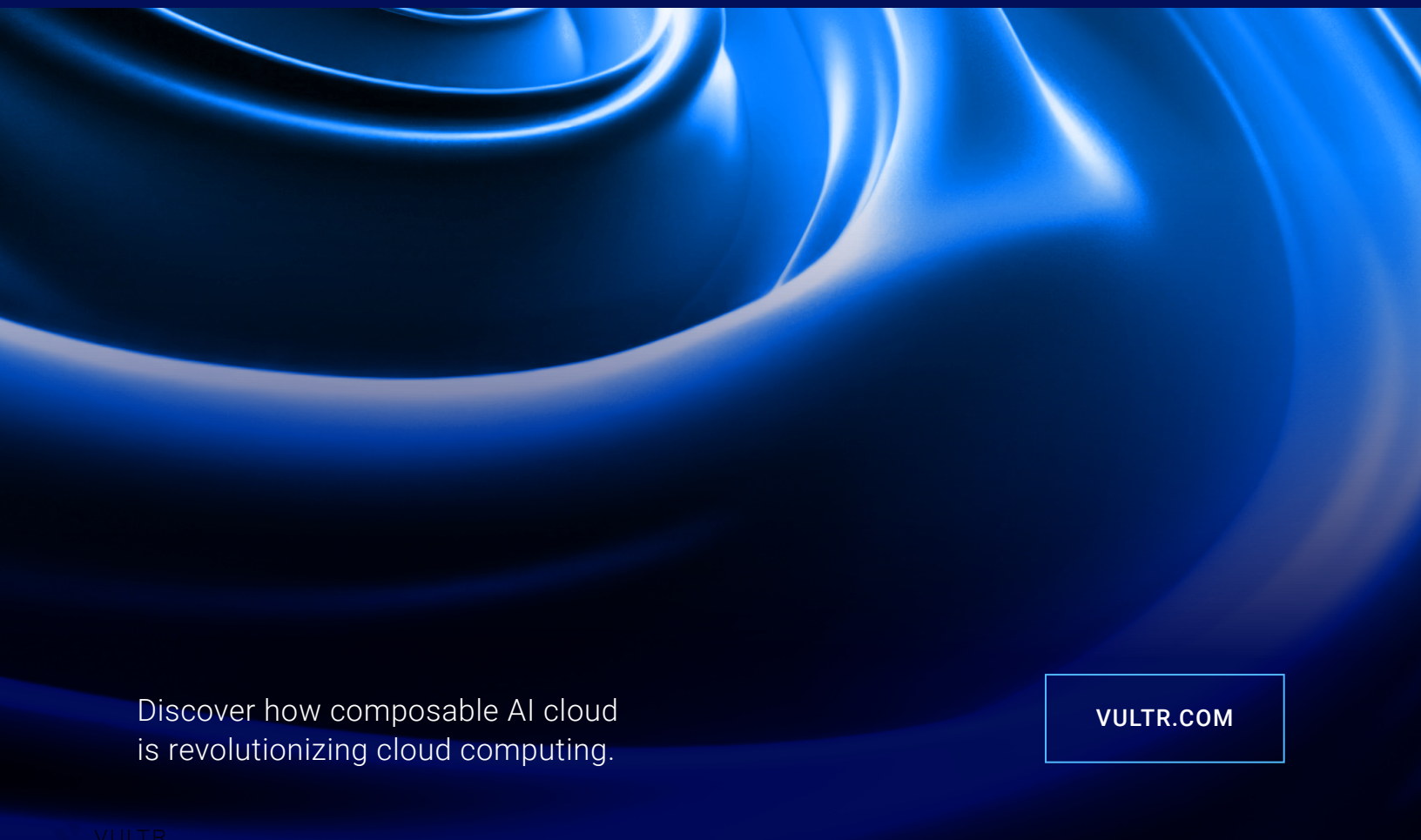
VULTR

WHITEPAPER

AI Without Boundaries

The Strategic Shift to Composable AI-First Cloud Infrastructure



Discover how composable AI cloud is revolutionizing cloud computing.

