Modernizing infrastructure for the age of GenAI

At NTT DATA, we see firsthand how generative AI (GenAI) is revolutionizing business for our global financial services customers. We also see that, to maximize the power of AI, they need the robust, scalable and cost-effective foundation a modern cloud infrastructure provides.

This eBook from our partner Google Cloud describes the how, why and value of rearchitecting your IT infrastructure for GenAI. Applying the learnings from this eBook is where NTT DATA's expertise is invaluable. Our Google Cloud consulting and integration services expertly manage GenAI workloads for critical financial systems, including core banking, payment processing, risk management, regulatory reporting and trading platforms. Working in close partnership, we guide you through every stage of your journey, offering:

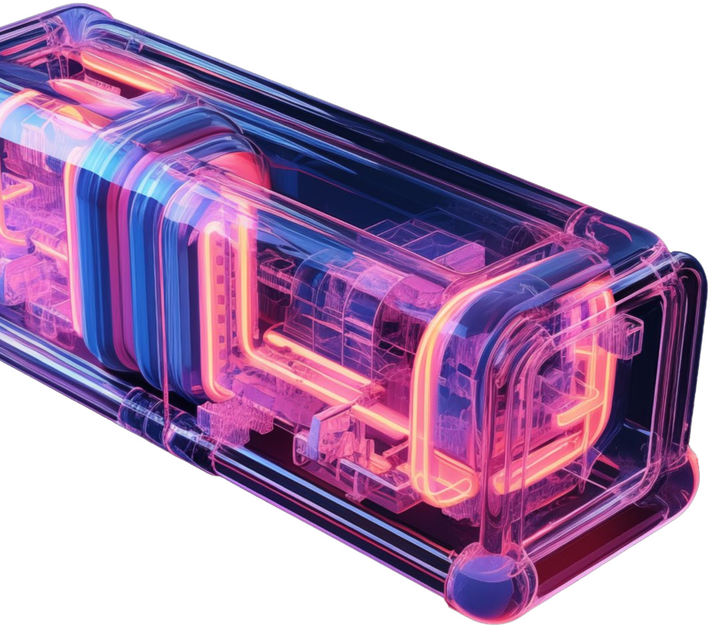
- **Strategic planning:** IT assessment and collaborative goal definition.
- **Solution design:** Optimal Google Cloud service selection.
- **Seamless implementation:** Smooth AI model integration.
- **Proactive support:** Ongoing monitoring and resource optimization.
- **Robust security:** Resilient, compliant security measures.
- **Proven expertise:** Best practices, leading platforms and strong partnerships.

Ready to optimize your infrastructure for GenAI to achieve breakthrough results? Visit our website to learn more about NTT DATA and Google Cloud's partnership and our capabilities for financial services.

Rearchitecting your infrastructure for generative AI

Google Cloud | NTT DATA

Contents



Executive summary

As generative AI propels organizations into the future, IT leaders must construct infrastructure able to withstand the performance requirements these revolutionary technologies bring. With the exponential growth in data generation, model size, and computation demands, existing infrastructure cannot handle the requirements of training and serving Large Language Models (LLMs) like PaLM2. To avoid AI efforts stalling out due to inadequate foundations, you'll need to architect for scalability and performance from the ground up.

This guide provides technology leaders a real-world roadmap for architecting robust generative AI systems, helping inform strategic decisions that transcend your current organization. We will dive into the infrastructure considerations that can determine AI success or failure—examining cost, scalability, security, and performance dimensions.

Within, you'll discover actionable strategies to evaluate AI platforms, optimize resources, and maximize the value of your AI tools.

For development and deployment, we outline paths leveraging managed machine learning offerings like Vertex AI and flexible container environments like Google Kubernetes Engine (GKE). To power the demanding generative workloads, this paper also outlines best practices for leveraging specialized virtual machines (VMs) optimized for AI and equipped with GPUs, and TPUs. Finally, how by tapping into Google Cloud's secure and interoperable virtualization resources, you can develop and run generative AI applications wherever they're needed.

Architecting infrastructure for performance, agility, and scale is imperative to unlocking generative AI's full potential. This paper offers technical leaders a roadmap to build a strong foundation for AI innovation powering the next generation of their business.

Chapter 1 | Introduction

Embracing the power of generative AI

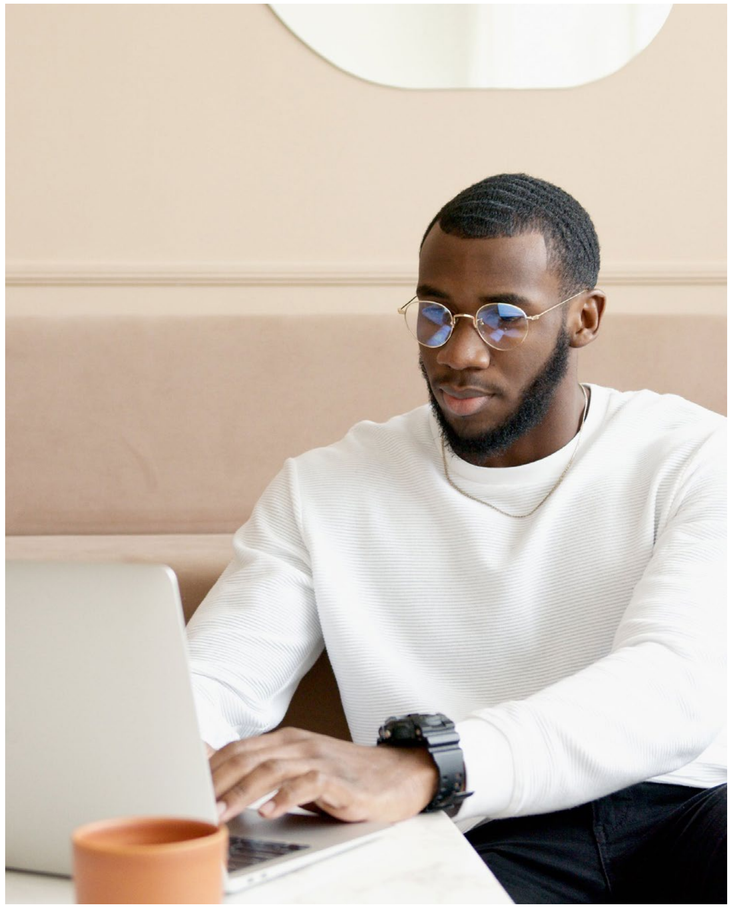
As we venture further into the era of digital transformation, there is one concept that is rapidly changing the landscape of technology, business, and society at large—generative AI. This powerful technology, with its ability to learn from data and generate predictive outputs, presents an unprecedented opportunity for technology leaders. It is not an overstatement to say that generative AI is both the future—and a revolution that is disrupting industries in real-time.

