cognizant®

Building an AI-ready infrastructure for agentic enterprise software

Contents

Executive summary 03

Building the runway for enterprise-scale agentic AI 04

From scripted automation to proactive orchestration 05

Why private AI becomes essential at scale 05

Platform foundation: Performance and governance, designed together 06

Multitenancy and GPU partitioning: MIG and vGPU 06

Cognizant AI Factory-layered reference architecture for agentic systems 07

- Infrastructure layer to strengthen the AI infrastructure foundation 07
- Management layer to operate the AI estate like a utility 07
- AI resilience platform for single pane of visibility, governance and recovery 08
- Monitoring, governing and managing the agentic layer 08
- Agentic layer spanning functional, industry and IT ops agents 08

What Cognizant adds beyond infrastructure 08

Delivery approach: From strategy to scale 09

Performance, economics and elasticity by design 09

Evidence from regulated environments 09

Positioning within the broader AI stack 09

Managing risk in agentic systems 09

Building the runway for scale 10

Appendix: Glossary 10

Executive summary

Agentic AI is moving enterprise software beyond task automation and into end-to-end work execution across an entire value stream. Rather than waiting on handoffs or simple triggers, agents can interpret intent, reason over enterprise context, apply policy and execute coordinated actions across business systems such as ERP and CRM. This shift reduces the distance from decision to execution, lowers exceptions and improves measurable throughput.

As organizations move from isolated proofs of concept to fleets of production agents, governance, security, auditability, cost predictability and operational resilience become decisive. Without a purpose-built platform and operating model, autonomy becomes risky and experimentation becomes uncontrolled spend.

Cognizant's approach is grounded in our AI Factory model, standardized on NVIDIA AI Enterprise, and implemented through a layered reference architecture and delivery system designed for enterprise agent fleets. This paper describes the shift to proactive orchestration, why private AI becomes essential at scale and why a platform and an operating model are required to run agent fleets safely and economically.

Building the runway for enterprise-scale agentic AI

Agentic AI is pushing enterprise software beyond isolated automation toward systems that can interpret intent, coordinate across applications and execute work end to end, raising new demands on infrastructure as organizations scale from pilots to production. This white paper explains why traditional AI setups fall short at enterprise scale and why platform design, governance and economics must be addressed together to run agent fleets safely and predictably. It outlines how private AI environments become essential as autonomy increases, examines the infrastructure and platform foundations required to balance performance with control and presents a layered reference architecture for operating agentic systems at scale. The paper also describes Cognizant's delivery approach and operating model, showing how organizations can build an AI-ready infrastructure that supports sustained and governed adoption of agentic enterprise software.

From scripted automation to proactive orchestration

Traditional enterprise automation performs well when work is predictable: a rule is fired, a workflow is executed and a task is completed. However, modern work spans applications, data domains, approvals and exceptions—exactly where simple automation degrades.
Agentic AI changes the operating model. Agents can:

- Interpret user and system intent
- Reason over enterprise data and knowledge
- Apply role-based policy and compliance rules
- Coordinate actions across applications and workflows

At scale, orchestration introduces new requirements. Hundreds of agents operating alongside business users and mission-critical systems demand strong governance, clear accountability and predictable operations. Infrastructure therefore becomes the runway that determines whether agentic AI remains a pilot capability or becomes a durable production platform.

Why private AI becomes essential at scale

Most agent initiatives begin as a proof of concept in the public cloud to accelerate experimentation and early validation. Once agents move into sustained production, the question shifts from “Can it work?” to “Can we trust it?”, control it, and run it economically— and reliably?”

At scale, organizations require:

