VULTR

The importance of a GPU strategy with Vultr

Graphics processing units (GPUs) are hardware accelerators capable of extremely fast computations. Think of them as a more specialized counterpart to the central processing unit (CPU). While CPUs are useful for various tasks in everyday computing, GPUs contain vastly more cores and can provide higher levels of parallelism on a subset of tasks.

Although GPUs were initially designed for rendering real-time graphics, developers have since found other applications for them, such as scientific computing, video editing, and more recently, machine learning (ML) and artificial intelligence (AI).

The recent advancements in modern AI are partly due to neural networks and deep learning. Deep neural networks can perform highly complex tasks when trained on huge volumes of data. However, this training is the most resource-intensive and time-consuming phase in the AI lifecycle. GPUs can significantly reduce training time thanks to their high capacity for parallel computing, which eliminates bottlenecks and other limitations.

Nevertheless, this same efficiency in AI and ML has caused a dearth in the availability of physical GPUs. There simply isn't enough supply to meet the rising global demand. Furthermore, building an in-house infrastructure for accelerated computing is a huge investment. It's difficult to procure all the resources needed for development, testing, and production, along with the installation, licensing, and maintenance of drivers.

This article covers what cloud GPUs are, their benefits, limitations, and use cases. It also provides a preview of the most prominent cloud GPU offerings from Vultr, so you can choose which GPU strategy to employ to ensure a successful AI/ML workload.

Cloud GPUs make testing easy

Traditionally, companies working on AI and ML projects install on-premises servers using physical GPUs. This infrastructure is not only difficult to procure but also requires dedicated IT teams for monitoring and maintenance. Regular hardware and software updates are also expected, as new technology is always emerging. In short, in-house infrastructure is not only significantly expensive but can also lead to bottlenecks, increased operational costs, and a reduction in developer productivity.

GPU time is a limited resource and must be shared between all members of an organization. To avoid resource contention, organizations usually impose policies for granting GPU access, which involves submitting ticket requests and long wait times. In this arrangement, production teams are usually given top priority, while development and testing workloads are pushed to the back of the queue.

This quickly becomes inconvenient, as developers need to request tickets for every build iteration when creating new features. Even for testing builds and updates, the setup needs to change to allow testing time on GPUs. This leads to delays in development and testing cycles, leaving developers both frustrated and disengaged.

Fortunately, the continued advancements in cloud technologies have opened up new avenues for infrastructure. Many organizations are moving toward cloud-based GPUs as a more flexible, sustainable, and scalable option.

Cloud GPU use cases

Cloud GPUs are exactly what the name implies: outsourcing GPUs to cloud vendors as a simplified alternative to on-premises infrastructure. Distributing GPUs over the cloud removes the challenges of limited GPU availability and the difficulties of setup and maintenance. This frees up local computing resources, and because they exist on the cloud, they can be scaled to meet demand without requiring additional infrastructure.

Cloud GPUs can be set up and configured quickly with no upfront investment. Some cloud vendors even provide ready-to-use GPU instances with everything preinstalled. This

means there's no hassle with software, drivers, or licensing, making it easy to solve heavy computational problems.

Since changes in resource demand are part of AI/ML workloads, developers often require frequent GPU access to experiment with new ideas to accelerate AI innovation. Furthermore, the GPU requirements themselves can drastically change depending on the nature of the project.

For instance, team-training a large ML model on a huge dataset requires considerable processing power. However, once the model is served in production, running predictions only requires a fraction of that power. Using the same powerful GPU for both workloads is a considerable waste of money and GPU time.

This is why some cloud vendors rent out on-demand GPUs on a fractional basis. You can rent a larger fraction for heavy workloads like model training, then reduce it to a smaller fraction for model inference. Fractional GPUs are even equipped with dynamic autoscaling to match the processing power of various workloads, and can efficiently scale across nodes and clusters. On-demand cloud GPUs also provide multiple offerings so that developers can experiment with different GPUs until they find the one that best fits their needs.

Modern-day cloud computing services have made deep learning far more accessible by offering GPUs on the cloud. They're far more flexible, affordable, and readily available than physical hardware, allowing developers to focus on their work without wait times or other distractions.

