VULTR

How to accelerate digital innovation with fractional GPUs

A graphical processing unit (GPU) is a hardware accelerator designed to run computations on large data volumes and handle parallel computations. Parallel computation is a problem you can break down into smaller, independent computations – each runs separately before recombining. GPUs have a highly parallel structure that enables processing large blocks of data in parallel, making them well-suited for graphic and non-graphic-related tasks, like building complex deep learning (DL) models, performing distributed training, and running inferences on GPUs to speed up workflows.

But GPUs are very expensive and difficult to procure. Installation is a challenge, and maintenance costs are high. To accelerate your digital innovations and explore new opportunities with artificial intelligence (AI) and machine learning (ML), you need to lower your costs. These initiatives require your developers to have immediate GPU access while keeping costs low.

While organizations once turned to on-premise servers, consisting of costly GPUs and license drivers, today, cloud GPUs – which allow you to quickly access preconfigured, up-to-date GPUs – are the way to go. Using on-demand GPUs in the cloud and renting them on a fractional basis allows you to optimize costs for every use case during building and testing.

The greatest barrier to accelerating AI and ML is the GPU bottleneck. Fractional GPUs enable you to move faster at the lowest possible cost, as they prevent wasting time and money on idle GPUs. By leveraging fractional GPUs for build and test workflows, your development teams can easily dedicate resources that match your workloads to the processing power you need.

Why choose fractional GPUs?

Modern GPUs are extremely powerful for accelerating DL workloads but can be overkill for tasks requiring only a fraction of GPU and memory resources. A DL project has varying workloads ranging from compute-intensive model training to light workloads like model inference, which only require a fraction of a GPU. A fractional GPU delivers the right amount of acceleration for your workload, maximizing GPU utilization.

How does a fractional GPU work?

A fractional GPU partitions a single GPU so that different workloads can run simultaneously. It allows you to take advantage of best-in-class GPUs by using only a fraction of them rather than dedicating an entire physical GPU for a single workload. You can match the performance level of your workloads, scaling GPUs up or down as needed.

GPUs are partitioned into virtual GPUs (vGPUs) – self-contained processors with dedicated computing, memory, and cache. This feature also enables guaranteed quality of service (QoS) and fault isolation – an application running on one fractional GPU instance doesn't affect other instances of applications running on the same physical GPU.

An example of this is the NVIDIA A100 Tensor Core GPU, which comes in 40GB and 80GB memory versions. It includes multi-instance GPU (MIG) technology, which allows you to partition the GPU into as many as seven independent instances. You can dynamically configure memory, compute, and bandwidth in response to changing requirements and demands.

You could create seven instances and run less demanding workloads on each instance, such as early scale development or low-throughput inference. Or you could dedicate the entire NVIDIA A100 to train a large deep learning model.

What are the benefits of fractional GPUs?

Physical GPUs can cost thousands of dollars a month and provide access to just a single cloud instance. With fractional GPUs, you can dynamically allocate part of the GPU to create multiple instances, each having a dedicated vCPU, RAM, storage, and bandwidth. Listed below are some of the benefits of fractional GPUs.

Freeing up resources to accelerate digital innovation

A dedicated GPU is often underused for the amount of processing power it can provide. For instance, a light workload like model inference uses only a few cores of a powerful GPU. If you were to use dedicated GPUs, running ten inference workloads concurrently would require you to pay the cost for ten separate GPUs.

With a fractional GPU, you can run multiple inference services on the same GPU and pay only for a fraction of the GPU used. Saving costs related to fractional GPUs allows you to invest in new digital innovation, allowing you to do more with less.

Immediacy of access

Fractional GPUs save the hassle of purchasing, installing, and maintaining physical GPUs on-premises. You can simply deploy a GPU instance on the cloud, skipping all the usual steps of setting up frameworks, libraries, and operating systems. This provides developers with on-demand access for building and testing workflows.

This immediate access to GPUs is well-suited to build and test workflows, where changes happen quickly and frequently. The on-demand nature of fractional GPUs means your developers can be sure they'll always have the resources they need for building and testing — without the costs associated with in-house GPUs.

Scale globally

A fractional GPU dynamically and automatically scales to match your workloads to the processing power required. You can use a small fraction of the GPU to train a neural network on a portion of a dataset and destroy the instance after saving your model. Meanwhile, you can use a larger fraction of the GPU to train a neural network on the whole dataset.

Fractional GPUs dynamically autoscale to run efficiently across multiple nodes and clusters. You can scale up from GPUs to multi-core GPUs on a worldwide basis, enabling you to match GPU utilization to your needs.

Fractional GPU use cases

Fractional GPUs are ideal for anyone wanting to get started with affordable ways to deploy and scale GPU-powered applications quickly. Here are some popular use cases for using fractional GPUs during build and test workflows.

Multiple teams running multiple workloads

Fractional GPUs use abstraction to enable a single physical GPU to be divided into many virtualized logical GPUs with dedicated memory and computing space. This enables multiple teams to run multiple GPU instances in parallel on a single, physical GPU.

Running lightweight workloads

Many tasks don't require significant processing power; for these, GPUs are simply too overpowered. For example, processing data or experimenting with model architectures require just a small instance of a GPU to complete the task. Here, a fractional GPU is the answer. It will significantly reduce the GPU cost and get the job done efficiently.

Deep learning inference

Although training deep learning models is highly compute-intensive and requires multiple GPUs and high GPU use, running inferences on the model is not compute-intensive. In an inference phase, a DL model is already trained on a dataset, and the weights are optimized. The model simply infers the output from new data. Using the same powerful GPU for training the model to run inference workloads wouldn't make sense. Here, fractional GPUs can be used to run multiple inference workloads on the same GPU in parallel.

Vultr Cloud GPUs, with Fractions of the NVIDIA A100 and A40

If you're looking for an affordable GPU solution that lets you use best-in-class GPUs at a fraction of the price, Vultr offers a choice of GPU models for every use case. Vultr is the first cloud provider to offer virtualization of NVIDIA A100 and A40. With Vultr Cloud GPUs, the entire NVIDIA AI software stack comes preinstalled.

Vultr offers two types of GPU servers:

- **Cloud GPU servers:** These contain dedicated vCPU and vGPU resources and are ideal for AI/ML, data analytics, HPC, visual computing, graphics rendering, and more.
- **Bare metal servers:** These are single-tenant dedicated hardware suited for larger scale machine learning training, analytics, and visual effects workloads.

Conclusion

Fractional GPUs let you use a part of GPU and resources to match your workloads to the processing power you need. They're affordable, convenient to set up, and easy to scale.

Vultr lets you deploy a fraction of NVIDIA GPUs instances while taking care of all the configuration steps. It offers Cloud GPUs and bare metal options that guarantee quotas of GPU resources and enable you to optimize billing and dynamically change resources.

Learn more about Vultr's fractional GPU offerings and get in touch with Vultr for help with your GPU needs.