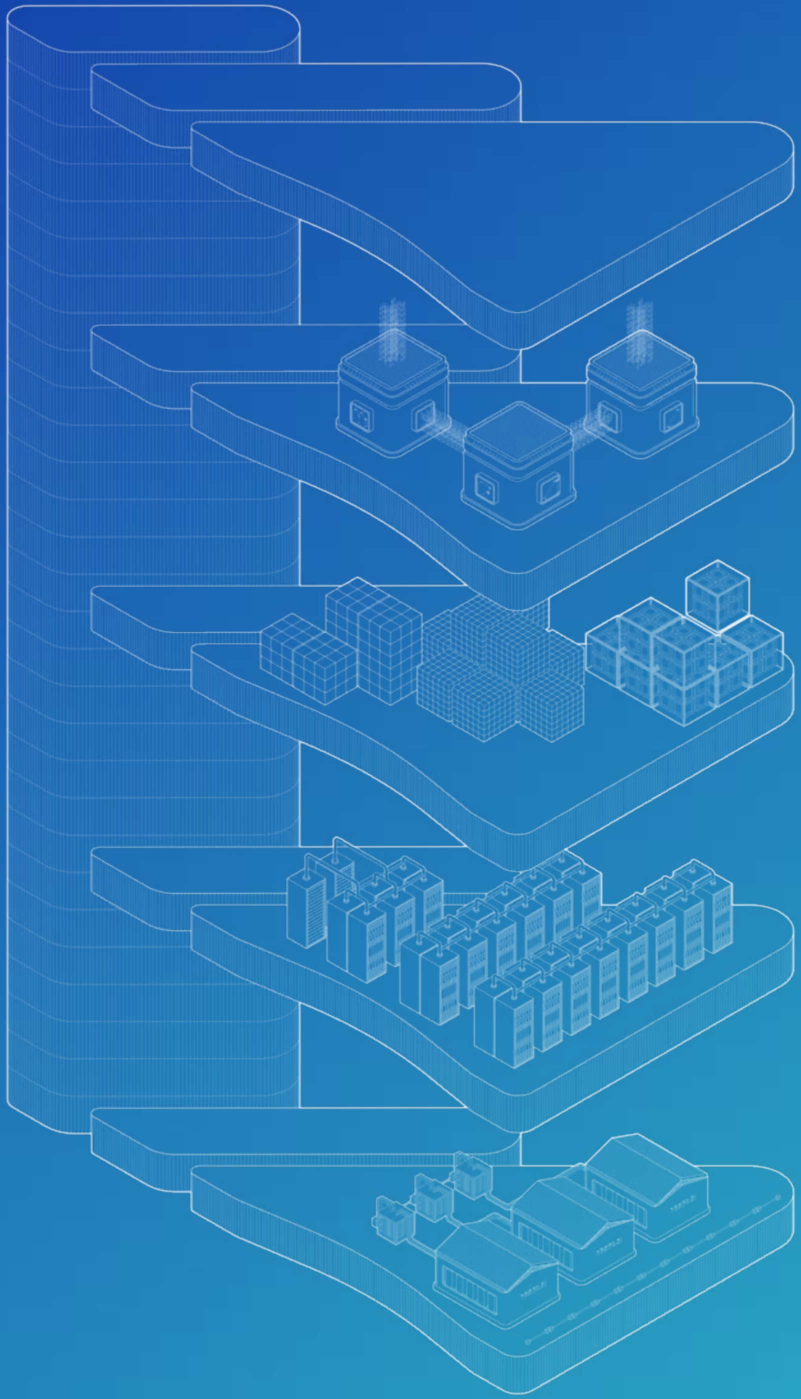
