niveus Part of NTT DATA

Google Cloud

Building the foundation for AI success

A strategic guide to AI infrastructure

Contents

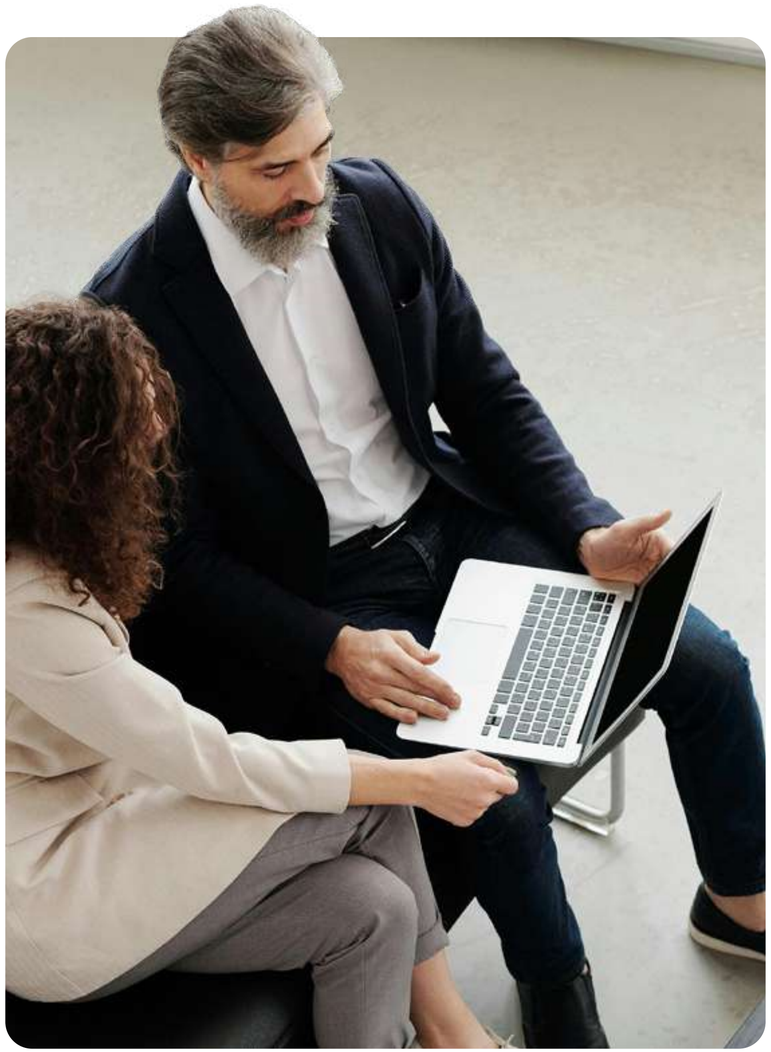
Summary

This ebook is your roadmap to identify and address the infrastructure challenges associated with successfully implementing generative AI (gen AI) at scale. Designed for technology leaders, architects, and engineering managers, it reveals the common infrastructure pitfalls hindering AI innovation.

You'll uncover how a purpose-built, integrated AI infrastructure, like Google Cloud's AI Hypercomputer, can unlock the full potential of gen AI, empowering you to navigate complexities, overcome challenges, and unleash transformative business impact now.

How AI is transforming business

Today, as organizations embrace the business potential of generative artificial intelligence (gen AI), they are reimagining their operations and customer experiences. The wave of gen AI-powered use-cases, from content creation to research summarization, customer service chatbots, and automated meeting notes, are just the start of what's possible.

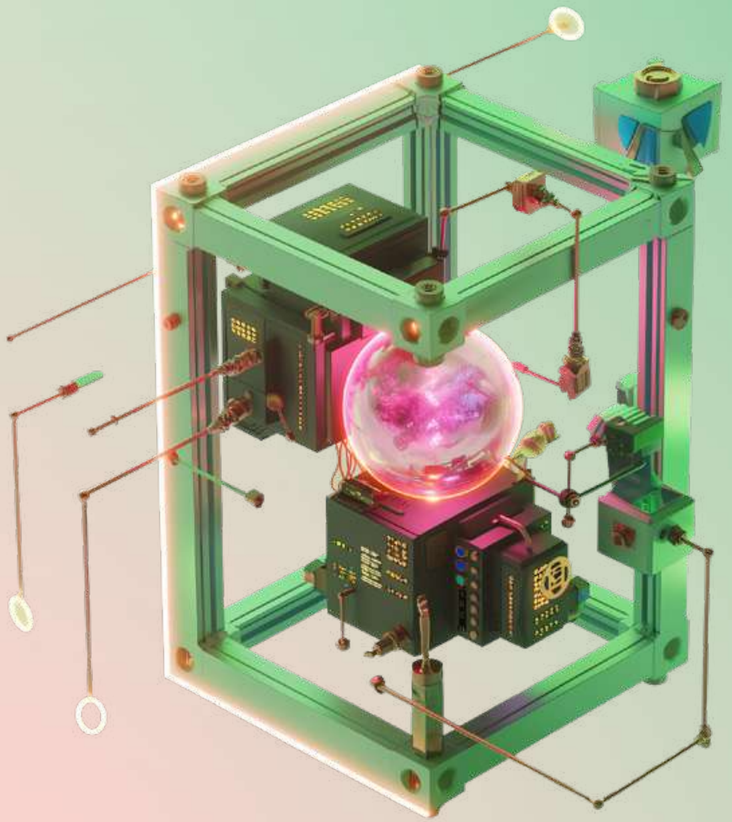


Business leaders across the globe are increasingly turning to generative AI, viewing it as a powerful catalyst for enhancing their products, services, and overall profitability. So it's no surprise that businesses are rushing to embrace AI's potential, with gen AI leading the way.

In September 2023, **71 percent of companies were using gen AI, and another 22 percent planned to implement it in the next 12 months**[1].

However, organizations exploring gen AI are quickly discovering that substantial infrastructure challenges stand in the way of fully realizing the models' potential.

