Deep Learning in the Cloud vs On-Premise

A Deep Learning Centric Performance and Cost Analysis

A Technical White Paper

Jerry Li, Thomas Zhu, and Rene Meyer, Ph.D.

AMAX Corporation

Publish date: January 18, 2018

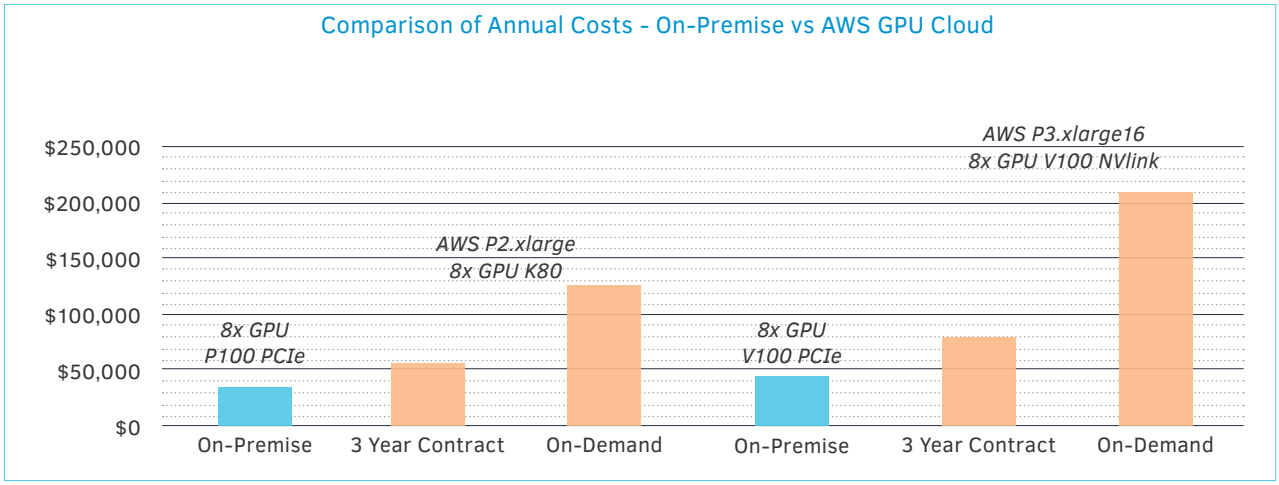
Executive Summary

This study provides a performance and cost comparative analysis of GPU-instances in the AWS public cloud against on-premise high-performance GPU servers. The objective is to determine the best performance and business options for enterprises and startups in need of accelerated computing solutions, particularly in the fields of AI and Machine Learning.

AWS offers cloud instances based on 2 hardware platforms, a high performance "Volta" platform (P3 instance) based on NVIDIA V100 NVLink GPU accelerator cards, and a general purpose "Kepler" platform (P2 instance) based on NVIDIA K80 GPU accelerator cards. AWS offers two payment options—on demand based on usage or signing a 3 year contract at a discount.

To represent comparative on-premise high-performance server solutions, we used the AMAX DL-E280 platform, a high-density 2U 8GPU server populated with either Volta-based V100 PCIe GPU accelerators or Pascal-based P40 or P100 PCIe GPU accelerators.

When calculating total cost of ownership over a 3 year period, we determined that the 8x GPU V100 on-premise solution—including operational and infrastructure costs—achieved an 80% savings when compared to an on demand AWS P3 instance, and a 45% savings when compared to an AWS P3 instance under a 3 year contract. The savings for an on-premise 8x GPU Tesla P40 solution were 73% compared to an on demand AWS P2 instance while outperforming the P2 instance, and 40% when compared to the P2 instance under a 3 year contract. Cost saving percentages were calculated assuming full utilization of AWS instances over 3 years and server cost amortized over a 3 year period with the addition of operational overhead.



Using GoogLeNet image recognition training session times as a benchmark, AWS P3 instances built on NVLink-enabled V100 hardware produced 3% better performance compared to on-premise V100-integrated GPU servers. On-premise servers with P100 GPUs outperformed AWS P2 instances by about 60%. In our benchmark, on-premise performance increased by 45% moving from the previous Tesla P100 GPU generation to the newly released Volta V100 generation.

