

How to accelerate AI innovation with containers: A guide for technology leaders

How Google Kubernetes Engine
can help you succeed with AI.



Introduction

AI is here. Is your infrastructure ready?

AI is fundamentally reshaping industries and competitive landscapes and tech leaders face immense pressure to deliver impactful AI solutions at lightning speed. Yet, the very infrastructure meant to enable this innovation often becomes the bottleneck—complex, costly, and struggling to meet the demands of modern AI.

The good news is that you don't need to reinvent your approach. Your existing containers and Kubernetes skills, and investments already give you a strong foundation for AI workloads. With Google Kubernetes Engine (GKE), your teams can turn infrastructure into a strategic advantage, empowering them to excel in this new era of AI.

This ebook shows you how to use your existing Kubernetes investments and rapidly transform your infrastructure into a high-performance engine for AI innovation.



Table of contents



1. AI's reality check: Five challenges	4
2. The opportunity: Transform vision into scale with AI and containers	7
3. Unleashing AI at scale is a top priority	9
4. Cost-effective AI inference is the path to profitability	13
5. Empower every AI role with seamless infrastructure for business success	19



Chapter 1

AI's reality check: Five challenges

Generative AI (gen AI) and large language models (LLMs) are revolutionizing industries, unlocking new capabilities, and accelerating innovation. But this explosive growth comes with a price—relentless pressure to balance costs, performance demands, and security. The core challenge? Keeping pace with AI's constant evolution while navigating growing infrastructure complexities. As AI models scale and workloads increase, key trends are driving technology leaders to critically re-evaluate their infrastructure strategy.

1 As gen AI adoption surges, platforms are becoming bottlenecks.

Despite 98% of organizations actively experimenting with or deploying gen AI, a staggering 45% of global business leaders struggle to build a robust, scalable AI platform.¹ This isn't just a technical glitch, it's a direct threat to missed opportunities for competitive advantage. It delays market entry for critical products and creates inefficient resource allocation that diverts valuable budgets away from core business initiatives. Delivering a scalable ML platform that serves diverse users—from developers and scientists to IT operators—demands precise planning and the right accelerators. It also requires infrastructure that can manage traffic at scale, deliver superior price performance, and ensure observability without slowing innovation or organizational agility.

¹ Google Cloud, 2025, [State of AI infrastructure research report](#)



2 Model sizes know no bounds and infrastructure is on the hook.

AI models are growing exponentially in size and intelligence, and companies must keep pace. 45% of global business leaders say that efficient scalability is one of the most important factors when evaluating a gen AI infrastructure system.¹ But this level of growth not only amplifies infrastructure costs and complexity, it demands massive computing power to train and run, forcing businesses to rethink how they build and scale their infrastructure. The relentless growth in AI model sizes can translate into spiraling costs and a substantial operational burden.

3 Cost efficiency is critical for delivering on AI initiatives.

83% of tech leaders say cost is a key factor when evaluating solutions.² Harnessing AI can become a minefield of unexpected expenses driven by scale, model complexity, data volume, and talent. Businesses need a cloud provider that can offer them cost-efficient solutions. Unforeseen AI costs can impact the bottom line and force difficult trade-offs between innovation and budget. Effective cost management is paramount for AI initiatives to deliver tangible ROI, not just consumption.

4 We're in a scarcity crisis of AI accelerators.

Right now, the demand for hardware accelerators such as GPUs is critically outpacing supply, creating a global hardware bottleneck. It is estimated that \$300B will be spent on software, hardware, and services for AI systems in 2026.³ This scarcity is causing prolonged development cycles, inflated costs, and an inability to meet the pace of rapid AI innovation. To do more with less, you need smarter ways to maximize your existing hardware and secure future capacity.

¹ Google Cloud, 2025, [State of AI infrastructure research report](#)

² Google Cloud, 2025, [State of AI infrastructure research report](#)

³ Tech Monitor, 2023, [AI spending to double to more than \\$300bn by 2026](#)



5 The open-source ecosystem is expanding —avoid lock-in.

76% of tech leaders expect increased adoption of open-source AI technologies in the years ahead.⁴ As new tools, frameworks, and APIs emerge almost weekly, businesses need seamless access, without getting locked into proprietary platforms. The right infrastructure should enable agility and collaboration, while safeguarding long-term investments as the open-source ecosystem keeps evolving.