"Nearly **80 percent of AI projects typically don't scale** beyond a proof of concept (PoC) or lab environment. Businesses often face challenges in standardizing model building, training, deployment, and monitoring processes."[2]



Infrastructure challenges pose risks

Today's infrastructure is not set up to handle the massive data and computational requirements of large-scale generative AI models. In this ebook, we will discuss the key infrastructure challenges that organizations face when trying to scale generative AI, and how Google Cloud's AI Hypercomputer addresses these challenges.

We will also explore the benefits of a purpose-built, integrated solution like AI Hypercomputer, and showcase real-world examples of organizations leveraging its power to drive innovation and achieve business success.

While there are a plethora of ready-to-use platforms, models, and tools, many customers have use cases that need more flexibility. From model training and tuning, to serving, having greater control of the implementation of their AI architecture – including down to the underlying infrastructure hardware – can often enable greater scale, efficiency, and cost effectiveness. While ready-to-use AI solutions, often built on the same powerful infrastructure, offer a simplified experience and cater well to certain needs, they may not always provide the granular control and customization required by some organizations.

Mega processing power

Over the past five years, **gen AI models have grown** tenfold each year in complexity. Training, fine-tuning, and serving of large-scale gen AI models involve trillions of complex calculations. Unlike traditional workloads that operate on structured data within predictable patterns, gen AI workloads involve intensive operations on unstructured data modalities – like images, text, and audio.

These operations, such as matrix multiplications, convolutions, and attention mechanisms, require high-throughput and low-latency infrastructure to efficiently process the vast amounts of data and model parameters involved. These operations place a considerable strain across infrastructure, requiring robust distributed computing capabilities, a high-bandwidth underlying network fabric for efficient data movement, and ample storage capacity to handle model checkpoints and intermediate results.

The demanding nature of training also necessitates exceptional reliability, as even brief interruptions can significantly disrupt progress.

Growing ecosystem of open-source software, frameworks, and platforms

The AI landscape consists of a complex mosaic of evolving models, libraries, and tools. Building a successful AI solution means navigating a patchwork of open and closed language models, each with their strengths and limitations.

While open-source tools and frameworks offer powerful capabilities, they often require significant resources to customize and integrate.

Challenges of orchestration and management across the model lifecycle

Organizations building and fine-tuning custom gen AI models for specific domain use-cases face a multifaceted challenge across the entire model lifecycle. This includes managing the training process, refining hyperparameters for peak performance, and ensuring efficient model deployment – with considerations like scalability, monitoring, and security. Balancing these diverse components often leads to intricate management complexities, impacting both direct costs (e.g., infrastructure) and indirect costs (e.g., delayed time-to-market).

Faced with these complexities, organizations are realizing that simply investing in the latest hardware isn't enough to unlock gen AI's full potential. Instead, they need an integrated system of performance-optimized hardware, open software (including frameworks and libraries like JAX and PyTorch), and effective orchestration. This helps empower developers with choice and flexibility, at every layer of the stack, to tailor solutions.

A unified AI supercomputing architecture is the ideal way to implement this strategy. It combines the necessary performance-optimized infrastructure with simplified development, deployment, and management of AI applications at scale. Google Cloud's AI Hypercomputer is a prime example of this approach, offering a cohesive platform on which customers can run their end-to-end model lifecycle while accelerating AI development and reducing costs.

Wayfair solved scalability and saved costs with AI

Wayfair, a leading online retailer for furniture and home goods, supercharged its e-commerce platform and data analytics capabilities with AI using Google Cloud TPU and AI Hypercomputer.

The company faced challenges with scalability during peak shopping events like Way Day and Black Friday.

By leveraging Google Cloud TPU, Wayfair achieved significant improvements in its e-commerce platform, resulting in fast processing, better performance, and cost savings.

Google Cloud's AI Hypercomputer also enabled Wayfair to harness real-time insights from its vast amounts of data, leading to improved customer and supplier experiences via personalized recommendations and accurate forecasting.

With Google Cloud's cutting-edge AI infrastructure, Wayfair is well-equipped to continue its growth and maintain its position as a leader in the competitive online retail landscape.

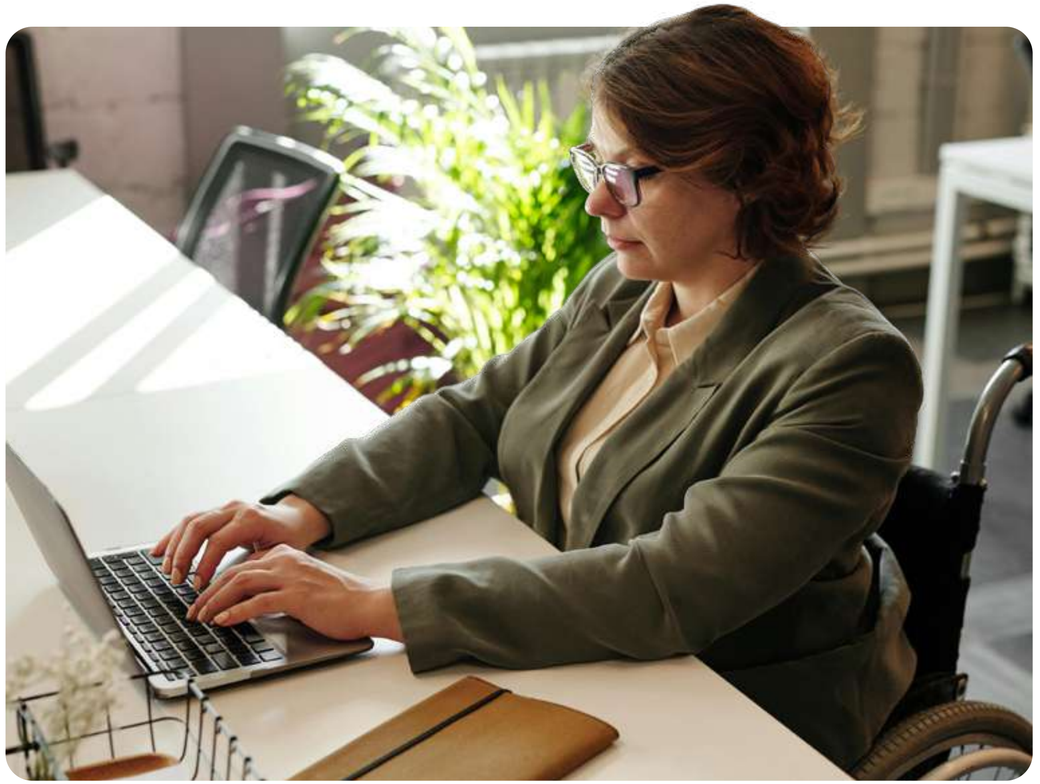
What you need to succeed

Put simply, you need a full-stack solution that's designed to work together rather than a disconnected set of components.