Infusing generative AI throughout an organization can significantly transform the way businesses operate and compete, enabling them to unlock innovative solutions, automate processes, improve decision making, and deliver personalized customer experiences.

However, to realize these benefits, a solid, reliable, and AI-optimized infrastructure is crucial. With the right infrastructure in place, organizations can implement an agile, flexible technology stack optimized for rapidly evolving AI models. And, with careful planning and execution, companies can lay a strong technical foundation to capitalize on the promise of AI.

“We’re offering Google Cloud’s industry-leading infrastructure, Google foundation models and AI tooling to companies across industries so they can build, train and deploy the future of AI creatively, reliably and at scale.

Thomas Kurian,
CEO of Google Cloud



Navigating the new age of generative AI

The onset of generative AI is rapidly driving innovation across sectors, from healthcare and finance to entertainment and education.

By automating rote tasks and content generation, these AI models significantly enhance efficiencies, reduce operational costs, and open up new avenues for user engagement. For instance, generative AI can help create personalized digital content based on user preferences, generate code from natural language, or effortlessly summarize and contextualize information for faster decision making.

However, as with any technological revolution, implementing such powerful AI models presents its own challenges. Traditional computing infrastructure, built for a pre-AI era, is unable to provide the computational resources and scalability needed by these AI models. The newest iteration of AI models requires radically different architecture to meet the exponential growth in computing demands.

The AI revolution: Why performance needs have skyrocketed

With billions of parameters and vast training datasets, generative AI and LLMs necessitate unprecedented computational resources. They already exceed the capabilities of traditional infrastructure, needing more processing power, faster memory, larger storage, and blazing fast network connectivity.

Purpose-built AI infrastructure can help to ensure the full power of generative AI, providing the robust, high-performance computing capability needed to support these advanced models. Moreover, it can allow organizations to scale their AI initiatives efficiently, enabling them to adapt to changing business needs and seize new opportunities in the vibrant AI landscape.

Chapter 2

Understanding infrastructure capabilities: Cost, performance, and scalability

Cost: How can I prioritize cost-effectiveness?

A successful AI infrastructure strategy must prioritize cost-effectiveness. As generative AI models and LLMs require considerable resources for training and development, these costs can quickly get out of hand without a carefully planned and managed infrastructure.

Choosing the right cloud provider, optimizing resource utilization, and leveraging AI-specific tools and features can significantly reduce these costs while ensuring your AI models run efficiently. The opportunity cost of inadequately investing in an AI-optimized infrastructure can be exceedingly high, resulting in poor performance, difficulty scaling, and missed opportunities

“Google Cloud is a true partner that gives us the stability and flexibility to support critical business applications needed to drive innovation and ensure business continuity. The migration to Google Cloud has been seamless and was a key project in our merger integration and modernization efforts as Keurig Dr Pepper.”

John Gigerich,
SVP and CIO for Keurig Dr Pepper

Performance: How can I deliver speed and efficiency?

Modern workloads are outpacing existing on-premises infrastructure, and with Moore's Law slowing, hardware alone cannot keep up. Instead of accepting limitations, you need an adaptable, software-defined approach to infrastructure that holistically optimizes every layer of the stack.

For data-driven companies, the solution is flexible cloud platforms designed to efficiently meet surging workload demands.

Scalability: How can I pave the way for growth?

In the rapidly evolving AI landscape, scalability is key. AI models are ever-growing, with some projected to surpass hundreds of billions of parameters. These models will require tens of ExaFLOPs (1018 FLOPs) of AI supercomputing to maintain training times of several weeks or less.

Achieving that performance will require tens of thousands of accelerators working efficiently together. However, traditional scaling solutions often prove insufficient, requiring complex manual tuning and resulting in suboptimal performance.

Thus, finding an infrastructure solution that seamlessly scales with your AI needs is vital to avoid these pitfalls and harness the full power of generative AI. Leveraging cloud-based solutions can provide the necessary scalability while ensuring cost-effectiveness and high performance.

Google Cloud ran the world's largest distributed training job for large language models, using 50000+ TPU v5e chips

“Cloud TPU v5e consistently delivered up to 4X greater performance per dollar than comparable solutions in the market for running inference on our production ASR model. The Google Cloud software stack is well suited for production AI workloads and we are able to take full advantage of the TPU v5e hardware which is purpose-built for running advanced deep learning models. This powerful combination of hardware and software dramatically accelerated our ability to provide cost effective AI solutions to our customers.”

Domenic Donato,
VP of Technology, AssemblyAI

Navigating your generative AI deployment: A checklist to drive success

1 Assess compute requirements:

Generative AI requires significant GPU/TPU performance, estimate workload needs to ensure adequate capacity.

2 Evaluate data pipelines:

Quality data is critical for training generative models. Audit data sources, ETL processes, labeling etc.

3 Implement MLOps:

To build, deploy and monitor generative AI models, MLOps processes like version control, experiment tracking, and model monitoring need to be in place.