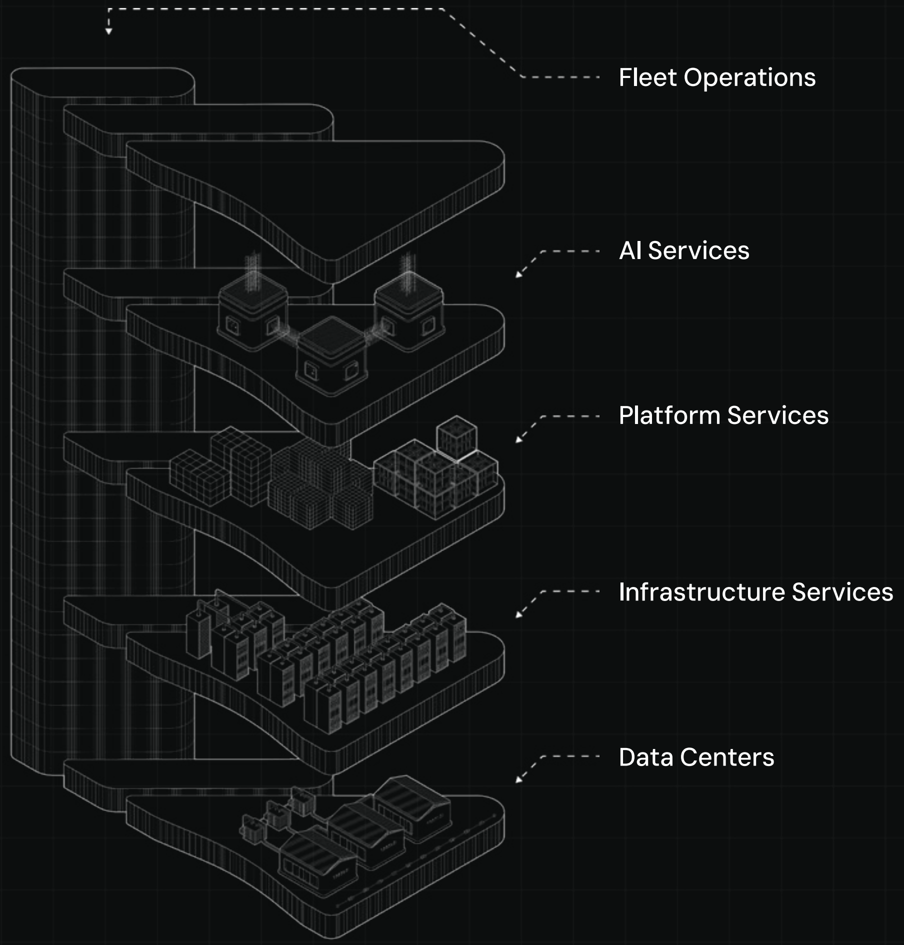
NSCALE

