ARCHITECTING FOR AI EVERYWHERE

HOW GOOGLE DISTRIBUTED CLOUD AND GEMINI BRING GENERATIVE AI ON-PREMISES

SUMMARY

AI is fundamentally redefining operational efficiency and value creation. When deployed at scale, it has the potential to alter nearly every aspect of business — from automating workflows to empowering business users who increasingly function as citizen data scientists.

Yet while many organizations can see the promise of an enterprise powered by generative and agentic AI, planning and activating this environment remains a major challenge. The seemingly countless issues around digital sovereignty, data locality, models, AI infrastructure, security, and integration across global operations can be daunting.

This research brief examines the unique needs for deploying, activating, and optimizing generative AI in the enterprise. It explores challenges associated with activating AI in environments where data residency/locality, privacy, security, and compliance are paramount concerns. More specifically, it details how solutions like Google Distributed Cloud (GDC) with Gemini can deliver the optimal generative AI platform. Finally, it provides practical guidance for organizations ready to operationalize AI innovation while maintaining control, cost, and operational efficiency.

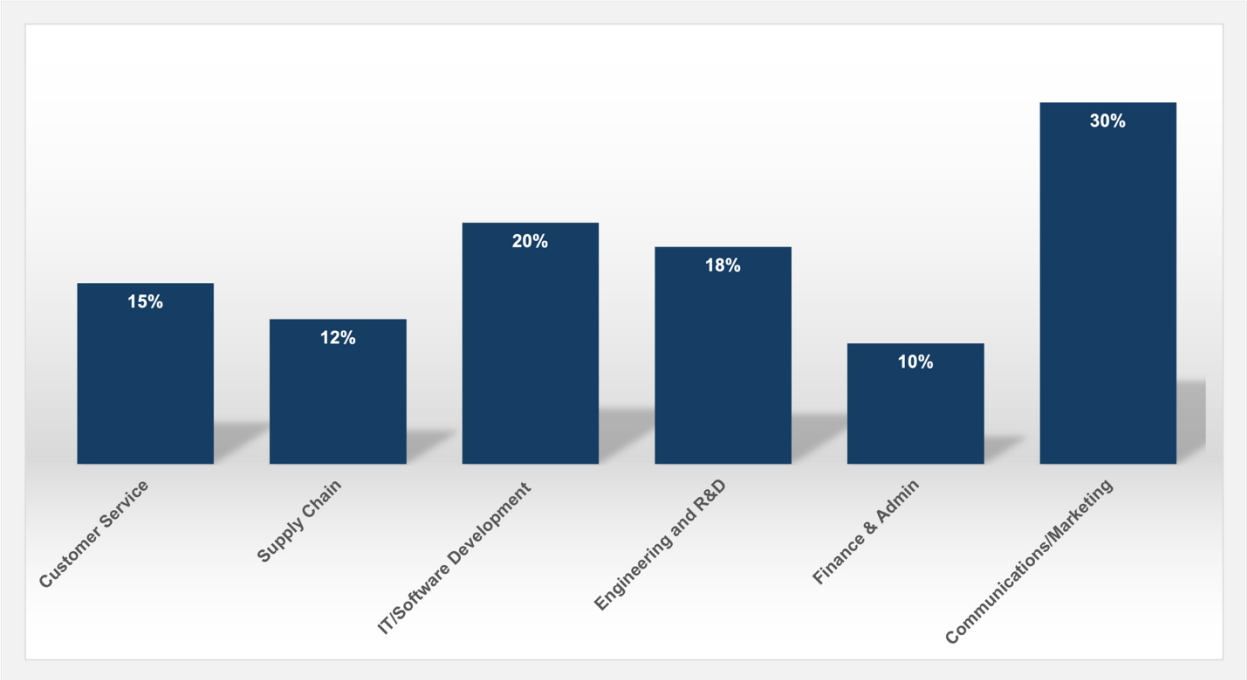
THE AI OPPORTUNITY

AI has expanded into every facet of our lives. On the consumer front, capabilities have shifted from novel to transformative. For example, consumers can use a Google Pixel phone to perform real-time voice translation in their own voice, tone, and inflection. Or consider the social media influencers who can record, edit, and amplify videos and other digital content all from a phone in real time — a process that not long ago would take hours or even days to accomplish.