Control over data residency and sovereignty

Consistent enforcement of enterprise security and compliance policies

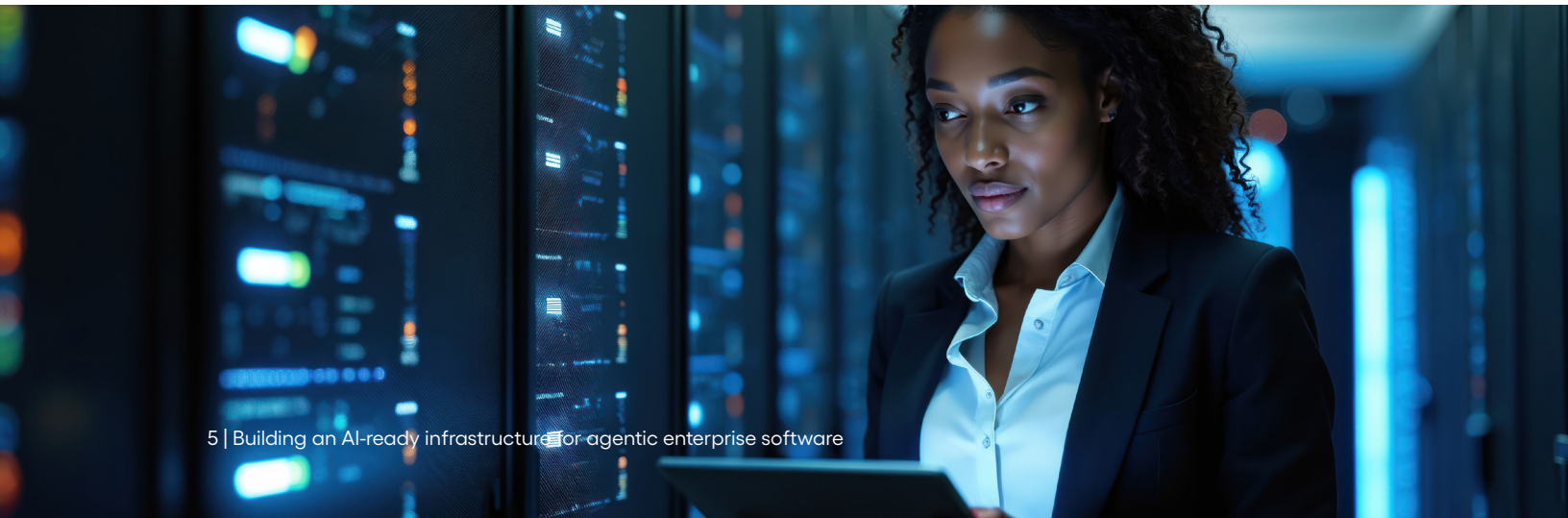
Predictable model behavior and lifecycle management

Full auditability of agent actions

These needs commonly drive programs toward private AI—deployed on-premises or in tightly governed virtual private clouds. So, sensitive content stays within trusted boundaries and material agent actions are captured for audit.

Private AI can also strengthen adoption economics as usage grows. Owned or reserved compute supports more predictable cost structures than pure pay-as-you-go models, enabling internal chargeback (tokenomics), forecasting and continuous optimization.

Once private AI becomes the destination, the next question is foundational, “What platform design can deliver performance and governance together without forcing tradeoffs?”.



Platform foundation: Performance and governance, designed together

In Cognizant's experience, NVIDIA AI Enterprise provides a strong foundation for private, multitenant AI environments that must deliver high performance while meeting enterprise governance expectations.

The platform combines GPU-accelerated infrastructure with an enterprise-validated software stack, including:

- CUDA and RAPIDS for accelerated data processing
- Triton for scalable inference
- TensorRT-LLM for optimized LLM performance

It integrates cleanly with enterprise platforms such as Kubernetes, VMware, and Red Hat, enabling organizations to extend familiar operational models into the AI estate.

