NVIDIA | Deloitte.

High Performance Computing in AI:

What GPUs mean for Deep Learning

There is no shortage of excitement over AI. Today, 93% of organizations report that AI is very important to remaining competitive over the next 5 years, [1]according to Deloitte's 2021 State of AI in the Enterprise. While organizations are eagerly embracing this technology, the AI platform and applications can be opaque to stakeholders outside the data science lab.

Unfamiliar terms like neural network, deep learning, and model training can seem esoteric. To be sure, underlying all AI is dense math, but you do not need a mathematical background to understand how accelerated AI computing is developed and why it is so valuable. The fundamentals of AI functionality can be probed by understanding three core elements—High Performance Computing (HPC), the graphics processing unit (GPU), and the field of machine learning. Looking through these lenses explains where AI is today and helps reveal how it can be developed going forward.

"93% of organizations state that AI is very important to remaining competitive over the next 5 years."

The Value of HPC and GPU in AI

Organizations are striving to realize the strategic goals of improving customer engagement, enhancing operational effectiveness, and finding costs savings. Using AI to meet these targets presents computational demands that have traditionally required an HPC cluster. Today, cloud technologies have made HPC capabilities widely available, with the capacity to manage a "virtual HPC" elastically and as needed. This is a powerful AI accelerator in as much as there is a low barrier to entry.

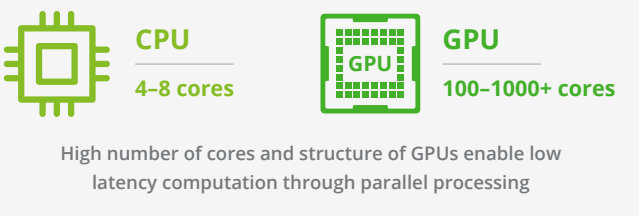
Yet, as data volumes grow and become more complex, there is an increasingly greater need for a higher level of compute for model training. Breaking new ground and discovering new solutions means model training needs to be faster, cheaper, and more nimble. To use machine learning to understand patterns in the data and increase insights beyond those achieved by traditional computing, HPCs require another ingredient.

Most computers rely on the central processing unit (CPU). From PCs to industrial systems, CPUs do the heavy lifting. At its most basic, a CPU uses transistors for calculations, and the number of transistors packed onto a chip determines the processor's computing power. There are powerful CPUs that can accomplish complex functions. Yet, by virtue of how the microprocessor makes calculations (sequentially), there are some applications for which a different kind of chip is much more effective.

The modern GPU reached mass accessibility in 1999 with the release of NVIDIA's GeForce 256, which offered a significant leap forward in video game graphics processing. Since then, owing in large part to NVIDIA's development of the chip and the infrastructure around it, the GPU has evolved to permit massively parallel processing. This is needed for machine learning. ***Why?***

Unlike a CPU, a GPU breaks apart a mathematical problem into smaller problems and solves them simultaneously (in parallel), doing in hours or minutes what would take a CPU days, months, or even years. Consider that the structure of an HIV protein that plays a key role in infection was not known until 1997, nearly two decades after HIV began proliferating around the world. Conversely, just a few months after the COVID-19 pandemic set in, researchers were reporting not just the virus's protein structure but identifying aspects of the protein that could be targeted for therapeutic treatments and vaccines. While a number of advances in genetic sequencing technology contributed to this acceleration in viral biology understanding, a key contributor was the use of parallel processing and AI, seen notably in the [2]partnership between research institutions and NVIDIA to run models on the U.S. Department of Energy Oak Ridge National Laboratory Summit supercomputer.

It is the addition of the GPU to computing infrastructure that makes this kind of game-changing progress possible. An HPC powered by GPUs provides more powerful computing capability with a smaller infrastructure footprint, using less power and cost, and making it highly available via cloud computing infrastructures. This permits the fast, practical use of neural networks and a type of machine learning known as deep learning.



