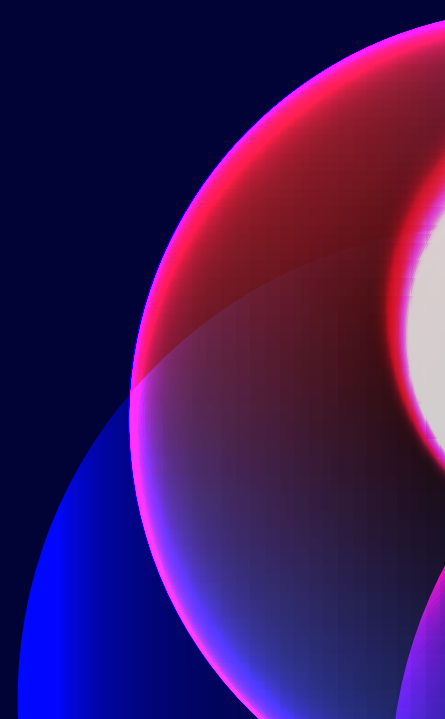


AI UNLOCKED

# AI READY INFRASTRUCTURE

BUILDING THE ENTERPRISE FOUNDATION FOR AGENTIC AI



# Contents

|                                                                 |           |
|-----------------------------------------------------------------|-----------|
| <b>Executive Summary</b>                                        | <b>03</b> |
| <b>Why GCC Enterprises Need a New Infrastructure Blueprint</b>  | <b>04</b> |
| <b>The GCC Market Context: Why This Matters Now</b>             | <b>05</b> |
| <b>What Cisco and NVIDIA Together Change for the Enterprise</b> | <b>06</b> |
| <b>GBM Point of View</b>                                        | <b>08</b> |
| <b>Conclusion</b>                                               | <b>09</b> |
| <b>References &amp; Validated Resources</b>                     | <b>10</b> |

## Executive Summary

Across the GCC, AI has moved from boardroom ambition to infrastructure imperative. Governments and enterprises are committing unprecedented capital to the compute, connectivity, and governance platforms that will define the next decade of economic growth. The question is no longer whether to invest in AI infrastructure — it is how to build it correctly.

**An AI-ready infrastructure is no longer just a GPU cluster.** For enterprises moving from experimentation to production, the stack must support continuous inference, retrieval-augmented generation (RAG), model routing, long-context memory, orchestration, security, and observability — across both data centre and edge environments.

At GTC 2026, NVIDIA framed the next wave of enterprise AI around inference factories and agentic systems. Simultaneously, Cisco announced a landmark expansion of its Secure AI Factory with NVIDIA — extending validated AI architecture from the core data centre all the way to local edge sites, compressing deployment timelines from months to weeks.

**From GBM's perspective, this is the defining architecture moment for GCC enterprises.** The organisations that align to validated, security-embedded, inference-optimised infrastructure today will lead their sectors tomorrow.

**\$30B+**

committed to GCC AI data-centre capacity through 2030

**84%**

of GCC organisations have integrated AI into at least one operation

**\$320B**

AI's projected contribution to the Middle East economy by 2030

“AI has made a giant leap — reasoning and agentic AI demand orders of magnitude more computing performance.”

— Jensen Huang, Founder & CEO, NVIDIA | GTC March 2025

# 1. Why GCC Enterprises Need a New Infrastructure Blueprint

Traditional enterprise data centres were optimised for virtualisation, transactional applications, and general-purpose traffic. The AI era changes the design centre entirely. Production AI must deliver token throughput, predictable response times, high-bandwidth GPU communication, data locality for retrieval, and security controls across models, agents, prompts, tools, and outputs.

NVIDIA defines agentic AI as systems that reason, plan, and act across multiple steps. Cisco’s GTC 2026 announcements explicitly position Secure AI Factory with NVIDIA for trusted agentic and physical AI applications — from core to edge. In practice, this means infrastructure must support model execution, orchestration, vector retrieval, policy, and runtime control simultaneously.

Industry analysts have gone further: IDC argues that agentic AI is becoming critical infrastructure because once systems can access proprietary knowledge and connect to enterprise tools, they move from a productivity layer into the operating core of the business. That is exactly why infrastructure decisions have become board-level decisions in the GCC.