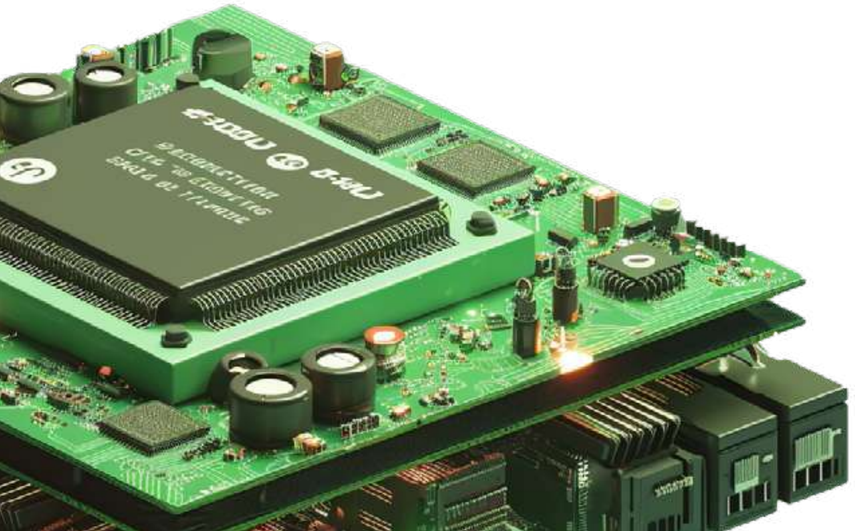
This holistic approach enables organizations to streamline AI development, simplify operations, and maximize the return on AI investments.

It provides the flexibility to scale resources, ensuring optimal performance and cost-efficiency throughout the entire AI lifecycle (from model training to deployment and inference).

So what should organizations consider when building infrastructure for their AI initiatives? Critical aspects include: affordability, flexibility, security, reliability, scalability, access to specialized expertise, and compatibility with the tools and services the organization already uses.

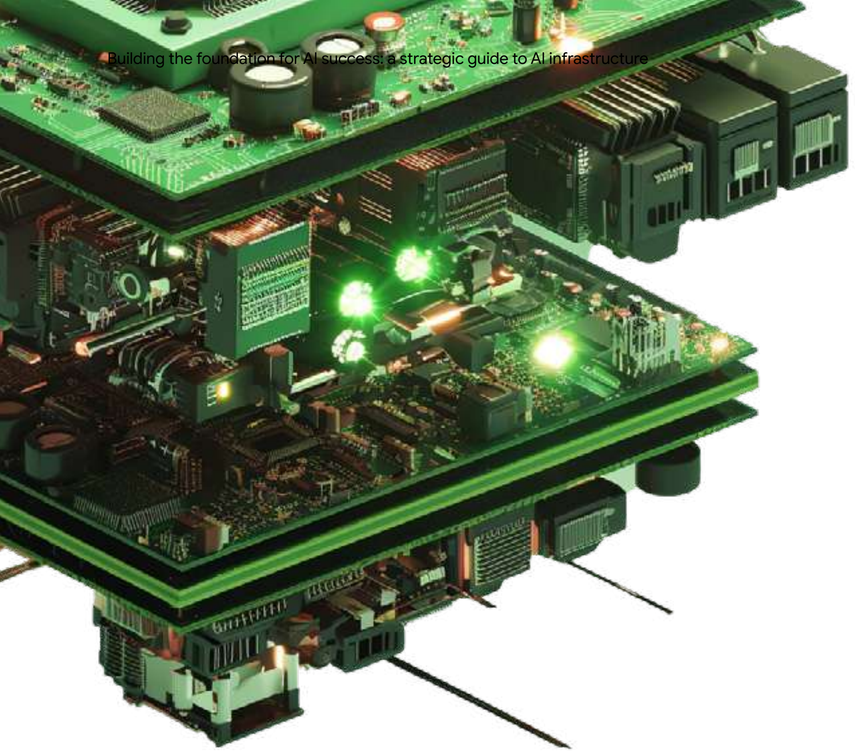


To zero in on the most promising options, begin your search with two key criteria in mind:



1 Your infrastructure needs to be purpose-built

to handle the computing demands of AI workloads, which are memory-intensive, bandwidth-intensive, and computation-heavy. The demands of AI technologies are constantly evolving, fluctuating as new models, techniques, and proofs of concept reveal additional ways of harnessing them. Purpose-built infrastructure is specifically engineered for exceptional performance in specific areas (like processing large datasets or running complex AI workloads), making it incredibly efficient and fast for those jobs.



2

Hardware is essential, but a chip is only as powerful as the stack that surrounds it.

When hardware and software come together into an integrated, easy-to-use, secure, and reliable computing system, it helps to avoid development bottlenecks, minimize manual integration time, and improve overall system productivity. While hardware provides the raw potential, you can't unleash it without a robust ecosystem of tools, frameworks, and resources such as:

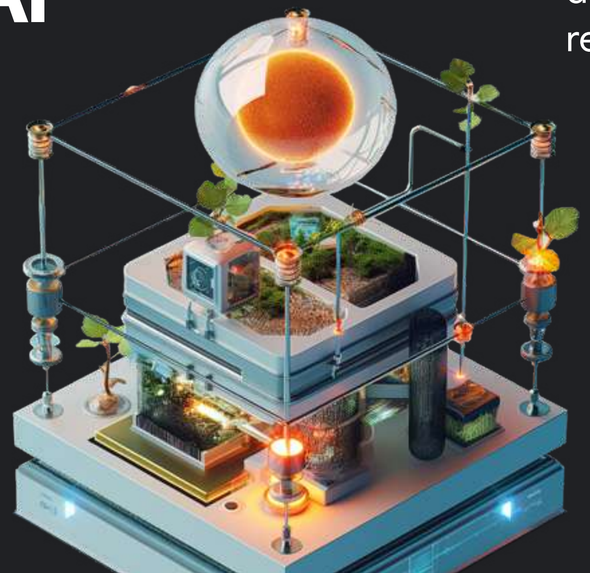
Open software for orchestration of AI workload:
ensuring efficient scaling, provisioning, checkpointing, and recovery –minimizing downtime and maximizing efficiency.