Today's AI challenges:



- As gen AI adoption surges, platforms are becoming bottlenecks
- Model sizes know no bounds and infrastructure is on the hook
- Cost efficiency is critical for delivering on AI initiatives
- We're in a scarcity crisis of AI accelerators
- Keeping pace with the expanding open-source ecosystem

⁴ McKinsey Digital, [Open source in the age of AI](#)



Chapter 2

The opportunity: Transform vision into scale with AI and containers

AI pioneers are pushing the boundaries of what's possible, and platform builders are making those innovations real at scale. For organizations with container expertise, success isn't about starting over—it's about using existing DevOps and container strategies as immediate advantages for AI initiatives. As AI workloads ramp up, balancing speed, scalability, and cost becomes non-negotiable.

Containers are well suited to serve the unique demands of AI. They provide the consistency, portability, scalability, and isolation needed to develop, deploy, and manage today's complex and resource-intensive AI/ML workloads. Recognizing these advantages and anticipating disruptive leaps of technology, Google Cloud has invested years into developing a robust managed container platform built to handle these big leaps. Given the evolving demands of AI workloads, GKE is ideally positioned to meet this opportunity and is built to meet the demands of modern AI workloads. It offers advanced orchestration and management, while reducing the operational challenges that come with scaling and deploying new AI solutions.



Rising above infrastructure

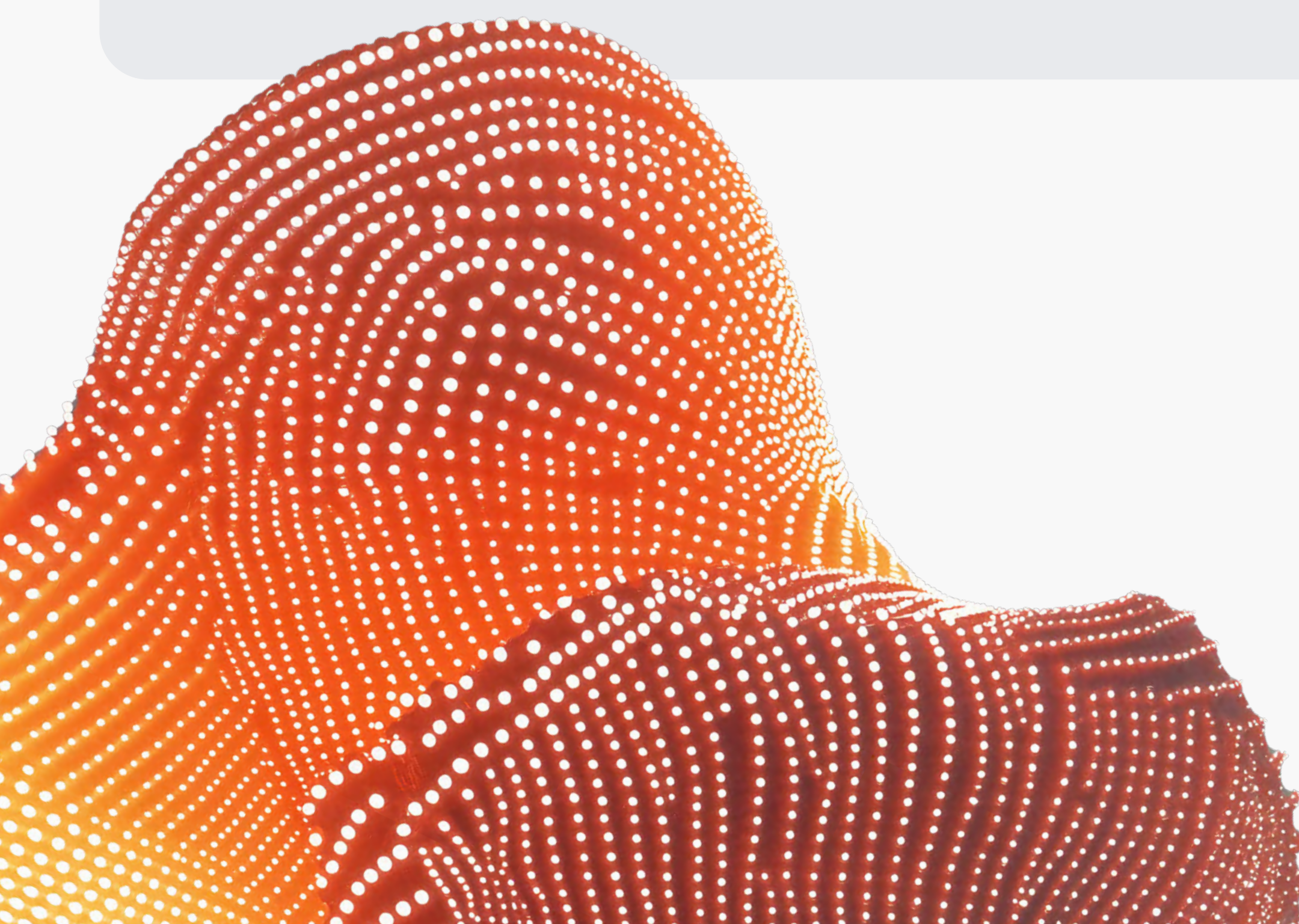
As infrastructure evolves—from raw compute and GPU time-sharing to TPUs—containers allow teams to instantly leverage new capabilities on their existing platforms.