Based on the cost and performance analyses presented in this white paper, using GPU-integrated cloud instances can be a convenient and potentially economical short-term option for companies engaged in initial Deep Learning exploration. However, companies with longer-term or larger-scale commitments to AI initiatives should consider the cost efficiency and performance advantages of an on-premise deployment.

Hardware and Software Setup

GPU Accelerator Cards

The testing utilized the NVIDIA enterprise GPU cards recommended by NVIDIA for optimal Deep Learning performance. These GPUs included: Tesla V100 in PCIe and SXM.2 form factors, Tesla P100 PCIe, Tesla P40 and Tesla K8o. Note that the Kepler-based K80 is generationally behind Volta, Pascal and Maxwell micro-architecture, but was included in this study due to being a current offering from AWS for GPU cloud compute. Table 1 shows a spec comparison between the cards.

Table 1: NVIDIA Tesla Specs Comparison

	K80	P40	P100 PCIe	P100 SXM.2	V100 PCIe	V100 SXM.2
Architecture	Kepler	Pascal	Pascal	Pascal	Volta	Volta
Chipset	2x GK210B	GP102	GP100	GP100	GV100	GV100
NVLink	-	-	-	Gen.1	-	Gen.2
CUDA Cores	4992	3840	3584	3584	5120	5120
Single-Precision	8.7 TFLOPS	12 TFLOPS	9.3 TFLOPS	10.6 TFLOPS	14 TFLOPS	15 TFLOPS
Double-Precision	2.9 TFLOPS	0.4 TFLOPS	4.7 TFLOPS	5.3 TFLOPS	7 TFLOPS	7.5 TFLOPS
Tensor Cores	-	-	-	-	640	640
Tensor Core Performance	-	-	-	-	112 TFLOPS	120 TFLOPS
GPU Memory	24GB	24GB	16GB HBM2	16GB HBM2	16GB HBM2	16GB HBM2
Max. Power Consumption	300W	250W	250W	300W	250W	300W

Due to their single and double precision performance, all but the P40 card are suitable for multipurpose workloads such as HPC, DL, and rendering. The P40 lacks double precision performance and is therefore not suited for HPC workloads. The native Tensor Core data type first introduced in the Volta micro-architecture is of particular interest for DL applications as it is optimized for DL specific calculations and is expected to significantly accelerate training and inference over the previous generation. The NVLink-enabled SXM.2 version of the V100 lists a 7% single and double precision compute performance increase over the PCIe version without NVLink, and at a 20% increase in power consumption.

GPU Compute Platforms

The table below lists the specifications of each platform—cloud and on-premise—used in the study.

Table 2: Cloud and On-Premise Platform Specs

	AWS P2	AWS P3	DL-E280 Pascal	DL-E280 Pascal	DL-E280 Volta
CPU	2x E5-2686v4	2x E5-2686v4	2x E5-2699v4	2x E5-2699v4	2x E5-2699v4
Cores per CPU	18 Cores	18 Cores	22 Cores	22 Cores	22 Cores
Clock	2.3GHz-3.0GHz	2.3GHz-3.0GHz	2.2GHz-3.6GHz	2.2GHz-3.6GHz	2.2GHz-3.6GHz
Memory	512GB DDR4	512GB DDR4	512GB DDR4	512GB DDR4	512GB DDR4
GPU	4x K80	8x V100 SXM.2	8x P100 PCIe	8x P40 PCIe	8x V100 PCIe
OS	Ubuntu 16.04	Ubuntu 16.04	Ubuntu 16.04	Ubuntu 16.04	Ubuntu 16.04
CUDA	CUDA 8.0	CUDA 8.0	CUDA 8.	CUDA 8.	CUDA 8.
CUDNN	v7	v7	v7	v7	v7
DIGITS	6.0	6.0	6.0	6.0	6.0

For the on-premise server platform, we chose the AMAX DL-E280 platform due to its advanced thermal design, the compact 2U form factor and its popularity and adoption by Deep Learning organizations. Memory and CPU were chosen to closely match AWS instance options.

Note: We would like to mention that the nomenclature used for AWS P2 and P3 instances may lead to some confusion about the number of GPUs available in an instance. For example, a p2.16xlarge instance has 8 physical GPUs attached and a p2.8xlarge instance has 4 physical GPUs attached. Similarly, a p3.8xlarge instance provides 4 physical GPUs. The reason for the labeling might be that a single K80 GPU card internally hosts 2 individual GPUs, even though within one card, diagnostics and orchestration tools like nvidia-smi recognize one physical card as 2 logical cards. For clarification, Table 3 lists number of physical and logical GPUs for the available P2 and P3 instances.