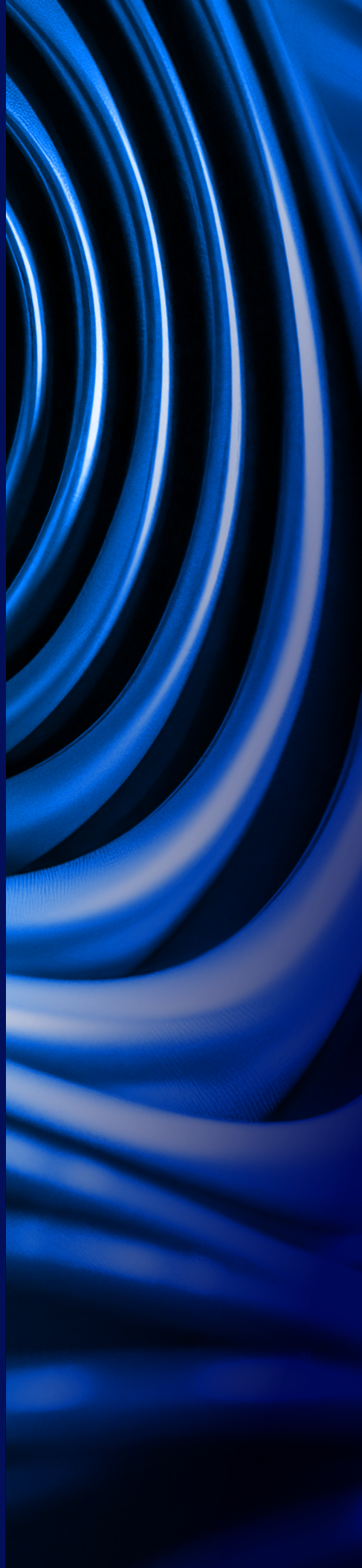
VULTR.COM

A new era of AI cloud composability

In 2023, we published a paper, *Composable Cloud: The Logical & Indispensable Evolution of the Modern Cloud Stack*, outlining the emergence of the composable cloud, a shift away from the dominance of traditional hyperscalers toward a more flexible, cost-efficient, and vendor-agnostic cloud ecosystem. At the time, we identified the limitations of hyperscalers – such as vendor lock-in, escalating costs, and a lack of adaptability to continuous innovation – as key drivers of this transition. The composable cloud emerged as a way for organizations to customize their cloud environments by selecting best-in-class IaaS, PaaS, and SaaS solutions across the stack, providing newfound freedom and control.

Fast forward a few years, and the landscape has evolved even further. AI adoption is at an all-time high, and innovation is moving even faster. Hyperscalers' dominance continues to wane as organizations demand greater flexibility, scalability, and cost-efficiency. We've officially entered the *era of the composable AI stack*. AI innovation requires openness, interoperability, and modularity. Organizations do not want to be confined to the rigid frameworks offered by hyperscalers. Instead, they are turning to AI-first architectures built on composable, open platforms.

This shift is reshaping the future of cloud computing. In this paper, we'll explore the next frontier of cloud computing: the composable AI stack, and the opportunities it presents for organizations.

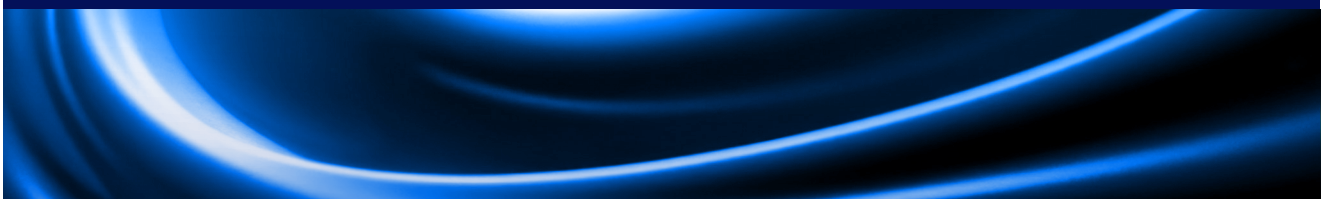


From monolithic to modular: The rise of composable infrastructure

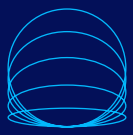
Cloud innovation will die unless walled gardens are broken open