Orchestrating complex workloads

Containers now support far more than just stateless workloads, they can also optimize a wide range of applications, helping both users and platform teams manage complexity. That's why Google uses GKE for our own breakthrough AI products like Vertex AI, and to power the next generation of AI innovation with DeepMind.

Supporting extensibility

Kubernetes' extensibility fuels a broad ecosystem of tools and frameworks, giving teams the flexibility to innovate quickly and adapt to AI's fast-changing landscape.





Chapter 3

Unleashing AI at scale is a top priority

Scalability with containers, particularly orchestrated by Kubernetes, is absolutely crucial for AI success. This flexible, efficient foundation allows you to develop, deploy, and manage AI solutions at scale, unlocking competitive advantages, faster innovation, reduced operational costs, and the ability to deliver high-performing, reliable AI products.

Scale has always been fundamental to the evolution of container-led technology. Kubernetes originated from Google's own need to automate the management of billions of containers across its vast infrastructure. Today's AI workloads are demanding, from training massive LLMs to serving real-time inferences for millions of users, computational requirements are immense and can spike unpredictably. Traditional infrastructure often struggles to keep up with these dynamic scaling needs, leading to bottlenecks, underutilization, or costly overprovisioning. Containers provide the essential foundation for scaling, and GKE amplifies this with intelligent, streamlined orchestration for even the most demanding AI workloads. Your scaling needs will only grow and GKE is built to keep pace.



GKE provides the foundation for AI at scale.

GKE empowers your teams to accelerate time-to-market, reduce training time, overcome hardware limitations, and optimize costs. As models become more complex, the ability to distribute workloads across thousands of accelerators becomes a competitive advantage. This is precisely what GKE was built for, with Autopilot mode enabling rapid innovation without the need for deep Kubernetes experience.

GKE is breaking records with support for up to

65,000

nodes, and counting⁵



Innovate faster with unparalleled scale.

The most advanced AI models demand unprecedented scale with tens of thousands of AI accelerator chips working in parallel. With its industry-leading scale, GKE allows organizations to accelerate training times and push the boundaries of multi-trillion-parameter models. Today, speed to market is everything and GKE gives you the scale to innovate faster and gain a critical competitive edge—not just for today, but also for your future AI needs.

“

GKE’s new support for larger clusters provides the scale we need to accelerate our pace of AI innovation.”

James Bradbury
Head of Compute, Anthropic

⁵ Google Cloud, 2024, [65,000 nodes and counting: Google Kubernetes Engine is ready for trillion-parameter AI models](#)



Boost performance, lower costs.

Unlock massive savings and faster throughput. GKE supports both NVIDIA GPUs and Cloud TPUs, giving you the flexibility to choose the right accelerator for every workload. Whether training with GPUs or TPUs, GKE simplifies optimizing price performance at scale. Combined with built-in automation and observability, these options deliver significant cost savings and faster training, directly improving your ROI.

“

By harnessing Google Cloud’s solutions, such as GKE and Cloud TPU, Moloco dramatically reduced ML training times by up to tenfold.”

Sechan Oh

Director of Machine Learning, Moloco

Enable smarter resource allocation.

AI training at scale means many teams and workloads are competing for the same limited accelerator pool. GKE eliminates bottlenecks with native support for Kueue and Dynamic Workload Scheduler (DWS). Together, these features enable teams to share expensive resources fairly, queue jobs efficiently, and tap into additional capacity. This drives more efficient training and cost reduction, without slowing innovation or causing staff burnout from resource contention.