Just as AI has redefined the consumer experience, Moor Insights & Strategy (MI&S) sees AI delivering value across many functions in the enterprise — from customer service to operations. Within software engineering, for instance, a Cornell University study found that using an AI coding assistant delivered speed improvements of up to

40% in repetitive coding tasks and unit test creation, and up to 50% time savings in code documentation and autocompletion.

FIGURE 1: INCREASING BUSINESS PRODUCTIVITY WITH GENERATIVE AI



Generative AI can drive substantial business value across a variety of functions.
Source: Generative AI surveys conducted by Accenture and McKinsey

As shown in the chart above, AI promises to deliver productivity gains across the entire enterprise — from software developers to marketing professionals. For example, consider the manufacturing company looking to automate its supply chain. In this case, AI agents can act as arbitrage mechanisms to poll suppliers, receive bids, and make selections based on prioritization and weighting of these considerations.

The impact on the bottom line is real: A study conducted by Accenture found that companies utilizing AI for supply chain automation achieved, on average, 23% greater profitability than their peers.

Bringing these trends together, we envision a future where AI-enabled devices integrate with enterprise systems to support workers in every aspect of their jobs. For example, a healthcare professional could use a locally hosted language model to help with patient care, delivering higher quality of care by keeping the model on-premises while adhering to patient privacy regulations.

The Distributed Computing Continuum

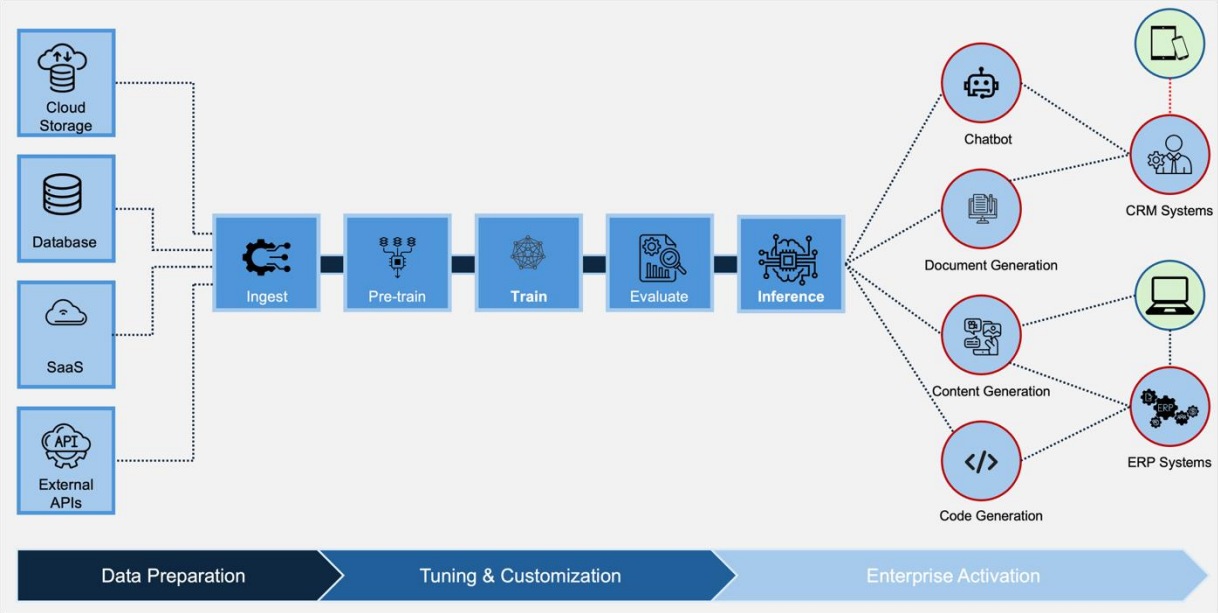
These use cases illustrate the AI continuum. On one end of it, consumers rely on devices like the Pixel phone augmented by, in some cases, Gemini running in Google's public cloud, delivering rich yet simple user experiences. Conversely, enterprises often require on-premises environments due to data gravity, privacy, or the need to minimize inference latency. This infrastructure must extend from the core datacenter to the edge, delivering the fastest, most accurate results at scale.

