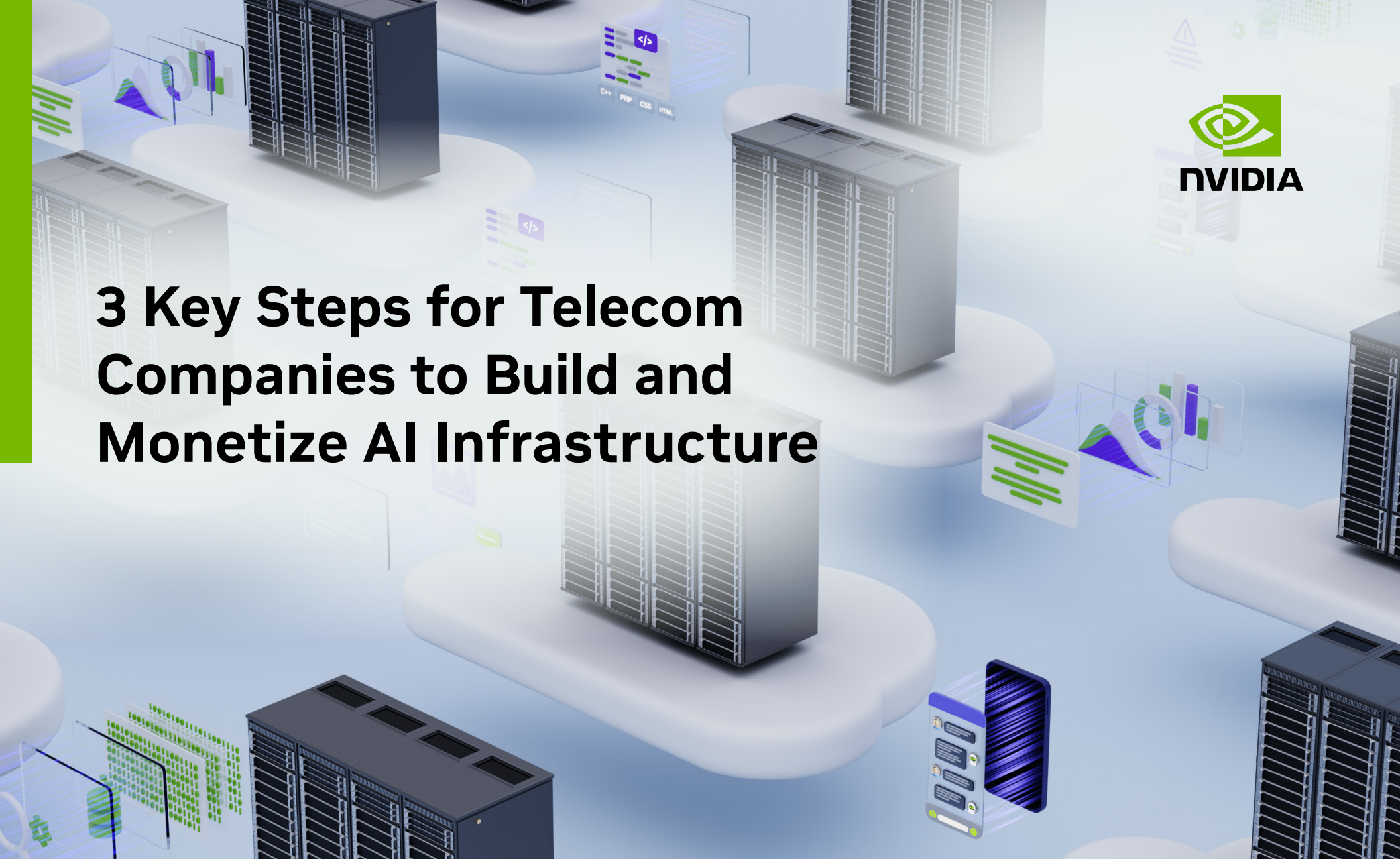


3 Key Steps for Telecom Companies to Build and Monetize AI Infrastructure

Why AI Factories Are a Natural Fit for Telecom Companies

By operating AI factories, telecom companies enable training and inference of AI models in-region, aligned with local language, culture, and customs.