Multitenancy and GPU partitioning: MIG and vGPU

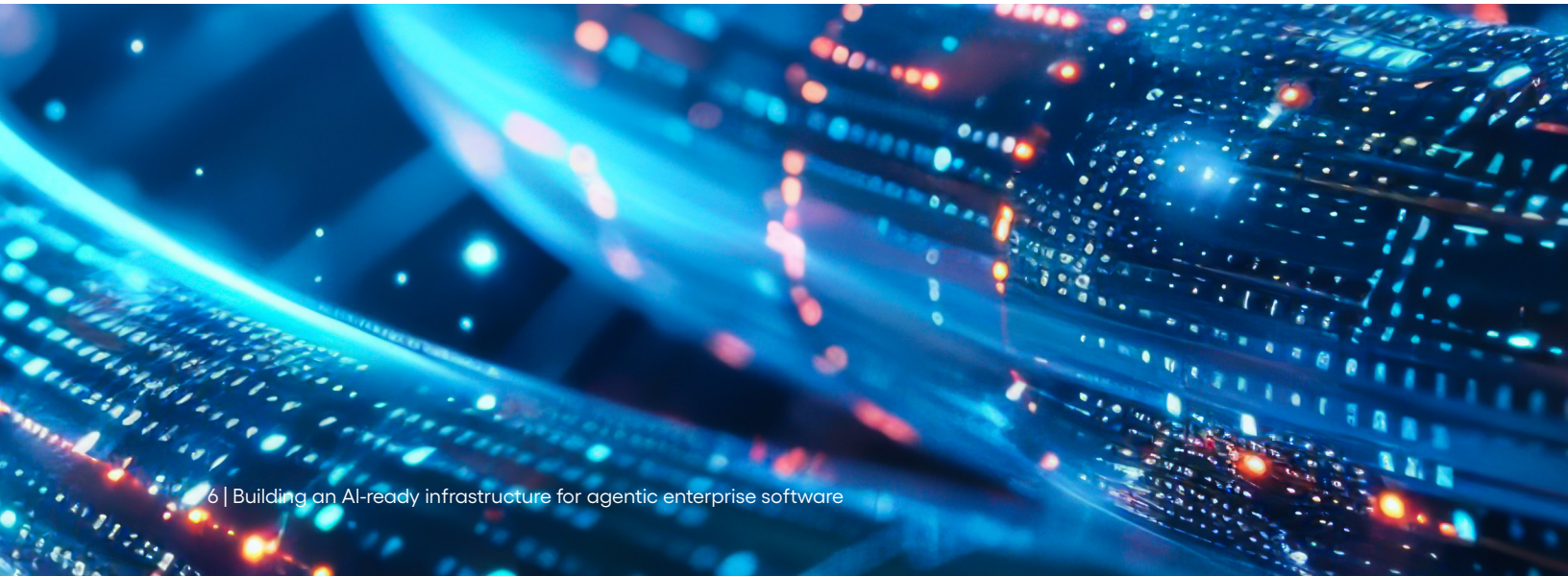
To operationalize multitenancy, GPU capacity must be shareable because dedicated GPUs per team quickly become cost prohibitive and underutilized as agent fleets grow. At the same time, sharing cannot come at the expense of isolation, predictable latency/throughput, or auditability—especially when agents are running alongside business critical systems and regulated data. In our delivery programs, this is exactly why we use MIG and vGPU. They let multiple teams and agent fleets safely share the same physical cards while preventing noisy neighbor behavior, delivering the isolation auditors can follow and the utilization levels finance expects.

MIG and vGPU are related, but distinct approaches:

- In vGPU mode, GPU memory is statically partitioned, while compute is time shared among virtual machines (VMs)
- In MIG mode, both memory and compute are statically partitioned into dedicated slices, providing stronger, hardware level isolation and more predictable performance per slice

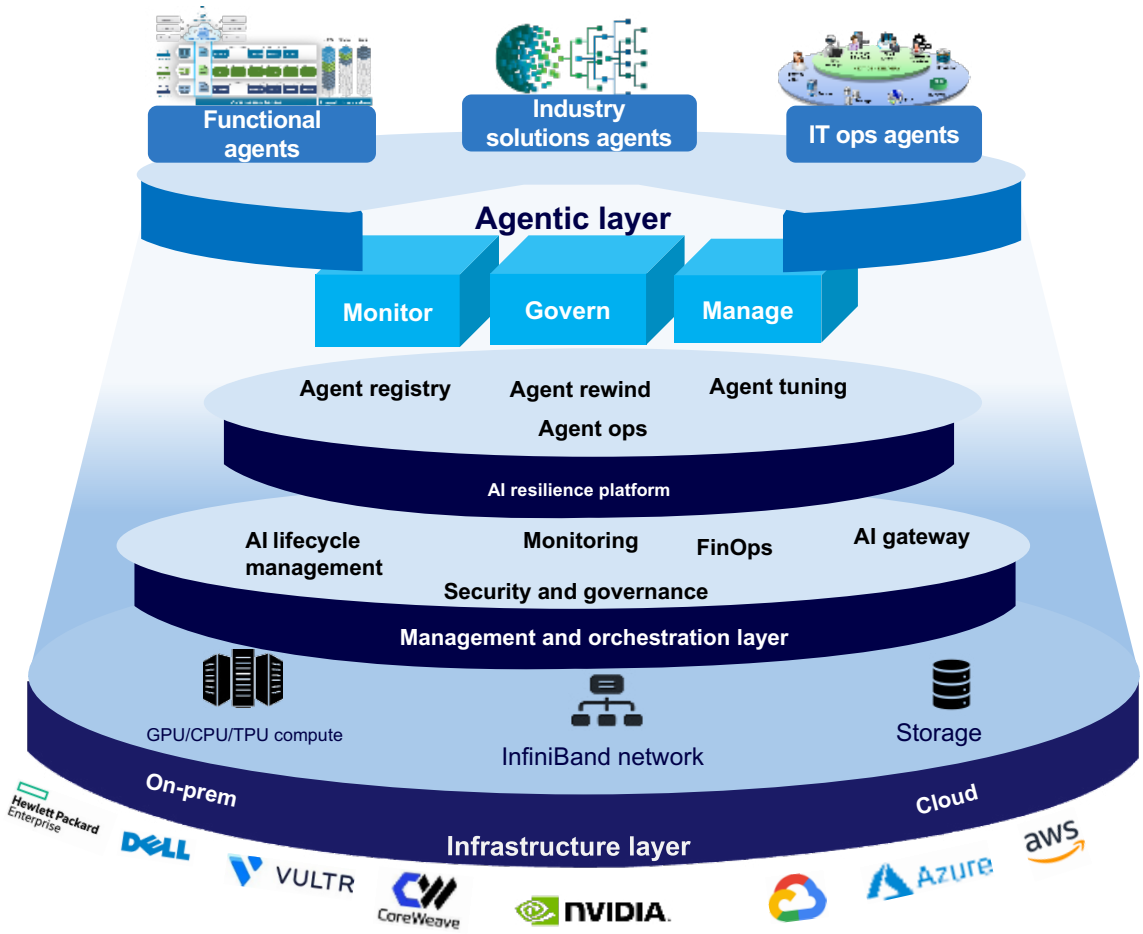
Used appropriately, these capabilities allow the platform to scale economically by increasing utilization without losing the governance properties such as isolation, predictability and traceability required for enterprise production.