Consider a use case from the real world. A global retailer was hampered by disjointed hardware and connectivity issues impacting customer service and operations. To deliver greater customer service and increase productivity, it partnered with Google Cloud. The modernization journey began with migrating in-store applications to microservices on Google Kubernetes Engine (GKE) on GDC, replacing these mostly unmanaged legacy systems.

This modernized environment delivered both scale and resilience to operations, ensuring an entire store could function in the event of internet outages that could last up to several days. Additionally, this modernized footprint laid the foundation for high-value AI at the edge.

This encapsulates the value of the distributed computing continuum: delivering the right amount of computational power at the correct location to overcome the latency introduced when resources are physically separated from data. In short, it means bringing the AI stack to the data, wherever that data resides.

FIGURE 2: ACTIVATING AI ACROSS THE ENTERPRISE



Activating generative AI across the enterprise requires significant resources.
Source: Moor Insights & Strategy

FINDING THE RIGHT BALANCE: CLOUD AGILITY AND ON-PREMISES CONTROL

As with most emerging technologies, early AI adoption has been best served from the public cloud. Indeed, MI&S has consistently observed that most enterprise AI journeys begin in the public cloud. This makes sense: The cost and complexity of standing up AI infrastructure, models, frameworks, and tooling is simply more than most organizations can take on initially; meanwhile, the cloud provides accessible infrastructure, models, and tools, lowering the barriers to experimentation. But the notion that AI will always live exclusively in the public cloud is fundamentally flawed.

Several forces are already pushing enterprises toward on-premises and hybrid AI. Latency concerns, the need for predictable cost models, regulatory obligations, and data-residency constraints make local AI environments unavoidable for many. Then there's the data challenge. AI delivers real value only when an organization can fully harness its data — hot, cold, dark, structured, unstructured, even physical archives that become meaningful once scanned and ingested through techniques such as RAG.

This is why, as organizations move from pilots to production, they will increasingly build on-premises AI environments to complement their cloud use and operate in a hybrid

model. This gives them the flexibility to use cloud resources when appropriate, while running sensitive or performance-critical workloads locally. MI&S expects this shift toward on-premises deployments to accelerate as generative and agentic AI mature.

Financial services is one of the clearest examples of this trajectory. Many large banks started their AI work in the public cloud, but the combination of enterprise-wide data use, regulatory constraints, and latency considerations will pull much of that activity back in-house. By tapping into historical and previously unused datasets, these institutions can deliver more personalized services and identify emerging trends earlier. Doing so requires analyzing sensitive data that cannot leave their environment, making on-premises AI not optional, but necessary.

The Cloud Exists for a Reason

This does not imply that the public cloud is a poor choice for customers. In fact, the cloud can deliver a fully integrated AI stack built on a rightly designed infrastructure that abstracts much of AI's complexity. This is exactly what many enterprise customers require to overcome the AI inertia that often stalls adoption. From toolchains to frameworks to large language models (LLMs), the public cloud allows organizations to innovate fast by managing the lower-level architecture and deploying a vertically integrated AI stack.

The question is, how can an enterprise IT organization deliver that same vertically integrated AI stack? How can it bring the cloud on-premises while avoiding cost and complexity challenges? MI&S believes Google has the answer with Gemini on Google Distributed Cloud.

