VULTR

SOLUTION BRIEF

Choosing the Right Cloud IaaS Provider for Your Compute-Intensive Workloads

How to source flexible, affordable, compute-intensive infrastructure anywhere your organization has business operations

VULTR.COM

What's the big idea?

As more compute-intensive workloads move into the mainstream of today's business operations, demand for GPUs is soaring. But different GPUs are engineered for different purposes, so organizations need to understand all their options to find the right GPUs and GPU provider for their business needs today and tomorrow.

Who should care and why?

CEOs/CDOs/CDAOs

Flexible, scalable, and affordable access to GPUs is foundational to optimizing your investment in infrastructure engineered for artificial intelligence (AI), machine learning (ML), and other compute-intensive workloads that use data to drive your business.

Heads of Finance

You face make-or-break choices today in how and where to source the infrastructure and services your organization needs to support the kinds of workloads that are becoming increasingly central to business operations.

Architects

You can achieve the composability that enables your business to rapidly deliver new products and services without the constraints of physical on-premises servers through the cloud GPU business model.

AI and ML Engineers

The flexible nature of Infrastructure as a Service (IaaS) combined with fractional GPU access right-sizes and simplifies your machine learning operations.

Compute-Intensive Workloads Are Causing GPU Demand to Skyrocket

Competition for GPUs is fierce today, as use cases that take advantage of their powerful compute capacity continue to expand well beyond the graphics processing and gaming applications for which they were first utilized en-masse. Rapidly eclipsing them all is machine learning operations (MLOps), the foundation of artificial intelligence applications. Large language models that power generative AI and other data models can have billions of parameters and consume massive data sets, requiring the parallel processing that GPUs supply to train these powerful data science initiatives.

Beyond data science, AI, and ML, compute-intensive applications in oil and gas exploration, design and engineering, government and defense, healthcare and medical imaging, media and entertainment, and manufacturing, among others, are pressing GPU manufacturers and infrastructure-as-a-service providers alike to produce more GPUs (and more powerful GPUs) to satisfy the raging demand.

Amid the GPU feeding frenzy, organizations in the market for GPU capacity should carefully assess their current needs and anticipate how these are likely to change over time before settling on an infrastructure provider. Given the short- and long-term significance of compute-intensive applications to your entire business, answering the questions below can help you choose the ideal mix of GPUs and, equally important, how your company will access them.

Amid the GPU feeding frenzy, organizations in the market for GPU capacity should carefully assess their current needs and anticipate how these are likely to change over time before settling on an infrastructure provider.

On-Premises vs. Cloud-Based GPU Infrastructure: How to Decide

The processing power required for different AI and ML models can vary greatly. The current rush to stake claims to the generative AI space can mean that AI start-ups that are constructing models need modest GPU capacity in the near-term building phase. But when the time comes to scale the operation and train the model on data sets that attempt to encapsulate the entirety of the Internet, capacity consumption (and the associated costs) will go through the roof.

Shirin Ghaffary wrote in Vox magazine in March of 2023, “The cost of training a single large AI model can be millions of dollars. Because of increasing volumes of data on the Internet, the average cost of training the kinds of machine learning models that generative AI runs on could grow as large as $500 million to train a single model by 2030, according to a recent report by advanced AI research group EpochAI.”

In such a plausible scenario, GPU access will change drastically over a relatively short period of time. Exceedingly few organizations can afford to overprovision in anticipation of these steep scaling phases. And, once complete, it’s equally important to be able to scale down just as quickly if and when capacity requirements diminish.

Moreover, the collision between how rapidly GPU technology advances and the time it takes to build your model before you can run it at scale creates another challenge. The fact is you’re likely to have only two years to recoup a GPU investment and avoid technical debt before having to upgrade the hardware. Therefore, for most businesses, renting GPUs through an infrastructure-as-a-service (IaaS) agreement makes more sense than buying them outright.

With that in mind, here is a set of questions each organization would be wise to answer before selecting which route to take to optimize their GPU infrastructure.

- Do we understand the extent of our current compute-intensive capacity needs? And do we understand our current mix of use cases requiring GPUs?
- Can we forecast how that mix and those needs will change over the course of the next one to three years? Five years? Do we understand what capacity we need during a build phase compared to a training phase, and then a production phase?
- Can we estimate the lifecycle duration of our AI initiatives vis-a-vis the lifecycle duration of the current fleet of GPUs available in the marketplace today? Meaning, will our needs outlive the practical lifespan of today's GPUs? And should we therefore expect to refresh our GPU infrastructure in the middle of our AI initiatives?
- Do we have the cash on-hand to invest in a GPU infrastructure that we would own? If we chose that course, will that prevent us from making other revenue-generating investments? And can we commit to refreshing that stock as new and more powerful GPUs become available over the course of the coming years?
- Do we currently have the expertise in-house to manage an on-premises deployment of GPUs? If not, what would it take to build such a team?
- Can we anticipate how ongoing supply chain uncertainty will affect our ability to procure more GPUs in the future? And what would be the impact if we have to compete with the IaaS providers that make enormous commitments to the hardware suppliers?
- Would a cloud-based IaaS investment versus an on-prem deployment make more sense for us?
- Do we know what options exist from different IaaS providers and how willing and able each will be to customize and optimize their offerings based on our needs today and what we may expect for the coming years?
- And, finally, do we have the expertise on staff to accurately assess our needs and requirements for each of the above? Or do we need the assistance of an outside organization to help us work through these questions?