## Traditional DC vs. AI-Ready Enterprise Infrastructure

| Dimension            | Traditional Enterprise DC                          | AI-Ready Enterprise Stack                                                              |
|----------------------|----------------------------------------------------|----------------------------------------------------------------------------------------|
| Primary Optimisation | VM density, application uptime, storage efficiency | Token throughput, latency, GPU utilisation, policy-controlled AI execution             |
| Compute Profile      | CPU-heavy, bursty accelerators                     | GPU-dense servers plus high-performance CPUs and DPUs (e.g. NVIDIA BlueField)          |
| Network Profile      | General east-west traffic                          | High-bandwidth east-west GPU fabric plus north-south flows to users, storage, and edge |
| Data Profile         | Application databases and file stores              | AI-ready pipelines, vector / RAG stores, KV cache, long-context memory                 |
| Control Plane        | Hypervisor and IT operations                       | Kubernetes, GPU schedulers, model serving, runtime routing, guardrails, observability  |
| Security Scope       | Endpoints, applications, network, IAM              | Models, prompts, agents, connectors, workloads, data paths, infrastructure, SOC        |

## 2. The GCC Market Context: Why This Matters Now

The Gulf Cooperation Council is executing one of the most ambitious AI infrastructure programmes in the world. Public commitments exceed \$30 billion in planned AI data-centre investment between now and 2030. The UAE's Stargate UAE project alone targets 1 gigawatt of AI compute capacity. Saudi Arabia's national AI champion HUMAIN is already building data centres in Riyadh and Dammam. Qatar, Kuwait, Oman, and Bahrain are each advancing sovereign AI infrastructure programmes aligned with their national visions.

The pace of enterprise adoption mirrors the ambition of governments. AI adoption among GCC organisations has accelerated from 62% in 2023 to 84% in 2025, with agentic AI, infrastructure development, and workforce upskilling now at the centre of corporate strategy. AI is projected to contribute up to \$320 billion to the broader Middle East economy by 2030.

This context is critical for GBM's positioning: our GCC customers are not evaluating AI infrastructure as a long-term project. They are making procurement decisions now, under regulatory, sovereign, and competitive pressures that make validated, secure, partner-led architectures the only viable path to production.

## 3. What Cisco and NVIDIA Together Change for the Enterprise

The market opportunity is not better hardware in isolation. It is the availability of validated, partner-led blueprints that reduce deployment risk and compress time to value. At GTC 2026, Cisco announced a major expansion of its Secure AI Factory with NVIDIA — delivering a unified framework for deploying AI across the entire enterprise infrastructure, from the central data centre to local edge sites.

### A. Inference at the Edge: Bringing AI to Where Decisions Are Made

AI inference now happens where data lives and where decisions cannot wait — whether on the hospital floor, at a port terminal, or within a financial trading environment. Cisco's GTC 2026 expansion introduces support for NVIDIA RTX PRO 4500 Blackwell Server Edition GPUs across the Cisco UCS and Unified Edge portfolios, enabling enterprises to run mission-critical AI workloads at edge sites without the energy cost and footprint of data-centre-scale hardware.

For GCC enterprises with distributed operations across retail, logistics, utilities, and government services, this is transformative. AI capability can now be deployed at the site where value is created — not only at the central data centre.

### B. Security Built Into the Architecture, Not Bolted On

Cisco differentiates Secure AI Factory through security embedded at every layer of the stack — models, agents, applications, workloads, and infrastructure — with observability integrated through Cisco AI Defense and Splunk. At GTC 2026, Cisco further extended Hybrid Mesh Firewall policy enforcement to NVIDIA BlueField DPUs and announced that Cisco AI Defense will govern NVIDIA's OpenShell runtimes — part of the NVIDIA Agent Toolkit — ensuring that every tool and action performed by an AI agent is continuously monitored and validated.

For GCC enterprises operating in regulated sectors — banking, healthcare, government, and critical national infrastructure — this is not a feature. It is a prerequisite.

## C. Validated Designs That Shorten Time to Production

Cisco's AI POD reference architecture packages prescriptive server, GPU, and switching guidance into a validated, modular blueprint. NVIDIA Enterprise Reference Architectures provide prescriptive building blocks for compute, networking, software, observability, and storage. Together, they transform what was previously a complex, multi-vendor integration exercise into a repeatable deployment model. Cisco confirmed at GTC 2026 that enterprises can now move from pilot to full-scale production without stitching together disconnected systems, compressing deployment timelines from months to weeks.