4 Assess model risks:

Generative models come with risks like bias, toxicity, and hallucinations. Put guardrails in place through testing and monitoring.

5 Evaluate AI ethics:

Consider potential harms from generative models and mitigate via ethics reviews procedures.

6 Audit security posture:

Generative models can create security risks. Review IAM, network security, user authentication, and access controls.

Navigating your generative AI deployment: A checklist to drive success

7 Plan for scalability:

Design infrastructure for rapid scaling of compute, storage and network to meet growing demands.

8 Enable collaboration:

Generative AI requires collaboration between data scientists, engineers, business teams, and technical leaders. Ensure tools are in place.

9 Consider platforms:

Leverage cloud-based AI platforms like Vertex AI to accelerate development with pre-trained models from Google Cloud and its partner ecosystem.

10 Develop responsible AI principles:

Create and commit to a responsible series of principles aligning to your organization's values.

11 Invest in skills development:

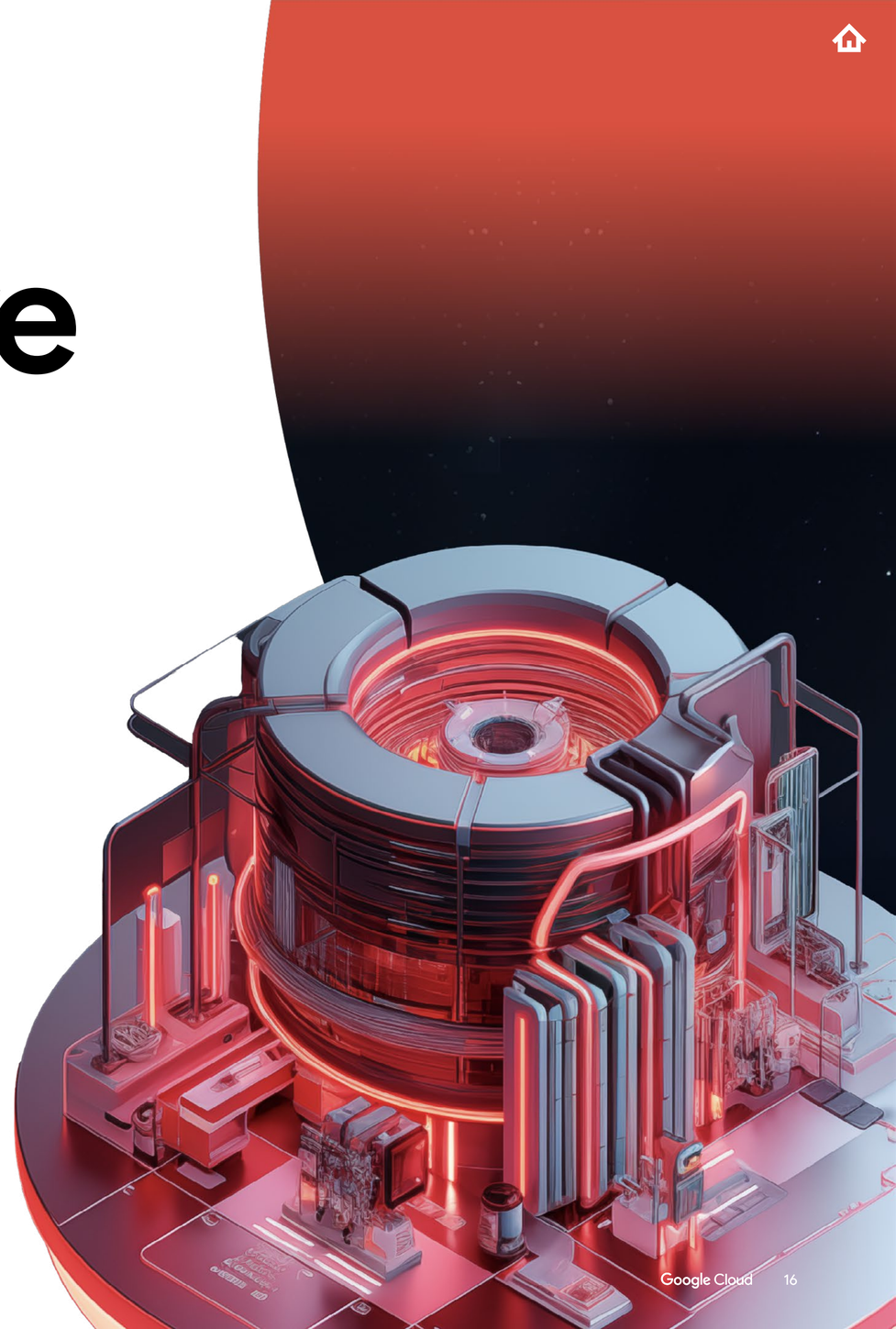
Sponsor training in MLOps, prompt engineering and learning paths about generative AI to increase familiarity among employees.

12 Connect with a trusted technology partner:

Talk to a technology provider like Google Cloud to leverage their technical expertise to ensure that your deployment is not only technically sound, but aligns with your organization's business needs and strategic objectives.

Chapter 3

Building a generative AI platform: Key components and strategies



Having successfully identified and addressed crucial infrastructure requirements, focus shifts to the choice of AI platform. This critical layer acts as an intermediary that not only integrates seamlessly with your existing infrastructure, but also provides essential access to advanced AI models and tools, pivotal for effective training and inference.

Your choice of AI platform is more than a technical decision, it's a strategic one, directly influencing the agility and adaptability of your AI projects. An ineffective AI platform can lead to cost overruns, model deficiencies, and a reduction in competitive edge.

Should I choose managed services, or build my own?

When developing AI applications, organizations have two primary options: managed ML services, or building their own applications.

Managed services, such as Google Cloud's Vertex AI, provide turnkey access to pre-trained models, enabling faster time-to-market and reduced operational overhead. However, custom applications allow for more flexibility to integrate proprietary models and data pipelines.