AI factories are specialized computing infrastructure that turn energy and data into intelligence, powering generative, agentic, and physical AI applications across nearly every industry. Countries around the world are racing to decide where their national intelligence will be produced and who will capture the value created by AI. For many governments and regulated enterprises, the strategic priority is not just to use AI models and applications that run on AI factories, but to have strategic autonomy over how AI is developed and deployed—using in-country AI infrastructure; fine-tuning models for local languages, cultures, and regulations; and building AI assets to drive long-term economic growth, narrow the digital divide, and empower local communities.

Telecom companies (telcos) are ideally positioned to deliver this vision and become national providers of AI infrastructure. As well-established technology partners, telcos already have data center space, power, and connectivity, operate complex critical infrastructure platforms in compliance with stringent regulations and service-level agreements, and maintain deep go-to-market motions with governments, regulated enterprises, and small to medium businesses. To further position themselves at the center of the emerging AI token economy, leading telcos across five continents are reorganizing around AI—through dedicated subsidiaries, new business units, and strategic partnerships—so they can scale AI capabilities with less overhead, faster decision-making, and the right expertise. These efforts span building AI factories, cultivating the local developers' ecosystem, and offering token-metered AI services for enterprise and government customers.

In this ebook, we outline three key steps for telcos to become successful AI infrastructure and service providers: building AI factories, developing tailored local-language and industry-specific models and applications, and driving sustained infrastructure consumption by packaging these capabilities into commercially viable, token-metered services for regulated and affinity customer segments.

By following this path, leading telcos around the world are turning AI factories into a growth engine—acting as regional ecosystem hubs that bring together governments, system integrators, ISVs, and startups on top of secure, full-stack AI infrastructure.

Step 1: Build AI Infrastructure

To become AI providers, telecom companies first need to align with the NVIDIA NCP reference architecture to maximize their accelerated computing capabilities.

AI can be understood as a "5-layer cake": energy, chips, infrastructure, models, and applications—every AI application at the top relies on every layer beneath it, all the way down to the power that keeps it running.

The first step for telcos is to build their AI factory—the in-country, full-stack AI infrastructure platform that turns energy, chips, and infrastructure into tokens, or units of intelligence.

NVIDIA is accelerating the deployment of AI factories through the NVIDIA Partner Network (NPN) Cloud Partner (NCP) program. The program leverages the NCP reference architecture, a full-stack, validated design blueprint for building high-performance, scalable, and secure AI factories, supported by NVIDIA Enterprise Support, deployed by NVIDIA Infrastructure Services (NVIS), and powered by the NVIDIA AI Enterprise software suite. Production-ready software within this suite adheres to controls of various standards from around the world.

Together, these offerings give telcos a proven path to move quickly from racked systems to production-grade AI services, de-risking deployment and reducing time-to-first-train and time-to-first-token. For example, Deutsche Telekom brought its Industrial AI Cloud powered by nearly 10,000 NVIDIA Blackwell GPUs into production in just four months using this approach.

