VULTR

WHITE PAPER

Scaling Generative AI Requires Specific Platform Engineering Considerations

VULTR.COM

Introduction

To be successful with Generative AI (GenAI), you need to adopt a platform engineering approach with a GenAI twist.

Platform engineering adapted to scaling GenAI and AI in general at the edge

How organizations that have adopted a platform engineering approach need to adapt it for GenAI

How organizations new to platform engineering can adopt this approach to scale GenAI

In the ongoing effort to adapt and simplify the DevSecOps lifecycle to better address the growing complexity around developing, deploying, and scaling traditional web applications at the edge, enterprises of all sizes are turning to platform engineering. Gartner predicts that by 2026, 80% of software engineering organizations will establish 'platform teams' as internal providers of components and tools for application delivery.

Concurrent with the rise of platform engineering is the rapidly growing investment in GenAI. Deloitte predicts that 2024 enterprise spending on GenAI will increase by 30% from an estimated US $16 billion in 2023.

These two trends are on intersecting courses. Among enterprises pursuing both GenAI and platform engineering, platform teams should ensure they are accounting for the unique requirements for scaling GenAI initiatives across the enterprise. In the next manifestation of digital transformation, these innovators have a head start – placing large language models (LLMs) at the core of business operations.

A far greater number of enterprises, however, have GenAI and/or platform engineering initiatives that are more nascent. For these organizations, platform engineering remains more of a greenfield opportunity. Regardless, at this point, collective wisdom around platform engineering for GenAI has emerged and is now available to every organization striving to improve its competitive posture.

As this guide outlines, the best practices and critical requirements for supporting GenAI with a purpose-built platform engineering solution will help enterprises scale and manage their GenAI initiatives more efficiently.

80%

of software engineering organizations will establish 'platform teams' as internal providers of components & tools for application delivery, by 2026

Gartner

30%

predicted increase in enterprise GenAI spending in 2024 from an estimated $16 billion in 2023

Deloitte

Requirements for supporting large language model operations at scale

As large language model operations (LLMOps) mature across the business landscape, it's becoming increasingly common for enterprises to develop and deploy multiple modestly-sized LLMs, each specialized for specific business processes, rather than deploying a single LLM trained to manage all processes (like the massive models that support ChatGPT and other popular GenAI applications). As such, enterprises need to provide machine learning engineers – the LLM developers – with the infrastructure, tools, services, and applications they need to optimize a complex LLMOps ecosystem.

The following have emerged as best practices for LLMOps at scale:

Establish a center of excellence

within the enterprise where LLMs and other machine learning models can be developed and trained centrally.

Leverage open-source models

from public repositories so your organization isn't starting from square one with model development.

Specialize models

and focus on developing multiple, smaller models to address specific business use cases.

Train models on proprietary data

and move trained models to a centrally-located private registry, so all models are accessible across the enterprise.

Tap a cloud-based edge architecture

that allows for tightly integrated CPU and GPU operations close to the geographic regions where your organization is doing business.

Fine-tune models at the edge

based on local data to account for regional and cultural considerations while maintaining data governance and privacy requirements.

Ensure responsible AI practices

by building observability into every phase of AIOps and LLMOps.

Must-have platform engineering capabilities for MLOps and LLMOps

The emerging best practices around LLMOps demand a platform engineering approach that can automate the provisioning and configuration of all the resources ML engineers need to build, train, deploy, and optimize LLMs and GenAI applications. Providing self-serve access to these resources frees ML engineers to focus on the high-value development work they were hired for, builds efficiencies into the workflows that support a multi-model GenAI strategy, accelerates the LLMOps development cycle, and reduces the overall time to value for the enterprise's GenAI investments.

Comprehensive platform engineering solutions designed for LLMOps at scale must address all of the following requirements:

Infrastructure optimization

Provide developers and ML engineers with easy access to edge infrastructure components optimized for GenAI workloads, allowing for tightly integrated CPU and GPU operations close to the geographic regions where the organization is doing business.

Model management and deployment

Establish a centralized model development and training environment and a Kubernetes-based private registry for trained models. This ensures all models are accessible across the enterprise, enabling efficient model management and deployment.