Table 3: AWS Physical and Logical GPU Count

Instance	GPU	Chipset	# of GPUs (logical)	# of GPUs (physical)
p2.xlarge	K80	2x GK210B	1	0.5
p2.8xlarge	K80	2x GK210B	8	4
p2.16xlarge	K80	2x GK210B	16	8
p3.2xlarge	V100	GV100	1	1
p3.8xlarge	V100	GV100	4	4
p3.16xlarge	V100	GV100	8	8

In both benchmark test and cost analysis, we compare AWS instances and on-premise systems with identical **physical** GPU counts.

Benchmark Tests

For the purpose of the tests, benchmarks were performed using NVIDIA's Digits 6.0 software running on the AMAX DL-E280 server with the following integrated NVIDIA Tesla GPU cards: P100 (PCIe), P40, and V100 (PCIe). As a representation of GPU cloud offerings, we tested Amazon's EC2 P2 instances (featuring Tesla K80 cards) and P3 instances (V100 SXM.2 with NVLink). The GoogLeNet image recognition training session times for 2,500,000 and 500,000 images were employed as a measure for the Deep Learning performance.

For an easier comparison, the lower performing system in each test setting was normalized to 100%. E.g., a system with 200% (=2X) performance would complete the task in half the time. For all practical purposes, results were identical for small and large data sets.

Benchmark tests were set up as follows:

Dataset Size

Dataset A: 2,500,000 images
Dataset B: 500,000 images

Dataset Settings

Dataset: Classification
Image Type: Color
Image size: 256x256
Resize transformation: Squash
% for validation: 25
% for testing: 0
DB backend: LMDB
Image Encoding: PNG

Model Settings

Model: Classification

Training epochs: 30

Snapshot interval (in epochs):1

Validation interval (in epochs):1

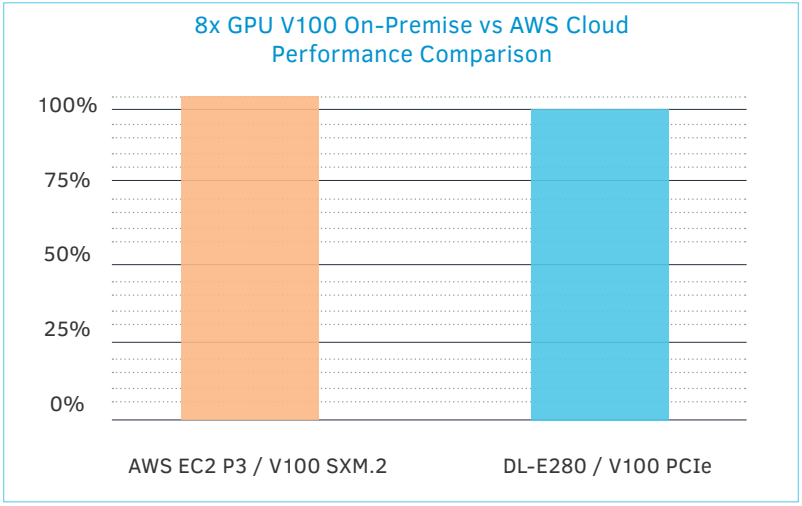
Base Learning Rate: 0.01

batch_size: 128 for one card, 256 for 2 cards, 512 for 4 cards, 1024 for 8 cards.

Results

Performance Comparison V100 On-Premise vs. Cloud (P3 Instance)

The results of the benchmarking revealed that the EC2 P3 8 GPU (V100 SXM.2) option offered by Amazon has a slight advantage over an on-premise 8 GPU system with 8x V100 PCIe GPUs. V100 SXM.2 are 2.8% faster than V100 PCIe models used in the on-premise server. The 2.8% increase in performance may be due to the higher TFLOP performance of the SXM.2 form factor over the PCIe form factor.