Extensive support for popular machine learning frameworks:
to lower entry barriers while boosting developer productivity.

Deep integration with tools (often also powered by gen AI):
to deliver efficient resource management, consistent ops environments, and proactively alert you to potential issues.

Built-in flexibility of compute resource management:
to optimize costs and ensure peak performance. Cloud-based solutions provide flexible resource management and scheduling services, enabling AI workloads to consume resources efficiently.

Wendy's taps Google Cloud to reimagine the drive-thru experience with AI



By creating a personalized and tailored drive-thru experience for customers to enjoy. Wendy's drive-thru accounts for nearly 80% of its business, and it's on the rise across the restaurant industry. Optimizing the business of the drive-thru is of utmost importance for all quick service restaurants to attract and retain customers.

Wendy's is working with Google Cloud on a groundbreaking AI solution, Wendy's FreshAI, designed to transform the quick service restaurant industry.

Using Google Cloud's AI Hypercomputer along with Google's generative AI and large language models (LLMs), FreshAI is able to discern the billions of possible order combinations on the Wendy's menu and accurately translates spoken orders into text for the restaurant crew and customers.

This streamlined process improves order accuracy, speeds up service, and has increased customer satisfaction, exemplifying Wendy's commitment to customer service, quality and continuous improvement.

Benefits of a purpose-built, integrated solution

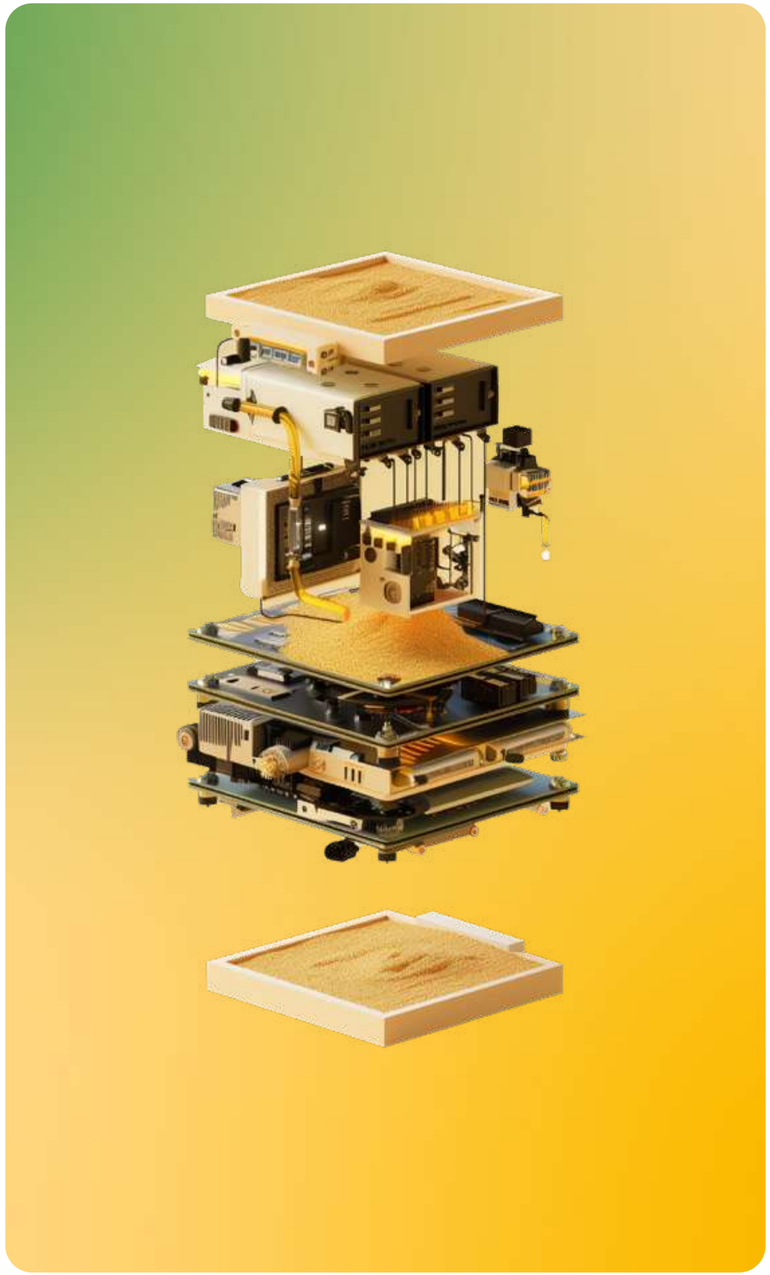
A purpose-built, integrated AI infrastructure offers several key benefits that address the unique challenges of AI workloads:

1 Optimized computational power

Purpose-built AI infrastructure leverages specialized hardware accelerators like GPUs (graphics processing units), and TPUs (Tensor Processing Units). They can break down complex computations into thousands of smaller tasks, processing them simultaneously and fast. With this parallelization, creation of today's complex models become possible.

However, parallel processing alone isn't the end of the story. Purpose-built AI silicon goes further. GPUs and TPUs are designed from the ground up for the specific mathematical operations that power AI.

This specialized design helps deliver superior performance and improved energy efficiency compared to general-purpose processors, pushing the boundaries of what's possible with AI. The flexibility to incorporate CPUs for handling less demanding tasks such as preprocessing and offline inference can allow for a tailored approach to resource allocation.

