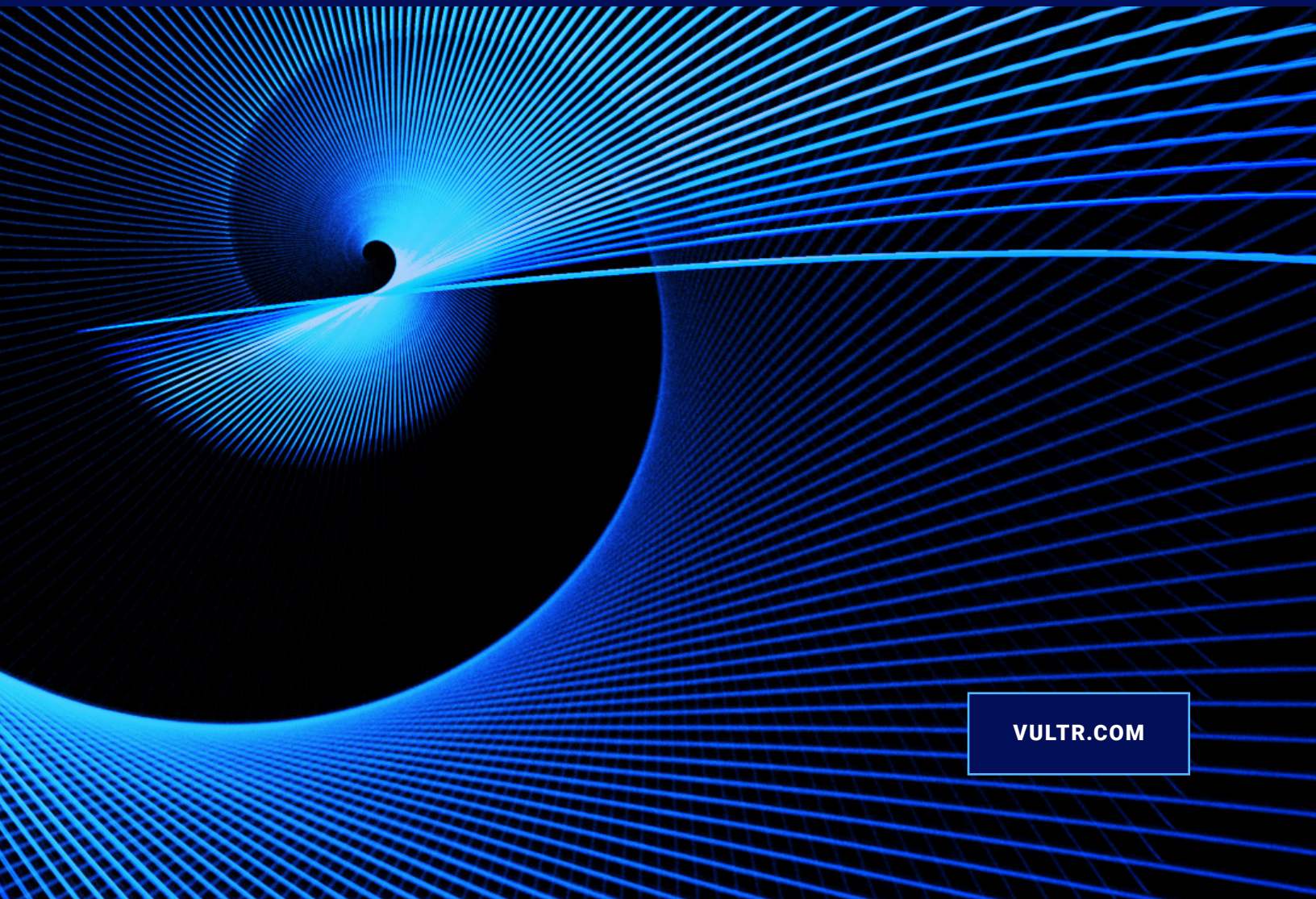
VULTR

WHITE PAPER

AI Inference at the Edge is the New Architecture for Apps



Introduction

The proof-of-concept phase is over, and the race to operationalize machine learning and artificial intelligence is on. In 2023, Generative AI (GenAI) pushed the broader category of machine learning/artificial intelligence from the margins of entrepreneurial innovation centers within enterprises to the top of the CIO's and CTO's agenda.

This innovation wave is washing over all industries, and enterprises of all sizes are pushing multiple machine learning models and large language models (LLMs) into production, scaling them for use in a distributed fashion at the edge. AI and GenAI use cases for internal operations and customer engagement are equally strong, making AI and GenAI the focal point around which all organizations will reinvent their businesses in 2024 and beyond.

According to a recent Menlo Ventures study, over 90 percent of funds invested in machine learning operations (MLOps) and large language model operations (LLMOps) across enterprises of all sizes are dedicated to inference rather than to the training of models.

From this, we can draw two conclusions:

01.

The hype is real; enterprises riding the innovation wave lean on AI and GenAI inference to drive business decisions and operations.

02.

Enterprises that can more efficiently manage inference at the edge stand to improve their bottom line significantly.

Efficiency in scaling inference at the edge starts with an understanding that managing multiple models and maintaining them in edge environments requires a fundamentally different approach to edge operations than is the generally accepted practice for managing web applications. For long-term financial viability, enterprises must think – and act – accordingly.

90%

of funds invested in machine learning operations (MLOps) and large language model operations (LLMOps) across enterprises of all sizes are dedicated to inference rather than to the training of models.

Menlo Ventures
November 2023

$443 B

The forecasted reach of the market for artificial intelligence services by 2027.

Gartner
November 2023

What makes scaling inference at the edge different from traditional or cloud-native DevOps?