Google Distributed Cloud with Gemini: The On-Prem AI Cloud

Google and Google Cloud have long been at the forefront of AI innovation. As early as 2001, Google was using machine learning to improve its search engine results. In 2006, the company employed AI to enable real-time multi-language translation. In 2016, the company deployed the first custom-made AI silicon, its Tensor Processing Unit (TPU), and released TensorFlow, an open-source, scalable deep learning framework.

In 2023, Google launched its Gemini multimodal generative AI model and assistant, which some consider to be the most comprehensive model for the enterprise. Now, with GDC as a delivery mechanism, customers can run Google's most powerful generative AI models on-premises. By bringing Gemini on-premises, Google will power the

enterprise agentic layer with a leading AI model. Google Cloud is currently the only major cloud service provider bringing its leading model on-premises.

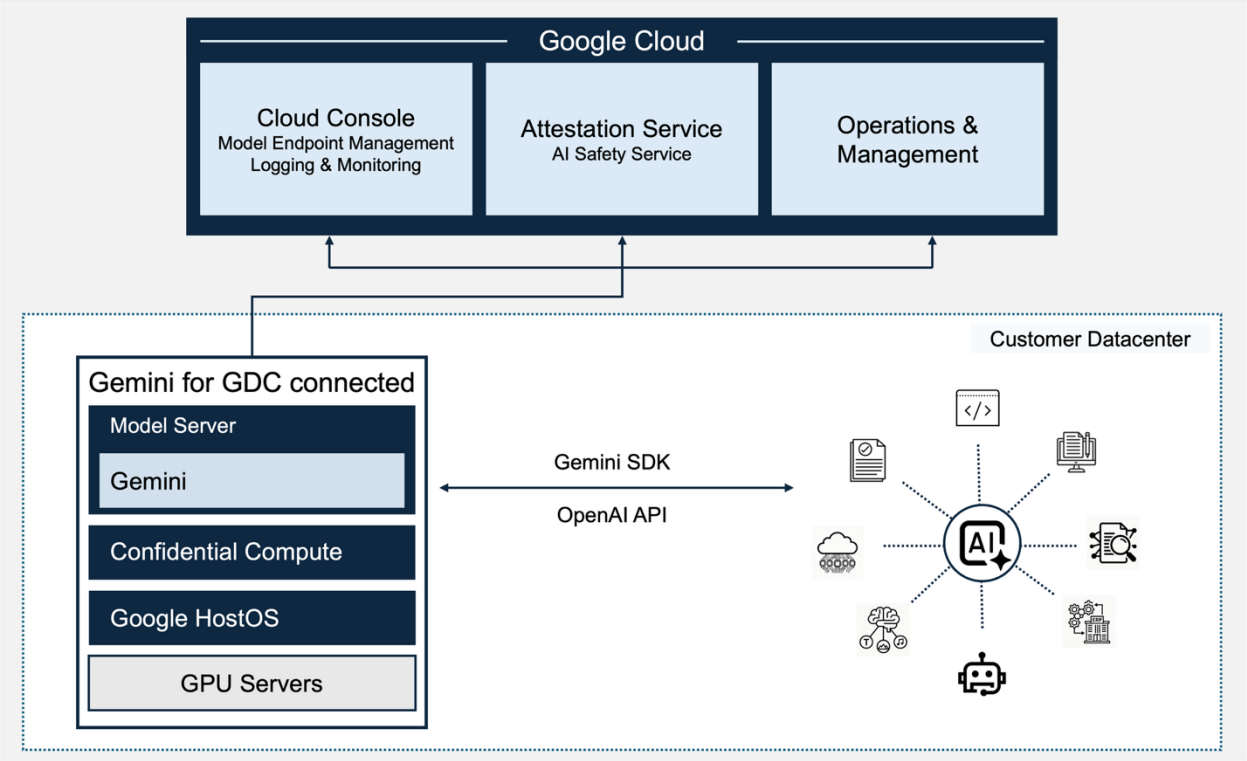
In short, Gemini on GDC brings the frontier model to the enterprise datacenter, enabling lower latency performance and on-premises control.