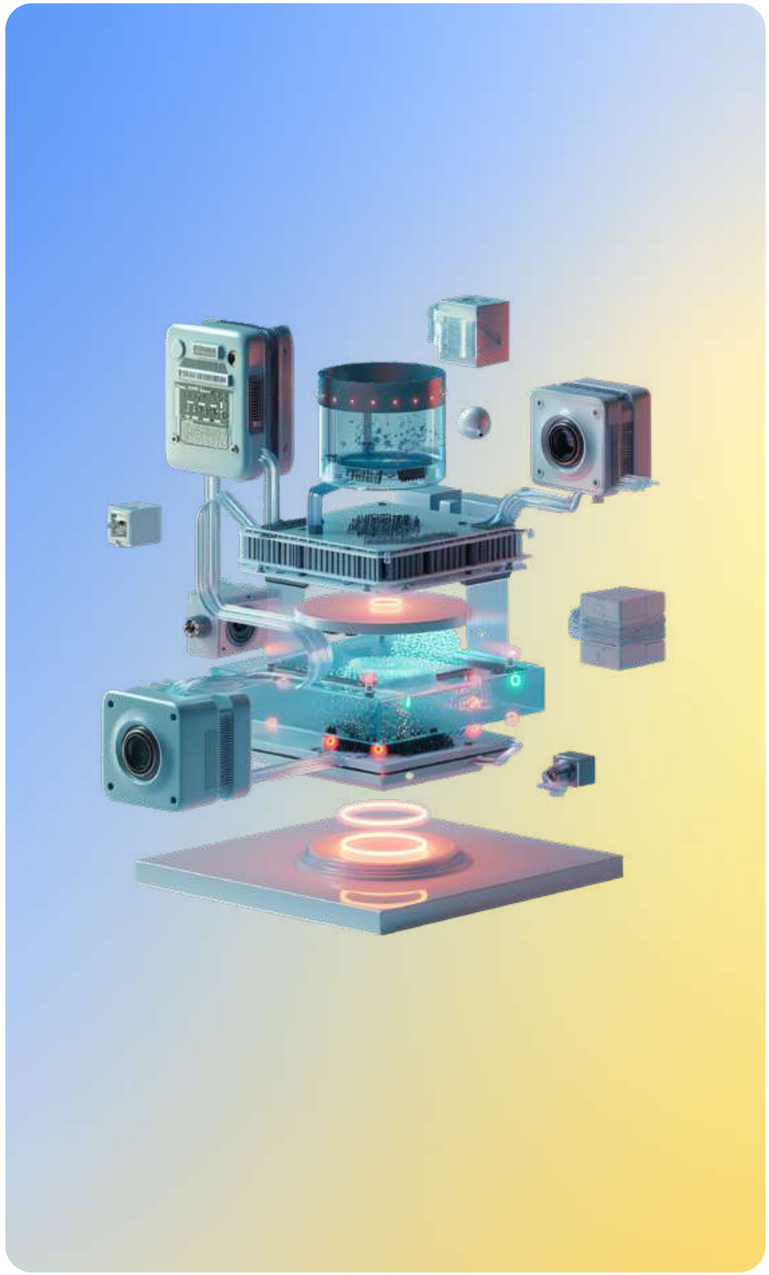


2 Massive data and networking capacity

AI models are data-hungry powerhouses that require vast quantities of information to learn from patterns and provide intelligent outputs. Traditional IT storage and networking solutions simply can't keep up with this demand.

To meet these challenges, the most effective purpose-built AI platforms incorporate advanced storage solutions offering high-bandwidth capabilities to manage petabytes of data efficiently.

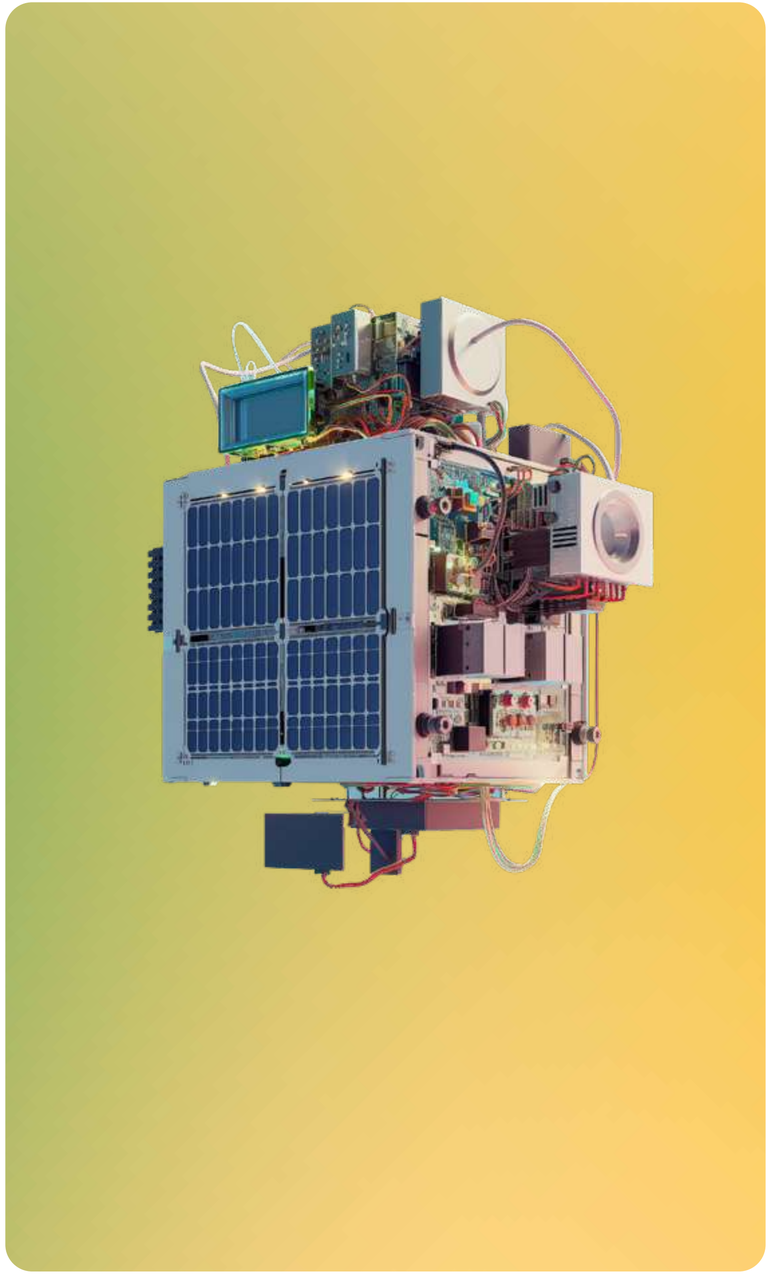
Low-latency networking ensures rapid communication between distributed systems, and fast data transfer speeds accelerate the entire AI workflow.