A platform foundation is only the base layer. Scaling agentic systems across domains and risk tiers requires an architecture that separates concerns and makes governance enforceable by design.



Cognizant AI Factory-layered reference architecture for agentic systems

Our reference architecture is intentionally layered as part of the Cognizant AI Factory because it mirrors how enterprises build and run production platforms. You start with a standardized infrastructure foundation and add a management layer to operate it at scale. Then, introduce a resilience and governance plane that can continuously monitor, govern and manage autonomous agent behavior. This separation of concerns is what allows Cognizant AI Factory to evolve over time without losing control as agent fleets grow.



Infrastructure layer to strengthen the AI infrastructure foundation

The infrastructure layer provides the physical and virtual backbone: GPU/CPU/TPU compute, high speed networking (for example, InfiniBand) and enterprise storage, spanning on prem, cloud and hybrid patterns. This is where the NVIDIA stack anchors the performance foundation of Cognizant AI Factory, enabling accelerated training and inference as a governed enterprise capability rather than an ad hoc project environment.

Management layer to operate the AI estate like a utility

The management layer turns raw infrastructure into an operable platform by standardizing the capabilities enterprises rely on for day to day operations such as infrastructure monitoring, orchestration and scheduling, security controls and guardrails, resource scheduling and lifecycle management. This is the layer that makes multitenancy and production reliability practical. Because it is where capacity is provisioned, workloads are scheduled, telemetry is collected and platform policies are enforced consistently.

AI resilience platform for single pane of visibility, governance and recovery
As agent fleets scale, enterprises need more than platform monitoring. They need a purpose built plane that governs agent behavior as a first class operational concern. Cognizant AI Factory positions an AI resilience platform as that plane with a single pane of glass that provides centralized visibility and control, enabling unified observability and governance and recovery for agent infrastructure. It supports core governance outcomes such as tracing agent behavior (agents registering into a central control point), managing risk in real time through visibility into higher access agents and collecting actions/logs for audit and compliance needs.

Monitoring, governing and managing the agentic layer
Cognizant AI Factory model explicitly treats monitor, govern and manage as an operational discipline applied to the agent ecosystem—not a set of after the fact controls. In practical terms, this governance plane is where the enterprise establishes continuous control over agent fleets through mechanisms such as centralized visibility, policy enforcement, risk and permissions oversight and auditability—so autonomy scales safely.

Agentic layer spanning functional, industry and IT ops agents
At the top of the stack sits the agentic layer, covering functional agents, industry solution agents and IT ops agents, coordinated to execute real business work. In Cognizant AI Factory model, these agents are not treated as apps running somewhere. They are treated as managed and governed assets that operate within defined policy boundaries and are observable through the resilience and management layers below. This is what enables enterprises to move from isolated pilots to fleets of production agents without losing operational control.

What Cognizant adds beyond infrastructure

Infrastructure is necessary but not sufficient. Cognizant's differentiation is the delivery system built around Cognizant AI Factory—the assets, patterns and operating disciplines that turn the layered platform into a governed production capability. In Cognizant AI Factory terms, this value shows up as accelerators and practices that strengthen the agent build and orchestration, agent operations and evaluation, and continuous governance and resilience.

Agent Foundry to build and orchestrate agents across the agentic layer
Cognizant Agent Foundry provides a practical framework to design, compose, orchestrate and govern agents across heterogeneous technology stacks. So, teams can move from experiments to production using repeatable patterns rather than one off builds.

