



Basic Performance Analysis of NVIDIA GPU Accelerator Cards for Deep Learning Applications

Frank Han, Thomas Zhu, and Rene Meyer

Executive Summary

The study provides a basic performance analysis of NVIDIA K40, K80 and M40 Enterprise GPU accelerator cards, and GeForce GTX Titan X and GTX 980 Ti (water-cooled) consumer grade cards for deep learning applications. As a benchmarking tool, we used the Caffe software suite running a 256x256 pixel image recognition training without further optimization. The study includes:

- Card specific performance analysis
- Performance scaling from single GPU system to up to 8x GPU nodes
- Performance impact of the CPU
- Single and dual CPU solutions
- Platform-specific performance differences

We find that the theoretical single precision TFLOP performance does not necessarily correlate well with the real life benchmark results and should not be used as a gating factor for system designers and architects. Cards based on the Maxwell architecture - M40, Titan X, and 980 Ti - outperform K80 and K40 cards based on the older Kepler architecture. The performance of the K80 is surprisingly low even though it has the highest single precision TFLOP performance spec. Configurations up to 8 cards per node show a linear performance increase with increasing number of accelerator cards independent of the type of cards. Doubling the number of accelerator cards leads to a performance increase around 70%. In multi-card configurations, the TESLA K80 scaled less favorable, which may be related to the internal dual-card design and the required PCIe bus multiplexing.

No significant impact on the performance is found for different CPUs except the entry level CPU. 4-card single CPU workstations and 4-card dual CPU server configurations perform similarly. Also, the tested systems performed equally within the precision of the test.

Based on the findings, we recommend the use of Maxwell-architecture based cards in combination with general purpose E5-2600 CPUs for Deep Learning (DL) applications. The water-cooled version of the GTX 980 Ti cards revealed the best benchmark performance and should be considered as a viable entry level card for R&D purposes.

Introduction

Deep learning is one of the fastest-growing segments of the artificial intelligence field [1]. Deep Learning, also known as deep machine learning and hierarchical learning, is a form of computer-simulated learning based on algorithms. It uses deep neural network, convolutional deep neural networks and recurrent neural networks to apply towards fields such as computer vision, automatic speech, facial and motion recognition, self-driving automotive technologies and other forms of artificial intelligence machine learning tools. Other enterprise use cases include fraud detection, security/surveillance and bioinformatics applications.

While most white papers on Deep Learning consider the computational expressions, implementations of algorithms and programing in the execution of Deep Learning, this white paper will address specific hardware findings as related to Deep Learning applications and optimal uses for hardware platforms. The goal of this white paper is to address and test specific hardware platforms, applications and best uses practices while targeting optimal performance of the system.

Hardware and Software Setup

GPU accelerator Cards

The test includes NVIDIA's enterprise grade GPU cards TESLA K40, TESLA K80 and NVIDIA GRID M40. The NVIDIA GRID M4 was not available at the time of the test. For consumer grade cards, we tested the ASUS GeForce GTX Titan X X-12GD5 and the GIGABYTE GeForce GTX 980Ti 6GB WATERFORCE. Table 1 displays the cards specifications.

Table 1

	K40	K80	M40	Titan X	980 Ti
Architecture	Kepler	Kepler	Maxwell-2	Maxwell-2	Maxwell-2
Chipset	GK110B	2xGK210	GM200	GM200	GM200
Base Clock (MHz)	745	562	948	1000	1241
Memory Clock (GHz)	6 GDDR5	5 GDDR5	6 GDDR5	7 GDDR5	7.2 GDDR5
Capacity (GB)	12	2x12	24	12	6
Memory BW (GB/sec)	288	2x240	288	337	337*
SP Base/ Turbo (TFlops)	4.3/5.0	5.6/8.7	5.8/6.8	7.0	6.4*
TDP (Watt)	235	300	250	250	250
Max GPU Temp (C)	90	90	NA	91	91
Cooling	Passive	Passive	Passive	Active	Water

**Base model, model tested has a 25% higher base clock*




Based on the single precision performance, we expected the K80 to come in first. We added Gigabytes GTX 980Ti WATERFORCE to the test, because the water cooling allows for the significantly lowering of the chip temperature even at a higher base and boost clock. At a clock 25% higher compared to the standard model, we expected the card to perform on par with the Titan X.

