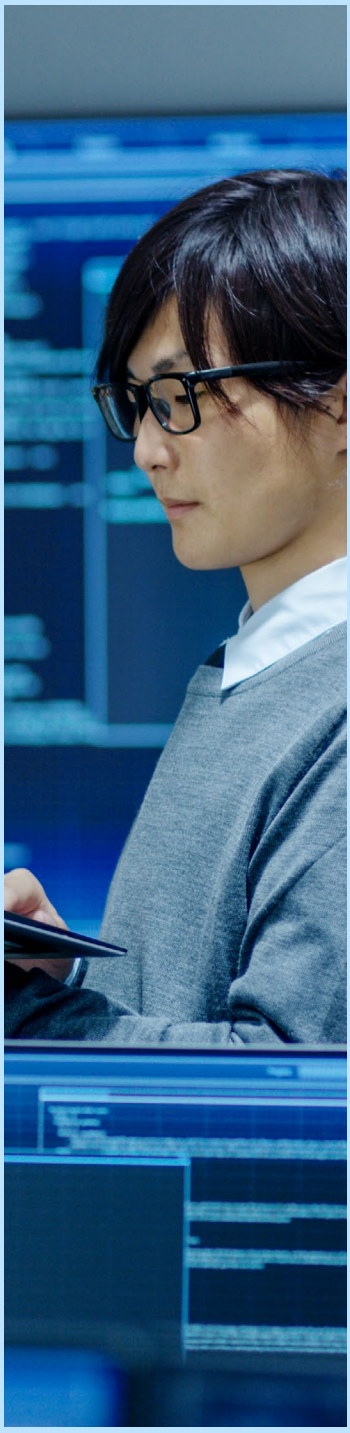
INFOSYS INFRASTRUCTURE FABRIC FOR AI

The foundation for Enterprise AI autonomy



Infosys
Navigate your next

Most organizations have proven that AI works.

The challenge now is scaling AI across the enterprise while maintaining control over data, models, infrastructure, governance, and costs.

The Infosys Infrastructure Fabric for AI provides the secure, scalable, and governed foundation enterprises need to move from AI experimentation to AI at scale.

The Enterprise AI challenge

As AI adoption accelerates, organizations must answer five critical questions:

- Who owns the data?
- Who owns the model?
- How do we avoid vendor lock-in?
- Which model should be used for which task?
- When does owning AI infrastructure become more economical than renting it?

From pilot to platform

MAXIMIZE

Get more value from existing AI investments

SCALE

Extend AI across the enterprise

EVOLVE

Move to an AI Factory operating model

Building Enterprise AI autonomy

Data Ownership

Maintain control of enterprise data with built-in residency, lineage, and governance, ensuring it stays within enterprise boundaries

Model Ownership

Retain ownership of AI models with built-in portability and auditability, free from vendor lock-in

Silicon Hedge

Avoid vendor lock-in with support for GPUs, CPUs, alternative accelerators, and hybrid infrastructure

Multi-Model Orchestration

Route every workload to the right model at the right cost through a unified governance layer

AI Factory Economics

Identify when owning AI infrastructure becomes more cost-effective than renting AI services

External Document © 2026 Infosys Limited

A layered foundation for AI at scale

Good AI Starts with a solid infrastructure foundation. Infosys Infrastructure Fabric for AI provides the platform capabilities required to support every stage of the AI lifecycle.

01

AI Factory as a Service

AI-ready infrastructure including compute, GPUs, storage, networking, and platform services

02

Model Building Services

Training data preparation, pre-training, fine-tuning, and model optimization

03

Model Alignment Services

Model evaluation, governance, safety, and performance alignment

04

Model Serving Services

Governed endpoints, lifecycle management, and operational support

Designed for Every AI Posture

Infrastructure Fabric for AI provides a common foundation that adapts to enterprise needs for every AI posture.



Sovereign AI



Regulated AI



Private AI



Hybrid-Governed AI

Built-in governance & FinOps

Responsible AI, auditability, lineage tracking, chargeback, cost visibility, and usage controls are embedded throughout the platform to help organizations scale AI responsibly and efficiently.

Why Infosys?

Partner intelligence: Strategic ecosystem relationships help organizations navigate evolving AI infrastructure, silicon, and platform landscapes

Enterprise-scale lifecycle expertise: Deep experience managing complex infrastructure environments and technology transformations at global scale

Proven platform & fabric: A platform-first approach built on proven architectures and operational experience

Governance & FinOps by design: Integrated governance, metering, chargeback, and cost optimization capabilities that help organizations scale AI with confidence

External Document © 2026 Infosys Limited

Deliver better AI economics with Infosys

- Increase GPU utilization
- Reduce idle capacity and waste
- Optimize infrastructure investments
- Improve AI cost efficiency
- Enable predictable unit economics

Proof Point

A governed, scalable AI infrastructure foundation designed to help enterprises accelerate AI adoption while maintaining control, transparency, and economic efficiency.

A production reference architecture for AI research, training and inference at scale — an NVIDIA DGX SuperPOD and RTX estate on an open-source stack.

202	4	3	2 PB	4
GPUs across the estate	NVIDIA generations	workload-turned tiers	all-NVMe storage	network fabrics

Workload-tuned estate

DEVELOPMENT
DGX A100 · 48 GPUs
24 × 80 GB + 24 × 40 GB
- Notebooks, dev and small-model hosting · OSS KubeFlow · S3-compatible object storage

TRAINING
DGX B200 · 48 GPUs
Continual pre-training & fine-tuning
- Slurm batch · NVIDIA Nemotron + OSS · deterministic gang scheduling

INFERENCE
H100 48 + RTX 5B · 106 GPUs
Large and small model serving
- NVIDIA Dynamo + TensorRT-LLM · Istio · Ray & Dask
- Roadmap: KAI, HPA, KEDA

Shared platform foundation

STORAGE · TIERED
- 2 PB Pure Storage all-NVMe
- Shared checkpoints, weights and training data with an archival tier behind it

NETWORK · 4 FABRICS
- InfiniBand plus a 200 Gbps storage fabric
- Three logically separate clusters with an isolated end-user plane

GOVERNANCE & OPS
- One LiteLLM gateway governs model access across every cluster
- BCM, Artifactory image provenance, Prometheus, Grafana and OTEL

A unified, governed platform delivering GPU-aaS and model lifecycle services, proven in production before client deployment.

For more information, contact askus@infosys.com

Infosys
Navigate your next

© 2026 Infosys Limited, Bengaluru, India. All Rights Reserved. Infosys believes the information in this document is accurate as of its publication date; such information is subject to change without notice. Infosys acknowledges the proprietary rights of other companies to the trademarks, product names and such other intellectual property rights mentioned in this document. Except as expressly permitted, neither this documentation nor any part of it may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, printing, photocopying, recording or otherwise, without the prior permission of Infosys Limited and/ or any named intellectual property rights holders under this document.