The production engine for scalable AI

How Nscale delivers the full AI lifecycle on one unified platform.



Nscale is building the engine of superintelligence. Today, that means delivering production-ready AI infrastructure that removes complexity, improves efficiency, and enables predictable deployment at scale.



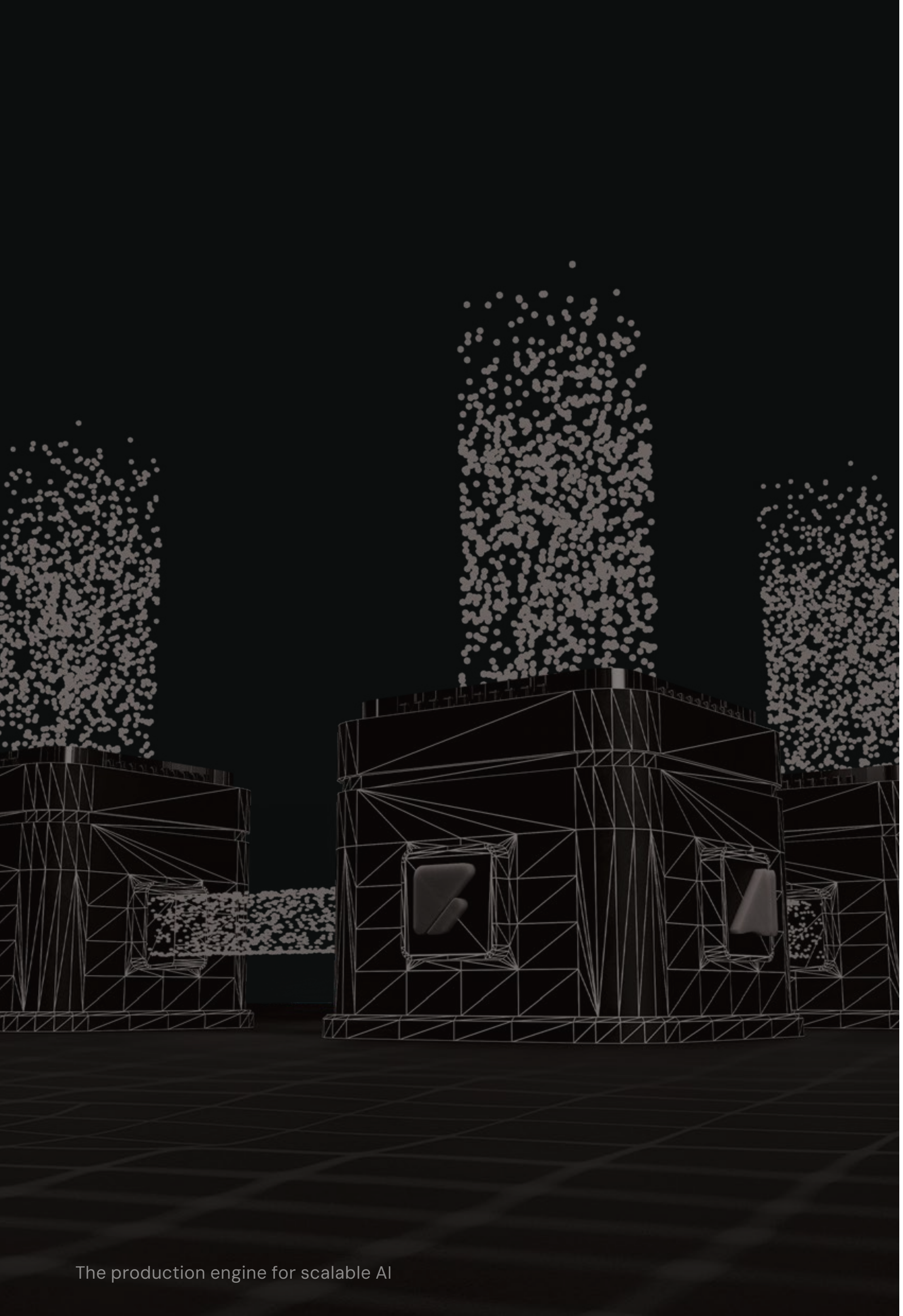
A full-stack AI cloud

Nscale is a vertically integrated AI cloud, purpose-built to run the full lifecycle of advanced AI workloads.

Unlike traditional cloud providers or software-only platforms, Nscale owns and operates the entire stack, from sustainably powered data centers to GPU infrastructure, orchestration platforms, and AI services.

This full-stack approach enables Nscale to co-design infrastructure, software, and operations as one system. The result is predictable performance, lower cost, and faster delivery of production AI, especially for large-scale, sovereign, and mission-critical workloads.

Nscale is trusted by telcos, enterprises, hyperscalers, and partners to run AI reliably, globally, and at scale.



Production AI demands a full system

Most AI cloud solutions are built from fragmented layers: separate vendors for compute, orchestration, tooling, and operations.

This fragmentation creates friction at every stage of the AI lifecycle, from teams facing long setup times, to inefficient GPU utilisation, and unpredictable costs. Reliability and compliance become operational burdens rather than built-in properties.

Nscale takes a different approach. By engineering the full stack as a single system, Nscale removes integration gaps, aligns incentives across layers, and delivers AI infrastructure that scales cleanly from first experiment to long-running production workloads.

Our full stack design also delivers fundamentally improved token economics. By maximising GPU utilization and reducing toolchain sprawl, Nscale enables customers to achieve more output per token over time. This makes production AI not just scalable, but economically efficient by design.

Nscale offers a wide range of models for serverless inference, ready to deploy for chat, vision, embeddings, and image generation.

Here are just a few:

GPT OSS 120B	OpenAI
Llama 4 Scout	META
Mistral 8x22B Instruct V0.1	Mistral
Qwen 4B Instruct	Alibaba
Stable Diffusion XL Base 1.0	Stability AI
DeepSeek R1 Qwen Distill 32B	DeepSeek