How LiveX AI reduces customer support costs by up to 85%

LiveX AI chooses GKE as a robust foundation for their AI agents.

The company:

LiveX AI builds custom, multimodal AI agents that can see, hear, and chat, delivering truly human-like experiences and allowing businesses to deliver strong customer engagement across a variety of platforms.



They needed:

A robust solution that would help them ramp up quickly, and deploy and operate containerized applications at scale on a secure and performant global infrastructure.

They chose:

GKE's platform orchestration capabilities to train and serve optimized AI workloads on NVIDIA GPUs, while also leveraging GKE's flexible integration with distributed computing and data processing frameworks. In particular, GKE Autopilot makes it easy to scale applications to different clients, especially when building multimodal AI agents for brands with high volumes of real-time customer interactions. GKE Autopilot manages the underlying compute of a Kubernetes cluster without LiveX AI needing to configure or monitor it.

They achieved:

85%

lower customer support costs

50%

lower TCO

25%

faster time-to-market

66%

lower operational cost

“

By combining the power of NVIDIA NIM with Google Kubernetes Engine, we've achieved a 6.1x increase in average token speed. This means, LiveX AI can deliver personalized experiences for customers in real-time, including seamless customer support, instant product recommendations, and reduced returns.”

Jia Li

Chief AI Officer & President, LiveX AI

Watch the video →

Read the case study →

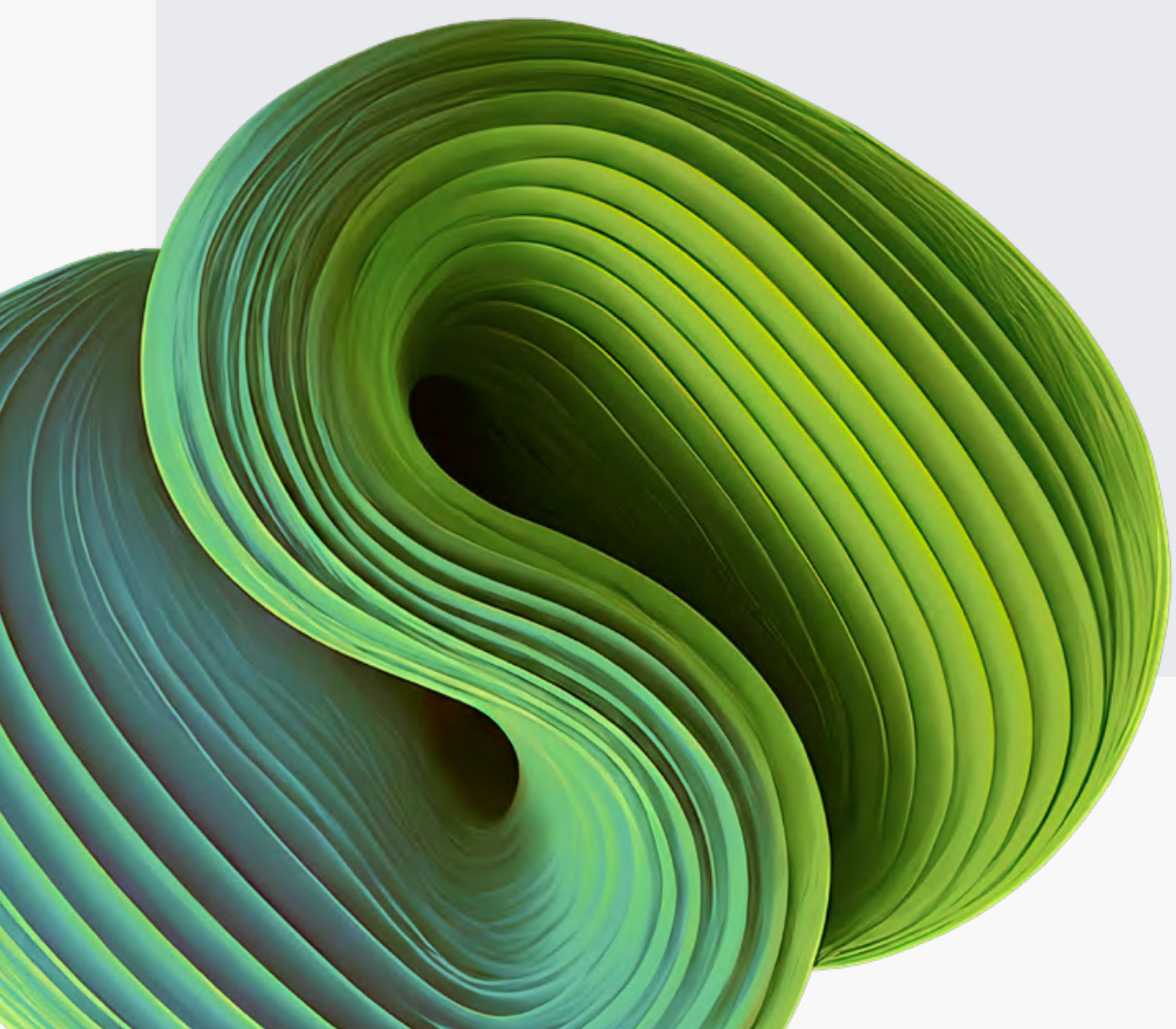


Chapter 4

Cost-effective AI inference is the path to profitability

While AI model training often gets attention for its significant computational demands, inference is where the ongoing, recurring costs are incurred. These costs can quickly spiral out of control as usage scales.

Containerization provides inherent advantages for efficient AI inference, including portability, consistency, dynamic scaling, elasticity, and resource efficiency. However, even with containers, massive AI models introduce new challenges. For example, huge container images can lead to cold start latency spikes. To compensate, teams often overprovision resources to handle unpredictable loads, which in turn exacerbates cold-start issues and leads to underutilization and unnecessary costs. Different phases of the inference journey pose different cost challenges.





GKE provides the foundation for cost-effective inference.

Deploying gen AI inference introduces unique complexities. First, you need to choose the right accelerator for your use case. The evaluation phase of the journey is where you should assess the pros and cons of all your accelerator options. Once you're in production, you'll need to load-balance traffic, manage price performance at scale and with real traffic, monitor performance, and debug any issues that arise. GKE's new inference capabilities are purpose-built to overcome these hurdles.