Neuro AI and multi agent accelerators for production grade AgentOps and evaluation
Cognizant Neuro® AI accelerators contribute reusable components for memory, tools, evaluation and AgentOps, plus orchestration patterns for mixed fleets, helping teams standardize how agents are tested, tuned and operated as production assets rather than prototypes.

Operating model blueprints to run the agentic layer with explicit objectives
Cognizant brings operating model patterns that align agent behavior to explicit service objectives and explainability requirements. So, the platform is managed like a governed utility, not a collection of unmanaged bots. These blueprints naturally connect into Cognizant AI Factory's management and resilience layers, monitoring, scheduling and providing security and operational controls.

DevSecOps and AIOps integration to embed governance into delivery and operations
Cognizant integrates policy as code, operational telemetry and enterprise operational tooling patterns, so AI workloads inherit the same discipline as the rest of the estate. This aligns directly with Cognizant AI Factory's monitor, manage and govern approach, connecting governance, monitoring and management into the platform lifecycle rather than treating them as add ons.

Delivery approach: From strategy to scale

Cognizant engagements emphasize cocreation rather than technology handoff. Programs typically follow four phases:

1. **Strategy and controls:** Define data boundaries, action policies and risk tiers to establish the policy envelope in which agents can operate.
2. **Design and pilot:** Set up the initial private AI runway and instrument early agents with evaluation harnesses, so quality is measured from day one.
3. **Build:** Scale agent fleets, integrate with ITSM and business systems, and activate AgentOps to monitor health, drift and ROI.
4. **Scale:** Introduce chargeback, capacity forecasting and cross-domain orchestration, so the platform operates as a governed utility.

Governance is embedded throughout, so compliance and resilience are engineered in from the outset rather than deferred until after early value is demonstrated.

Performance, economics and elasticity by design

Agent responsiveness is central to trust and adoption. A private, GPU-accelerated backbone keeps inference close to data and users, while maximizing throughput through batching and concurrency. Fractional allocation and pooling reduce idle capacity. Token-level accounting and FinOps practices provide transparency into cost per use case, enabling continuous tuning of prompts, model selection and routing strategies with clear economic feedback.

Evidence from regulated environments

In highly regulated industries, agentic patterns have delivered efficiency gains under strict governance. In one pharmaceutical program, orchestrated agents improved data anomaly resolution efficiency and reduced manual effort across structured and document-based sources. The same platform reduced data gaps across thousands of policies, demonstrating the combined impact of private AI patterns and domain-specific agent design.

Managing risk in agentic systems

The following three risks dominate enterprise deployments.

- Safety and misuse: Addressed through layered controls, identity enforcement, policy-checked tool access, audit logging and human-in-the-loop thresholds
- Operational fragility: Mitigated through AgentOps, sandboxed rollouts, evaluation-gated promotion and graceful degradation
- Change adoption: Addressed through codesign with business owners and embedded outcome metrics in executive dashboards

Agent-first thinking becomes practical when organizations build not only agents, but the operational fabric that keeps the ecosystem visible, accountable and resilient at scale.

Preparing enterprises for scalable agentic AI

Agentic AI is no longer experimental. When implemented with discipline, it compresses decision latency, optimizes work and turns enterprise data into continuous improvement. Achieving these outcomes requires more than models. It requires a secure, multitenant and economically transparent platform.

Our approach combines an NVIDIA-backed Cognizant AI Factory with Cognizant Agent Foundry, Cognizant Neuro AI accelerators and an operating model that embeds governance into speed. It is the same approach we use to modernize our own enterprise and the one we bring to clients, scaling from pilots to portfolios of production agents.

Appendix: Glossary

Agentic AI: AI systems composed of agents that can interpret goals and take coordinated actions across tools and systems.

Private AI: AI deployed within governed boundaries (on-premises or tightly governed virtual private cloud) for control, auditability and predictable operations.

MIG: Multi-Instance GPU partitions a physical GPU into dedicated slices (instances) with statically partitioned compute and memory.

vGPU: Virtual GPU presents GPU resources as virtual devices to VMs; memory is statically partitioned while compute can be time-shared among VMs.

Tokenomics: Internal chargeback/accounting model that ties AI usage costs to business value.

References

- NVIDIA. Multi-Instance GPU (MIG) Programming and Deployment Guide
- NVIDIA Agentic RAG guide
- Cognizant AI lab NeuroSAN Studio

Authors

- Arun Kumar, arun.skumar@cognizant.com
- Hemant Patade, hemant.patade@cognizant.com
- Kirk Beaumont, kirk.beaumont@cognizant.com

WF4413901