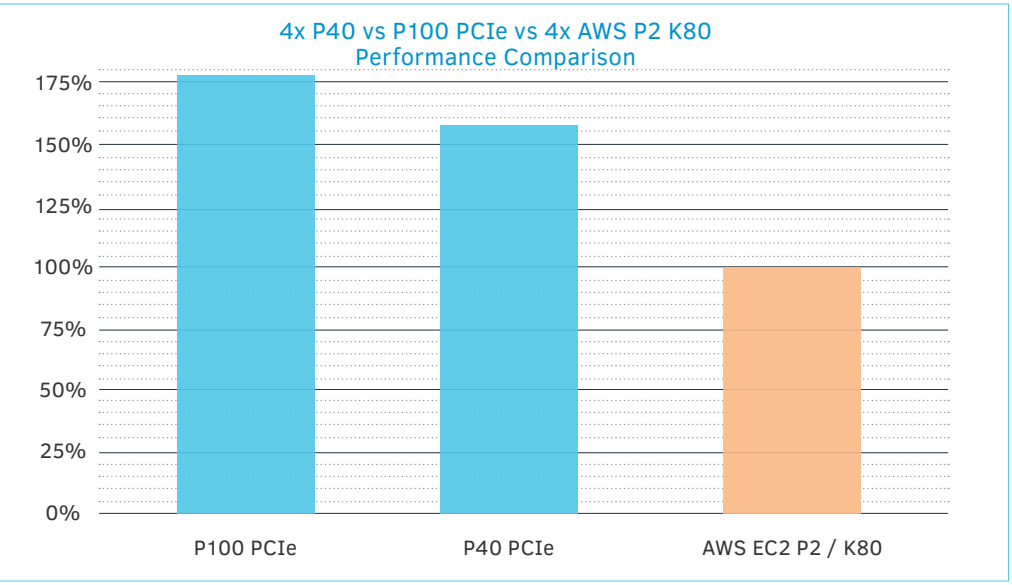


Graph 1: V100 DL-E280 vs. AWS P3

Performance Comparison P40/P100 On-Premise vs. Cloud (P2 Instance)

Both the P40 and P100 on-premise servers completed the DL training sessions significantly faster than the K80 GPUs available in EC2 P2 instances. While the P40 card is listed as having a better single precision performance on paper, for the HW configuration listed in Table 1, the P100 system outperformed the P40.

The comparison shows the on-premise DL-E280 server populated with 4x P40 Pascal GPU cards offers about twice the performance of the AWS EC2 P2 instance with 4x physical K80 (8x logical) Kepler GPU cards. Benchmark results of the on-premise P40 and P100 solution and the AWS P2 instance are summarized in Graph 2.

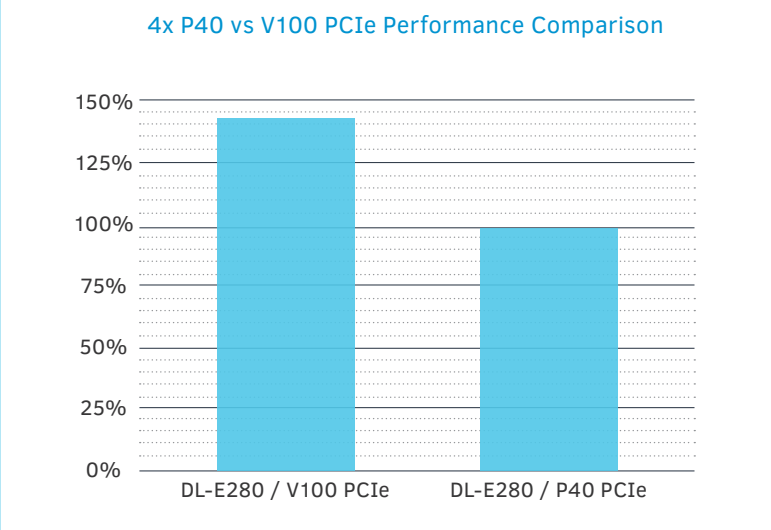


Graph 2: 4x P40 vs P100 PCIe vs AWS P2 K80

Performance Comparison On-Premise P40 vs. V100

Comparing V100 with P40 performance is of particular interest because it provides insight into how effective the design improvements made in the Volta micro-architecture over Pascal accelerate real-world DL workloads. Comparing the spec sheet in Table 1, a 30% theoretical performance increase can be derived based on the single precision TFLOP performance.

