Announcements

NVIDIA Tesla Accelerators on IBM Cloud Demonstrate Advanced Performance for Training Deep Learning Models

- New performance benchmarks for NVIDIA Tesla P100 GPU accelerators on IBM Cloud can reduce deep learning training time by up to 65 percent compared to NVIDIA Tesla K80 GPU accelerators on IBM Cloud.
- New benchmarks reinforce that IBM Cloud is cognitive at the core and tailored for running AI and cognitive workloads.

ARMONK, N.Y., May 8, 2017 /PRNewswire/ -- IBM (NYSE: IBM) announced that benchmarks conducted by IBM engineers found that the NVIDIA® Tesla® P100 GPU accelerators on the IBM Cloud can provide up to 2.8 times more performance than the previous generation NVIDIA® Tesla® K80 for certain test cases. A second benchmark conducted by Rescale also found significant performance gains for the NVIDIA Tesla P100 GPU on IBM Cloud. This reduces the corresponding training time required for deep learning models and helps enable organizations to quickly create advanced artificial intelligence (AI) applications on the cloud.

Demand for AI applications is growing rapidly. According to Research and Markets, the AI market is expected to be worth 16.06 billion dollars by 2022[1]. Deep learning techniques are a key driver behind the increased demand for and sophistication of AI applications. However, training a deep learning model to do a specific task is a compute-heavy process that can be time and cost-intensive.

The availability of the NVIDIA Tesla P100 GPUs on the IBM Cloud is making it faster and more cost-effective to leverage deep learning techniques to train AI systems. According to a recent performance benchmark conducted by IBM, certain deep learning workloads running on the IBM Cloud with the NVIDIA Tesla P100 GPUs outperform the previous-generation NVIDIA Tesla K80 GPUs by a factor of 2.8 times. The combination of NVIDIA Tesla P100 GPUs on the IBM Cloud reduced the corresponding training time for a deep learning image classification model by 65 percent from the NVIDIA Tesla K80 GPUs.

To conduct the benchmark, IBM engineers trained a deep learning model for image classification using two NVIDIA Tesla P100 GPU PCIe cards (a total of two P100 GPU cores) on Bluemix bare metal servers and compared the results to the same deep learning model running two Tesla K80 GPU PCIe cards (a total of four K80 GPU cores) on Bluemix bare metal servers. The engineers conducted the ILSVRC image classification challenge using the VGG-16 deep neural network on the Caffe framework. The goal of the ILSVRC is to teach a deep neural network model to correctly classify images; models are trained on approximately 1.2 million images with an additional 50,000 images for validation and 100,000 images for

testing.

The benchmark also found that the NVIDIA Tesla P100 GPUs on IBM Cloud can deliver greater performance for the cost. According to the benchmark, the NVIDIA Tesla P100 GPU on IBM Cloud can process more than 116,000 images per US dollar spent – 2.5 times higher than the previous generation NVIDIA Tesla K80 GPUs on the cloud for the same test case.

"Innovation in AI is happening at a breakneck speed thanks to advances in cloud computing," said John Considine, general manager, cloud infrastructure services, IBM. "As the first major cloud provider to offer the NVIDIA Tesla P100 GPU, IBM Cloud is providing enterprises with accelerated performance so they can quickly and more cost-effectively create sophisticated AI and cognitive experiences for their end-users."

A second performance benchmark conducted by Rescale using its ScaleX™ platform also demonstrated deep learning training time reductions. Rescale is a global leader for high-performance computing simulations and deep learning in the cloud. ScaleX features capabilities for deep learning SaaS, including interactive notebooks, enabling data analysis in-browser and turnkey delivery of deep learning libraries for training on large datasets. When training the InceptionV3 deep neural network on the ILSVRC dataset using TensorFlow 1.0, Rescale stated that they found that the deep learning model could be trained in approximately half the time when using NVIDIA P100 GPUs on the IBM Cloud over the NVIDIA K80 GPUs.

"Rescale is excited to be working with IBM to push the boundaries of Deep Learning and AI research." said Joris Poort, CEO of Rescale. "Our ScaleX platform provides a highly accessible and easy to use environment for hardware benchmarking, allowing testing as soon as new hardware is deployed."

To learn more about GPU computing on the IBM Cloud, please visit: https://www.ibm.com/cloud-computing/bluemix/gpu-computing

To learn about Rescale's cloud HPC platform for IBM, please visit http://www.rescale.com/ibm/

About IBM Cloud
For more information, visit: http://www.ibm.com/cloud-computing.

[1] **Research and Markets:** Artificial Intelligence Market by Technology (Deep Learning, Robotics, Digital Personal Assistant, Querying Method, Natural Language Processing, Context Aware Processing), Offering, End-User Industry, and Geography - Global Forecast to 2022, November 2016

The IBM benchmark results were achieved using the following criteria:

- *Two NVIDIA Tesla P100 GPU PCIe cards (a total of two P100 GPU cores) on Bluemix bare metal servers (with Dual Xeon E5-2690v4 processors) running the VGG-16 deep neural network on the*

Caffe-1.0.0-rc5 framework, CUDA version 8.0.61, NCCL version 1.3.4, cuDNN version 6.0.20, and CUDA driver version 375.51. The ILSVRC image classification challenge dataset was used. The training batch size was maximized to 102 images per GPU core to exploit the larger available memory capacity on the NVIDIA Tesla P100 GPU cards.

- *Two NVIDIA Tesla K80 GPU PCIe cards (a total of four K80 GPU cores) on Bluemix bare metal servers (with Dual Xeon E5-2690v4 processors) running the VGG-16 deep neural network on the Caffe-1.0.0-rc5 framework, CUDA version 8.0.61, NCCL version 1.3.4, cuDNN version 6.0.20, and CUDA driver version 375.51 . The ILSVRC image classification challenge dataset was used. The training batch size was maximized to 67 images per GPU core to use all available memory capacity on the NVIDIA Tesla K80 GPU cards.*

The results of the Rescale performance benchmark were achieved using the following criteria:

- *Two NVIDIA Tesla P100 GPU PCIe cards (a total of two P100 GPU cores) on Bluemix bare metal servers running the InceptionV3 deep neural network on the ILSVRC dataset using TensorFlow 1.0.*
- *Two NVIDIA Tesla K80 GPU PCIe cards (a total of four K80 GPU cores) on Bluemix bare metal servers running the InceptionV3 deep neural network on the ILSVRC dataset using TensorFlow 1.0.*

Performance may vary based on any variables in these configurations.

Contact:
Sarah Murphy
IBM Media Relations
srmurphy@us.ibm.com
336-337-7584

SOURCE IBM

Web Site: http://www.ibm.com