"A handful of big tech companies in the United States have historically controlled most of the world's cloud computing infrastructure. This structure encourages lock-in, consolidating access to AI by building walls around the infrastructure needed to utilize it at scale. In 2025 we're going to see a shift away from "all in one" commercially available models and towards lightweight, open-source, purpose-built deployments. This will do three things: lower the bar to entry for startups and scaleups, improve accessibility in regions traditionally underserved by the hyperscalers, and make enterprise workloads more efficient. If we don't see this, I'm afraid for cloud, and particularly AI, innovation will stagnate and AI adoption will become prohibitively costly. Closed platforms inherently lack the flexibility needed for businesses to rapidly adapt their AI tech stacks to capitalize on the latest innovations, building innovative latency into the foundations of their tech stacks."

Kevin Cochrane, CMO of Vultr



Composable cloud architectures represent the logical evolution of IT infrastructure, building on decades of advancements from on-premises data centers to software-defined systems. Traditional hyperscalers like AWS, Microsoft Azure, and Google Cloud initially dominated the market by offering end-to-end solutions. However, their monolithic models often led to vendor lock-in, unpredictable costs, and limited flexibility. Gartner reports that 77% of enterprises have faced unexpected cost spikes with hyperscalers, while only 22% of IT leaders feel confident about controlling cloud spending.



So just what is composable cloud, anyway?

Composable infrastructure makes data center resources as readily available as cloud services and is the foundation for private and hybrid cloud solutions. While this may sound like infrastructure as a service, the difference is that composable infrastructure has no dependencies between infrastructure components, allowing customers infinite flexibility to assemble and reassemble components and capacity to meet their unique and changing needs.

Composable cloud addresses these challenges by decoupling infrastructure components – compute, storage, networking – and enabling dynamic resource allocation via software-defined orchestration. This modular approach allows organizations to "assemble" bespoke cloud environments using best-in-class tools at each layer (IaaS, PaaS, SaaS).

Unsurprisingly, AI-mature enterprises – those that have embedded AI across all aspects of their operations – are moving away from the traditional hyperscalers. Rather than investing further in their clouds, a choice made by only 15% of enterprises[1], AI leaders are turning to alternative, AI-first composable cloud platforms that offer the scalability, flexibility, and silicon diversity required for larger AI deployments.

[1]S&P Global Market Intelligence 451 Research, "The New Battleground: Unlocking the Power of AI Maturity with Multi-Model AI," commissioned by Vultr, September 2024.

Composable, AI-first clouds: the new AI innovation and scale-out engines

Hyperscalers have shaped the cloud landscape for years, providing a straightforward entry point for AI adoption. However, as AI accelerates, their closed ecosystems, once considered a convenience, are now a constraint. Proprietary lock-in, rising costs, and restricted access to best-of-breed tools make it clear: the traditional hyperscaler model alone is no longer the optimal path for AI innovation.

Enter the era of composable AI infrastructure stacks – an open, flexible approach that empowers organizations to leverage AI-first cloud environments on their own terms. By breaking free from restrictive ecosystems, enterprises gain the agility to integrate cutting-edge AI technologies, optimize costs, and scale innovation without compromise.

Today's developers and operators demand seamless, adaptable infrastructure for building and deploying complex AI workloads. Training and inference require purpose-built compute power, backed by a silicon-diverse foundation that can support both today's AI and future innovations. Meanwhile, IT leaders are focused on cost efficiency, avoiding vendor lock-in, and ensuring scalable architectures can keep pace with AI's rapid evolution.