Step 2: Adapt AI to Local Language, Culture, and Policy

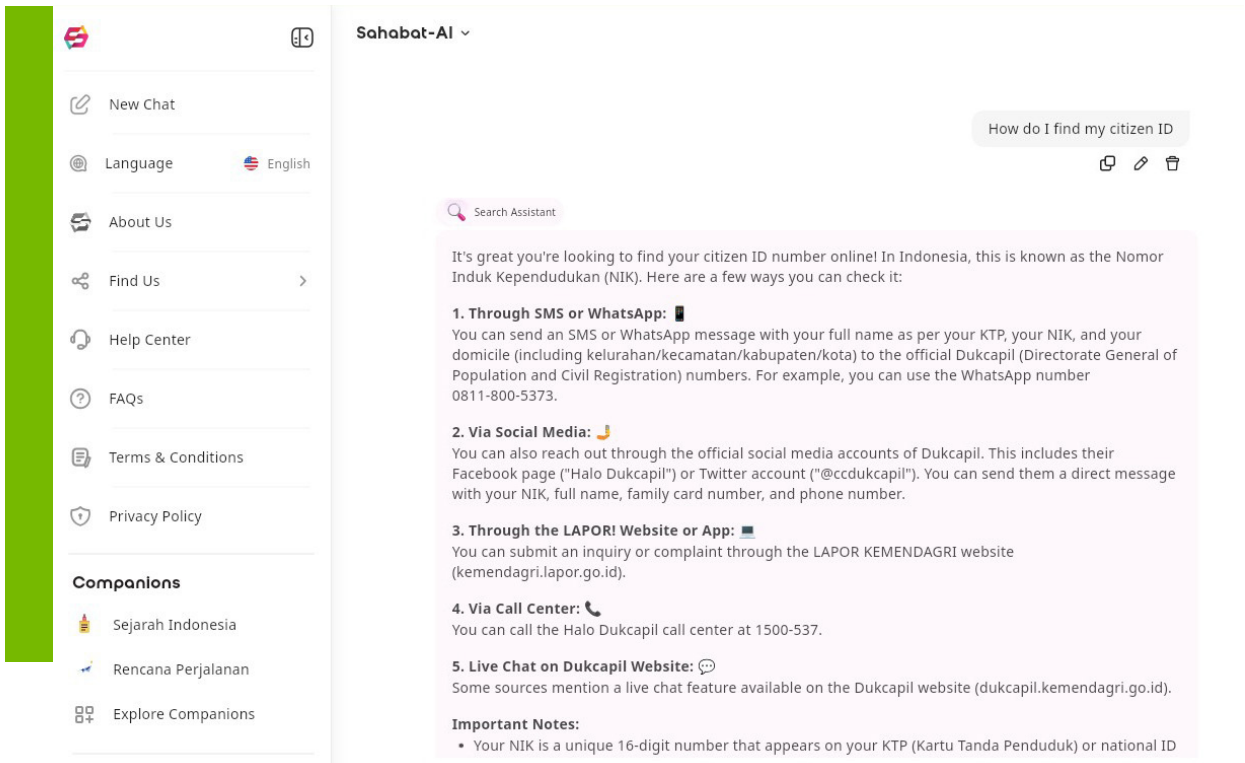
Once the AI factory is built, telcos have the energy, chips, and infrastructure foundations to develop and deploy AI models and applications that meet local needs. A key pillar is foundation models that are fine-tuned to align with local language, culture, regulations, and domain-specific data. These models significantly improve accuracy and relevancy for local industries and citizen services, becoming shared AI assets that governments, enterprises, startups, and universities can all adapt.

Open source models are a key catalyst, helping democratize access to intelligence by giving developers a transparent, adaptable starting point. When countries run and adapt these models on telco AI infrastructure, they can grow local expertise and applications without giving up control over how their data and models are used.

NVIDIA AI Foundry and NVIDIA AI Enterprise provide the tools and runtimes needed to build, adapt, and operate these models and agents end-to-end. Telcos can use NVIDIA NeMo™ for data preparation, fine-tuning, evaluation, and guardrailing, then deploy models as secure NIM™ microservices on their AI factories with built-in monitoring, scaling, and lifecycle management. This lets operators and their ecosystems turn local data into reliable AI services more quickly and with less risk, while keeping model customization, inference, and telemetry data aligned to local regulations.

Indosat Ooredoo Hutchison (IOH) worked with the Indonesia Ministry of Communications and Digital Affairs as well as other technology partners to launch Sahabat-AI, a collection of Indonesian-language LLMs and AI services built with NVIDIA NeMo and NVIDIA NIM microservices.

Built by Indonesians, for Indonesians, Sahabat-AI models understand local language, culture, and context, enabling organizations and developers to build AI products and services for more than 277 million people who speak Bahasa Indonesia and other local languages. Running on the GPU Merdeka AI factory operated by IOH's subsidiary Lintasarta and powered by NVIDIA AI Enterprise, these models form the foundation of an open national LLM ecosystem focused on preserving local languages and dialects, powering citizen-service platforms, and enabling AI applications across sectors such as agriculture, financial services, healthcare, and public services—keeping sensitive data, models, and value creation within Indonesia



Indosat Ooredoo Hutchison launched Sahabat-AI, a collection of LLMs, to provide generative AI services for Indonesia residents in Bahasa and multiple local languages.

