NVIDIA | Deloitte.

The Secret to rapid and insightful AI—GPU-accelerated computing

Thanks to AI, breakthrough solutions to long-standing challenges are proliferating, and entirely new products, services, and business models are emerging.

In retail, AI is helping companies reduce theft and personalize the customer experience with autonomous real-time recommendations. Financial services firms are using AI to do things like spot suspicious activity in transactions and simplify risk assessments for insurance underwriting. And in manufacturing, companies are leveraging AI to reveal deeper visibility into the supply chain, manage defects during production, and conduct product design and testing in metaverses.

The potential use cases for AI are near infinite. Accessing this value requires a vision for how AI can advance the enterprise, as well as access to an accelerated AI platform that allows data scientists to innovate, iterate, and deploy models at a faster clip. Taking a deeper dive into accelerated computing reveals how the pieces fit together and what that means for AI endeavors.

New requirements for a new era

The traditional computing infrastructure used for standard enterprise applications is just not enough for large-scale AI. Indeed, AI is not an enterprise application or platform. It requires a distinct suite of talent, resources, and technologies. Consider some of the component characteristics of how AI models are created today.

Scale of the data

AI model's effectiveness is measured in part by how well it makes generalizations to deliver accurate outputs. AI that is highly accurate on training data but brittle and error-prone in the real world is not terribly valuable. Machine learning developers often attempt to improve model generalization performance by increasing the size of the training data. It has been proven that model accuracy improves with more training cycles. Thus, effective AI requires enormous volumes of data.

Nature of the data

The types of data fueling AI training are diverse, complex, and often unstructured. Datasets are composed of texts, images, video, and audio, and their volume is ever increasing. The amount of information within collected data (i.e., images) and the pace at which it is generated makes data processing extremely complex. The algorithms must make sense of this data in an unsupervised way, the computational demands for which are predictably significant.

Deep neural networks

Many of the breakthrough capabilities in AI today stem from a subset of machine learning called Deep Learning. In this, Deep Neural Networks are made up of layers of connected nodes. Training data is consumed, calculations are performed between layers, and algorithms tune the network such that it mathematically represents the real world. When a Deep Neural Network is supplied with large datasets, data scientists can create more complex pattern-based models that deliver extremely accurate AI.

The data, the complexity of the models, and the computational demands for performing calculations all lead to a sobering reality for AI. The standard CPU-powered computing infrastructure that permits most technologies today will not be suited to enabling large-scale AI. To yield any meaningful impact on business functions, training times need to be collapsed and model processing and iteration needs to happen quicker than is possible with traditional computing.

Moreover, scaling models can become cost-prohibitive on CPU-based infrastructure. To power dozens or even hundreds of AI solutions using traditional computing, an enterprise would need to continuously invest in CPUs. At some point, the business will encounter a diminishing return with CPU-only computing.

To move faster and seize cost efficiency, a higher level of compute is required. Enter NVIDIA Accelerated computing: an end-to-end software and hardware platform powered by the Graphics Processing Units (GPU).

What accelerated computing means for AI

Innovation is never a straight line and creating something new necessarily requires experimentation and iterative improvement. In AI, the key is accuracy and continuous improvement. Traditional computing infrastructure can inhibit or slows down this approach for complex AI models. Integrating accelerated computing has the potential to unleash efficiency and speed in AI experimentation, model training and AI inference.

What is a GPU? It is a type of processor designed for parallel computing. While a CPU has four to eight cores that conduct mostly linear calculations, a GPU is composed of hundreds or even thousands of cores that conduct numerous calculations in parallel. It does this by taking a large task, breaking it into smaller tasks, and solving them all simultaneously. AI models present substantial parallel computing tasks, which accelerated computing can solve orders of magnitude faster than CPUs.