Answering the above questions can be tricky for most businesses, but doing so can help ensure your organization is set up for success.

Established businesses and startups alike are likely to decide that remaining flexible by procuring GPU capacity from an IaaS provider is financially and operationally preferable to buying their own.

Identifying the Right Cloud for Your GPU Infrastructure Strategy

The largest enterprises with the deepest pockets may feel comfortable pursuing an on-premises or private cloud strategy for their GPU infrastructure. However, most companies will be inclined to choose a different path.

According to **Crunchbase**, generative AI startup investments in 2021 funded 291 deals totaling more than $3.9 billion. The following year saw nearly the same level of investment – just over $3.7 billion – spread across 211 deals. It's unclear whether

or to what extent economic uncertainty is going to take a bite out of that level of commitment moving forward. Given that, established businesses and startups are likely to decide that remaining flexible by procuring GPU capacity from an IaaS provider is financially and operationally preferable to buying their own. The important considerations then become understanding:

- What type and volume of workloads you want to run that require GPUs
- How broad their GPU portfolio is to match specific machines to your specific use cases
- How the provider structures its GPU pricing, including:
    - the granularity of access they offer
    - how their GPU portfolio fits into their broader suite of offerings
    - how likely they will be to work with your organization to continually optimize their offerings to your changing needs.

When asked to think about different cloud IaaS providers, many start and end their lists with the three hyperscalers – Amazon, Google, and Microsoft. They think of cloud as being monolithic and don't understand that there are viable, global alternatives to the Big Three hyperscalers. And there is great variation in the range of infrastructure products and services each offers and their pricing structures, as well.

Developing an understanding of how each cloud IaaS provider can meet your needs is critical to forming your cloud and GPU infrastructure strategy. Only once you've established that strategy does it make sense to select a cloud and GPU IaaS provider as well as the right mix of specific GPUs.

Vultr is the only independent cloud GPU provider that can provision as little as a fraction of a single GPU or multiple interconnected GPUs across 30+ global cloud data centers, offering a flexible IaaS plan customized for you and your business.

Matching GPUs to Your Use Cases

Different GPUs are engineered and powered to perform different tasks. Choosing the right mix of GPUs depends on your business needs. NVIDIA is the leading provider of GPUs today, with about **80% of the market**. More organizations, including OpenAI's ChatGPT chatbot, trust their most valuable intellectual property to NVIDIA infrastructure. The range of NVIDIA GPUs available through Vultr offers companies flexibility to optimize their GPU spend based on the combination of performance and price they are looking for:

- **NVIDIA H100:** Delivers unprecedented acceleration to power the world's most advanced AI, data analytics, and HPC workloads
- **NVIDIA A100:** Empowers developers and innovators to leverage a seamless integration of simulation, data analytics, and AI
- **NVIDIA A40:** Combining professional graphics with powerful compute and AI to meet today's design, creative, and scientific challenges
- **NVIDIA A16:** Enabling virtual desktops and workstations with the power and performance to tackle any project from anywhere

Any of NVIDIA's GPUs can support resource-intensive use cases. But for more demanding AI and ML workloads, the A100 and H100 are the preferred units. With the H100, scaling GPU clusters provides orders of magnitude advantage over previous generations of NVIDIA GPUs. When scaling, the interlinkages within the H100 GPU clusters are highly optimized and offer 30x or more faster processing than similar A100 clusters. This makes the H100 the ideal power plant for extremely compute-intensive workloads like generative AI, where models may have as many as 10s or even hundreds of billions of parameters.

By articulating current and near-future use cases, you and your team can select the ideal mix of GPUs that the business will need. And you can rest easy knowing that Vultr will work with you to reconfigure your stack as your business needs change.

Why Vultr is the ideal GPU IaaS provider

Vultr enables businesses at any stage and size to access NVIDIA GPUs in more than 30 data centers around the world. Vultr also provides fractional GPU access to handle your compute-intensive workloads at a price point that enables you to provision every developer, anywhere in the world they work, with the resources they need to experiment and build models without breaking the budget and without bottlenecks to access.

Vultr is the only independent cloud GPU provider that can provision as little as a fraction of a single GPU or multiple interconnected GPUs, offering a flexible IaaS plan customized for you and your business. More importantly, your business and its data scientists and developers can use NVIDIA GPUs to train, operationalize and monetize ML models sooner, as well as access Vultr's full suite of IaaS offerings to drive other business initiatives.

Ready to see how Vultr can give you flexible, affordable, global access to NVIDIA GPUs engineered for your compute-intensive workloads? Get in touch to learn how.