Step 3: Package and Monetize AI Services

After building an AI factory and deploying localized models, telcos should focus on enabling AI adoption at scale. The objective of this step is to turn AI infrastructure into usable services for local governments, enterprises, startups, and universities, so more workloads run on the factory and more value is created in-country.

To do this, operators package AI into clear service offerings that move from raw compute capacity to outcome-aligned solutions. At the base, Compute-as-a-Service exposes secure, in-country accelerated computing that customers can rent by the hour. Above that, Token-as-a-Service delivers managed models, tools, and applications as APIs, so enterprises can adopt AI without operating infrastructure themselves. Increasingly, these AI services are packaged and priced around tokens, workflows, and other measurable outcomes rather than GPU-hours, so telcos monetize AI in the same units that customers use to track value.

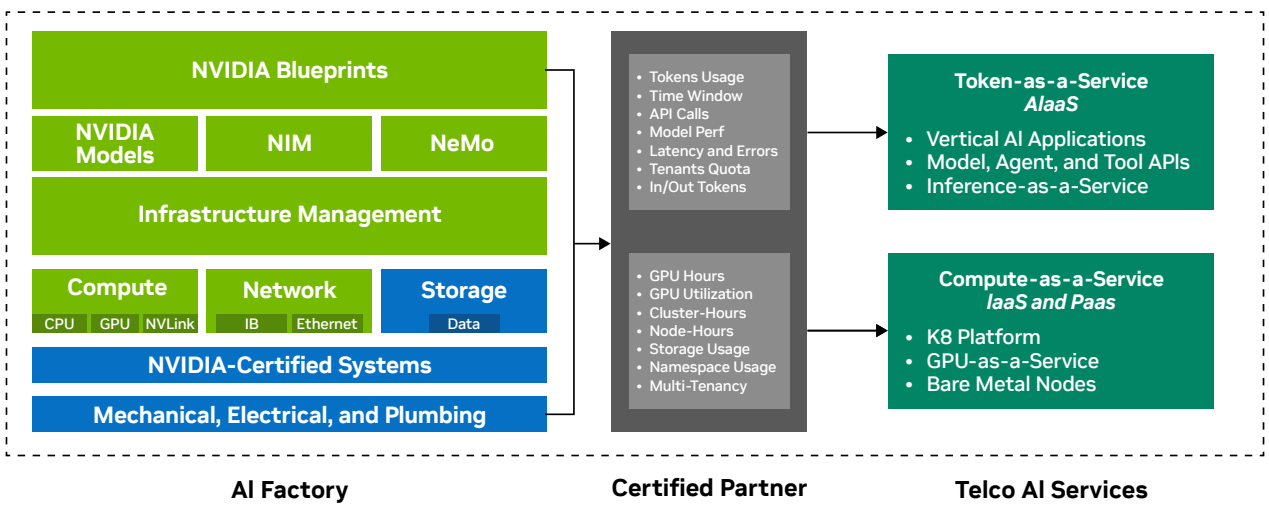
Additionally, many telcos work with partners to develop AI developer studios and marketplaces to simplify service creation and consumption at the Token-as-a-Service layer. An AI studio gives technical teams a workspace to turn models into services: Data scientists and developers select foundation models, adapt them with enterprise data, wire them into RAG or agentic workflows that can call tools and systems on behalf of users, and expose them as secure inference endpoints. Those endpoints, agents, and blueprints become reusable building blocks that feed Token-as-a-Service offers. AI marketplaces make those building blocks easy to discover and buy.

Business and application owners browse a catalog of models, agents, and vertical-specific applications, choose what they need, and launch it with a few clicks. For example, Swisscom's Swiss AI Platform combines GPU-as-a-Service, managed model inference, and self-service AI tooling — enabling Swiss enterprises to build and deploy AI applications on infrastructure that is operated, governed, and legally anchored in Switzerland.