Enterprises today are focused on developing, maintaining, and operationalizing an inventory of multiple machine learning models and LLMs. As such, enterprises are establishing AI/ML centers of excellence from which data science teams can collaborate to develop and train their models centrally.

At this point, efficient enterprises are not building these models from the ground up; instead, they are drawing on and customizing open-source models available in public registries and/or repurposing existing models from their own repositories for different use cases. The workflow calls for central development and training, followed by wide-scale distribution to edge-based cloud environments built for inference, where models are tuned to local data that accounts for regional and/or cultural differences.

Aside from the data governance requirements by which all distributed applications and edge operations must abide, the volume of data flows for AI, particularly for GenAI, makes transferring data across regions operationally infeasible (if not illegal). Due to the need to process massive data volumes with ultra-low latency, it is essential to generate inference at the edge, close to where workers, customers, and IoT devices interact with the models.

As such, AI can't be an add-on to existing IT and cloud operations. Enterprises must treat the infrastructure and supporting tools, services, and applications for developing, maintaining, and scaling models as their primary IT imperative. They must seek out cloud partners with the same agenda.

The hyperscalers are not the answer

In partnering to enable inference at the edge, enterprises might consider the biggest cloud providers to be natural choices. But there are probably better options. Why? The hyperscalers developed business models long before today's AI/ML era, and they are focused on catering to the cloud infrastructure needs of the world's largest enterprises with the deepest pockets. They weren't architected from the ground up to address the unique MLOps and LLMOps workflows, and their begrudging efforts to bolt on stack components to accommodate these unique workflows are marginal, inflexible, and inefficient.

In short, if your organization isn't among the top one percent of enterprises worldwide, you'll likely find the hyperscalers unwilling or unable to offer your organization the customization needed to address your business' unique profile of needs. And they're not focused on helping enterprises save money on their inference-at-the-edge initiatives.

When looking for complete AI stack solutions (as opposed to only the compute infrastructure), prospective customers often find, too, that the hyperscalers offer restrictive technology stacks featuring components that maximize their revenue at the expense of customer flexibility. For companies that contract with the hyperscalers, this can result in overpaying for tools and/or services they don't need, supplied by vendors they don't want to work with.

This rigidity, vendor lock-in, and expense make choosing the hyperscalers as cloud partners operationally and financially impractical for most enterprises.

The new architectural approach: attributes of the right cloud stack

Enterprises need a global provider of full-stack, cloud-based AI and MLOps resources, including the infrastructure, platform, and application layers. But they shouldn't consent to paying the unnecessarily high prices charged by the hyperscalers.

Instead, they should seek out a cloud partner that can offer all of the following:

A global cloud data center footprint

32 cloud data center locations to enable edge operations wherever the organization is doing business.

A cloud-based architecture purpose-built for AI

A global architecture designed for the AI era, whose operations are built around supporting centralized development, training, and management of multiple models and operationalizing models in edge environments.

A composable architecture

A composable architecture to ensure that each enterprise can assemble the ideal AI stack for its particular requirements.

Public and private container registries

Public and private container registries to simplify development, deployment, and scaling of multiple machine learning models and LLMs:

A public registry: offering best-in-class open-source models, providing a means to rapidly develop, train, and repurpose specialized, domain-specific LLMs

A private registry: enabling enterprises to push proprietary models to edge environments where they can be tuned to local data sources

Affordable availability

A partner committed to keeping cloud GPU and CPU resources affordable so that all can adopt inference at the edge.

Integrated CPU and GPU stacks

Tightly integrated CPU and GPU stacks that can offload to CPUs less compute-intensive tasks that would otherwise be run on GPUs.

A robust partner ecosystem

A broad ecosystem of best-in-class providers of tools, services, and applications committed to the interoperability espoused by composability.

Optimize AI Data Security with Retrieval-Augmented Generation (RAG)

RAG isolates enterprise data from LLMs, enhancing security by ensuring AI assistants follow access control parameters. It leverages IAM tools for secure data retrieval and processing, reducing exposure and misuse while meeting performance demands[1].

Identifying the right IaaS provider can help ensure their inference-at-the-edge operations are manageable, sustainable, and economically viable.

1. *Implementing Data Access Controls for Generative AI Assistants,* Gartner

Vultr leads GenAI inference at the edge

Vultr was built to enable inference at the edge. Vultr's global array of cloud data center locations, distributed across six continents, supplies enterprises of all sizes with all necessary components to scale AI inference in edge environments.

Global availability of AI resources

Our expansive suite of GPU and CPU resources includes a full range of state-of-the-art NVIDIA GPUs, including the the NVIDIA GH200 Grace Hopper Superchip and H100 Tensor Core GPUs, to power multiple machine learning models and LLMs in each of our 32 cloud data center locations worldwide.

NVIDIA Elite Partner

NVIDIA GPUs are the centerpiece of the Vultr GPU stack, a finely tuned and integrated operating system and software environment where data science and engineering teams can initiate model development and training with a single click.

Vultr Container Registry

Vultr Container Registry comprises public and private components, offering a seamless solution for sourcing performance-optimized NVIDIA AI Foundation models from the NVIDIA API catalog and provisioning them to Kubernetes clusters through our global network of cloud data center locations.

Vultr Cloud Alliance

The Vultr Cloud Alliance brings together world-class vendors committed to the principles of composability to offer best-in-class tools, services, and applications to customize an AI stack that perfectly fits your organization's current needs and adapts as those needs change.

VULTR

A final word

The AI innovation wave offers unprecedented potential for growth and efficiency for businesses. However, as companies face organizational and industry pressures to launch and scale AI apps, they require a new architectural approach that allows them to scale AI and ML at the edge. By leveraging a global cloud infrastructure, enterprises can efficiently scale AI inference in edge environments.

To learn more about Vultr visit **vultr.com** or **contact sales.**

VULTR.COM **CONTACT**