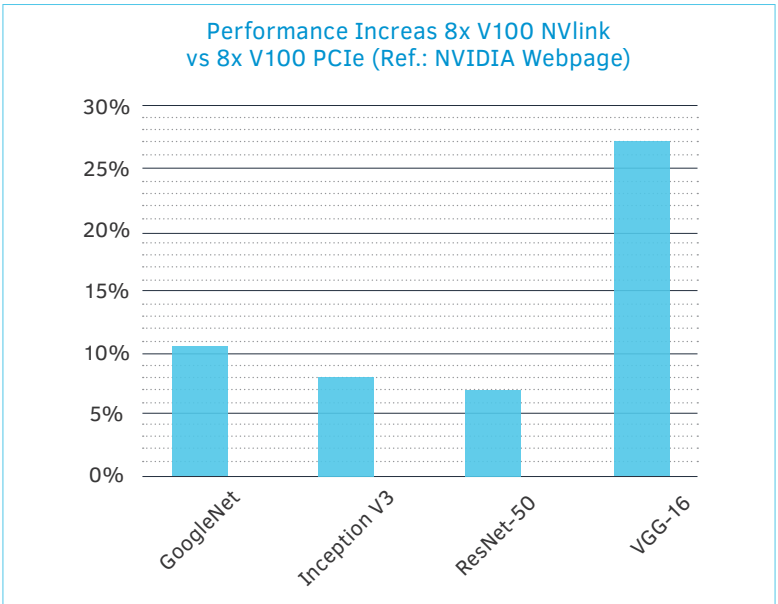
Graph 3 shows a significant performance increase of the V100 system over the P40 system. For a 4 GPU configuration, the performance increased by 45%.



Graph 3: V100 DL-E280 vs. P40 On-Premise

NVIDIA V100 NVLink vs. PCIe Performance Comparison

To provide a more complete picture, we added performance data published by NVIDIA for GoogLeNet, Inception V3, ResNet-50 and VGG-16 [1]. The test platform was a dual Intel® Xeon® Processor E5-2698 v4 machine. OS was Ubuntu 14.04.5, driver versions were CUDA 9.0.176, NCCL 2.0.5, and CuDNN 7.0.2.43 – driver 384.66. Graph 4 displays the performance delta between NVLink and non-NVLink systems. NVIDIA reports a performance increase between 7% and 10% with the exception of VGG-16 with a performance increase of 27%.



Graph 4: Performance Comparison for V100 PCIe vs V100 SXM.2

As mentioned earlier, system performance can largely vary with DL framework, size of the data set, implementation, and other factors. Comparing the performance delta of this study with NVIDIA, our results are lower but in line with NVIDIA's results.

In the following section, we will discuss a cost comparison between AWS instances and on-premise options.

Costs Analysis

The cost analysis for an on-premise deployment presented in this white paper is based on CAPEX costs and an OPEX cost estimate. The cost overhead was estimated at 50% of the hardware costs of the GPU compute server which includes power, cooling, network, and infrastructure amortization as well as maintenance costs. OPEX costs may vary based on location, energy costs, PUE of the data center, etc.

The cost calculation for AWS GPU cloud computing only includes the costs for P2 and P3 instances. Any additional costs for bandwidth, storage, data egress, etc. that may occur and may vary depending on the particular use case have not been included. It is assumed that the instances are used continuously. AWS provides 2 payment options, one for on-demand and one for a 3-year contract with partial upfront RI. Table 4 lists the per year cost of P2 and P3 instances.

Table 4: Cost analysis of AWS GPU Instances

Instance	GPU	Form Factor	#GPUs (physical)	#GPUs (logical)	NVLink	3-Year Contract (per year)	On-demand (per year)
p2.xlarge	K80	PCIe	0.5	1	-	$3,500	$7,884
p2.8xlarge	K80	PCIe	4	8	-	$27,997	$63,072
p2.16xlarge	K80	PCIe	8	16	-	$55,994	$126,144
p3.2xlarge	V100	SXM.2	1	1	yes	$13,403	$26,806
p3.8xlarge	V100	SXM.2	4	4	yes	$53,611	$107,222
p3.16xlarge	V100	SXM.2	8	8	yes	$107,222	$214,444

For comparison, Table 5 shows the estimated costs for on-premise deployment assuming capital is amortized over a three year period.