Data governance and privacy

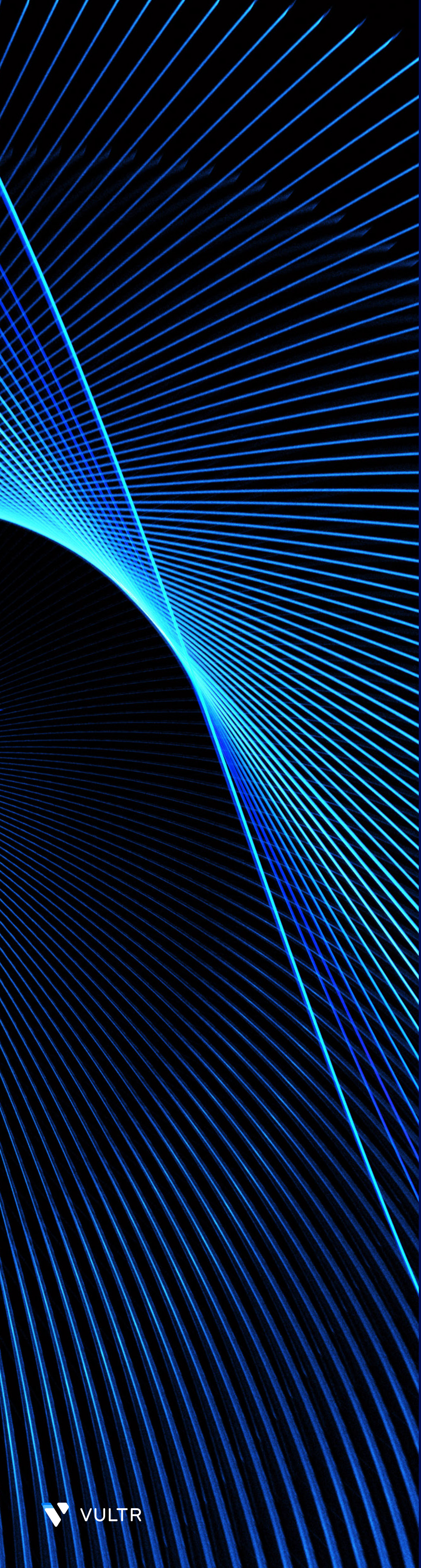
Provide edge-based data storage and security measures for maintaining data governance and privacy when training models on proprietary company data and fine-tuning models at the edge based on regional data.

Model observability

Build observability into every phase of LLMOps to ensure responsible AI practices. This involves integrating monitoring and observability tools into the platform engineering solution to track the performance of GenAI models and ensure that they adhere to ethical and operational standards.

Automating tasks and self-service

Automate code builds, testing, and deployments through CI/CD pipelines, as well as infrastructure provisioning and management using Infrastructure as Code (IaC) tools. Self-service capabilities enable the development of new software and models in less time and accommodate a range of workflows.



SPOTLIGHT

The four core tenets of building and maintaining a platform engineering approach

Think product, not project

Platform engineering must be managed as a product rather than a project. This requires treating users – the internal developers and ML engineers interacting with the platform – as customers and assigning a dedicated product support team to assist them. The platform engineering product team must continue to improve the solution as requirements change and technology evolves.

Build self-service and automation into the platform

To accelerate deployments of LLMs and GenAI applications, platform engineering solutions must make routine processes easily repeatable and consistent. This approach enhances the developer experience and increases productivity by offloading ancillary tasks that distract from their highest-value work.

Demand uncompromising reliability

Establish SLAs for availability and other attributes as part of the product mindset. This approach helps ensure timely resolution of issues and helps improve productivity.

Think minimally viable product (MVP) versus full-fledged finished product

In the age where products are never “finished” and continuous improvement is expected, platform engineering teams should focus on getting the platform up and running. Platform teams can maintain a backlog of additional features and decide what enhancements to introduce as users’ needs change.

What “good” looks like in enterprises that have adopted platform engineering

While the history of platform engineering for GenAI is not yet long, there are specific attributes that define platform engineering excellence.

Composability

Full interoperability enables developers to assemble flexible tech stacks that address specific business functions. Composability also allows platform engineering teams to swap components to ensure that platform engineering solutions remain relevant and are optimized as needs and capabilities change.