This capability speeds up the process for training and updating AI models, which is necessary for rapid development and managing overall costs. It also dramatically accelerates AI inference, where AI models are deployed in applications to extract insights from new data. Accelerated computing AI models can reach greater accuracy, deliver more timely insights, and allow an enterprise to

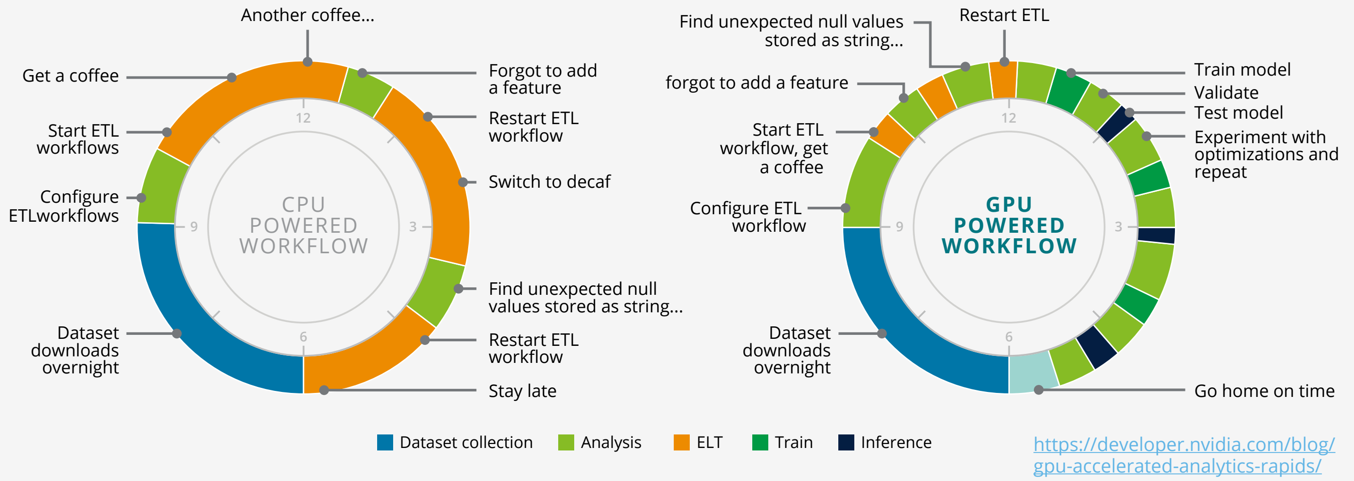
Improve speed to execution by minimizing AI development and deployment latency;

Reduce operation costs due to lower infrastructure footprint; and

Transform products, operations, and improve real-time engagement with customers.

Using accelerated infrastructure also has a valuable impact on the data science team. In a traditional, non-accelerated AI development process, the data scientist's workflow is dominated by a lot of waiting for the computations to complete. what's more challenging is that problems that are discovered in the process perpetually lead to restarting the training and testing. This is hugely inefficient, wasting valuable employee time that might better be focused on ideation and innovation. It also slows the AI lifecycle overall. With a GPU-accelerated workflow, computations take place dramatically faster, improving both productivity and the time to model production and deployment.

DAY IN THE LIFE OF A DATA SCIENTIST



https://developer.nvidia.com/blog/gpu-accelerated-analytics-rapids/

The value in GPU-fueled AI is evident. What is needed is a fine balance of platform engineering, model engineering, and research, all guided through MLOps, which features automated development pipelines, processes, and tools that permit efficient management of the AI lifecycle. Taking a closer look at how GPUs accelerate all phases of the AI lifecycle reveals precisely where the value lies.

Acceleration across the AI lifecycle with optimized AI software

The pace of model creation directly impacts the value an organization can realize from its AI efforts. With accelerated computing, enterprises can speed up AI development across the areas of data preparation, model training, deployment, and ongoing maintenance. NVIDIA AI Enterprise includes best-in-class AI frameworks and tools from NVIDIA including NVIDIA RAPIDS and Triton Inference Server and is licensed and supported by NVIDIA

Data preparation

Data sets are growing, increasing in size and complexity. Extracting, cleaning and preparing big data to be used for analytics is prohibitively slow without accelerated computing. Data preparation can be sped up by using the RAPIDS suite of open-source software libraries, also part of NVIDIA AI Enterprise. RAPIDS accelerates data preparation for analytics and data science by parallelizing tasks on NVIDIA accelerated computing. RAPIDS also includes support for multi-node, multi-GPU deployments, enabling vastly accelerated processing and training at scale.