Pro tip:

To avoid overspending by paying for cloud components and services that the hyperscalers force customers to invest in whether or not they use them, organizations should look to partner with vendors that are certified by the MACH Alliance, which identifies vendors that embrace composability.

The modern composable AI stack defined

New, AI-first composable clouds fully embrace silicon diversity. This is because general-purpose chips alone can no longer keep up with modern AI models' mounting complexity, scale, and diversity. Increasingly, specialized silicon – including high-performance CPUs, domain-specific accelerators, and energy-efficient Arm-based processors – underpins composable, open, and AI-first clouds. Silicon diversity enables data science and engineering teams to optimize cloud-native and AI workloads, enabling faster model training, improved inference speeds, and more responsive, scalable AI systems. (To learn more about the role of silicon diversity in AI inference and scale, access the Trend Advisory: *Silicon Diverse Clouds: The New Foundation for Modern, Scalable and Sustainable AI*.)

Containerized AI workflows and open-source AI orchestration come in next: Containerization has become central to composable AI, enabling portable, reproducible environments. Purpose-built container registries offer public and private repositories for open-source models, transformers, and custom-built containers. This, in turn, facilitates seamless collaboration across global teams while ensuring compliance with data sovereignty regulations. Latest developments on the orchestration front include Vultr's collaboration with dstack, which simplifies AI orchestration by replacing Kubernetes with a lightweight open-source alternative, offering intuitive job scheduling, cost tracking, and hybrid-cloud portability.

Key components of a modern, composable AI-first cloud

Beyond the core tenets of composability (modularity, atomicity, independence, and orchestratability), a composable AI stack must also address the specific requirements of AI workloads:

AI-optimized infrastructure:

Composable infrastructure must offer access to diverse hardware and specialized networking capabilities optimized for training and inference.

MLOps integration:

Composable AI stacks must seamlessly integrate with MLOps platforms for model training, deployment, monitoring, and management. This includes supporting automated workflows, version control, and model explainability tools.

Data orchestration:

Efficient data management is crucial for AI. Composable stacks must integrate with data lakes, warehouses, and pipelines to provide seamless access to training data.

Data sovereignty and privacy, security, and governance:

As AI becomes more critical, security and governance are paramount. Composable stacks must incorporate security best practices and provide tools for model governance and compliance. By offering enhanced control over data privacy and compliance, alt clouds ensure enterprises can meet stringent regulatory requirements across globally distributed environments without compromising operational agility.

Open ecosystem AI frameworks and tools:

The stack should support a wide range of open-source AI/ML frameworks and tools, allowing developers to choose the best technology for their specific tasks.

Flexibility and open ecosystems:

Unlike monolithic hyperscaler solutions, global, AI-first alt cloud platforms support open ecosystems, enabling enterprises to customize and adapt their AI infrastructure more freely. This composability allows for faster, more efficient deployment of AI models.

Cost optimization:

With significantly lower costs for CPU resources and overall infrastructure, alt clouds offer budget-friendly scalability without sacrificing performance. Enterprises can manage costs by eliminating unnecessary fees and optimizing resource allocation and then reinvest their CPU savings into GPUs for enhanced processing power.

Support for generative and agentic AI:

With flexible and affordable solutions, alt cloud platforms are uniquely positioned to power GenAI innovations, from training large language models (LLMs) to deploying creative AI applications at scale.

The benefits of a composable AI-first cloud approach

There are numerous advantages to working with cloud vendors that embrace composability:

Flexibility, agility, and no vendor lock-in:

Business requirements can change quickly. The most agile companies thrive when market conditions change and can adapt rapidly. To maintain that agility, organizations must be able to reconfigure their cloud stack on demand. Composability ensures they can add or drop components and services and work with new vendors as they need to without incurring burdensome financial penalties.