3 Elastic scalability

AI initiatives allow organizations to experiment, rapidly scale successful projects, and optimize costs by paying only for the resources they use. Managed solutions like Google Kubernetes Engine (GKE) can further streamline this process by offering auto-scaling capabilities.

This reduces management overhead by automatically adjusting resources to meet demand, ensuring optimal performance and cost-efficiency.



4 Cost efficiency

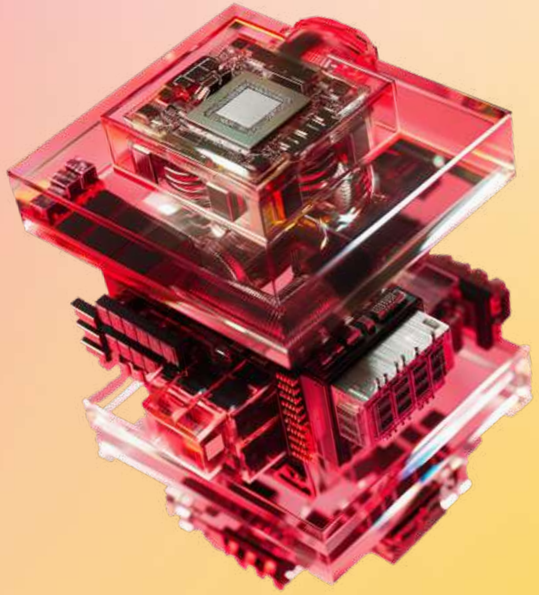
Purpose-built infrastructure that optimizes resource utilization for your AI workloads, ensures you only pay for what you use. Full-stack solutions also come with tools for monitoring and analysis, making sure every bit of computational power is effectively utilized.

This not only prevents wasteful spending but also helps track the return on investment for AI projects. Areas for cost optimization range from efficiently using GPU and TPU resources with Dynamic Workload Scheduler, to leveraging cloud-native services and containerization technologies such as Kubernetes.

By providing IT leaders with a granular level of control and visibility into their resources that was previously unattainable, AI tools reveal new avenues of cost savings across their environment.

The choice of hardware accelerator also plays a crucial role in cost optimization. While GPUs and TPUs excel in certain AI workloads, CPUs offer a cost-effective and readily accessible solution for a wide range of AI tasks, especially those that are less latency-sensitive like model development, experimentation, and pre-processing.

They're also worth considering for classical machine learning tasks like recommendation systems and image classification.



5 Accelerated developer productivity

AI optimization is a comprehensive strategy addressing each layer of the stack, from hardware and software to workflows and management. The most effective, purpose-built platforms prioritize streamlining the AI development process.

They offer simplified workflows, comprehensive toolkits, tailored for model creation, and effortless integration with popular open-source AI and machine learning frameworks (like JAX, TensorFlow, and PyTorch). They also incorporate specialized managed infrastructure services that reduce the burden of setting up and scaling AI-centric environments.

By integrating with the tools and systems developers already know and use, these platforms help empower them to focus on innovation rather than grappling with complex infrastructure setups.

Google Cloud's experience building and deploying AI at a global scale has proven that purpose-built infrastructure is essential for achieving high-performing levels of system productivity and resource optimization, which in turn helps lead to cost optimization.

While AI may be accessible to many, achieving a true competitive advantage requires a unified AI infrastructure that improves efficiency across the entire development lifecycle.

Our AI infrastructure is one of the reasons nearly 90 percent of generative AI unicorns and more than 60 percent of funded gen AI startups have chosen Google Cloud as an infrastructure partner for developing their own models and services.

Forrester Research has recognized Google as a leader in The Forrester Wave™ - AI Infrastructure Solutions, Q1 2024. Google received the highest scores of any vendor evaluated in both the Current Offering and Strategy categories in this report. We believe this is a testament to our vision and strong track record of delivering continuous innovation and leading AI infrastructure products for our customers.

Download the report to learn more.

"Google has strengths across the board with the highest scores of all the vendors in this evaluation."

- The Forrester Wave™: AI Infrastructure Solutions, Q1 2024

Choosing the right AI infrastructure provider

After narrowing down your AI infrastructure search to focus on solutions that are both fit for purpose and full-stack, there's another important element to consider. The rapid evolution of AI technology, with ever-increasing scale and complexity of foundational models, creates a **critical need for specialized expertise**.

Collaborating with a trusted infrastructure partner that offers both extensive experience and a deep understanding of AI is crucial for bridging the existing skills gap and unlocking AI's transformative power.

Through a continuous process of understanding and overcoming AI challenges, from chip design to transformer architecture, Google Cloud has cultivated a knowledge base spanning the entire AI development lifecycle.