180x speedup in data preparation

Mid-market specialty retailer

My previous bottleneck was I/O. ...15 seconds to pull in data for 10 stores (about 1 Million rows). With RAPIDS, we can pull in data for about 600 stores (60 Million rows) in less than 5 seconds. ... just plain awesome

Training

Model training takes time and experimentation as the model learns patterns from data to form accurate representations. Iterating and improving quickly are ways to expediting development time to production. At the training phase, the imperative is reducing the time required to arrive at an accurate model. There is empirical evidence demonstrating that increasing the volume of datasets and models size historically results in better results, as observed with recent advances in NLP and language generations.

GPUs and GPU-optimized software and libraries significantly accelerate training, allowing many more training cycles in the same period as traditional training while also reaching higher accuracy in less time than CPU-only computing. The result is faster model delivery and more time for exploration and AI tuning. This is especially important when moving from the prototype dataset to the production-scale data, where the computational demands require acceleration for faster time to accuracy

Acceleration to save lives

The COVID-19 pandemic was the most significant public health crisis in a generation. Vaccine development for a novel virus can take years, even decades. The Argonne National Laboratory combined AI and accelerated computing to simulate COVID-19 protein drug interactions using the NVIDIA DGX A100, a GPU-powered supercomputer. With 120 petaflops of AI performance, scientists accomplished years' worth of research in months.

Model deployment

When it comes to deploying models, there are two essential criteria. The first is scale as it relates to how many model requests can be served at once. The second is performance in terms of how quickly a model can be executed. Importantly, accelerated computing can be available across numerous types of infrastructure, including on premise, in the cloud, or in devices and systems at the edge. They can also operate on diverse software platforms, giving the flexibility needed to speed deployment in varying technology ecosystems. NVIDIA Triton Inference Server, available as open source that is part of the NVIDIA AI Enterprise suite, helps with fast and scalable multi-framework model deployment for any AI application

Scale and speed in practice

An example use case is real-time fraud detection. In one instance, a financial services company improved accuracy and met a two millisecond latency requirement using Triton and accelerated computing on GPUs, a speed far faster than what a CPU-based configuration could achieve.

Model maintenance

AI models do not run perfectly indefinitely. As data and circumstances in the real-world environment change, the model can drift from its desired accuracy. Monitoring and managing model drift is a necessary component of AI use, and with accelerated computing, the task of retraining or updating a model can be accomplished faster, ensuring the AI tool remains accurate over time. Likewise, as new algorithms are developed to improve deployed solutions, retraining using NVIDIA accelerated computing AI Platforms can be accomplished faster than with traditional computing. Some AI models are even retrained several times a day.

Deloitte's ReadyAI offering is specifically designed to engage with clients on MLOps including maintenance of models to help ensure continuous validity.

Summary of accelerated AI

Succeeding with AI requires computational power. Using GPU-accelerated computing for model creation and deployment in application delivers essential time savings, higher accuracy, and a greater capacity for experimentation. As enterprises refine and expand their AI strategies, the clear call is to identify where accelerated computing can be used to enhance existing capabilities and accelerate the entire AI lifecycle.

Get in touch

Christine Ahn

Principal
Deloitte Consulting LLP
chrisahn@deloitte.com

Anthony Abbattista

Principal
Deloitte Consulting LLP
aabbattista@deloitte.com

AUTHORS

Shankar Chandrasekaran

Sr. Product Marketing Manager
NVIDIA
shankarc@nvidia.com

Tanuj Agarwal

Senior Manager
Deloitte Consulting LLP
tanuagarwal@deloitte.com

As used in this document, "Deloitte" means Deloitte Consulting LLP, a subsidiary of Deloitte LLP. Please see www.deloitte.com/us/about for a detailed description of our legal structure. Certain services may not be available to attest clients under the rules and regulations of public accounting. This publication contains general information only and Deloitte is not, by means of this publication, rendering accounting, business, financial, investment, legal, tax, or other professional advice or services. This publication is not a substitute for such professional advice or services, nor should it be used as a basis for any decision or action that may affect your business. Before making any decision or taking any action that may affect your business, you should consult a qualified professional advisor. Deloitte shall not be responsible for any loss sustained by any person who relies on this publication.