Compute Platforms

Two server platforms, DL-E800 and DL-E380, and one workstation platform, DL-E400, from AMAX's Deep Learning Solutions product line, have been tested. The primary difference between the servers is the form factor, the number of supported drive bays and power redundancy. For example, while both systems support up to 8x GPU cards, the DL-E800 system supports 24x 2.5inch drive bays and comes with a 2+2 redundancy on the power supply while the DL-E380 supports 6x 2.5inch drive and 2+1 power supply redundancy.

Also tested was the DL-E400, a 4U single socket workstation.

Table 2

Hardware	DL-E800	DL-E380	DL-E400
Form Factor	4U (5Uw/cover) 	3U (4Uw/cover) 	4U Workstation 
Sockets	Dual Socket	Dual Socket	Single Socket
Memory	256 GB	256 GB	32 GB
HDD	KINGSTON SV300S3	KINGSTON SV300S3	KINGSTON SV300S3

Processors from the Intel E5-2600v3 family were selected to investigate the effect of core count and CPU clock on the DL benchmark test. A basic work load and an entry level CPU were also added to the test (see Table 3).

Table 3

CPUs	Core #	HT Core #	Frequency
E5-2669v3*	12	24	2.3 GHz
E5-2637v3	4	8	3.5 GHz
E5-2620v3	6	12	2.4 GHz
E5-2609v3	6	6	1.9 GHz

*Note: E5-2669v3 is a special edition of CPU, it has same core count and frequency with E5-2670v3.

Software

The software configuration for the test system is listed in Table 4.

Table 4

Software	DL-E800	DL-E380	DL-E400
OS	Ubuntu 14.04.3 LTS	Ubuntu 14.04.3 LTS	Ubuntu 14.04.3 LTS
CUDA	7.5	7.5	7.5
Selinux	Disabled	Disabled	Disabled
NVidia Digits	2.0	2.0	2.0
Caffe	0.13	0.13	0.13
cuDNN	3	3	3
CNMeM	1.0.0	1.0.0	1.0.0

Benchmark Test

We used the Caffe suite version 0.13 for internal benchmark testing. The test constitutes of a 256x256 pixel color image recognition training phase followed by validation. As a reference system, we choose the DL-E380 3U server with dual E5-2669v3 CPU and 256GB memory and one Titan X card. The speed of which the task was completed was used as performance index for the test. No further optimization was performed. The sample database contained about 10000 images. A benchmark test with a larger sample size is planned for the next iteration of the test. The test time for the reference system is about 15min. Results are displayed normalized to the test time of the reference system. Larger number indicates better performance compared to the reference. As an example, a system with performance index 2 completed the test in half the time.

