Technology | DOI:10.1145/2967979 Samuel Greengard

GPUs Reshape Computing

Graphical processing units have emerged as a major powerhouse in the computing world, unleashing huge advancements in deep learning and AI.

As researchers continue to push the boundaries of neural networks and deep learning—particularly in speech recognition and natural language processing, image and pattern recognition, text and data analytics, and other complex areas—they are constantly on the lookout for new and better ways to extend and expand computing capabilities. For decades, the gold standard has been high-performance computing (HPC) clusters, which toss huge amounts of processing power at problems—albeit at a prohibitively high cost. This approach has helped fuel advances across a wide swath of fields, including weather forecasting, financial services, and energy exploration.

However, in 2012, a new method emerged. Although researchers at the University of Illinois had previously studied the possibility of using graphics processing units (GPUs) in desktop supercomputers to speed processing of tasks such as image reconstruction, a group of computer scientists and engineers at the University of Toronto demonstrated a way to significantly advance computer vision using deep neural nets running on GPUs. By plugging in GPUs, previously used primarily for graphics, it was suddenly possible to achieve huge performance gains on computing neural networks, and these gains were reflected in superior results in computer vision.

The advance proved revolutionary.

"In only a few short years, GPUs have emerged at the center of deep learning," says Kurt Keutzer, a professor in the Electrical Engineering & Computer Science Department at the University of California, Berkeley. "The use of GPUs is now moving into the mainstream, and by applying dozens to hundreds of processors to a single application, they are on a trajectory to radically change computing."

NVidia's Titan X graphics card, featuring the company's Pascal-powered graphics processing unit driven by 3,584 CUDA cores running at 1.5GHz.

Adds Wen-Mei W. Hwu, Walter J. Sanders III–Advanced Micro Devices Endowed Chair in Electrical and Computer Engineering at the University of Illinois at Urbana-Champaign, "GPUs are remarkable throughput computing devices. If you only have one task, it doesn't execute particularly fast on GPUs. However, if you have a large number of independent tasks, it works really well."

A Deeper Vision

GPU architectures have their roots in basic graphical rendering operations like shading. In 1999, Nvidia introduced the GeForce 256, which was dubbed the world's first GPU. Simply put, the specialized circuits—which may be built into a video card or on a motherboard—manipulate and optimize computer memory in order to accelerate rendering on displays. Today GPUs are used in a wide array of devices, including personal computers, tablets, mobile phones, workstations,

IMAGE COURTESY OF GEFORCE.COM/NVIDIA

electronic signage, gaming consoles, and embedded systems.

However, "Many emerging applications in computer vision and deep learning are memory bandwidth-limited," Keutzer explains. "In these applications, the speed of the application is often ultimately dictated by the time it takes to draw data from memory and stream it through the processor."

A big advantage of a GPU implementation, and something that is frequently overlooked, is its superior processor-to-memory bandwidth. As a result, "In bandwidth-limited applications, the relative processor-to-memory bandwidth advantage transfers directly to superior application performance," Keutzer points out. The key is that GPUs provide greater floating-point operations per second (FLOPs) using fewer watts of electricity, and they actually extend this energy advantage by supporting 16-bit floating point numbers, which are more power- and energy-efficient than single-precision (32-bit) or double-precision (64-bit) floating point numbers.

The manycore approach to GPUs relies on a larger number, such as 32 to 64, of simpler processor cores implemented in larger numbers. By contrast, multicore approaches use smaller numbers of conventional microprocessors, typically 2 to 4 to 8. The upshot? "GPUs deliver superior performance and better architectural support for deep neural networks. The performance advantages of GPUs on deep neural nets are transferred onto an increasingly broad variety of applications," Keutzer says.

Today, a typical cluster is comprised of 8 to 16 GPUs, though researchers such as Keutzer are now pushing the numbers into the hundreds to simultaneously train deep neural nets on extraordinarily large datasets that would otherwise require weeks of training time. The training consists of running massive amounts of data through the system in order to get it to a state where it can solve problems. At that point, it may run on a CPU or hybrid processor. "This is not an academic exercise." Keutzer notes. "We need this kind of speed in training neural nets to support emerging applications such as self-driving cars."