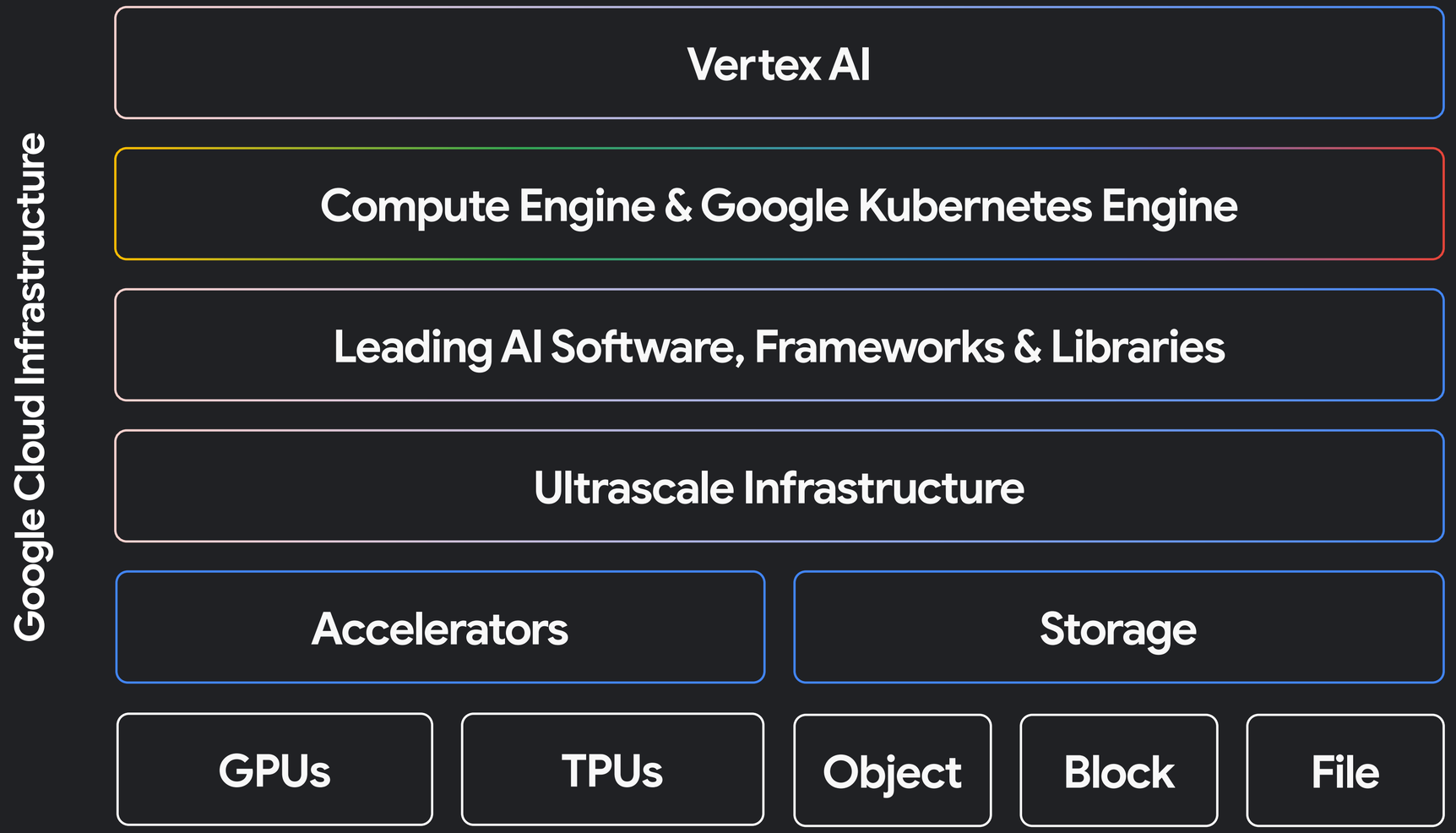
Each approach empowers organizations in different ways depending on their generative AI workloads, and each method has their benefits and drawbacks.

In this guide, we explore the considerations you need to build your own custom solution.

“Easily one of the most exciting features for our development team is the unified toolset. The amount of wasted time and hassle they can avoid by not having to make different tools fit together will streamline the process of taking AI from idea to training to deployment. For example: Configuring and deploying your AI models on Google Kubernetes Engine (GKE) and Compute Engine along with Google Cloud TPU infrastructure enables our team to speed up the training and inference of the latest foundation models at scale, while enjoying support for autoscaling, workload orchestration, and automatic upgrades.”

Yoav HaCohen,
Core Generative AI, Team Lead - Lightricks

Building a generative AI platform: Key components and strategies



DIY solutions: Unlock the full benefits of generative AI

AI models thrive when given flexible, scalable environments to call home, and solutions such as virtual machines (VMs) and containers provide customizable and optimizable platforms to run your generative AI workloads.

With Compute Engine's scalable virtual machines, your data science team can spin up environments tailored for large-scale model training in seconds. While GKE's dynamic containers enable seamless deployment of resource-hungry models across hybrid, multi-cloud, and edge environments.

Purpose-built AI infrastructure like TPUs optimize and accelerate machine learning workloads through parallel processing and tensor computations

For example, Google Cloud demonstrated the world's largest distributed AI training job for large language models across 50000+ TPU v5e chips that are capable of achieving 10 exa-FLOPs (16-bit), or 20 exa-OPs (8-bit), of total peak performance.

How do I choose the right tool for generative AI?

When it comes to gen AI, both VMs and containers can play a role, depending on the specifics of your use case. For instance, if you are working with highly sensitive data, a VM's superior isolation might be preferable. Alternatively, for large-scale deployments with many replicated tasks, or apps needing to scale up or down dynamically and quickly, the lightweight and scalable nature of containers could be more suitable.