GDC: The AI Foundation

GDC is a fully managed service that brings hardware and software together to deliver Google Cloud's foundational and high-value capabilities in an on-premises environment. Additionally, it can be consumed through a variety of means and form factors to meet the security and management needs of enterprise IT.

- **GDC connected** is an appliance-like delivery model in which finely tuned and secured hardware is deployed across the enterprise to deliver the cloud operating model with seamless integration to Google-managed services. The GDC environment is monitored for performance and support, making this model well-suited for organizations requiring the lowest latency for local processing, such as telco edge computing, smart factories, or retail.
- **GDC air-gapped** brings the Google Cloud environment on-premises while remaining fully disconnected from the internet and public cloud. It enables strict compliance and sovereignty for organizations in industries such as defense, government, healthcare, and public utilities.
- **GDC software only** enables customers to install the GDC platform on hardware procured by the customer, guided by Google's reference architectures. In this consumption model, users can deploy GDC as a connected service with the ability to integrate into existing datacenter operations.

FIGURE 3: GEMINI ON GDC CONNECTED



Gemini for GDC delivers the full Gemini environment on-premises, accelerating innovation.
Source: Moor Insights & Strategy

MI&S finds GDC's flexibility particularly compelling, as customers can define their own cloud experience and connectivity. This model delivers the benefits of the cloud, including agility and the ability to quickly set up an enterprise-to-edge AI environment, along with predictable costs and secure operations.

GEMINI: THE AI STACK

While GDC is the delivery mechanism for Google's on-premises AI solution, Gemini is the model that delivers a turnkey-like experience for administrators and rich functionality to developers and business users alike. Gemini is deployed as a whole, non-quantized model, enabling the highest accuracy, context retention, and nuanced understanding required for complex reasoning and in-depth research.

One of Gemini's strengths is its native support for multimodality, enabling coherent interpretation and synthesis across various data formats, including text, images, audio, and more. The model is designed with extremely long context windows, enabling it to parse, recall, and reason over extensive documents or datasets in a single session.

This capability is particularly advantageous for use cases such as legal research, scientific literature analysis, and other domains where comprehensive context is critical.

Gemini also offers robust language capabilities and advanced reasoning functions, extending utility to global enterprises operating in multilingual environments. Local AI control is a critical capability; RAG, multimodal operations, and context handling are governed by enterprise policies on data access, residency, and compliance. For enterprise IT, this creates confidence that sensitive workloads and data remain managed under local policies.

Importantly, Gemini is tightly integrated with Google's Vertex AI platform, running as a locally managed service. This architecture supports end-to-end workflows — including RAG, endpoint connection, and inference — without requiring external API calls. It also supports multi-model and custom model deployments, benefiting organizations with diverse workloads and evolving requirements.

MI&S finds Gemini and Vertex AI compelling for their unwavering focus on user and developer experience — without sacrificing the control that IT requires to secure and manage the environment. For enterprises beginning their generative and agentic AI journey, Gemini running in Vertex AI can be a valuable environment, where even multi-model support can be stood up with straightforward point-and-click workflows.

GDC and Gemini — The Cloud AI Stack Comes to the Enterprise

Deploying Gemini on Google Distributed Cloud delivers tangible technical benefits that are particularly relevant for enterprise IT environments that prioritize performance at scale, security, and control, as well as advanced AI capabilities. The architecture supports the enterprise use cases that MI&S consistently hears about in its engagements with enterprise IT leaders, including native multimodal information understanding, code generation, large-context understanding, agent-driven AI, RAG, multi-turn Q&A, and language translation.

GDC's flexibility in deployment further enhances operational agility. Enterprises can connect to Google Cloud for managed services integration or run in air-gapped, fully isolated environments when tighter compliance and operational autonomy are required. Edge support enables real-time AI inference at remote sites with minimal latency, powering distributed and mission-critical applications across manufacturing, retail, healthcare, and more. In practical terms, GDC and Gemini bring AI to where the enterprise data resides, delivering the lowest latency and highest levels of security.