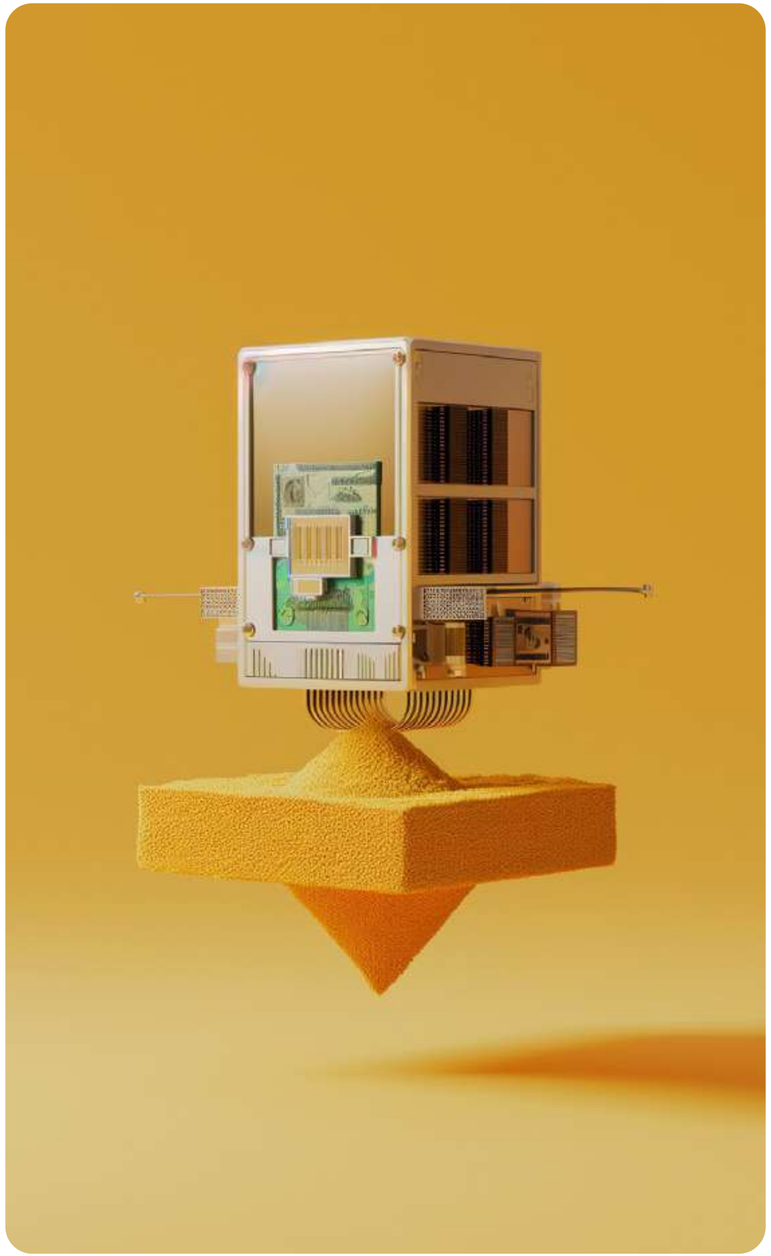
This has led to unparalleled insights into optimization and improvement. For example, the recently announced sixth generation of Trillium TPUs provide a 4.7X increased peak compute performance per chip and 2X HBM Capacity & Bandwidth improvement over the previous generation of TPUs.

Google Cloud's AI Hypercomputer architecture is built based on these insights, with the goal of making AI accessible and useful to everyone, everywhere.

Google Cloud continues to explore strategies and technologies that make gen AI powerful, responsible, and sustainable throughout its lifecycle.

We do this by developing new metrics for evaluating AI systems, moving beyond traditional measures to account for factors like energy consumption and carbon emissions, all within the boundaries of our responsible AI principles.

Additionally, we invest in hardware like TPUs which are designed to be energy efficient and reduce the carbon footprint of AI workloads, with the average Google-owned and-operated data center being approximately 1.8 times as energy efficient as a typical enterprise data center.



Enabling next-generation AI workloads with AI Hypercomputer

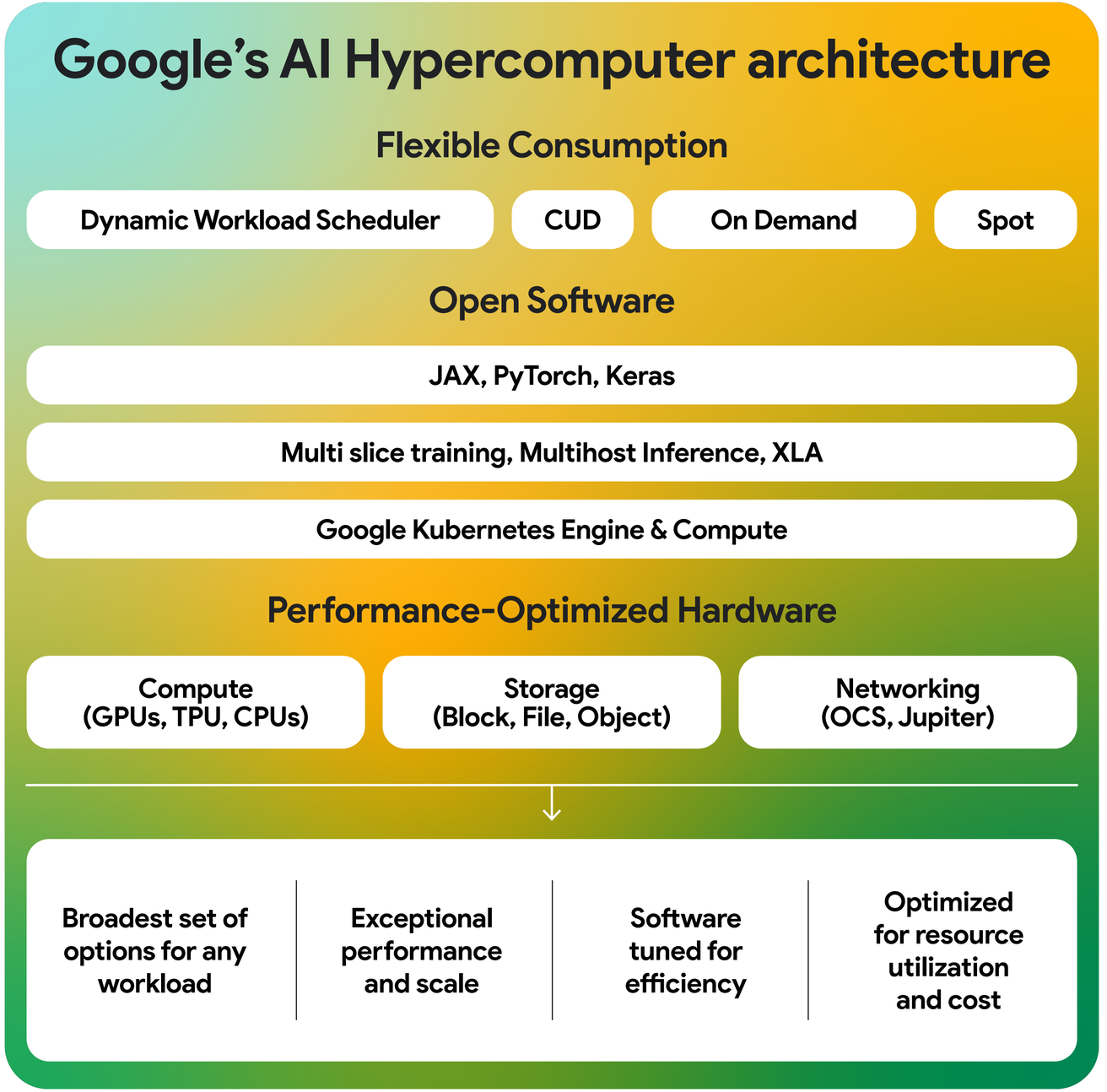
AI Hypercomputer combines performance-optimized hardware, open software, leading machine learning frameworks, and flexible consumption models into a groundbreaking supercomputer architecture.