The choice between VMs and containers comes down to your specific needs in terms of security, isolation, resource utilization, and scalability. In some cases, a hybrid solution containing both VMs and containers may be optimal. Understanding these considerations will help guide your decision and enable efficient and effective deployment of gen AI.

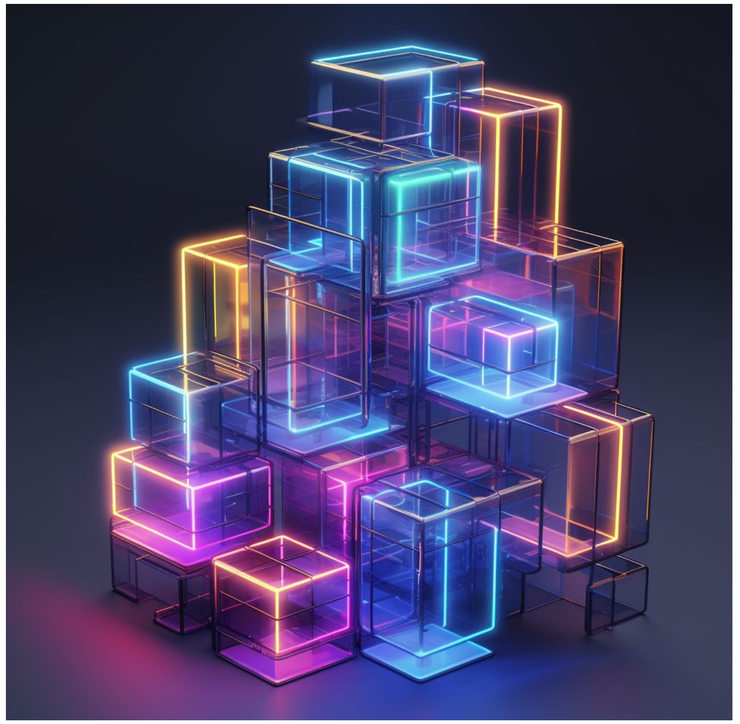
"Increasing the productivity and velocity of our product and engineering teams was paramount. We needed to be more flexible for customers and make it easier for them to search, price, and book faster."

Martin Brodbeck,
CTO, Priceline

Read the full story

Virtual machines: Building blocks of computation

VMs are emulations of physical computers, each with its own operating system (OS) and resources. They provide a high degree of control, making them ideal for running applications that require specific OS or hardware configurations.
One of the key benefits of VMs is their isolation. Because every VM operates independently, they provide an extra layer of security by isolating the application and its dependencies. This makes VMs a good fit for running large-scale, complex applications with multiple dependencies.



Containers: The lightweight alternative

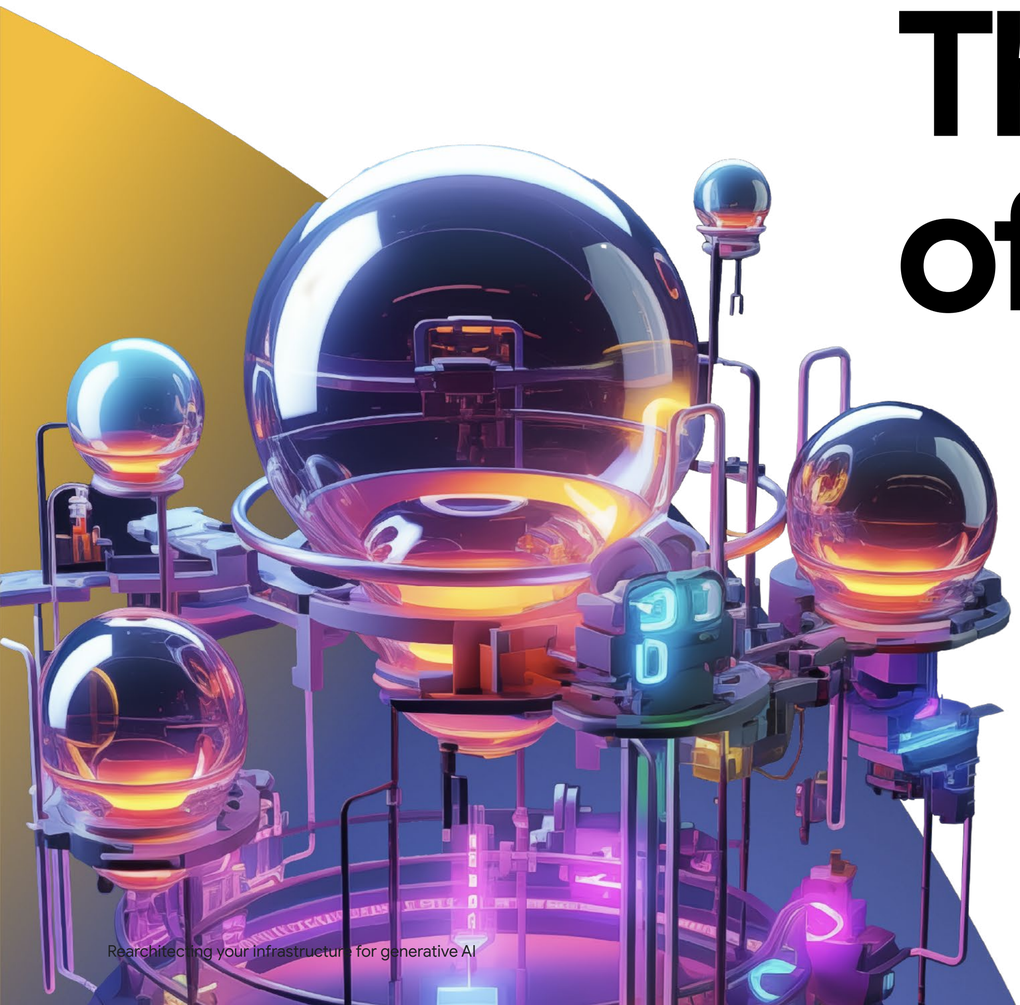
Containers, on the other hand, are a more lightweight compute option. They package up the code and its dependencies so the application runs quickly and reliably from one computing environment to another. Unlike VMs, multiple containers can run on the same OS kernel, sharing resources and reducing overhead.