Cloud GPU benefits

This section outlines some advantages that cloud GPUs provide over on-premises GPUs, which organizations can leverage to develop their global strategies.

Agility

Cloud GPUs remove the biggest barrier to accelerating AI/ML workloads – the GPU bottleneck. Organizations can set up guaranteed quotas of GPU resources and alter resource allocation to ensure each team gets what they need. This means that teams can move faster by avoiding the hassle of procuring and installing GPUs locally.

Ease of setup

With cloud-based GPUs, developers can set up and tear down GPU instances without needing to know the underlying infrastructure. Cloud GPUs come with everything preinstalled, including licensed drivers and toolkits.

Furthermore, instead of needing to learn about DevOps, developers can offload DevOps tasks to the cloud vendor and focus on development and testing workloads. This also frees up local computing resources.

Scalability

Cloud GPUs provide teams with the right-sized resources for testing and development, which can be easily scaled to meet demand without any long-term commitments or additional infrastructure. Development teams can also use fractional GPUs to build and test workflows on a pay-as-you-go basis.

Cost-effectiveness

Cloud GPUs can be rented on a fractional basis to optimize costs for different use cases, meaning there's no need to pay for an entire physical GPU. Fractional GPUs dynamically allocate parts of a physical GPU to create multiple instances. This means that multiple teams can run multiple inference services on the same GPU instance and pay only for what they use. Cloud GPUs save costs and provide an affordable way to deploy and scale GPU-powered applications quickly.

Accessibility

Cloud GPUs are useful when teams are spread out across different geographical areas. With on-premises equipment, it can be difficult to procure GPUs close to where team members are located.

Cloud-based solutions offer a large worldwide network, enabling teams to program and easily scale a low-latency infrastructure no matter where they may be. Teams can develop and test on the GPU instance closest to them and deploy on the ones closest to the customers.

Adaptability

A cloud-based model is great for experimenting with different GPUs to find the architecture best suited for your use case. For instance, the NVIDIA A100 Tensor Core GPU is great for training large neural networks, while the NVIDIA A40 is well known for visual effects rendering and virtual workstations.

Instead of purchasing and installing new GPUs on the premises to test different architectures, it's easier to run experiments on GPUs on the cloud. Cloud vendors are also usually the first to upgrade GPU models when they become available.

Cloud vs. On-Prem GPUs

While cloud GPUs provide a low financial barrier to entry, they can be more expensive on a per-hour basis than on-premises GPUs, especially if you use them extensively. Ultimately, the choice between cloud and on-premises GPUs will depend on your specific needs and budget.

For all their strengths, cloud-based GPUs are unlikely to replace on-premises production GPUs entirely. This is particularly true when running enterprise AI/ML systems. Enterprises often have a business case for moving heavy GPU workloads to on-premises servers. For instance, iteratively training complex ML models for long-term projects requires continuous computing power. In this case, investing in an on-premises GPU infrastructure is more beneficial since you don't have to keep track of GPU usage hours on the cloud.

Some enterprises also depend on low-latency GPU-powered workloads or must comply with strict standards for data sovereignty and privacy. They need infrastructure to create quality customer experiences or handle sensitive data such as financial information or health care records. In such cases, on-premises GPU infrastructure is essential for enterprises to keep data behind firewalls.

Production-grade or data center GPUs such as NVIDIA DGX systems provide a robust infrastructure that can scale in an enterprise production setting. They incorporate features such as enterprise-grade orchestration, an optimized OS for AI workloads, and a network infrastructure that operationalizes AI at scale.

They're built for running enterprise AI/ML workloads with low latency, real-time processing, and high reliability. They provide solutions for mission-critical applications or systems with specific requirements.

Cloud GPUs cannot match the performance or reliability of on-premises production

GPU systems. However, they do provide a sufficient entry point for development and testing workloads. Cloud GPUs are generally more scalable than on-premises GPUs, as you can easily add or remove GPU instances as needed. Cloud GPUs should be viewed as complementary services that can be used for developing, debugging, or serving applications where latency and reliability are not critical requirements. They provide a substantial starting point to explore new ideas and experiment without a long-term commitment to a specific GPU configuration.