Build, tune, and deploy models without running infrastructure

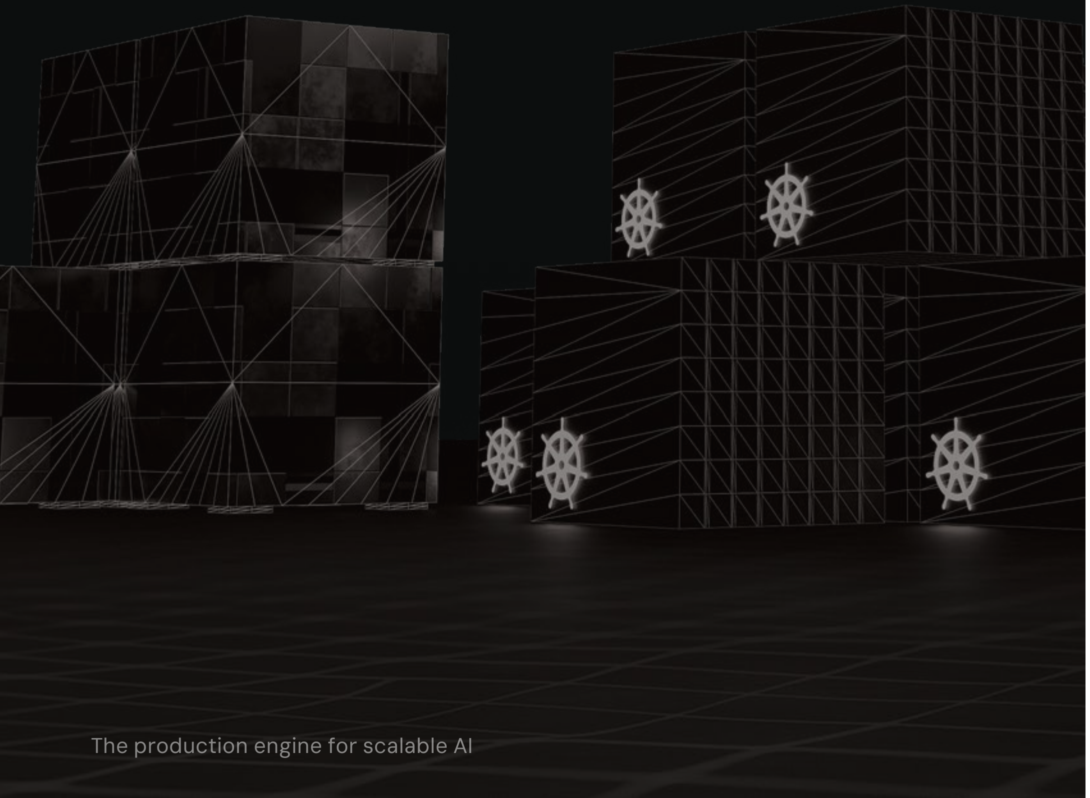
Nscale AI services provide the fastest path from experimentation to production-grade AI.

Inference Endpoints allow teams to deploy and scale production inference without managing clusters or GPUs. Endpoints autoscale for both real-time streaming and high-throughput batch workloads, with strict data boundaries.

Fine-Tuning enables low-friction model adaptation using customer data. A streamlined workflow handles dataset validation, job orchestration, and model output, allowing teams to customise foundation models and move tuned variants directly into production.

Prompt Workbench brings structure and reproducibility to prompt engineering. Teams can version prompts, compare models and parameters, and export validated configurations directly into inference workflows.

Together, our AI services reduce experimentation cost, accelerate iteration, and turn validated work into reliable production deployments.



Production-grade orchestration for AI workloads

Platform services provide the control layer that turns raw GPU capacity into predictable, production-ready systems.

Slurm Training delivers HPC-grade batch scheduling for large-scale GPU training on Kubernetes. Queueing, prioritisation, and GPU-aware scheduling ensure predictable throughput for distributed jobs.

Nscale Kubernetes Service offers managed, AI-optimized Kubernetes with GPU-aware scheduling, autoscaling, and enterprise security. Virtual Kubernetes provides instant, isolated environments for development and experimentation.

Instances give teams flexible access to lifecycle-managed bare metal or GPU VMs, with prebuilt AI images and optional VPC isolation.

These services can also help reduce operational risk while supporting mixed workloads across research, development, and production.