True multicloud and hybrid capability:

The openness of the composable cloud approach means organizations can choose the vendors, components, and services they want to use, which could even mean using a hyperscaler for some of their cloud operations while using cloud providers committed to composability for everything else.

Cost efficiency:

When organizations adopt a composable cloud, they can be sure they are only paying for the components and services they need when needed. This stands in stark contrast to the cloud models hyperscalers offer, as cloud components from these vendors include interdependencies that may require the organization to use (and pay for) components from vendors they aren't interested in supporting.

Future-proofing of cloud investments:

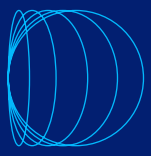
This one could be filed under "flexibility," because a composable cloud approach gives organizations the flexibility to reconfigure their cloud stack whenever needed. This means customers are free from the risk of making rigid investments in components and services that are no longer useful as their business evolves.

The strategic advantages of Vultr's composable AI-first cloud

The future of AI requires more than just powerful compute; it demands flexible, composable, and cost-effective cloud platforms that can adapt to the unique needs of enterprise-wide AI deployments. Vultr provides the expertise and infrastructure necessary to accelerate innovation, giving CTOs and developers the confidence to focus on their projects without worrying about AI pitfalls.

As the only AI-first independent cloud platform, Vultr helps future-proof AI deployments with unmatched performance. Vultr empowers enterprises to extend their existing cloud investments and seamlessly scale AI models and applications across globally distributed environments.

Trusted by 1.5 million customers across 185 countries and 32 cloud data center regions, Vultr powers enterprise-grade businesses across industries, including financial services, telecom, healthcare, retail, media, manufacturing, and energy.



The Vultr composable cloud difference

Silicon diversity

Vultr is advancing AI infrastructure with a silicon-diverse ecosystem featuring cutting-edge chips, such as AMD Instinct™ MI300X and MI325X Accelerator, and the NVIDIA GH200, HGX H100, A100, L40S, A40, and A16 GPUs. Vultr is also the first cloud provider to offer AMD Instinct™ MI325X Accelerator access. This flexibility empowers enterprises to choose the optimal hardware for their AI workloads, maximizing performance and cost efficiency.

Cost efficiency and predictable pricing

Hyperscalers' opaque pricing models often lead to budget overruns, particularly for GPU-intensive workloads. Composable stacks like Vultr's offer transparent pricing, with cloud GPU instances priced up to 50% lower than competitors.

Global scalability and low-latency inferencing

With 32 cloud data center regions spanning six continents, Vultr reaches 90% of the world's population within 2-40ms, a critical capability for real-time AI applications like fraud detection and IoT analytics.

Serverless inference

The recent launch of Vultr's cloud inferencing services repurposes edge GPUs for localized model deployment, reducing bandwidth costs and improving responsiveness. With Vultr's serverless model, compute is automatically optimized for edge and inference workloads – so you can focus on innovation, not infrastructure. Burst into the cloud with unlimited scalability and control and stay ahead with the latest technologies.

Vendor neutrality and ecosystem integration

MACH Alliance-certified, Vultr has built a robust Cloud Alliance program featuring partnerships with NetApp, AMD Console Connect, DDN, dstack, and Qdrant, among others, to provide turnkey integrations for AI model acceleration, data management, private networking, and vector databases, eliminating reliance on proprietary hyperscaler ecosystems.

API-first architecture and Kubernetes-native flexibility

Vultr's API-first design and fully managed Vultr Kubernetes Engine (VKE) enable seamless orchestration of AI workloads, allowing enterprises to deploy, scale, and manage containerized applications with ease. With Vultr's Container Registry, users can securely store and retrieve container images, ensuring smooth integration with the technologies of their choice. Terraform automation further enhances this flexibility, enabling teams to define and provision their cloud infrastructure as code, streamlining deployment and ensuring consistency across environments. This composability empowers teams to build AI stacks tailored to their needs without vendor lock-in or hyperscaler constraints.