Table 5: AMAX On-Premise Offerings

System	GPU	Form Factor	# GPUs	NVLink	HW Cost* (per year)	HW + Overhead** (per year)
DL-280 Pascal	P40/P100	PCIe	8	no	$22,333	$33,499
DL-280 Pascal	P40/P100	PCIe	4	no	$16,599	$24,898
DL-280 Volta	V100	PCIe	8	no	$29,866	$44,799
DL-280 Volta	V100	PCIe	4	no	$20,399	$30,598

**Price may vary based on configuration.*
***We assume a 50% overhead for power, cooling, infrastructure and other operational costs.*

Over a three year period, the total cost of ownership (TCO) of the DL-E280 on-premise platform with 8x Volta V100 PCIe GPUs—including operational and infrastructure costs—is 45% less than the comparable AWS P3 (p3.16xlarge) instance under a signed 3 year contract (for a savings of approximately $36,437/year), and 79% less than an on-demand P3 instance (for a savings of approximately $169,645/year). The DL-E280 featuring 8x Tesla P40 PCIe GPUs is faster than the comparable AWS P2 (p2.16xlarge) instance, and has a TCO of 40% less than the AWS instance under contract (saving approximately $22,495/year), and 73% less compared to on demand (saving approximately $92,645/year).

A DL-E280 featuring 4x Volta V100 GPUs presents a TCO of 71% less than the comparable AWS P3 (p3.8xlarge) on demand instance (saving approximately $76,624/year), while a 4x P100 on-premise solution offers 61% less TCO than the AWS P2 (p2.8xlarge) on demand instance (saving approximately $38,174/year). A comparative summary of the TCO savings is given in Table 6.

Table 6: On-Premise vs AWS Total Cost of Ownership (TCO) Comparison

AWS Instance	AWS Cost per Year (3-Year Contract)	AWS Cost per Year (On-Demand)	AMAX On-Premise Solution	Total Cost per Year (On-Premise)	Cost Savings On-Premise vs. 3-Year Contract	Cost Savings On-Premise / On-Demand
p3.16xlarge	$107,222	$214,444	DL-280 8X V100	$44,799	- 45%	- 79%
p3.8xlarge	$53,611	$107,222	DL-280 4X V100	$30,598	- 25%	- 71%
p2.16xlarge	$63,072	$126,144	DL-280 8X P100	$33,499	- 40%	- 73%
p2.8xlarge	$31,536	$63,072	DL-280 4X P100	$24,898	- 11%	- 61%

We found that all on-premise solutions provide substantial cost savings if the total costs are distributed over three years. Therefore, it is recommended to consider an on-site deployment if a company projects the length of the DL initiative and commitment to be one year or longer. Table 7 presents the breakeven point for purchasing on-premise hardware versus continuous use of AWS on demand. For example, the cross-over for the 8 GPU V100 configuration is seven and a half months to break even with the comparable AWS cloud instance, after which it becomes the more cost-efficient solution.

Table 7: On-Premise vs AWS Total Cost of Ownership (TCO) Comparison

AWS Instance	AWS On-Demand Cost per Year	AMAX On-Premise Solution	3 Year TCO On-Premise	Breakeven (Months)
p3.16xlarge	$214,444	DL-280 8X V100	$134,397	7.5
p3.8xlarge	$107,222	DL-280 4X V100	$91,794	10.3
p2.16xlarge	$126,144	DL-280 8X P100	$100,497	9.6
p2.8xlarge	$63,072	DL-280 4X P100	$74,694	14.2

Conclusion

Based on our testing of an array of both on-premise GPU configurations and corresponding GPU-integrated cloud instances, we have determined that on-premise GPU servers have performance and cost advantages worthy of consideration for companies engaged in Deep Learning/Machine Learning initiatives.

Due to the high cost of cloud GPU instances and little to no advantage in performance, on-premise GPU servers are significantly more cost efficient if the use case extends beyond a year. For example, the AWS P3 V100 SXM.2 options represent the most advanced cloud option. In our benchmarking, they performed 3% better than the on-premise DL-E280 V100 PCIe solutions, yet cost over 40% more over three years of service. The AWS P2 K80-based instances performed approximately half as well as on-premise P40 servers, but cost more over a span of three years. While this calculates only the capital expense, other factors to consider—negative impact or delay to development cycles and time to market, as well as personnel expense—should be included in a company's decision of if and when to utilize cloud services for Deep Learning and AI development.

In conclusion, while using GPU-integrated cloud instances are a good option for companies in initial explorations of Deep Learning or with limited usage, any company with AI initiatives beyond an exploratory phase with longer-term or larger-scale use cases should consider an on-premise deployment for better cost efficiency and performance.

References:

[1] http://images.nvidia.com/content/pdf/v100-application-performance-guide.pdf