Behind the scenes, a usage and billing layer meters consumption in tokens, API calls, or workflow runs, enforces quotas and SLAs, and rolls that data up into AI-native KPIs such as tokens per second, time-to-first-token, latency percentiles, and tokens per GPU-hour. That metering layer is what lets telcos design price-per-million-token plans, align service tiers to cost-per-token, and continuously improve unit economics as each new NVIDIA platform generation drives higher throughput and lower cost per token.

Compute-as-a-Service: In many regions, demand for GPU capacity far exceeds supply. AI factory operators often start by offering GPU-as-a-Service, typically as bare-metal or virtualized GPU clusters fronted by a Kubernetes-based platform, where customers can access high-performance, securely isolated AI infrastructure and run their own software stack for training and inference. This is a practical first step: It addresses urgent capacity needs and has a lower learning curve for telecom operators. Over time, however, it leaves value on the table if operators do not also help customers design, build, and operate AI applications on top of that compute capacity.

Token-as-a-Service: Token-as-a-Service builds on Compute-as-a-Service with developer tools, managed services, and ready-to-use AI applications metered by token consumption and backed by service-level agreements. It includes model-customization offerings that enable enterprises to adapt open source or proprietary foundation models with their own data, plus managed inference: Instead of provisioning raw GPU instances, providers expose models, agents, and vertical applications through inference API endpoints. This lets enterprises scale usage without operating the underlying infrastructure, while giving telcos an outcome-aligned monetization model that matches how customers actually consume AI. In this model, revenue scales with token throughput and price per token, while margin improves with each new NVIDIA platform generation and each software optimization, not just with higher hourly rental rates.



Telco AI architecture showing an NVIDIA-powered AI factory, cross-stack metering and billing, and telco AI services delivered as token-metered offerings.

Where Telecom Companies Win

AI factories are already changing how telcos around the world create and sell value. Operators are using them to modernize their own operations, build reusable AI assets adapted to local needs, and power AI services for government and regulated enterprise workloads.

For most telcos, the winning play is not to compete in foundation-model training, but to focus on where they have a structural advantage: localizing models for their markets, offering model-customization services, and operating secure, high-availability inference at scale on in-country AI infrastructure. Successful operators package these capabilities as Compute-as-a-Service and Token-as-a-Service offerings, often bundled with connectivity and professional services so customers can adopt AI through familiar commercial models and treat telcos as end-to-end technology partners.

The go-to-market motions that gain traction combine regulatory expertise, partnerships, and industry focus. Many telcos are aligning their AI factories with national AI programs, then partnering or co-selling with global system integrators and ISVs to bring complete solutions to market. They are concentrating first on regulated and affinity sectors such as financial services, public sector, healthcare, and manufacturing, where their existing networks, data flows, and established relationships create a natural right to play.

Some telecom operators start by using their AI factories for internal workloads such as customer-care copilots, autonomous network operations, and back-office automation. Running these use cases on their own platforms helps operators prove performance and ROI, harden day-to-day operations, and build reference stories before taking the same capabilities to market. In Europe, Orange Business, the B2B subsidiary of telecom company Orange, followed this approach by first testing and scaling its Live Intelligence gen AI platform internally with more than 100,000 active users across the company, then making it available externally as a trusted platform for businesses and public-sector organizations across Europe to adopt AI securely with data hosted in the region.

Also, some telecom operators build or co-create vertical-specific solutions on top of their AI factories, capturing more of the end-to-end solution value.

Finally, citizen and government services are becoming a natural focus area. Many public-sector organizations face constraints in AI skills and investment, yet urgently need trusted, in-region platforms for workloads such as citizen-service copilots, license and permit processing, and case-management automation. Telcos can help close this gap by providing turnkey compliant AI infrastructure plus domain-specific solutions.

TELUS has launched Canada's first AI factory, ranking as Canada's fastest and most powerful supercomputer on the TOP500 list. TELUS is positioning its AI factory as a platform for governments, enterprises, and startups to train, fine-tune, and run AI models on NVIDIA full-stack AI infrastructure while keeping data, compute, and governance entirely on Canadian soil. On top of this infrastructure, TELUS runs its Fuel iX enterprise AI platform, providing governed access to AI tools and models—with orchestration, observability, and safety controls—so developers can easily build and scale enterprise-grade AI applications.