Technical architecture

Tight integration of cloud CPU and GPU infrastructure and automated, self-serve access to these resources allow enterprises to scale compute, inference, and ultra-low latency globally at the edge while optimizing resource utilization to minimize operating costs.

Vertical use cases

Developing precomposed stacks and Integrated Development Environments (IDEs) geared toward developing models for different business services can shorten the path to model-based operations and accelerate GenAI adoption.

Extensibility

Building in the flexibility to incorporate new capabilities or functionalities can future-proof the platform engineering solution against obsolescence while optimizing the enterprise’s investment in the solution and the platform engineering team supporting it.

Looking ahead

The future of platform engineering

As more enterprises organize their business practices around GenAI and other AI and ML initiatives, we anticipate that serving inference at the edge will become an even greater priority within enterprises. At the same time, new efficiencies in MLOps and LLMOps will continue to emerge, so platform engineering teams will continue to evolve their platform engineering products to better accommodate the developers and ML engineers who will focus on delivering the best user experiences for their internal and external customers.

Already, we see that companies are looking at platform engineering as a professional competency. The 2023 State of DevOps Report by Perforce found that 71% of respondents stated that their employer plans to hire people with platform engineering experience in the near future. As this competency grows into an established career field, platform teams will introduce even greater innovation into the products and processes that facilitate the move to GenAI-based business operations.

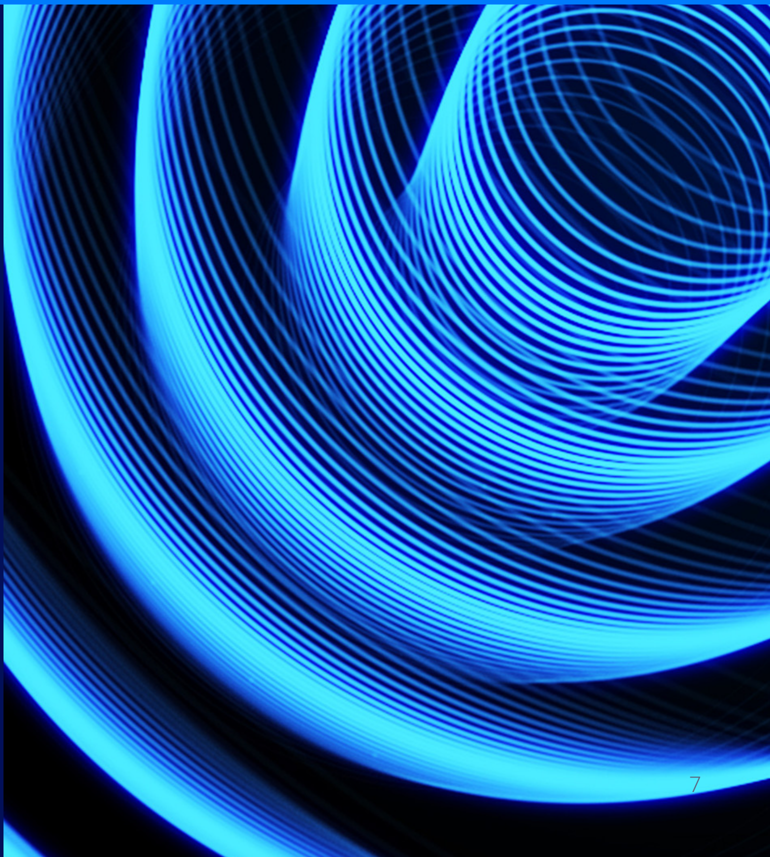
The platform engineering field is rapidly evolving.

71%

of organizations stated that they plan to hire employees with platform engineering experience in the near future.

2023 State of DevOps Report by Perforce

Enterprises that start addressing the critical considerations of platform engineering for GenAI today, and implement the best practices in this guide, will be best positioned to not only keep up with the pace of change but also contribute to the innovation in platform engineering and LLMOps that is sure to come.



VULTR

A final word

The most direct path to success in effectively scaling GenAI across the enterprise lies with a tailored platform engineering approach. Organizations that prioritize this will put themselves in the best position to future-proof their AI operations, and establish a framework for sustainable innovation.

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

VULTR.COM	CONTACT