VULTR

EBOOK

Closing the GPU Divide

Democratizing infrastructure access to support transformational AI and ML

VULTR.COM

What's the big idea?

The fractional GPU model for cloud access to graphics processing units (GPUs) offers a new, affordable means to powering artificial intelligence (AI) and machine learning (ML) workloads. ML engineers can deploy fractional GPUs to begin building everywhere around the globe through on-demand provisioning of full-scale systems for testing and production deployment of ML models. Fractional GPUs accelerate time-to-model and the lifecycle of building, testing, and deploying of AI and ML initiatives.

Who should care and why?

CDOs/CDAOs
Flexible, scalable access to GPUs is foundational to placing ML models at the core of business operations to achieve AI transformation.

Heads of Finance
Fractional GPU access relieves companies of the high costs of overprovisioning while enabling them to operationalize AI and ML throughout the business.

Architects
The fractional GPU model fosters composability so the business can rapidly deliver new products and services without the constraints of physical on-premises servers.

AI and ML Engineers
The flexible nature of infrastructure-as-a-service (IaaS) that comes with fractional GPU access simplifies machine learning operations and makes ML available wherever data resides.

Snapshot

AI and ML are the drivers of breakthrough innovation. Until recently, access to GPUs, the infrastructure that makes AI and ML possible, has been out of reach for most companies due to this disruptive technology's high price tag. Only the wealthiest organizations have been able to afford to purchase and maintain their own GPUs or pay the exorbitant prices charged by AWS, GCP, and Azure – the hyperscale cloud providers – for access to cloud GPUs.

Fortunately, a new paradigm has emerged that features affordable access to cloud GPUs and democratizes access to this vital AI infrastructure. Cloud GPUs allow organizations of all sizes to rent just the amount of GPU compute power needed to run their AI and ML workloads, avoiding the costly overprovisioning of GPU resources. Cloud GPUs give all organizations a seat at the breakthrough innovation table, closing the GPU divide and putting everyone on a level playing field with even the giants of the tech industry.

By partnering with the right independent cloud provider, organizations can get affordable access to the latest and best-performing GPUs for AI and ML, VFX, and VDI, available on demand as either virtual machines or bare metal. Companies can provision as little as a fraction of a single GPU for developers to start building, or deploy multiple interconnected GPUs for test and production deployment, choosing just the right amount of compute resources to fit their business needs and budget.

Artificial Intelligence (AI) and Machine Learning (ML) Are Driving the Future of Innovation

> ***"The parallel processing power of GPUs boosts performance for AI use cases ranging from natural language understanding (NLU) – such as speech recognition, text analytics, and virtual agents – to computer vision – such as defect detection, object recognition, and facial analysis. Indeed, [GPUs] are critical for nearly every AI application built on unstructured and semi-structured data."***
>
> -Kjell Carlsson, head of data science strategy & evangelism, Domino Data Lab

Sensational applications for AI and ML are all over the news as generative AI tools like ChatGPT and other deep learning apps capture the world's attention. But where do these drivers of breakthrough innovation come from? In short, *everywhere*. Or at least it should.

Why *should*? The current reality is that only a select few organizations – those with the deepest of pockets – hold an advantage over all others in their ability to afford the foundation for operationalized AI and ML. These players – the well-known tech giants – have made massive investments in GPUs to keep the tools of breakthrough innovation out of the hands of their competitors. And so, the path to AI transformation has been theirs alone to travel.

We all recognize that we are better off as a global community when the means of breakthrough innovation are not restricted to an incredibly small but powerful set of global enterprises. Access to the AI and ML infrastructure that makes breakthrough innovation possible must be democratized.

Central processing units (CPUs) – even those with the most advanced processors – are not suitable for advanced AI and ML workloads. They are unable to process in a reasonable timeframe the immense volumes of data that train and maintain the deep learning models that power AI applications and enable breakthrough innovation. However, just like with CPUs, GPUs similarly need a cloud-based delivery model to enable development teams everywhere to rapidly build, test, and deploy new applications.

Fortunately, there are visionaries in the marketplace today who understand the need to provide affordable access to GPU-powered compute resources. They are taking steps to make AI transformation attainable by all organizations, changing the way access to GPUs is delivered to the masses. And as more organizations learn about the new, affordable model for cloud-based GPUs, barriers to innovation will fall, paving the way for all of us to prosper.

NVIDIA and Other GPU Manufacturers Bring Parallel Processing Innovation to Market

GPUs have a fundamentally different architecture than CPUs, which enables them to process a far greater number of concurrent mathematical and geographical calculations with greater accuracy. GPUs' parallel processing breaks complex problems into millions of separate tasks and solves them simultaneously instead of relying on a sequential approach that a CPU would take. As a result, GPUs can manage the processing of billions of data points in a machine learning model in a fraction of the time it would take CPUs to accomplish.

Originally, GPUs were designed to offload computationally expensive tasks like graphics rendering from CPUs. As a result, GPUs found a specific niche in the gaming industry. Now, with the Internet of Things and organizations collecting more data to analyze and deliver new business insights and tailored services to their customers, GPUs have exploded into prominence as the critical infrastructure needed to support many practical applications of resource-intensive computing.

> ***Now, with the Internet of Things and organizations collecting more data to analyze and deliver new business insights and tailored services to their customers, GPUs have exploded into prominence as the critical infrastructure needed to support many practical applications of resource-intensive computing.***

Manufacturers such as NVIDIA, AMD, and Intel have produced and sold GPUs for years. As the pioneer of GPUs, NVIDIA has continued to advance this critical infrastructure platform to enable a wholesale revolution in worldwide compute. NVIDIA's complete line of GPUs, including their groundbreaking HGX H100 platform, are now providing all organizations the compute power to handle the most compute-intensive workloads including any AI or ML training model or application.