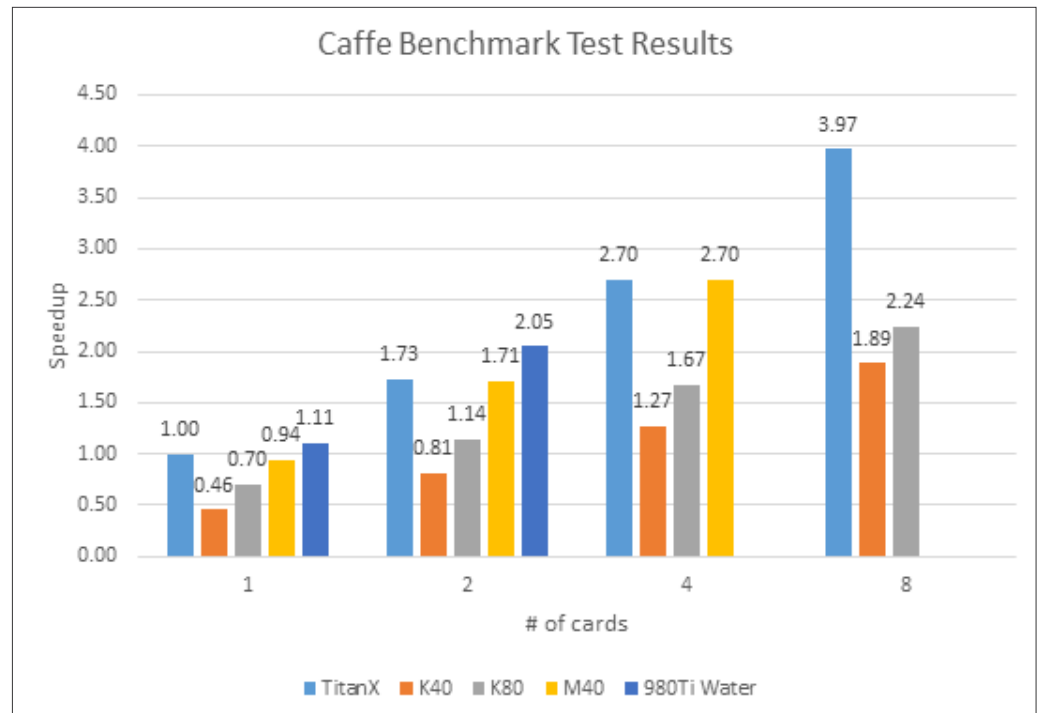
Results

GPU Comparison

The GPU test includes NVIDIA's enterprise grade Tesla K40 and Tesla K80 as well as the M40 Grid card. We added the GTX Titan X as well as the GTX 980 Ti as consumer grade cards to the test pool. Based on the single precision performance shown in table XX, we expect to see the highest performance from the Tesla K80, closely followed by M40 Grid and Titan X. As a new addition to the test, we added the water-cooled version of the GTX 980Ti. The card comes factory set with the highest clock rate and the fastest memory clock. The water cooling enables low chip temperatures and low noise levels, which provides an interesting option for under-the-desk workstation settings for R&D environments potentially attractive for development use cases.

Figure 1 illustrates the benchmark test results of the DL-E380 reference system for one, two, four and eight Titan X accelerator cards. We observed an increase in system performance of 70% for two cards, 60% for four cards and 50% for eight cards. The performance is comparable to that of the enterprise class M40 Grid card. As the test shows, Tesla K40 and Tesla K80 based on the old Kepler architecture performed significantly lower compared to the cards based on the newer Maxwell-2 architecture. Especially in the case of the Tesla K80, it performed surprisingly low in the test. Furthermore, the performance gains for 2x K80 card, 4x K80 card, and 8x K80 card systems is smaller compared to all other cards. The frontrunner in this test is the GTX 980Ti water-cooled card. Further testing is being performed and will be published as data becomes available.