Containers shine in their portability and efficiency. They start almost instantly, use fewer resources, and can run anywhere, making them ideal for microservices-based applications or deploying applications across different environments.

While containers share the host's kernel, this flexibility allows for customized security configurations and tailored resource allocation for specific applications.

Chapter 4

Infrastructure: The backbone of generative AI



“The optionality of GPUs and TPUs running on the powerful AI first infrastructure makes Google Cloud our obvious choice as we scale to deliver new features and capabilities to millions of users.”

“It’s exciting to see the innovation of next generation accelerators in the overall AI landscape, including Google Cloud TPU v5e and A3 VMs with H100 GPUs. We expect both of these platforms to offer more than 2X more cost-efficient performance than their respective previous generations.”

Noam Shazeer,
CEO, Character AI

The infrastructure underpinning generative AI solutions forms the backbone of reliable and efficient AI operations.

It is a comprehensive composition of various components including networking, operations, accelerators, and storage, each playing a crucial role in the successful deployment and operation of generative AI. Understanding how each of these components contribute to the overall AI ecosystem can enable a more informed discussion about improving and optimizing your infrastructure.

Accelerators: The power boosters

Generative AI accelerators, such as GPU and TPU chips, are optimized to perform the intense parallel processing required for deep learning, tuning, and AI inference, something CPUs were never designed for. GPUs and TPUs dramatically reduce the time and cost associated with compute-intensive generative AI workloads.

Whether deploying on-premises servers or leveraging cloud-based accelerator instances, integrating high-performance hardware accelerators is critical for implementing generative AI at scale.

up to **8** **GPUs**

per instance for your individual workload

Storage: The reservoir of knowledge

Accessible, adaptive, and scalable storage is paramount for generative AI use cases. When you're training a model you may need high IOPS, low latency storage, and when you're running a model you may want to shift to higher latency and lower IOPS to control cost. It's critical to choose a storage option that can flex to your needs throughout the generative AI cycle of preparing, training, serving, and archiving data, so you avoid having to rearchitect your application each step of the way.

Google's AI optimized storage options include Cloud Storage Fuse, which adds a file system interface layer over your Cloud Storage buckets and cuts down compute idle time, so data is streamed and the training job can start right away. Parallelstore and Filestore are also designed for with different tiers that meet bandwidth, IOPS, and latency needs. As generative models grow more data-intensive, storage becomes the pillar supporting their endless hunger for information, so it's critical to set yourself up for storage success at the start.

Data: The foundation of generative AI

Databases are the foundation of generative AI, storing and retrieving massive amounts of data. The choice of database impacts performance, scalability, and reliability in AI.

Google Cloud's BigQuery is a robust, flexible, and fully-managed data warehouse. It enables super-fast SQL queries and real-time analysis of large datasets, making it ideal for generative AI.

When choosing your database, consider factors such as compatibility with AI frameworks, performance under high-load scenarios, handling structured and unstructured data, and security measures.

Networking: The pipeline of data

Adopting a service-centric, any-to-any connectivity networking is an important architectural consideration for deploying generative AI. A cloud service that provides high-performance, low-latency interconnectivity for best-in-class application services across clouds is ideal for optimizing AI algorithm efficiency.

To address this challenge, Google Cloud offers dedicated networking capabilities like the Cross-Cloud Network to provide reliable high throughput. Cross-Cloud Network is a global network platform that is open, secure and optimized for applications and users across on-prem and clouds. It uses Google Cloud's planet-scale network for multicloud connectivity and to secure applications and users.

As generative AI models grow in size and complexity, examining and addressing network connectivity constraints and contingencies can help ensure infrastructure scalability.