GKE's new inference capabilities deliver:

30%

reduction in
serving costs

60%

reduction in
tail latency

40%

increase in
throughput

These improvements translate directly to superior user experiences, significant operational savings, and faster delivery of AI-powered features.

Quickstart your inference journey in days, not months.

Choosing the right accelerator and model server, and scaling setup for your inference workloads can be complex and time-consuming. [GKE Inference quickstart](#) simplifies this process with benchmarks, compatibility guides for GPUs and TPUs, and pre-configured scaling recommendations. Soon, an integration with Gemini Cloud Assist will offer faster, smarter deployment and optimization, accelerating your time-to-value for new AI-powered applications.





Elevate the user experience and reduce costs with AI-aware Inference Gateway.

LLMs exhibit highly variable request patterns, often leading to high-tail latencies and inefficient compute utilization. This can directly impact the user experience and unnecessarily inflate inference costs. The traditional Gateway does not support routing infrastructure for popular Parameter-Efficient Fine-Tuning (PEFT) techniques like Low-Rank Adaptation (LoRA)—which are critical for optimizing GPU efficiency through model reuse during inference.

The new GKE Inference gateway provides AI-aware load balancing for optimal routing. It also supports LoRA, enabling you to map multiple models to the same underlying service for superior efficiency, reduced inference costs, and a superior user experience.



Optimizing price-performance for generative AI inference is key for our customers. We are excited to see GKE Inference Gateway with its optimized load balancing and extensibility in open-source. The new GKE Inference Gateway capabilities could help us further improve performance for our customers' inference workloads."

Chaoyu Yang
CEO and Founder, BentoML

Reduce costs while simplifying operations.

GKE's Autopilot mode empowers you to build and deploy faster without deep Kubernetes expertise. Google SREs back the resources your Kubernetes pods request, significantly reducing day 2 operational burden. GKE autoscaling and node auto-provisioning capabilities seamlessly leverage Spot GPUs and TPUs at up to 90% cheaper,⁶ and fall back to on-demand instances when needed.

⁶ Google Cloud, [Spot VMs pricing](#)



A white double quote icon inside a green circle.

Scaling our infrastructure as Moloco's Ads business grew exponentially was a huge challenge. GKE's autoscaling capabilities enabled the engineering team to focus on development without spending a ton of effort on operations."

Sechan Oh
Director of Machine Learning, Moloco

How HubX brings superior experience to their customers

HubX balances inference cost, latency, and scale with GKE for their gen AI mobile apps.