GPUs at the edge

With more organizations tapping into real-time processing, the concept of edge computing has gained a lot of traction. Real-time processing helps gain faster insights from data, improving services and streamlining operations. Edge computing is the practice of capturing and processing data physically closer to where it was originally generated or at the user's end. Hence, it's called computing "at the edge."

Think of the information passed between servers as electricity across a wire. A current passing through a longer wire takes more time and energy to reach its destination. By moving the source of the electricity closer to the destination, the current can pass much faster and more effectively. This is what edge computing aims to do by distributing its data storage across servers in different geographic locations.

Imbuing edge computing with AI creates an intelligent computing model that can serve applications sensitive to latency issues or with limited internet connectivity. Many AI applications can benefit from edge computing to implement real-time data processing with low network overhead and reduced latency.

For instance, in the case of Internet of Things (IoT) applications, ML models can be deployed on edge devices and run inference on Cloud GPUs at the edge network. This drastically reduces the overhead of transmitting data over the network back to the cloud for processing.

Cloud GPUs can provide the power and performance needed for AI and ML workloads when deployed closer to the data source or end user. GPUs at the edge provide low latency, reduced bandwidth, and higher data sovereignty.

Companies are increasingly going global, with users spread out worldwide. Edge GPUs enable organizations to scale and eliminate costs. The infrastructure for transferring, storing, and processing large volumes of data can be very costly to manage.

By leveraging unused or underutilized resources at the edge, organizations can save money on expensive infrastructure while getting the best performance.

Vultr Cloud GPUs

Vultr is a cloud provider that offers virtual machines (VMs) and bare metal powered by a variety of GPUs. It enables optimal resource utilization of GPUs and makes accelerated computing affordable and easy for all.

Vultr offers cloud GPU instances partitioned into virtual GPUs (vGPUs). It lets you pick the performance level that matches your workload and budget. vGPUs look just like physical GPUs – each has its own dedicated computing, memory, and cache, making them self-contained processors. They provide VMs simultaneous access to physical GPUs hosted on hypervisors. vGPUs allow a physical data center GPU to be shared across multiple VMs without affecting other instances running on the same physical GPU.

Vultr offers two types of GPU servers:

- **Cloud GPU servers –** These contain dedicated vCPU and vGPU resources and are well suited for ML workloads and high-performance computing (HPC) applications, such as video encoding, cloud gaming, and graphics rendering.
- **Bare metal servers –** These are single-tenant, dedicated hardware and are ideal for applications with high performance demands, such as large-scale model training, real-time processing, or mission-critical applications.

Fractional GPUs

Vultr also offers fractional GPUs that are ideal for development use cases where much of the work involves testing and iterating, with inconsistent GPU usage. Vultr's fractional GPUs provide an affordable way for organizations to deploy and scale GPU-powered applications quickly.

Cloud GPUs come in affordable fractions ranging from 1/20th of a card to a fully dedicated NVIDIA A100 GPU. Bare metal servers also support multiple cards for the most demanding applications. The vGPUS are not shared among customers – each cloud server has a dedicated vGPU core attached, so the instance never waits for other processes. Workloads can run at peak speed and efficiency.

Vultr at the edge

Vultr's edge GPU availability provides a series of cloud computing options for clients across the world. Vultr optimizes routing and peering agreements to ensure that VMs experience a low-latency, high-performance network. It offers superior geographical coverage with servers in more locations than any of its competitors, reducing latency for users and allowing organizations to develop a global strategy.

Cloud GPUs are a clear choice

GPUs have diverse applications ranging from graphic processing, scientific computing, video editing, and 3D rendering, to AI and ML. Organizations can benefit from moving to cloud GPUs, especially when training and testing ML workloads.

Cloud GPUs drastically reduce the complexity of setting up testing and development environments. Developers can easily provision new GPU instances on the cloud and get started with experimentation without needing much DevOps knowledge. Organizations can set up guaranteed quotas of GPU resources to avoid resource contentions. Teams can move faster with minimal cost and deploy applications on VMs closer to the users.

Vultr offers cloud GPUs that make accelerated computing affordable and VMs with fractional GPUs to scale as needed. With Vultr, organizations get guaranteed quotas of GPU resources that come preconfigured with licensed NVIDIA drivers and ML/AI frameworks, making it easier to get started.