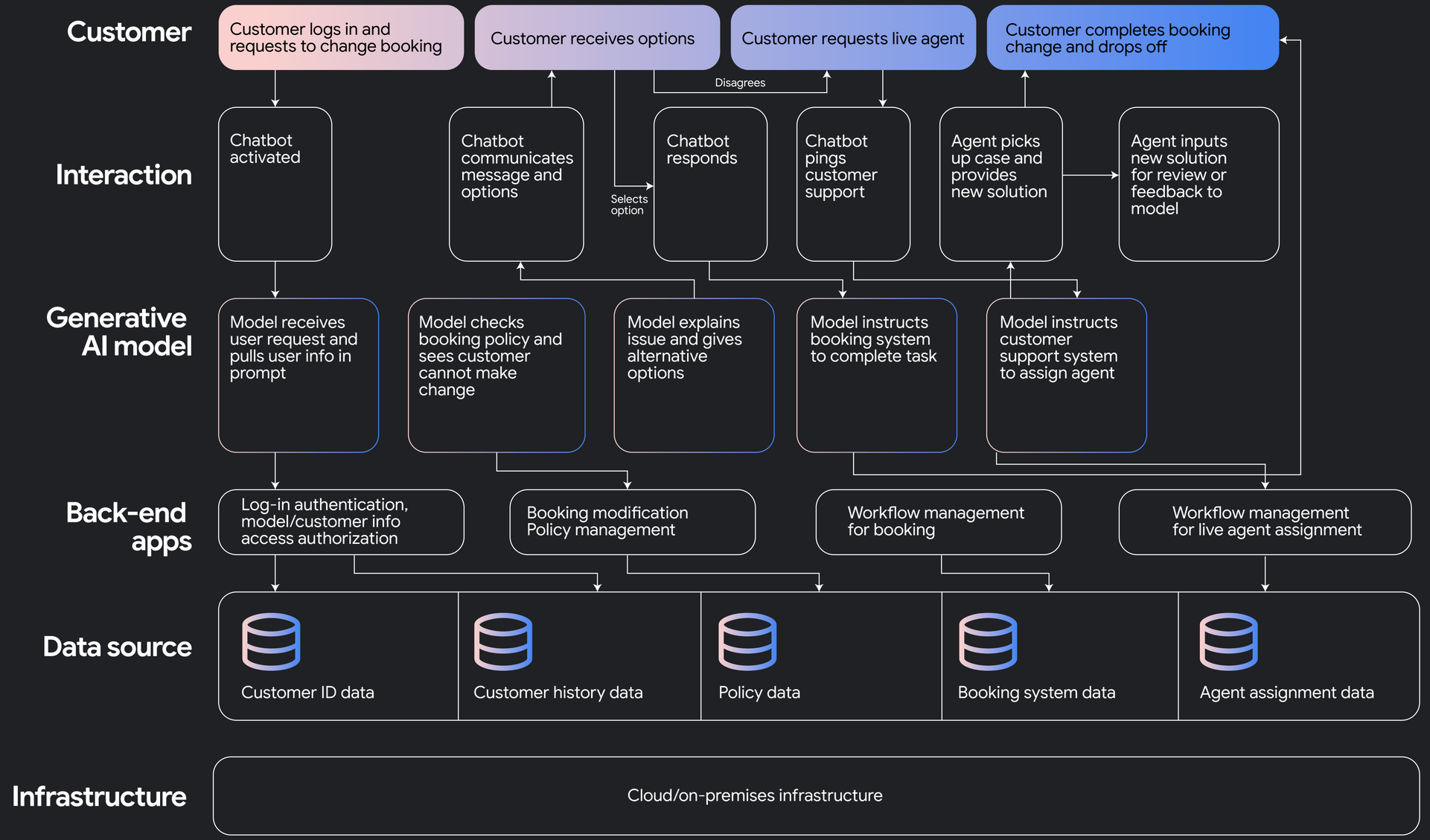
Operations: The heartbeat of AI systems

Operational tools are critical for the day-to-day running of AI systems. These tools help to monitor, manage, maintain, and optimize the performance of AI solutions, providing insights into system behavior and alerting potential issues before they escalate. This constant vigilance is key to achieving the desired output from generative AI systems, enabling timely interventions and adjustments that keep the system performing optimally.

Gemini for Google Cloud provides your teams with an AI-powered collaborator to fast-track troubleshooting, debug code with conversational assistance, and act as a subject matter expert on best practices.

“Cloud TPUs have been a game-changer for Craiyon. Cloud TPUs have allowed Craiyon to train and serve AI models much faster and more efficiently, which has led to a significant improvement in the quality of our AI-generated content. For example, we were able to gain the same performance on Cloud TPU v5e using only half the cores as that of the Cloud TPU v4 generation. We are also in the process of revisiting the model configurations and scaling them up for increased efficiency across training and inference on Cloud TPU v5e.”

Boris Dayma,
Founder, Craiyon


Customer
Customer logs in and requests to change booking
Customer receives options
Customer requests live agent
Customer completes booking change and drops off
Disagrees
Interaction
Chatbot activated
Chatbot communicates message and options
Selects option
Chatbot responds
Chatbot pings customer support
Agent picks up case and provides new solution
Agent inputs new solution for review or feedback to model
Generative AI model
Model receives user request and pulls user info in prompt
Model checks booking policy and sees customer cannot make change
Model explains issue and gives alternative options
Model instructs booking system to complete task
Model instructs customer support system to assign agent
Back-end apps
Log-in authentication, model/customer info access authorization
Booking modification Policy management
Workflow management for booking
Workflow management for live agent assignment
Data source
Customer ID data
Customer history data
Policy data
Booking system data
Agent assignment data
Infrastructure
Cloud/on-premises infrastructure