## D. Sovereignty and Regional Relevance

Cisco frames Secure AI Factory as a modular architecture for enterprises, neoclouds, sovereign clouds, and service providers. The GCC's requirements for sovereign AI deployment models — data residency, regulated-sector controls, low-latency edge use cases, and carrier-grade reliability — are directly addressed by this architecture. The Cisco AI Grid with NVIDIA reference design, announced at GTC 2026, further enables service providers to deliver managed edge AI services with carrier-grade sovereignty.



## 4. GBM Point of View

**GBM frames the AI-ready infrastructure as an enterprise platform decision — not a server refresh.**

The strongest commercial message for GCC enterprise customers is that AI success depends on aligning Cisco’s networking, security, observability, and pod-based architecture with NVIDIA’s accelerated compute, enterprise reference architectures, and AI factory design principles — as a single, validated, managed platform.

For GBM customers in government, healthcare, banking, financial services, energy, and large enterprise, the winning conversation is not ‘which GPU?’. It is: How do we deploy trusted AI at scale — on-premises or in hybrid form — with the performance, governance, and operational discipline required for regulated and mission-critical environments?

**That is where GBM differentiates:** translating validated Cisco and NVIDIA architectures into deployable, secure, and managed AI platforms designed specifically for GCC enterprises.

### GBM Design Principles for GCC AI Infrastructure Conversations

| GBM Design Principle                        | Why It Matters for GCC Enterprises                                                                                                                                                    |
|---------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Design for inference first</b>           | Reasoning and agentic workloads drive continuous token demand. Infrastructure economics must reflect inference — not one-time training.                                               |
| <b>Treat the network as an AI asset</b>     | Poor fabric design reduces GPU utilisation and undermines latency-sensitive multi-node inference. Network is not a commodity layer — it is a performance multiplier.                  |
| <b>Bring data into architecture early</b>   | RAG, context retrieval, and long-context memory are throughput and governance problems, not software features. Storage and pipelines must be co-designed with compute.                |
| <b>Embed governance into the stack</b>      | Models and agents require runtime controls, inherited permissions, auditability, and security telemetry — from silicon to software. This is non-negotiable for GCC regulated sectors. |
| <b>Standardise on validated pod designs</b> | Validated Cisco AI POD and NVIDIA Enterprise Reference Architectures reduce deployment risk and shorten time to value for enterprise procurement teams.                               |
| <b>Measure in business terms</b>            | Token cost, response time, model quality, and incident risk are executive metrics. GBM should lead customer conversations at this level — not at the infrastructure level alone.      |

## Conclusion

An AI-ready infrastructure is a full-stack operating environment for enterprise intelligence. It must combine accelerated compute, AI-optimised networking, data pipelines, orchestration, observability, and security into a coherent, validated architecture. Cisco and NVIDIA are now presenting this as a modular foundation for agentic AI, inference at scale, and sovereign deployment models.

For GCC enterprises, the window for advantage is open — and narrowing. Organisations that build AI-ready infrastructure correctly today will compound that investment across every application, agent, and workflow they deploy tomorrow. Those that defer, or that build on unvalidated architectures, face compounding technical debt and governance risk.

**GBM is positioned to be the trusted partner that turns global AI architecture into local, deployable, sovereign, and commercially compelling reality for the enterprises that will define the GCC's digital decade.**