The company:

Using world-class expertise and cutting-edge tech, HubX builds highly-scalable mobile apps—some of which include Nova, Davinci, and Momo. Three of HubX's applications are listed among the top gen AI mobile apps with the most active users worldwide, and all of these applications run on Google Cloud.

They needed:

A platform that could help them quickly scale, deploy, and manage a large number of AI powered applications. For HubX, optimization was a balancing act, where inference latency, scale, and cost all came into play.



They chose:

A familiar platform like GKE—allowing their teams to build and deploy AI-powered applications with agility, speed, and scale. They also needed to ensure their developers could pivot strategies and scale quickly to stay relevant in a fast-paced digital world. That’s why they built their infrastructure on Google Cloud, leveraging GKE as the core orchestrator, and tapping into the power of Trillium TPU.

They achieved:

40%

reduction in costs

2.5x

faster inference speeds

20–30x

faster boot-up times

How it works:

01

A user request hits our front-end, hosted on the Cloud Run service

02

The request containing the user images for training is passed via Pub/Sub for asynchronous processing

03

For personalization, a training job is triggered within our GKE cluster

04

Training nodes generate user-specific LoRAs, which need to be quickly available for inference

05

HubX uses FileStore, mounted directly to the GKE nodes, for fast, low-latency access to these LoRAs

06

Immediately after training, and potentially later on, an inference request is routed to the inference nodes within the same GKE cluster

07

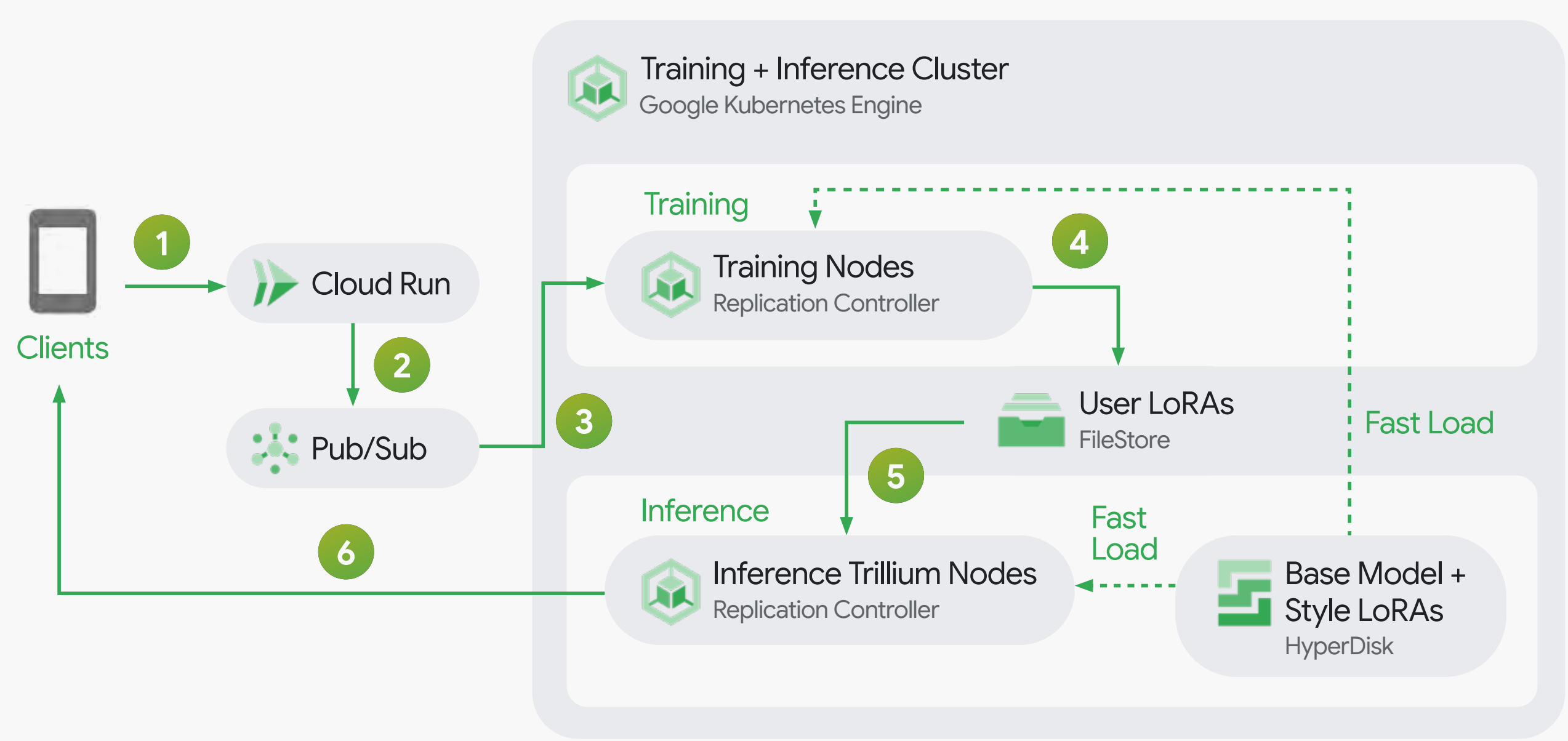
These nodes, that use Trillium TPUs, load the necessary base model and the specific user and/or style LoRAs