Chapter 5

How Google Cloud empowers leading AI companies

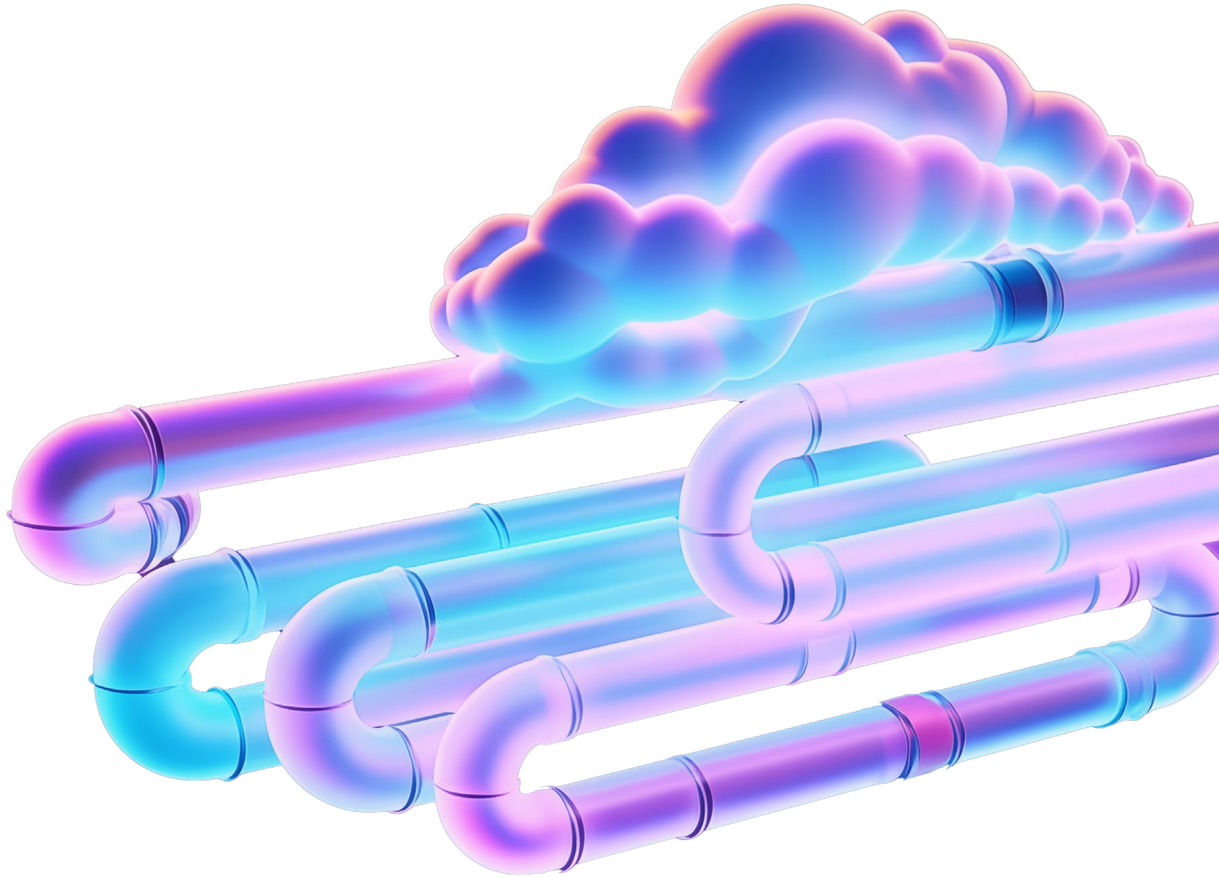
Pioneering new technologies

With the recent merger of our two world-class research teams, the Brain team from Google Research and DeepMind, a single team at Google is now collectively responsible for some of AI's most significant breakthroughs. Over the last decade this team spearheaded technologies like AlphaGo, Transformers, word2vec, WaveNet, and AlphaFold which continue revolutionizing machine learning, drug discovery, customer experience, and speech to text.

These technologies have brought a rapid paradigm shift in how organizations drive value, and deliver services. By empowering teams with the ability to automate tasks, enhance decision making, and personalize experiences, businesses enjoy reduced costs, increased productivity, and improved customer satisfaction.

How can these technologies benefit me?

Being able to adopt these state-of-the-art technologies and talents into your own AI infrastructure via Google Cloud can significantly increase the efficiency and effectiveness of your AI models, providing a competitive edge in the AI landscape.



Enterprises like Wendy's, Canva, Uber, Instacart, Deutsche Bank, and more leverage Google Cloud for generative AI, in order to create amazing content, synthesize and organize information, automate business processes, and build engaging customer experiences.

Google Cloud has also emerged as the preferred platform for startups building generative AI, and there's a compelling reason for this. Google Cloud's success stems from its unique team structures and collaboration, where cross-disciplinary teams work together to develop, implement, and refine AI-optimized infrastructures. This approach allows Google Cloud to provide a platform that is not only robust and scalable, but also tailored to the needs of generative AI.

With a clientele that includes over half of all funded generative AI startups and 70% of generative AI "unicorns," it is clear that Google Cloud's team dynamics and innovative technology offer a distinct competitive edge. Startups like Bending Spoons, Faraday, Jasper, Replit, and Typeface leverage this platform to integrate generative AI into everyday tasks, enhancing efficiency and productivity.

Moreover, Google Cloud's does more than provide a platform for generative AI startups. It is also a trove of vital AI tools and resources. From AI21 Labs and Anthropic to Character.AI, Cohere, Midjourney, and Osmo, many significant model-builders and researchers have developed and trained their models using Google Cloud's infrastructure, GPUs, and TPUs.

Additionally, essential AI tooling applications from Aible, Anyscale, GitLab, Gretel, Labelbox, Snorkel AI, and Weights & Biases are readily available on Google Cloud, making it easier for other companies to build and refine their own generative AI experiences. This comprehensive suite of tools and resources, combined with Google Cloud's optimized infrastructure, makes Google Cloud the ideal platform for organizations looking to scale and innovate.

“Google Cloud’s generative AI technology creates a huge opportunity for us to deliver a truly differentiated, faster, and frictionless experience for our customers, and allows our employees to continue focusing on making great food and building relationships with fans that keep them coming back time and again.”

Todd Penegor,
President and CEO, The Wendy’s Company

Get the executive’s guide to generative AI to read the full story

How can Google's global-scale capabilities help me scale?

Google Cloud's high-availability, low latency, and vast planet-scale capabilities have been instrumental in powering some of the world's most popular platforms and tools, from YouTube and Gmail to the ubiquitous Google Search. Each of these tools serves billions of people daily, managing enormous volumes of data and complex computations with unrivaled speed and efficiency.

Drawing from this wealth of experience, Google Cloud architects infrastructure to scale to serve billions of users worldwide. The same infrastructure that supports Google's mission-critical services we provide to our Google Cloud customers. This means you can leverage the same robust, secure, and high-performance infrastructure that powers Google's own products, enabling you to scale your AI applications smoothly and efficiently.

“Using G2 VMs has allowed us to considerably lower latency times for processing by up to 15 seconds per task. Google Cloud has also been instrumental in helping us seamlessly scale up to 32,000 GPUs at peak times like when our Remini app soared into the No. 1 overall position on the U.S. App Store and down to a daily average of 2,000 GPUs.”

Luca Ferrari,
CEO and Co-Founder, Bending Spoons

Conclusion

The generative AI revolution is here—delivering unprecedented capabilities in content creation, personalization, and more. Now is the time to leverage this technology and gain competitive advantage. But realizing generative AI's full potential requires infrastructure ready for production deployment. With specialized networking, accelerators, and storage, Google Cloud provides an optimized foundation to run AI efficiently at scale. This enables you to achieve key business benefits: faster delivery of insights through accelerated training, reduced costs with optimized workloads, and the ability to engage customers in new ways through AI-generated content and recommendations. Prepare today to deliver business results powered by responsible, scalable AI that solves real-world problems. The future of AI is generative—embrace it and thrive.