"The use of GPUs is now moving into the mainstream, and by applying (many) processors to a single application, they are on a trajectory to radically change computing."

GPU technology is advancing far faster than that of conventional CPUs. The scalability of GPUs, along with their sheer floating point horsepower and lower energy consumption, is turbocharging deep learning and machine learning tasks, says Bryan Catanzaro, senior researcher at Chinese-based Internet services, search, and data firm Baidu. "Deep learning is not new. GPUs are not new. But the field is taking off because of huge advances in computational capabilities and the availability of richer datasets."

Much of the thrust has come from Nvidia, which has introduced increasingly sophisticated GPUs, including the new Pascal architecture that is designed to tackle specific tasks, such as training and inference. Its latest GPU system, the Tesla P100 chip, packs 15 billion transistors on a piece of silicon, twice as many as previous processors.

Baidu, for example, is pushing into new frontiers in speech recognition. Its "Deep Speech" initiative, which relies on an end-to-end neural net, provides speech recognition accuracy that rivals humans on short audio clips in both English and Chinese. The company is also venturing into the autonomous vehicle space with GPU technology; it has developed a self-driving vehicle that has navigated the streets of Beijing, with maneuvers including changing lanes, passing other vehicles, and stopping and starting.

Meanwhile, researchers at Microsoft Asia have used GPUs and a variant of Deep Neural Nets, called Residual Neural Nets, to achieve superior accuracy on the computer vision problems of object classification and object recognition.

Google, too, is using these techniques to continually improve its image recognition algorithms. Says Ilya Sutskever, a former Google AI researcher who is now research director at Open AI (a non-profit artificial intelligence research company, https://openai.com), "Neural networks are enjoying a renaissance. The core ideas of neural networks and deep learning have been discussed and pondered for many years, but it was the development of the general-purpose GPU that was a key enabler of their success."

One Step Beyond

While GPU technology is pushing into new frontiers in the deep learning space, plenty of computing challenges remain. For one thing, "Programming individual manycore devices such as GPUs for high efficiency is still very difficult, and that difficulty is only compounded when these devices are gathered together in multi-GPU clusters," Keutzer says. Unfortunately, he adds, "Much of the expertise for effectively programming these devices is housed in companies and many details of the techniques that have been developed are not widely shared."

Similarly, the design of Deep Neural Nets is widely described as a "black art," Keutzer says; creating a new Deep Neural Network architecture is as complicated, in many ways, as creating a new microprocessor architecture. To make matters worse, once the Deep Neural Network architecture is created, "there are many knobs, known as hyperparameters, used during training, and accuracy is only achieved when these knobs are set appropriately. All of this adds up to a knowledge gap between those 'in the know' and others.

"Individuals with expertise in either Deep Neural Nets or GPU programming are scarce, and those who know both well are very rare."

Another challenge is understanding how to use GPUs most effectively. For example, Baidu requires 8 to 16 GPUs to train one model, achieving 40% to 50% of peak floating point math throughput on the entire application.

"This means there is very little performance left on the table," Catanzaro says. "There are things that we would like to do to scale to more GPUs, so rather than using 8 or 16 GPUs, we would like to use 128 GPUs, for example." This translates into a need for better interconnects, as well as the ability to move from 32-bit floating point support to the throughput of 16-bit floating point support. Nvidia's next generation GPU, codenamed Pascal, may address some of these issues.

Still another obstacle is better integrating GPUs with CPU/GPUs. Hwu says those two types of processors are not often integrated together, and they usually do not have high bandwidth communication between the two. This translates into a limited number of applications and capabilities that run well on these systems. "You really need to be able to give the GPU a kind of a very big task and with some amount of data and then let the GPU crank on it for a while to make this offloading process worthwhile," Catanzaro explains.

Current Nvidia GPUs are located on separate chips. They are usually connected to the CPU via an I/O bus (PCIe). This is the reason one needs to send large tasks to the GPU. Future systems will integrate GPUs and CPUs in one tightly coupled package that supports higher bandwidth, lower latency, and cache coherent memory sharing across CPUs and GPUs.