# References & Validated Sources

- [1] NVIDIA GTC 2026 Event Page — <https://www.nvidia.com/gtc/>
- [2] NVIDIA Blackwell Ultra AI Factory Platform Paves Way for Age of AI Reasoning (March 2025) — <https://nvidianews.nvidia.com/news/nvidia-blackwell-ultra-ai-factory-platform-paves-way-for-age-of-ai-reasoning>
- [3] NVIDIA Agentic AI Platform Overview — <https://www.nvidia.com/en-us/ai/>
- [4] NVIDIA Enterprise Reference Architectures — <https://www.nvidia.com/en-us/technologies/enterprise-reference-architecture/>
- [5] NVIDIA Enterprise AI Factory Validated Design — <https://www.nvidia.com/en-us/solutions/ai-factories/validated-design/>
- [6] NVIDIA ERA White Paper — Building AI Factories for the Enterprise — <https://docs.nvidia.com/enterprise-reference-architectures/white-paper/latest/ai-factories-for-enterprise.html>
- [7] NVIDIA BlueField-4 Inference Context Memory Platform — <https://developer.nvidia.com/blog/introducing-nvidia-bluefield-4-powered-inference-context-memory-storage-platform-for-the-next-frontier-of-ai/>
- [8] NVIDIA Vera CPU — Purpose-Built for Agentic AI — <https://nvidianews.nvidia.com/news/nvidia-launches-vera-cpu-purpose-built-for-agentic-ai>
- [9] Cisco Secure AI Factory with NVIDIA — Solution Page — <https://www.cisco.com/site/us/en/solutions/artificial-intelligence/secure-ai-factory/index.html>
- [10] Cisco Secure AI Factory with NVIDIA — Technical FAQ — <https://www.cisco.com/c/en/us/solutions/collateral/artificial-intelligence/secure-ai-factory-nvidia-faq.html>
- [11] Cisco Expands Secure AI Factory with NVIDIA from Core to Edge (GTC 2026, March 16) — <https://newsroom.cisco.com/c/r/newsroom/en/us/a/y2026/m03/cisco-secure-ai-factory-with-nvidia-GTC-2026.html>
- [12] Cisco AI POD Infrastructure for NVIDIA 400-9-8-2 Enterprise Reference Architecture — <https://www.cisco.com/c/en/us/products/collateral/servers-unified-computing/ai-pod-ucs-c885a-servers-nexus-936-4e--sg2-switches.html>
- [13] Cisco Secure AI Factory Unlocks Enterprise Data for Agentic AI (September 2025) — <https://newsroom.cisco.com/c/r/newsroom/en/us/a/y2025/m09/cisco-secure-ai-factory-with-nvidia-unlocks-enterprise-data-for-agentic-ai.html>
- [14] IDC: Agentic AI is Critical Infrastructure — <https://www.idc.com/resource-center/blog/agentic-ai-is-critical-infrastructure/>
- [15] IDC FutureScape 2026 — Predictions on the Rise of Agentic AI — <https://my.idc.com/getdoc.jsp?containerId=prUS53883425>
- [16] Gartner Hype Cycle — Top AI Innovations in 2025 (August 2025) — <https://www.gartner.com/en/newsroom/press-releases/05-08-2025-gartner-hype-cycle-identifies-top-ai-innovations-in-2025>
- [17] Gartner Predicts %60 of Brands Will Use Agentic AI by 2028 (January 2026) — <https://www.gartner.com/en/newsroom/press-releases/15-01-2026-gartner-predicts-60-percent-of-brands-will-use-agentic-ai-to-deliver-streamlined-one-to-one-interactions-by-2028>
- [18] Analysys Mason: GCC AI Data-Centre Investment to Reach \$7–5B in 2026 — <https://www.analysismason.com/about-us/news/predictions-2026/prediction-ai-investment-gcc/>
- [19] Precedence Research: How GCC Countries Are Transforming AI Investment (March 2026) — <https://www.precedenceresearch.com/insights/gcc-ai-economic-growth-strategy-middle-east>
- [20] Wamda: AI Capital Flooding the World, GCC Emerging as Hub (March 2026) — <https://www.wamda.com/03/2026/ai-capital-flooding-world-gcc-emerging-hub>

## About GBM

With more than 35 years of experience, 7 offices, and over 1500 employees across the region, Gulf Business Machines (GBM) is a leading end-to-end digital solutions provider, offering the region's broadest portfolio, including industry-leading digital infrastructure, digital business solutions, security, and services. We have nurtured partnerships since 1990 with the world's leading technology companies and invested in a talented, skilled workforce to implement solutions that cater to customer's specific, complex, and diverse business needs.

Some of our strategic partners in the Gulf include IBM as their sole distributor throughout the GCC (excluding Saudi Arabia and selected IBM products and services) and Cisco as a Gold and Master Partner (the highest level of certification at Cisco).

### GBM

Emarat Atrium Building, Block B, 3rd floor,  
Sheikh Zayed Road, P.O. Box 9226, Dubai, UAE

 [Gbmme.com](https://gbmme.com) |  [linkedin.com/company/gbm](https://www.linkedin.com/company/gbm) |  [gbmmiddleeast](https://www.instagram.com/gbmmiddleeast)