High-performance infrastructure built for AI at scale

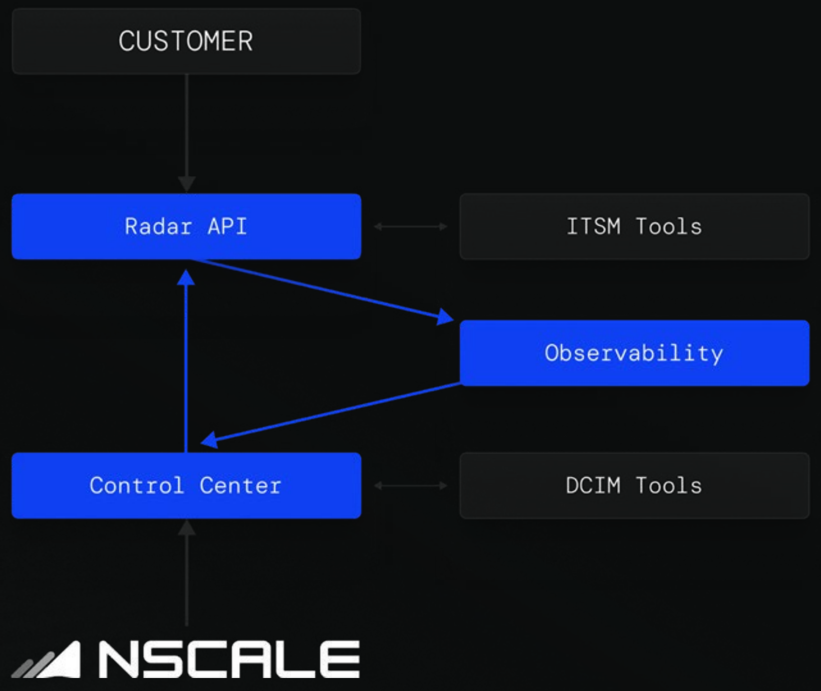
Infrastructure services form the foundation of the Nscale platform, engineered for maximum efficiency and predictable performance.

Compute delivers high-performance NVIDIA GPUs as bare-metal resources, enabling teams to extract peak performance with full hardware control.

Storage provides GPU-optimized, distributed file systems designed to prevent I/O bottlenecks during training and inference. Predictable throughput ensures GPUs stay productive.

Networking uses low-latency, high-bandwidth fabrics including RDMA, InfiniBand, and NVLink to support synchronous, multi-node training across racks and pods.

By tightly integrating compute, storage, and networking, Nscale maximises GPU utilisation and lowers cost per run.



Operate AI infrastructure like a system

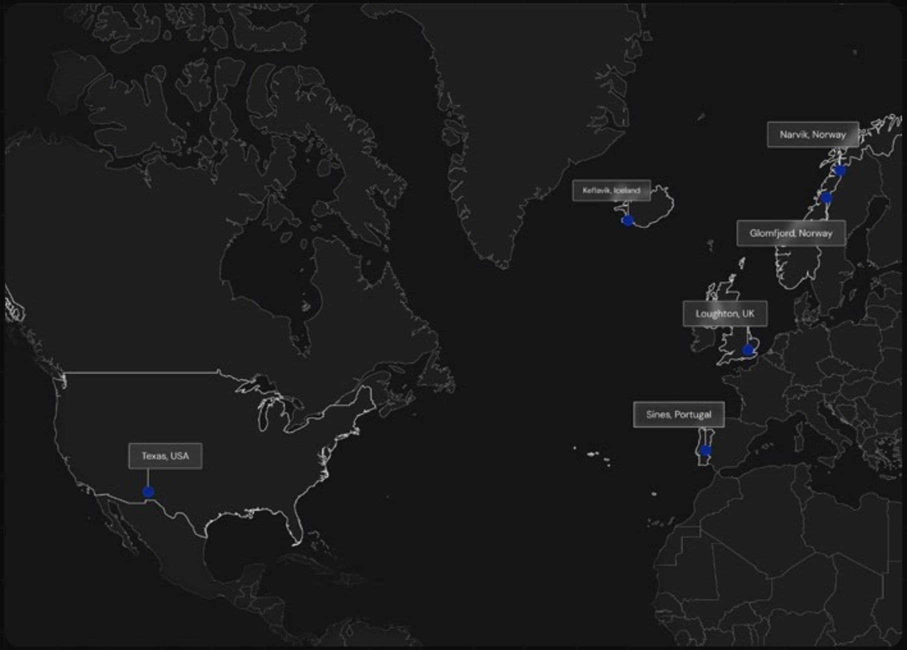
Fleet Operations unify lifecycle automation, observability, and capacity control across Nscale-managed infrastructure.

Control Center automates provisioning, scaling, patching, and retirement of nodes and clusters, reducing operational overhead and improving GPU utilisation.

Observability delivers platform-grade telemetry across infrastructure and workloads, providing real-time insight into performance, health, and cost.

Radar API exposes real-time GPU availability, repair status, and maintenance signals, enabling data-driven capacity planning and integration with enterprise systems.

Fleet Operations turn complex GPU fleets into reliable, high-yield infrastructure.



Engineered for regulated and mission-critical AI

Sovereignty and reliability are built into the Nscale platform by design.

Nscale operates AI data centers in low-cost and renewable power markets, delivering structural efficiency advantages while supporting regional data residency and compliance requirements.

Lifecycle management, observability, and contractual controls ensure predictable SLAs for enterprise and sovereign workloads.

