel deep learning framework.

The experiment shows that Tegra X1 with FP16 is an order of magnitude more energy-efficient than CPU-based inference.

**For Powerful Infrastructure**

For the Titan X GPU and the Xeon E5-2698 v3 server-class processor the results are as follow:

| Network: AlexNet | Batch Size | Titan X(FP32) |
|---|---|---|
| Inference Performance | 1 | 405 img/sec |
| Power | 1 | 164 W |
| Performance/ Watt | 1 | 2.5 img/sec |

Table 7: Performance of Titan X(FP32)for batch size 1

Table 7 describes the performance of Titan X(FP32) when the experiment is done using batch size 1. The performance is measured in terms of performance in inference phase.
Only the 32 bit versions is used to do this analysis.

| Network: AlexNet | Batch Size | Xeon ES-2698 v3 (FP32) |
|---|---|---|
| Inference Performance | 1 | 76 img/sec |
| Power | 1 | 111.7 W |
| Performance/ Watt | 1 | 0.7 img/sec/W |

Table 8: Performance of Xeon ES-2698 v3 (FP32) for batch size 1

Table 8 describes the performance of Xeon ES-2698 v3 (FP32) when the experiment is done using batch size 1. The performance is measured in terms of performance in inference phase.
Only the 32 bit versions is used to do this analysis.
Xeon ES-2698 v3 (FP32) made use of intel deep learning framework.

| Network: AlexNet | Batch Size | Titan X(FP32) |
|---|---|---|
| Inference Performance | 128 | 3216 img/sec |
| Power | 128 | 227.0W |
| Performance/ Watt | 128 | 14.2 img/sec/W |

Table 9: Performance of Titan X(FP32)for batch size 128

Table 9 describes the performance of Titan X(FP32) when the experiment is done using batch size 128. The performance is measured in terms of performance in inference phase.
Only the 32 bit versions is used to do this analysis.

Table 10 describes the performance of Xeon ES-2698 v3 (FP32) when the experiment is done using batch size 1.

| Network: AlexNet | Batch Size | Xeon ES-2698 v3 (FP32) |
|---|---|---|
| Inference Performance | 48 | 476 img/sec |
| Power | 48 | 149 W |
| Performance/ Watt | 48 | 3.2 img/sec/W |

Table 10: Performance of Xeon ES-2698 v3 (FP32) for batch size 48

The performance is measured in terms of performance in inference phase.
Only the 32 bit versions is used to do this analysis.
Xeon ES-2698 v3 (FP32) made use of intel deep learning framework.

Comparing both the cases we can see, powerful GPUs like Titan X where in table

## *[5] Conclusion*

In conclusion we infer that the inference phase is benefitted as training by use of GPU. The high-performance and small consumer computing spaces both show significant increase boost when used GPUs ranges from most powerful GPU Titan X to the smallest GPU(Tegra X1). Overall if we observe the growth of Deep Learning popularity, considering its intensive requirement of high infrastructural hardware, was majorly supported by developments of GPU.

## *[6] References*

[1]. Deep Learning on GPUs
Conference- GPU technology Conference
Seminar Link -
http://on-demand-gtc.gputechconf.com/gtcnew/on-demand-gtc.php
Date of Publish- March 2016
Access Date- 10 Dec 2016
URL- Link

[2]. The Breadth of the GPU Accelerated Computing Platform and Its Impact on Deep Learning
Conference 2015 Stanford HPC Conference
Seminar Link -
http://www.hpcadvisorycouncil.com/events/2015/stanford-workshop/agenda.php
Name of speaker - Sumit Gupta from Nvidia
Date of Publish- 2-3 February 2015
Access Date- 10 Dec 2016
URL- Link

[3]. DEEP LEARNING WITH GPUS
Conference GEOINT 2015
Seminar Link -
http://www.nvidia.com/object/geoint-2015.html
Name of speaker - Larry Brown from Nvidia
Date of Publish- May 2015
Access Date- 10 Dec 2016
URL- Link

[4]. GPU-Based Deep Learning Inference: A Performance and Power Analysis
WhitePaper by Nvidia
Date of Publish- Nov 2015
Access Date- 10 Dec 2016
URL- Link

[5]. NVIDIA, "NVIDIA cuDNN - GPU Accelerated Deep Learning,"
Link: https://developer.nvidia.com/cudnn

[6].On Optimization Methods for Deep Learning
By http://machinelearning.wustl.edu/
Authors:Quoc V. Le quocle@cs.stanford.edu
Jiquan Ngiam jngiam@cs.stanford.edu
Adam Coates acoates@cs.stanford.edu
Abhik Lahiri alahiri@cs.stanford.edu
Bobby Prochnow prochnow@cs.stanford.edu
Andrew Y. Ng ang@cs.stanford.edu
University: Computer Science Department, Stanford University, Stanford, CA 94305, USA
URL: Link

[7]. NVIDIA GPUs - The Engine of Deep Learning
URL: Link

## *[7] Appendix*

| | |
|---|---|
| HPC | High Performance Computing |
| GPU | Graphics Processing Unit |
| CPU | Central Processing Unit |
| ML | Machine Learning |
| DDT | Debugger Development Tool |
| 3D | 3 Dimension |
| NN | Neural Networks |

| | |
|---|---|
| CNN | Convolutional Neural Networks |
| DNN | Deep Neural Network |