08

They process the request, generating the images which are then returned to the client



GKE provides the flexibility to manage distinct node pools for training and inference, handle scaling, and integrate seamlessly with other GCP services like Pub/Sub, FileStore, and Persistent Disks like HyperDisk for base models.



“

By leveraging Google Kubernetes Engine and its cutting-edge features, we’ve not only optimized performance but also enhanced cost efficiency —allowing us to innovate faster and serve our customers better.”

Kaan Ortabaş
Co-founder, HubX



Chapter 5

Empower every AI role with seamless infrastructure for business success

Platform engineers trust Kubernetes for its powerful container orchestration, resource isolation, and auto-scaling capabilities. With increased AI usage, these same platform teams now need to serve a new user base—data scientists and AI/ML engineers. However, most data scientists and AI/ML engineers aren't familiar with Kubernetes and need a simpler, more approachable way of interacting with distributed infrastructure. To curb the steep learning curve, many organizations have widely adopted Ray, an open-source framework that gives AI/ML engineers the ability to develop Python code on their laptops and then scale it elastically across a Kubernetes cluster. To enable these users, Google and Anyscale partnered on KubeRay to support OSS Ray deployments on Kubernetes. With KubeRay, data and ML scientists can focus on computation, while infrastructure engineers can focus on Kubernetes.





Google’s commitment: Making Kubernetes the premier platform for Ray

Committed to making AI applications as straightforward to build and run as writing Python code, Google and Anyscale deepened their partnership to make Kubernetes the best platform for Ray with RayTurbo on GKE. RayTurbo is an optimized version of open-source Ray that boosts task execution speed, throughput, and GPU and TPU utilization. And now on GKE it will feature enhanced performance, better resource efficiency, and higher performance storage options.

Ray and GKE—better together

Together, RayTurbo and GKE form a distributed operating system purpose-built for AI. This enables teams to build, deploy, and scale applications by rising above the infrastructure. The advantages are transformative, delivering optimized performance and efficiency, and significant cost reductions. These results translate directly to faster experimentation, lower operational costs, and accelerated time-to-value for your most critical AI initiatives.

- ✓ **4.5x** faster multimodal data processing⁷
- ✓ **54%** higher throughput in model serving⁸
- ✓ **50%** fewer nodes for online model serving⁷



How IPRally is innovating patent search

IPRally leverages AI on GKE and Ray for customization and efficiency.

The company:

IPRally uses the power of AI to handle the complexity of patent searches and analysis, resulting in an intelligent search engine that operates like a patent professional and delivers results in a few seconds.

They needed:

A solution that would help them resolve a case faster than traditional patent search tools.

They chose:

GKE and Ray to help them build a customized ML platform for performance and efficiency—one that would streamline their MLOps, make technical knowledge more accessible, and scale with them.

They achieved:

70%

lower ML RnD costs with spot instances

20%

faster start-up time for a training job⁹

“

We have seen that container image streaming in GKE has a significant impact on expediting our application startup time. Image streaming helps us accelerate our start-up time for a training job after submission by 20%.”

Juho Kallio
CTO and Co-founder, IPRally

Google Kubernetes turns 10

It's been 10 years since Kubernetes began empowering organisations and developers to manage and scale containerised applications.

To mark the anniversary, read the 'GKE turns 10' ebook and discover the journey from its inception to empowering modern applications.

[Read it here →](#)

Watch how customers are using GKE to achieve remarkable results with their AI/ML workloads.

[Watch the video →](#)



Ready to build your infrastructure on GKE?

Get started for free today →