TELUS's path to commercialization began with its own operations: More than 70,000 employees leverage Fuel iX, alongside 53,000 custom AI copilots, demonstrating that AI can reliably improve productivity and decision-making inside a complex, regulated business.

Building on that foundation, TELUS is onboarding tenants operating in regulated sectors. For example, healthcare organizations are running AI-powered clinical and patient experience solutions for Canadians, improving patient outcomes while keeping sensitive health data within Canadian borders. In addition, Canadian software companies are building next-gen AI-powered developer tools, and Canadian startups and early-stage companies are training and deploying AI models on the same class of infrastructure used by the world's largest technology companies.

Image courtesy of TELUS

With the initial AI factory in Rimouski, Quebec fully sold out due to unprecedented demand, TELUS is planning a significant expansion to build three additional AI factories in Canada.

IOH is developing applications for industry-specific use cases based on its new AI cloud, Sahabat-AI, and the NVIDIA AI Enterprise software platform. The collaboration taps into the Accenture AI Refinery platform to help Indonesian enterprises build AI solutions tailored across key industries, while delivering sovereign data governance. Initially focused on financial services, IOH's work with Accenture and NVIDIA is delivering pre-built enterprise solutions that can help Indonesian banks more quickly harness AI.

Healthcare AI company Hippocratic AI is using the Sahabat-AI models, the NVIDIA AI platform, and IOH's sovereign AI cloud to develop digital agents that can have humanlike conversations, exhibit empathic qualities, and build rapport and trust with patients across Indonesia. IOH's sovereign AI cloud lets Hippocratic AI keep patient data local and secure and enables extremely low-latency LLM inference.

GoTo's generative AI-powered digital assistant, Dira, taps into Sahabat-AI to deliver customized and culturally relevant interactions for its customers to book rides, order food delivery, transfer money, pay bills, and more.

Image courtesy of IOH.

Deutsche Telekom (DT) has launched its Industrial AI Cloud in Germany as one of Europe's largest AI factories, providing an end-to-end stack powered by NVIDIA so manufacturers, automakers, robotics companies, and other industrial and public sector players can train and run AI applications while meeting strict German and EU security and compliance requirements.

Built with an initial 10,000 GPUs, the Industrial AI Cloud combines DT's connectivity, data-center, and security capabilities with T-Cloud compute and SAP's Business Technology Platform, with T-Systems leading integration and adoption for industrial, healthcare, and public-sector customers.

From launch, DT's Industrial AI Cloud is anchored by tenants such as Siemens, which will use the platform for AI-powered digital-twin and simulation services for automakers, with manufacturers like Mercedes-Benz and BMW expected to run complex vehicle simulations, and robotics innovators like Agile Robots using the same sovereign infrastructure to train and operate robotic foundation models for deployment on real factory floors.

