VULTR

SOLUTION BRIEF

How Generative AI Will Transform Your Business Operations

And why democratized GPU access makes it affordable

VULTR.COM

What's the big idea?

When trained on a company's own data, generative AI has the potential to revolutionize how that company retains institutional knowledge, builds intellectual property, and outcompetes its rivals. Companies today are already pursuing internal-facing generative AI. If yours isn't one of them, you're already well behind your competitors. One essential aspect of developing an internal-facing generative AI solution is affordable access to graphics processing units (GPUs) that power the machine learning models that make generative AI possible.

Who should care and why?

CEOs/CDOs/CDAOs

Training generative AI on your organization's data will radically improve team enablement, provide a much more consistent user experience for external parties interacting with your brand, and accelerate the sales cycle for your business development teams.

Heads of Finance

An internal generative AI tool can lower overall OpEx by accelerating team enablement and increasing retention of institutional knowledge. Partnering with an independent cloud provider on generative AI can add to the overall OpEx savings by significantly reducing the cost of access to the infrastructure needed to train and maintain the generative AI model.

Architects

Internal generative AI will give architects unparalleled insights into their organization's specific and optimal infrastructure configurations to capitalize on unique advantages and opportunities within its marketplace.

AI and ML Engineers

Internal generative AI will help ML engineers develop code more efficiently, freeing them to focus on the organization's highest-value initiatives.

The Generative AI Use Case Brings New Urgency to ML Operations

While the world speculates on how ChatGPT and other generative AI tools will disrupt how we search the internet, or how students write essays, or how people create art and literature, there's a less obvious yet more significant opportunity for businesses: training generative AI on your company's own data to fundamentally reinvent business operations and outcompete rivals.

Your employees' experiences and expertise, your documentation, every company email and ticket that's been written, every invoice sent and received, every press

release, every presentation, every page of your intranet and website, every blog post, every contract, every piece of intellectual property that makes your business so much more than just a collection of products and services – all of these can fuel a transformative, internal generative AI solution.

That wealth of information would endure employees retiring or leaving to join your competitors. It could not be accidentally deleted or simply forgotten. Instead, your company could use that AI-augmented knowledge base to:

- Rapidly onboard new hires
- Uniformly up-level existing employees
- Retain, innovate, and iterate on earlier ideas
- Develop new solutions, products, and services
- Anticipate new market needs
- Optimize pricing strategies
- Predict events and anticipate opportunities
- Ensure a consistent customer experience across digital touchpoints
- Acquire new customers and upsell current ones

In short, by tapping into the power of your internal data, you will create the ultimate team enablement tool. The good news is the technology exists to make it happen. The issue is access to affordable graphics processing units (GPUs) needed to process structured and unstructured data while handling a variety of AI and ML tasks.

Training generative AI on your core operational data enables organizations to turn discrete bits of information into an accessible and unassailable knowledge base.

Generative AI Demands the Processing Power of GPUs

If only the world's biggest companies with the deepest pockets can afford to build a generative AI-enabled knowledge base, they will exercise an unfair advantage over all other businesses. Those organizations that don't have the top 0.1% of cloud compute budgets will be dependent on the tech giants for access to this transformational technology or barred from using it and similar leading-edge technologies.

Most companies are unable to attain these benefits, largely due to the high costs traditionally associated with the infrastructure needed to train and deploy data models used for generative AI.

The workloads that process such vast volumes of data require the latest GPUs. Yet accessing GPUs produces challenges that most businesses, including many large enterprises, may be unable to overcome.

Supply chain challenges through the pandemic made GPUs scarce. Even now as inventory problems have largely subsided, purchasing GPUs is beyond the means of most companies. The primary alternative to purchasing has been leasing GPUs from the hyperscale cloud providers, but that route to GPUs is also problematic.

Businesses of all sizes need the flexibility to operationalize AI and ML to develop the ultimate knowledge base and empowerment tool to remain competitive.

Hyperscaler Business Models Prevent All but the Largest Organizations from Pursuing Generative AI

Renting GPUs from the likes of AWS, GCP, or Azure can be prohibitively expensive for the vast majority of businesses. The offerings from the hyperscalers tend to include complex, unnecessary services. And unless your organization is making one of the largest financial commitments to these providers, you can expect a take-it-or-leave it posture of inflexibility from the hyperscalers when it comes to the potential for customized services. Conversely, if the hyperscalers do express willingness to customize, you can expect the surcharges to be steep.

Once you decide to use AWS, GCP, or Azure, you may quickly find that vendor lock-in can make it exceedingly difficult to move workloads to other environments. The cost of transferring data from the hyperscale cloud providers and the difficulty of porting complex cloud applications can serve as strong disincentives to making a switch. Further, lack of transparency on pricing can lead to inflated invoices and other bad surprises.

The phenomenon of cloud sprawl can cause a range of problems, too. For example, the same ease of spinning up a cloud instance for a test environment makes it easy to leave it running, causing surprise billing as unnecessary charges accumulate. Cloud sprawl can also produce security risks as forgotten, unmonitored workloads provide ingress opportunities for hackers.

All of these factors contribute to an emerging trend revealed in SlashData's 21st Developer Nation Survey: providers such as AWS, Microsoft Azure, and Google Cloud Platform (GCP) have experienced just an 18% growth rate over the past four years – while the alternative cloud market has nearly doubled during this same period.

Fractional GPUs: Democratizing GPU Access through a New Cloud Delivery Model

There's an affordable alternative for companies that require GPU access for generative AI (or any other AI or ML applications): fractional GPUs. The fractional GPU model allows individual developers to rent a portion of a GPU without having to commit to the overprovisioning that the hyperscalers require through their full-GPU pricing models. They don't have to buy, and they don't have to play the hyperscalers' game.

Pricing for fractional access provides organizations with a far more flexible OpEx spend to enable individual developers to experiment, prototype, and build. This significantly reduces the cost barrier as organizations explore developing their own generative AI-powered knowledge base. Combined with the ability to provision on demand full-scale systems for test and production, fractional and cloud GPUs allow organizations to meet variable demands and scale in the most cost-efficient manner globally.

Fractional GPUs unlock new potential for organizations to get started building next-generation generative AI solutions by empowering developers to innovate faster and more affordably than ever.

The internal-focused generative AI movement is already underway. A handful of the biggest spenders in cloud compute are already exercising unfair advantage based on their unique access to generative AI providers that train such solutions on their own data. Fractional GPU access, however, now opens the door to all organizations looking to arm their developers with the platform they need to get started on their own internal generative AI applications. Fractional GPU access levels the playing field and recasts the balance of power.

Enabling More Businesses to Create their Ultimate Knowledge Base via Generative AI

Cloud access to fractional GPUs closes the generative AI gap. But fractional GPU access isn't available from the hyperscalers. For that you have to turn to an alternative to the hyperscalers – the independent cloud provider. You have options when it comes to cloud infrastructure-as-a-service (IaaS) providers, but there's only one cloud provider that can offer 30+ global data center locations, access to virtualized cloud compute, bare metal servers, managed Kubernetes, managed databases, object and block storage, and fractional GPUs – everything you need to run every aspect of business operations for an AI start-up or established company.

You get all that and more with Vultr. We're disrupting the cloud IaaS market so you can disrupt your market. It's the forward-looking businesses that will tap into Vultr's fractional GPU access to close the generative AI gap and take down all rivals in their domain. Doesn't that sound like your company?

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