Figure 1

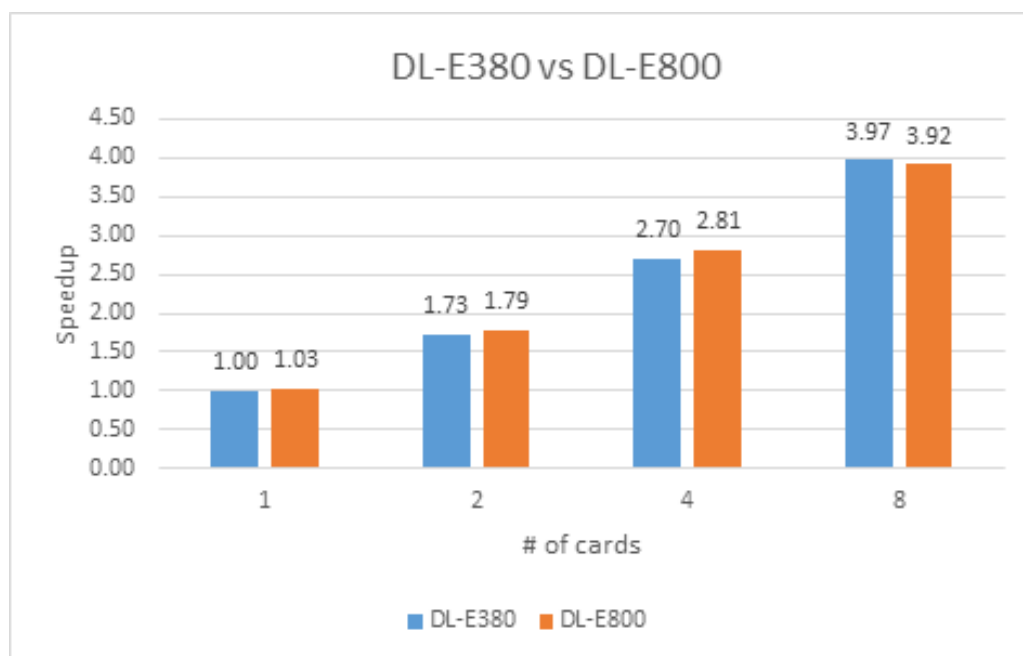


In summary we found that the on-paper, single precision performance does not necessarily translate into better DL benchmark test results. We recommend testing DL systems with application-specific workloads.

Dependence on Platform

Figure 2 shows a performance comparison of the 3U DL-E3800 reference system and the 4U DL-E800 server. The performance test results were, as expected, comparable within the error of the test. The decision for one or another platform should be based on form factor tolerance (4U versus 5U with cover), the PCIe resource requirement, and the storage requirement.

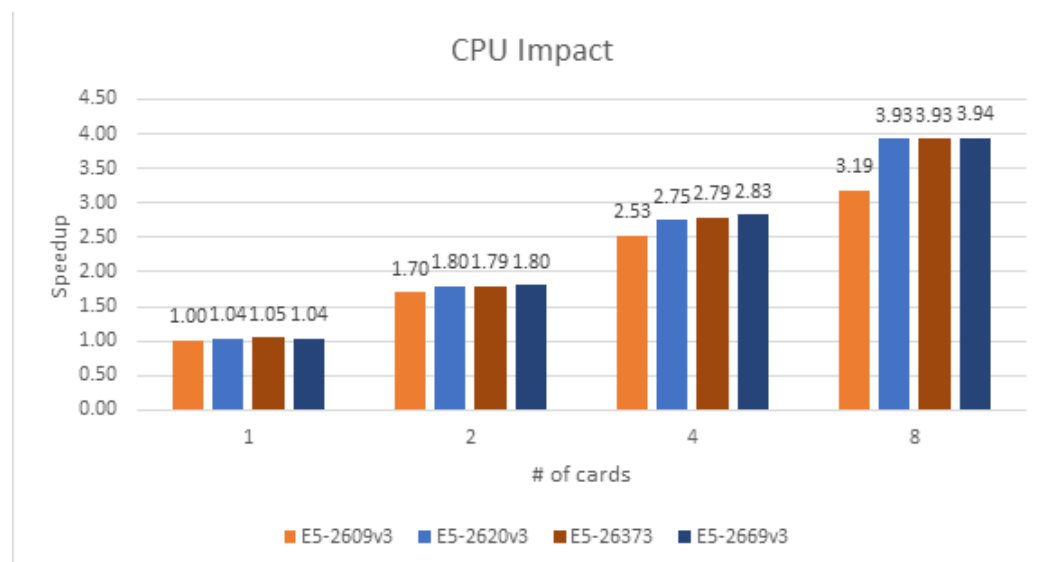
Figure 2



Impact of Core Count and CPU Clock

Figure 3 shows the impact of different E5-2600v3 CPUs on the benchmark test results. All tests were performed on the DL-E380 8-way reference system (dual socket, 256GB memory). With the exception of the E5-2609v3, all configurations performed identically. Besides the lower CPU clock, reasons for the lower performance of the E5-2609v3 might be slower memory bus, slower QPI, or the lack of hyper-threading.