It is purpose-built to boost efficiency and productivity across AI training, tuning, and serving. Businesses and developers are using AI Hypercomputer to achieve more than 2X efficiency relative to just buying the raw hardware and chips.

AI Hypercomputer is designed to meet the diverse and evolving demands of your AI projects.

Built on Google Cloud's fundamentals of operational excellence, cost efficiency, and robust security, it extends these benefits to AI workloads.

Its system-level design can ensure maximum efficiency and unparalleled flexibility and choice, allowing you to tailor your AI infrastructure to your specific needs.



Google Cloud and NVIDIA partner to advance AI computing

The Google Cloud and NVIDIA partnership provides a comprehensive platform for AI, We've seamlessly integrated NVIDIA accelerated computing into AI Hypercomputer, providing customers with access to a broad portfolio of the latest NVIDIA GPUs, such as the NVIDIA H100, L4 and A100 Tensor Core GPUs.

To maximize application performance on NVIDIA GPUs and boost developer productivity, we've optimized popular frameworks like PyTorch, JAX and TensorFlow and offer enterprise-grade solutions and support through NVIDIA AI Enterprise on the Google Cloud Marketplace.

Additionally, we've optimized NVIDIA TensorRT-LLM and NVIDIA Triton Inference Server on the AI Hypercomputer to accelerate AI inference performance and simplify the path from prototype to production deployments at scale.

Google Cloud and NVIDIA also collaborate closely on integrations that bring the power of the full-stack NVIDIA AI platform to a broad range of Google Cloud services, including Google Kubernetes Engine (GKE), Dataproc, Dataflow and Vertex AI giving developers the choice to develop and deploy AI applications at the level of abstraction they need. By partnering at every layer of the AI stack, we empower organizations to build, deploy, and scale AI applications with greater efficiency and flexibility.

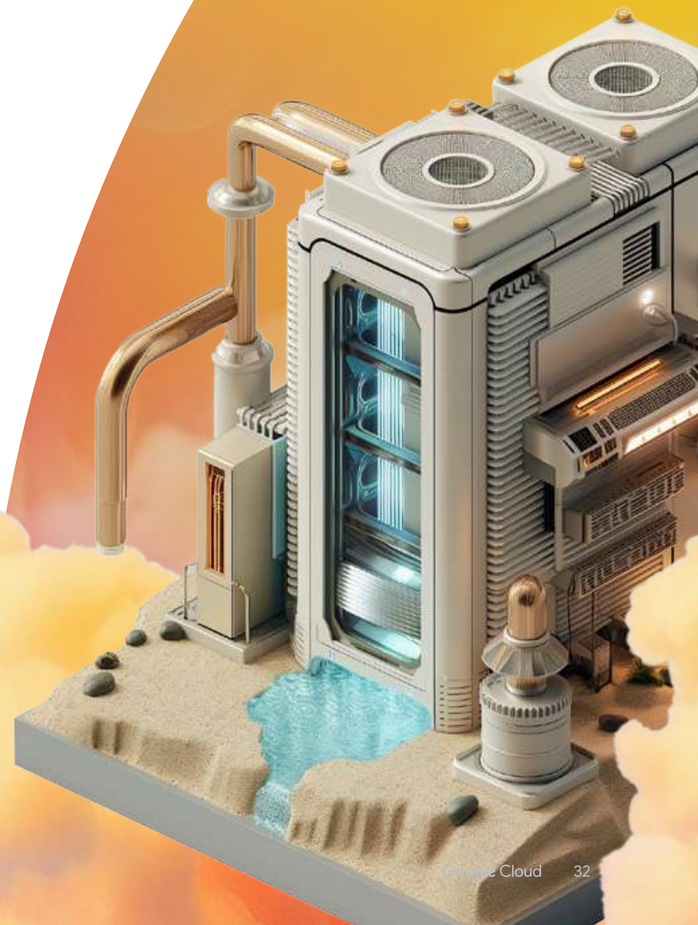
Putting it all together

Generative AI is redefining industries, transforming businesses, and improving customer experiences. Yet, traditional IT infrastructure is ill-equipped to handle the intense demands of these advanced technologies, hindering organizations from fully realizing AI's potential.

The complex nature of generative AI and large language models (LLMs) requires a new approach to infrastructure, one that is purpose-built for the unique demands of these workloads.

Google Cloud's AI Hypercomputer bridges this divide. It helps organizations drive faster, more effective innovation, bringing new products and services to market faster than ever before.

By driving efficiency and cost savings both in terms of IT infrastructure and the productivity potential of gen AI, organizations can enjoy enhanced decision-making and insights without breaking the budget.



Ready to learn more about Google Cloud's AI Hypercomputer?

CONTACT NIVEUS SOLUTIONS

biz@niveussolutions.com | www.niveussolutions.com

niveus Part of NTT DATA

Google Cloud