High number of cores and structure of GPUs enable low latency computation through parallel processing

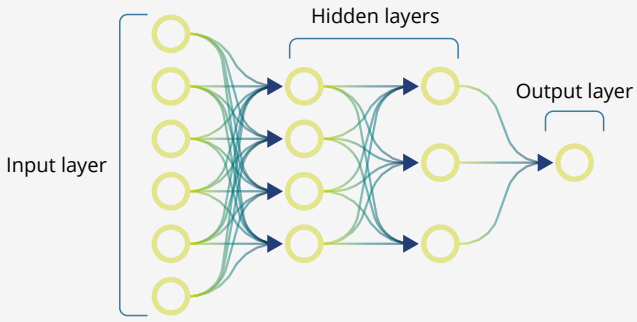
Deep Learning in Simple Terms

AI does not "think." The neural networks used in deep learning are just mathematical models that excel at narrow tasks. By connecting multiple narrow tasks, engineers can design complex systems that can do powerful things, such as drive a car. Tuning a neural network to accurately describe the real world takes large datasets, algorithms, and the processing power of GPUs.

Take as an example an AI model that is intended to recognize school buses in images. It starts with a neural network, which is composed of connected "nodes" that form a series of layers. An input layer takes in training data, such as a massive dataset of unlabeled images (some of which are school buses). Calculations are performed between layers, and the last layer outputs the solution.

Did it recognize all the school buses in the training data? If not, data scientists instruct algorithms to adjust the network such that it can better classify buses. This is known as "training the model." At the end of this deep learning journey, what results is an AI image recognition tool that generalizes beyond the training data to the real world as well.

The AI models being deployed today are trained in the same way but at much greater complexity. There are AI tools that process plain language for customer engagement, recognize thousands of objects for self-driving cars, and find valuable patterns in tumor pathology images to support delivering the right medicines to patients. The increased complexity of the algorithms, combined with ever-increasing data volumes, requires more computational horsepower to train them. GPU-enabled computing makes these computational challenges tractable by computing in parallel, reducing time to compute and overall cost.



Deep Neural Network processes data in several layers to analyze multiple facets of data and their relation to other characteristics to improve output accuracy

Looking Over the Horizon

A GPU microprocessor on its own is not enough for AI. Around the GPU is critical hardware and software designed to permit AI training and management. Collectively, an accelerated AI platform offers the computational power to push the envelope on what is possible with deep learning.

The bleeding edge of AI development requires some of the most powerful GPU-enabled computers available. It also demands data, expertise in data science, and a deep understanding of the industry landscape. The combination of these essentials might be currently out of reach for many organizations, but that should not stand in the way of innovation.

To help push accelerated AI forward, we created the Deloitte Center for AI Computing, which is powered by NVIDIA's DGX™ A100 systems (i.e., a GPU-enabled supercomputer). With it, we can innovate with clients to prove out new use cases, create new frontiers in the marketplace, and capture growth by selling new products and services. With an accelerated AI platform and the expertise and knowledge needed to use it, the potential in AI truly is limited only by the imagination.

This is the [3]Age of With™, where humans working with AI develop world-changing achievements that make business and life better. The math is complicated, and the jargon can sound exotic, but the value of AI is evident. And with the power of GPU-enabled computing, it is more within reach than ever.

Get in touch

Christine Ahn
Principal
Deloitte Consulting LLP
chrisahn@deloitte.com

Anthony Abbattista
Principal
Deloitte Consulting LLP
aabbattista@deloitte.com

Tanuj Agarwal
Senior Manager
Deloitte Consulting LLP
tanuagarwal@deloitte.com

1–https://www2.deloitte.com/us/en/insights/focus/cognitive-technologies/enterprise-artificial-intelligence-4th-edition.html?id=us:2el:3dc:4diUS144384:5awa::MMDDYY:&pkid=1008396

2–https://phys.org/news/2020-11-collaborative-ai-effort-unraveling-sars-cov-.html

3–https://www2.deloitte.com/global/en/pages/strategy-operations/solutions/welcome-to-the-age-of-with.html