With NVIDIA GPUs enabling massive parallel processing and seamless scaling to accommodate even the most demanding workloads, data scientists and developers using NVIDIA GPUs will be able to train, operationalize, and monetize ML models sooner.

Barriers to GPU Access and the Unfair Advantage Mega Enterprises Hold

In the early 2020s, GPU demand outstripped supply, driven by, among other factors, the significant but now-diminished spike in crypto mining operations, which coincided with the semiconductor shortage stemming from the global supply chain challenges of COVID-19.

As a result, until recently, access to GPUs and their incredible power had largely been limited to the smallest fraction of large enterprises that have the biggest budgets for cloud and data center infrastructure. For years, this tiny cadre of enormous companies with the financial resources to purchase and deploy GPUs in private data centers and private cloud environments has enjoyed the advantage of market exclusivity.

This market exclusivity propagated a myth that they alone are empowered to lead the AI transformation movement. Illustrations from the hyperscalers are readily apparent:

- Microsoft's investments in OpenAI to transform its Bing search engine
- Alphabet's stable of Google, YouTube and Waymo
- Amazon's ecommerce algorithms and Alexa virtual assistant technology

What we see here is not that these tech giants introduced the innovation; rather, they bought (or bought into) the companies that brought forth market disruption. These acquisitions came about in part because the economics of scaling an AI company are prohibitively expensive, largely due to the cost of the GPUs needed to provide the necessary compute power.

As they did originally with CPU-based data center resources, the major cloud providers began offering cloud access to GPUs. At first it looked like they were beating a path to GPU equity. Instead, however, their customers ended up with more of the same hyperscaler bloat that exists with their other cloud offerings: too much capacity, too little flexibility, too costly. And even as the use cases for GPU-based infrastructure grow, access to the infrastructure via the hyperscalers grows further out of reach, creating the GPU divide.

Common GPU Use Cases

Let's take a look at the various use cases for deep learning models that are driving the exploding demand for GPUs today.

Gaming
GPUs were initially developed to support the graphics-rich compute-intensive workloads involved in rendering a satisfactory user experience for gamers. Today, GPUs go far beyond the gaming industry to support high-quality graphics used in myriad disciplines.

Deep Learning and AI
As mentioned, GPUs are widely used for training and running deep learning and artificial intelligence models due to their parallel processing capabilities and large memory bandwidth.

Scientific Research and Computing
Scientists and researchers employ GPUs for simulation, numerical analysis, modeling and other high-performance computing tasks in various fields, including physics, chemistry, and biology.

Virtual Reality and Augmented Reality
GPUs enable a host of VR and AR applications, providing a more immersive and interactive experience. Critical applications such as training simulations and digital twins require the massive parallel processing capabilities GPUs enable.

Video Editing and Rendering
GPUs are foundational to the architecture assembled for video editing and rendering applications that accelerate video processing and reduce rendering times.

Virtual Desktop Infrastructure (VDI)
With GPUs, organizations can now finally cash in on the promise of VDI, which has been dangled before them for more than a decade. The performance of even the most modest GPU can deliver a seamless user experience with minimal latency, even over a significant geographic distribution.

Computer-Aided Design (CAD) and 3D Modeling
GPUs are used in CAD and 3D modeling applications to provide real-time rendering of complex graphics and visualizations.

Medical Imaging

GPUs greatly accelerate the processing of medical imaging data, which is often large and complex. This leads to quicker diagnoses, more effective treatments for patients, and faster therapy development.

Web 3.0 and Blockchain

These technologies involve massive workloads, which can be addressed through GPUs, to deliver new benefits for a wide range of sectors.

Other High-Performance Computing

GPUs power supercomputers and clusters for compute-intensive applications, such as weather forecasting, oil and gas exploration, and financial modeling.

As the above use cases demonstrate, businesses must recognize the growing importance of positioning AI and ML at the core of their operations. To achieve this they need to consider how infrastructure choices facilitate the massive parallel processing that powers AI and ML.

Fractional GPUs: A Breakthrough Innovation to Democratize Access for Developers Everywhere

Without democratized access to GPUs, innovation will stagnate. Fortunately there's a new paradigm for giving organizations the access they need: fractional GPUs. Fractional GPUs partition physical GPUs into discrete virtual GPUs, each with its own memory and compute. Fractional access is designed to provide just the GPU compute power an individual developer needs to begin building.

Organizations no longer have to over-provision expensive GPUs for development. They can now gain cost-efficiency in scaling AI and ML by leveraging fractional GPUs for individual development tasks and general cloud GPUs to provision production scale test and deployment environments, all at an affordable cost. Fractional access for developers comes with fractional pricing, meaning AI-based start-ups and mature organizations pursuing AI initiatives now have a much more cost effective path to the GPUs they need.

> ***The fractional GPU business model enables organizations to leave GPU management – from procurement and deployment to maintenance and security – to the cloud provider and focus instead on development, training, and deployment of their models and other AI initiatives.***

Fractional GPUs are perfect for AI inference, NLP, voice recognition, computer vision, and other AI and ML workloads critical for innovation. The fractional GPU business model enables organizations to leave GPU management – from procurement and deployment to maintenance and security – to the cloud provider and focus instead on development, training, and deployment of their models and other AI initiatives.

The Future of Innovation

With democratized access to fractional and cloud GPUs, every organization can take a seat at the table of breakthrough innovation. This change will drive a future in which the tech giants will no longer maintain exclusive access to the infrastructure that makes innovation possible.

How will affordable access to GPUs empower your company's AI transformation journey? Find out for free.

Sign up for a Vultr account and get free credits to try fractional GPUs.