The result is AI infrastructure that organisations can trust to run critical systems today and at scale.

Bringing AI closer to where enterprises operate

The Telco AI distribution network extends Nscale's full-stack AI platform into metro, edge, and in-country environments operated by trusted telecommunications partners.

This model enables enterprises to run AI workloads closer to users, data, and regulated environments, without sacrificing performance, control, or operational consistency. For enterprises, the network combines low-latency access, sovereign deployment, and a unified control plane across distributed sites.

Finance & Regulated Enterprise

- Sovereign, real-time AI for sensitive financial workloads
- Stop fraud in real time with sub-30ms edge inference and data residency preserved

Healthcare & Life Sciences

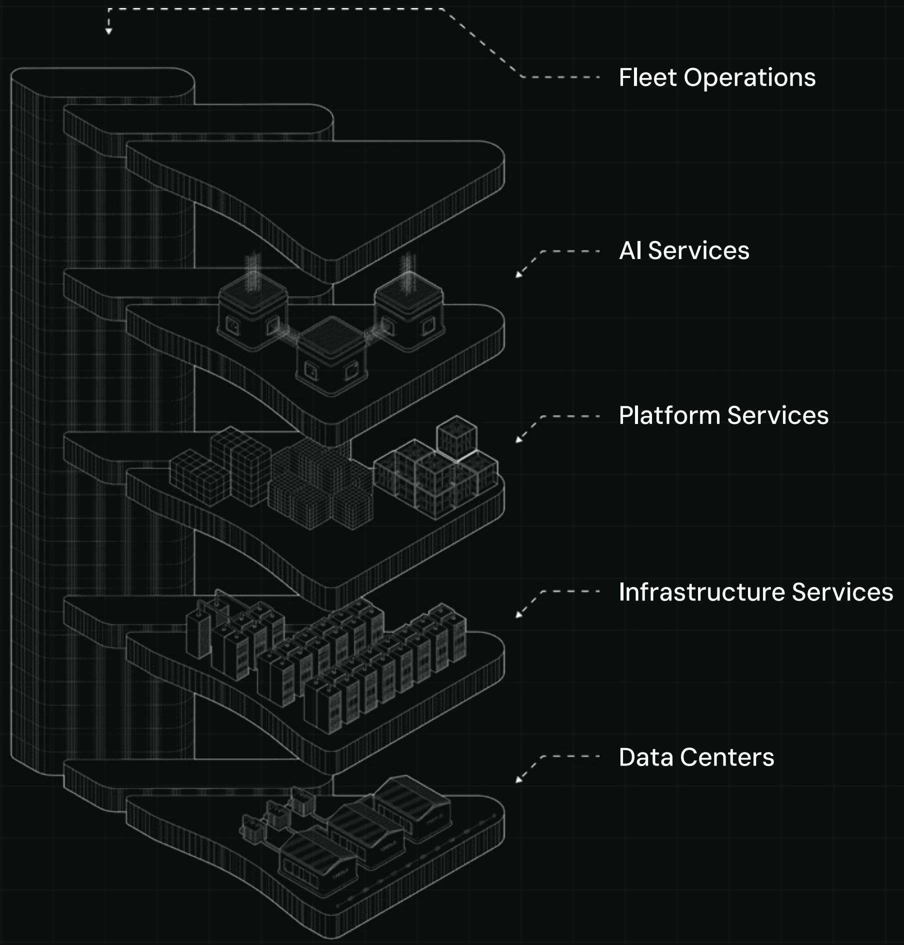
- Low-latency AI for clinical decision-making and operations
- Always-on patient monitoring and decision support at the edge

Public Safety & Government

- Trusted AI for mission-critical national services
- Secure AI for crowd management, border control, and critical infrastructure

Retail

- Real-time, in-network AI for physical and digital commerce
- Sovereign AI infrastructure aligned with regional data and consumer privacy regulations



Infrastructure for advanced intelligence at scale

Nscale delivers faster iteration, lower total cost of ownership, and predictable scaling across the full AI lifecycle.

By unifying infrastructure, platforms, operations, and partners, Nscale removes complexity and enables organisations to deploy AI with confidence.

The Nscale way:

- Designed to deliver scale
- Architected for efficiency
- Engineered for resilience and compliance
- Optimized for rapid execution
- Proven through partnerships

NSCALE

Nscale is building the engine of superintelligence by delivering reliable, scalable AI infrastructure today.

www.nscale.com