Across the board, it seems that Google Cloud has focused on removing friction from enterprise AI activation. For organizations seeking to deploy a fully integrated AI stack and accelerate time to innovation — the time to first token — Google AI services merit consideration. For enterprises requiring greater control, performance, or regulatory adherence, GDC with Gemini is especially compelling.

Activating Generative AI — A Practitioner's Guide

Based on interactions with IT and business executives, MI&S has identified three strategic pillars for successfully activating AI across the enterprise:

1. Establish Strategy and Governance

- Define an enterprise-wide vision: Generative AI will impact the entire enterprise. Gather stakeholders from across business units to guide the strategy.
- Establish a governance framework: Ensure all deployments are guided by shared principles, ethics, and regulatory obligations from day one.
- Work with a trusted partner: Select a technology partner with an established record of deploying generative AI at scale to navigate the complexity.

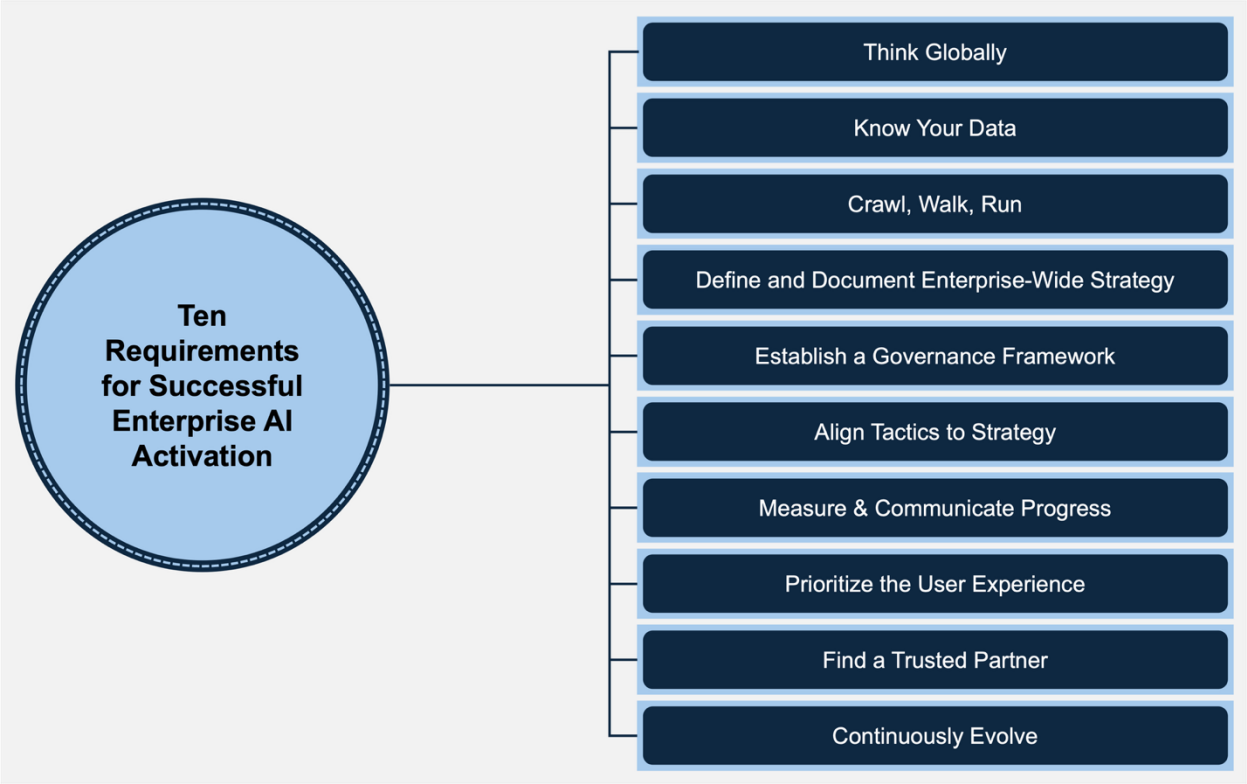
2. Architect for the Hybrid Reality

- Consider the entirety of your environment: Account for the unique needs of distributed environments, from the core datacenter to the edge.
- Understand your data estate: Know where data resides and how it is governed. This is critical to success in a distributed enterprise.
- Align deployment to strategy: Determine which workloads belong in the public cloud versus on-premises or air-gapped environments based on latency and sovereignty needs.

3. Execute, Measure, and Evolve

- Prioritize high-impact use cases: Adopt a "crawl, walk, run" approach. Start with high-impact, low-risk use cases to demonstrate value.
- Focus on user experience: Begin the development of every use case with the user experience in mind to ensure adoption.
- Measure and communicate: Track progress frequently and share updates with stakeholders to maintain support.
- Continuously evolve: Treat AI activation not as a destination but as a foundation for ongoing innovation.

FIGURE 4: ACTIVATING GENERATIVE AI IN THE ENTERPRISE



Enterprise-wide AI activation is an evergreen process requiring organization-wide coordination.
Source: Moor Insights & Strategy

CALL TO ACTION

AI is not a passing trend — it will reshape how businesses operate and how work gets done. For consumers, this shift will feel seamless as AI becomes embedded into devices, mobile operating systems, and everyday applications. The underlying complexity is significant, but platforms like Google minimize that complexity for end users.

Enterprises face a different reality. Achieving cloud-like simplicity across distributed environments is difficult, and building, deploying, and managing an AI stack at scale is a major undertaking. This is why the public cloud has become the preferred starting point for enterprise AI: it accelerates activation and abstracts much of the operational burden, even if costs can vary.