Keutzer expects that over time, as CPUs and GPUs become better integrated, better cache coherence and synchronization between the two types of processors will result. In fact, Nvidia and Intel are both focusing on this space. Keutzer notes a new Intel chip dubbed Knight's Landing (KNL) offers unprecedented computing power in a Xeon Phi 72-core supercomputing processor that integrates both CPU and GPU characteristics. This chip also offers 500 gigabyte-per-second processor-to-memory bandwidth that will erode GPU's advantage in this area, he says.

Hwu notes each of the KNL chip's 72 cores can execute "a wide vector instruction (512 bytes). When translated into double precision (8 bytes) and single precision (4 bytes), the vector width is 64 and 128 words; in that sense, it has a similar execution model to that of GPUs."

Keutzer expects as CPUs and GPUs become better integrated, better cache coherence and synchronization between the two types of processors will result.

The programming model for the KNL chip is the traditional x86 model, Hwu says, so programmers "need to write code to either be vectorizable by the Intel C Compiler, or use the Intel AVX vector intrinsic library functions." The programming model for GPUs is based on the kernel programming model, he adds.

Also, X86 cores have cache coherence for all levels of the cache hierarchy, Hwu says, "whereas GPU's first-level caches are not coherent. It does come with a cost of reduced memory bandwidth." However, he says, "For deep learning applications, cache coherence for the first level cache is not very important for most algorithms."

Over the next decade, a big wildcard in all of this will be how development cycles play out, Hwu says. He believes Moore's Law can continue at its present rate for about three more generations. At the same time, he says, it will likely take about three generations for system designers and engineers to move away from mostly discrete CPU and GPU systems to true hybrid designs.

"If Moore's Law stalls out, it could dramatically impact the future of these systems, and the way people use hardware and software for deep learning and other tasks," Hwu points out. "Yet, even if we solve the hardware problem, certain deep learning tasks require huge amounts of labeled data. At some point, we will need a breakthrough in generating labeled data in order to do the necessary training, particularly in areas such as self-driving cars."

Over the next few years, Sutskever says, machine learning will tap GPUs extensively. "As machine learning methods improve, they will extend beyond today's uses and ripple into everything from healthcare and robotics to financial services and user interfaces. These improvements depend on faster GPUs, which greatly empower machine learning research."

Adds Catanzaro: "GPUs are a gateway to the future of computing. Deep learning is exciting because it scales as you add more data. At this point, we have a pretty much insatiable desire for more data and the computing resources to solve complex problems. GPU technology is an important part of pushing the limits of computing." C

Further Reading

Raina, R., Madhavan, A, and Ng. A.Y.
Large-scale Deep Unsupervised Learning using Graphics Processors, *Proceedings of the 26th International Conference on Machine Learning*, Montreal, Canada, 2009. http://www.machinelearning.org/archive/icml2009/papers/218.pdf

Wu, G., Greathouse, J.L., Lyashevsky, A., Jayasena, N., and Chiou, D.
GPGPU Performance and Power Estimation Using Machine Learning. Electrical and Computer Engineering, The University of Texas at Austin, 21st IEEE International Symposium on High Performance Architecture, 2015. http://hgpu.org/?p=13726

Coates, A., Huval, B., Wang, T., Wu, D.J., Ng, A.Y., and Catanzaro, B.
Deep learning with COTS HPC systems. *Proceedings of the 30th International Conference on Machine Learning*, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28. http://cs.stanford.edu/~acoates/papers/CoatesHuvalWangWuNgCatanzaro_icml2013.pdf

Chen, X., Chang, L., Rodrigues, C.I., Lv, J., Wang, Z., and Hwu, W.
Adaptive Cache Management for Energy-Efficient GPU Computing, MICRO-47 *Proceedings of the 47th Annual IEEE/ACM International Symposium on Microarchitecture*, 343-355, IEEE Computer Society, 2014. http://dl.acm.org/citation.cfm?id=2742190

Samuel Greengard is an author and journalist based in West Linn, OR.