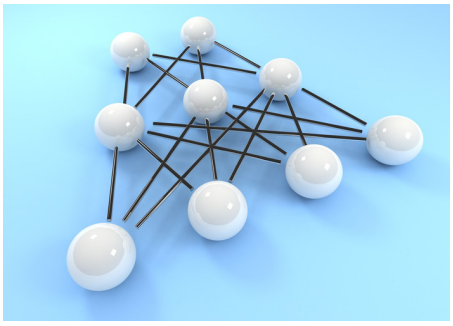
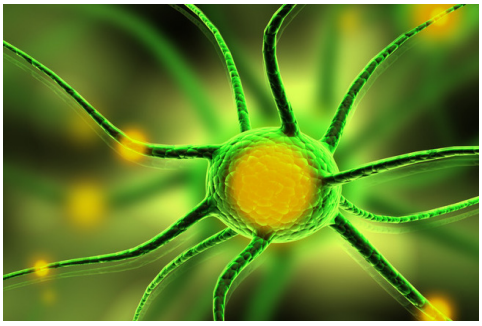
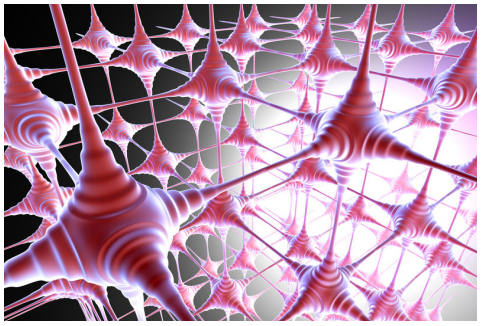
Deep Learning Multi-GPU Server Case Study

Accelerated Computing Using Multi-GPU Servers with NVIDIA® Tesla® GPU Accelerators

The Customer

The NYU Center for Data Science is a focal point for New York University's university-wide initiative in data science and statistics. The Center was established to help advance NYU's goal of creating the country's leading data science training and research facilities, and arming researchers and professionals with tools to harness the power of big data. The Center's faculty members and scientists are established experts in the field of data science. Their interests are primarily in mathematical statistics, computational statistics and machine learning, optimization and large-scale computation, system design for large-scale data science, and several application areas, such as artificial intelligence, computational biology, computational economics, or quantitative methods in social science. The center was founded by deep learning pioneer Yann LeCun.

The Challenge

Next generation advances in deep learning will rely on advanced and significantly more sophisticated algorithms as well as the most advanced computing technologies. The challenge is to enable the computers to be able to achieve and even surpass human capabilities.

In order to accomplish this, advanced GPU compute and server technology is needed. In the early stages of deep learning, CPU compute power was used but never scaled the way it was necessary in order to address these challenges. Recently, GPUs have been identified as the preferred technology for deep learning, reducing the time it takes to train deep neural networks by days or even months, and in some cases years.

Until recently, many organizations and researchers relied on systems that had one or maybe two GPUs. This resulted in limits being placed on the size of models and edges that could be developed and discovered. With the right hardware and software combination though, one could effectively increase the size of the models that can be trained and tested resulting in significant advancement of deep learning accelerated computing models.

The Solution

NYU recently installed a new deep learning computing system — called "ScaLeNet." It consists of eight Cirrascale RM4400 Series servers with 32 top-of-the-line NVIDIA Tesla K80 dual-GPU accelerators. Each server contains eight GPUs, a significant increase in GPU compute resources from what was previously available.

(continued on reverse side)

These servers enable NYU researchers take on bigger challenges and create deep learning models that let computers do human-like perceptual tasks for research projects and educational programs at the NYU Center for Data Science by a large community of faculty members, research scientists, and graduate students.

Most of the hardware configurations available today only provide maximum performance between specific pairs of GPUs; and since GPUs are paired up, jobs requiring communication between arbitrary GPUs experience a performance impact. Additionally, there can be a significant performance impacts with trying to scale more than four GPUs on multi-socket systems. These have been persistent problems for customers who are pushing the limits of GPUs with large, complex data-sets and calculations, or where data must be streamed between GPUs. Cirrascale has been able to overcome these issues, and achieve near linear performance scaling with its design. Cirrascale servers are unique and highly desired for deep learning due to several key factors including their ability to peer large numbers of multiple GPUs on a single root complex. This is why these servers were selected by NVIDIA for the NYU installation.

In a recent NVIDIA blog post regarding the NYU installation, Yann LeCun said, "Multi-GPU machines are a necessary tool for future progress in AI and deep learning. Potential applications include self-driving cars, medical image analysis systems, real-time speech-to-speech translation, and systems that can truly understand natural language and hold dialogs with people."

Discover More

Discover more about NYU, NVIDIA, and Cirrascale through one of NVIDIA's blog posts located at: *http://blogs.nvidia.com/blog/2015/04/30/nyu-to-advance-deep-learning-research-with-multi-gpu-cluster/*.

Accelerate Your Applications Even Further

Cirrascale now sells the GB5600 Series blade server that contains an amazing 16 GPUs on a single root complex making it the only advanced Multi-GPU offering of this magnitude. It's all made possible with the power of the NVIDIA Tesla K80 Dual-GPU Accelerators. To learn more about this incredible server, visit http://www.cirrascale.com.

CM076 - REV A - 09/2015

Cirrascale 12140 Community Road, Poway, CA 92064 USA **Phone** 858-874-3800 or 888-942-3800 **Web** www.cirrascale.com