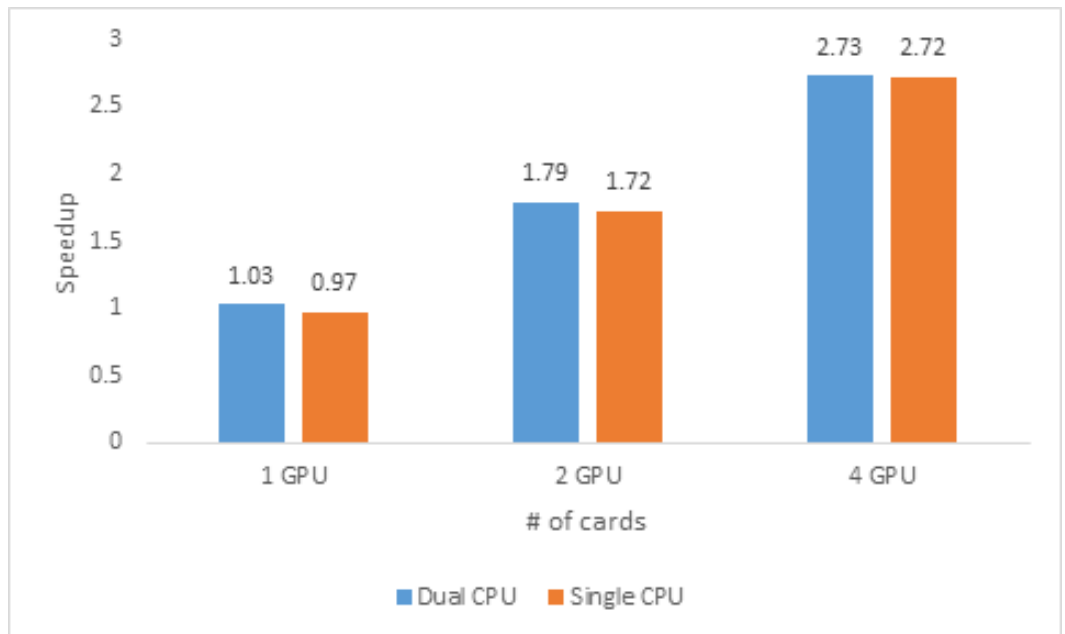
Figure 3



Single vs. Dual CPU Systems

In order to investigate the effect of PCIe multiplexing on the system performance, we benchmarked the DL-E400 workstation with single CPU E5-2620v3 and 32G memory against the DL-E800 with dual CPU E5-2620v3 and 256G Memory. From the 8-card servers we already know that there is no performance lag going from a 4x GPU to an 8x GPU configuration. We found a similar result comparing a single and dual socket system. The test results are listed in the graph below.

Figure 4



There is no significant performance difference between a single CPU workstation and dual CPU server. Therefore, if 4 GPU cards are sufficient for the DL application, we recommend considering a workstation as a more cost effective solution without a performance penalty. However, due to the PCIe resource sharing between CPUs on a dual socket board, it is best practice to always populate a dual socket board with 2 CPUs.

Impact of System Memory

Testing the effect of the memory size on the benchmark results was not within the scope of the current study and will be - in combination with larger data sets - part of future investigation.

Conclusion

M40 Grid, GTX TitanX and GTX 980Ti displayed comparable benchmark results obtained from a Caffe based benchmark test. Tesla K40 and Tesla K80 showed significantly lower performance.

Up to 8 cards, which is the maximum number due to limitation on the PCIe resources, the system performance increased between 50% and 70% when doubling the number of GPU cards.

Only little impact was found on the CPU choice. E5-2620v3 is the CPU recommended for deep learning 4x and 8x card platforms based on cost per performance.

AMAX and DL

As a leading innovator in cutting edge HPC and Data Center technology, AMAX has been an early developer in Deep Learning platforms, supporting leading enterprise and research organizations in this field. AMAX's Deep Learning Solutions feature both workstation and high-performance servers specifically designed and optimized to fast track Deep Learning algorithm development. Most recently, several teams using AMAX platforms placed first in their categories at the 6th annual ImageNet Large Scale Visual Recognition Challenge (ILSVRC 2015).

For more information please visit **www.amax.com** or contact Dr. Rene Meyer at rene_meyer@amax.com.

Reference:

[1] NVIDIA GTC: NVIDIA Bets Big On Deep Learning, Patrick Moorhead. <http://www.forbes.com/sites/patrickmoorhead/2015/03/24/nvidia-bets-big-on-deep-learning/#19fd792f5327>