However, the public cloud cannot meet every requirement. Performance, privacy, and regulatory constraints mean portions of the AI stack must move closer to where data is

created and governed. MI&S believes that developing a clear, enterprise-wide activation plan — one that accounts for distributed data, infrastructure, and workloads — is essential.

After evaluating the market, MI&S sees Gemini on GDC as well-suited to deliver performant, secure, and scalable AI with predictable economics. Data-centric enterprises beginning their generative AI journey should consider GDC with Gemini.

For more information, please visit cloud.google.com/distributed-cloud.

Important Information About This Paper

Contributor

Matt Kimball, Vice President and Principal Analyst, Datacenter Compute and Storage

Publisher

Patrick Moorhead, CEO, Founder and Chief Analyst at Moor Insights & Strategy

Inquiries

Contact us if you would like to discuss this report, and Moor Insights & Strategy will respond promptly.

Citations

This paper can be cited by accredited press and analysts but must be cited in-context, displaying author's name, author's title, and "Moor Insights & Strategy." Non-press and non-analysts must receive prior written permission by Moor Insights & Strategy for any citations.

Licensing

Disclosures

Google commissioned this paper. Moor Insights & Strategy provides research, analysis, advising, and consulting to many high-tech companies mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

Disclaimer

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. Moor Insights & Strategy disclaims all warranties as to the accuracy, completeness, or adequacy of such information and shall have no liability for errors, omissions, or inadequacies in such information. This document consists of the opinions of Moor Insights & Strategy and should not be construed as statements of fact. The opinions expressed herein are subject to change without notice.

Moor Insights & Strategy provides forecasts and forward-looking statements as directional indicators and not as precise predictions of future events. While our forecasts and forward-looking statements represent our current judgment on what the future holds, they are subject to risks and uncertainties that could cause actual results to differ materially. You are cautioned not to place undue reliance on these forecasts and forward-looking statements, which reflect our opinions only as of the date of publication for this document. Please keep in mind that we are not obligating ourselves to revise or publicly release the results of any revision to these forecasts and forward-looking statements in light of new information or future events.