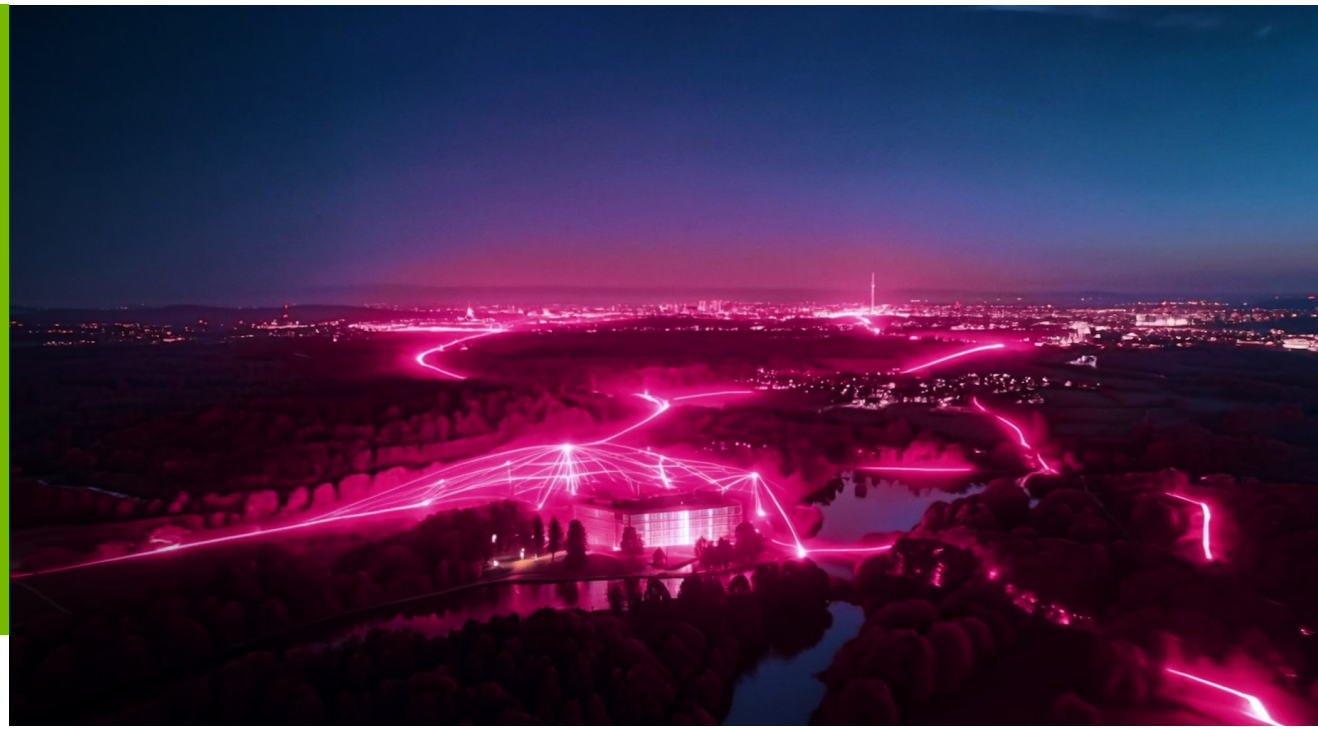


Image courtesy of Deutsche Telekom.

Image courtesy of SK Telecom.

SK Telecom (SKT) is turning its “Haein” AI cluster into a core pillar of South Korea’s AI strategy. Haein has been selected by the Ministry of Science and ICT to provide in-country sovereign compute for the national “Sovereign AI Foundation Model” project, providing GPU resources for the development of sovereign AI foundation models. These models are optimized for industrial AI, focused on agentic task execution, industrial reasoning, and workflow automation.

Building on this foundation, SKT plans to integrate A.X K1’s reasoning capabilities into A., its B2C AI service with 11.74 million cumulative subscribers. SKT has also signed an MOU with the Ministry of National Defense to develop defense-specific AI models based on these sovereign foundation models.

SKT exposes Haein to customers through its proprietary platforms—Petasus AI Cloud, AI Cloud Manager, and GPUaaS Service Orchestrator—as Compute-as-a-Service and AI-as-a-Service offerings. Developers and enterprises can rent slices of the cluster, spin up sandbox or production environments, and run full model lifecycles—data preparation, training, fine-tuning, and inference—without moving data or workloads outside Korea.

Singtel has launched RE:AI, an AI cloud service hosted in its Nxera AI-ready data centers and powered by NVIDIA GPUs, giving enterprises and public-sector customers in Singapore and Southeast Asia access to high-performance AI infrastructure while keeping data and governance under local control. RE:AI combines GPU-as-a-Service and AI-as-a-Service with AI workspaces and model services orchestrated by Singtel's Paragon platform, so enterprise customers can build and run AI applications across mobile and fixed networks without standing up their own stacks.

Above the infrastructure layer, Singtel uses Paragon to aggregate an ecosystem of AI applications, models, platforms, tools, and security services for sectors such as healthcare, manufacturing, fintech, and public services.

To accelerate AI innovation in Singapore, Singtel and NVIDIA have launched a Centre of Excellence for Applied AI. The Centre unites AI infrastructure, model developers, solution providers, and global expertise to address real-world challenges while advancing talent through hands-on experience. It enables enterprises and government agencies to run AI pilots faster and at lower cost, build capabilities, and scale solutions into production to support long-term AI goals.

This shows up in offerings like AI services built with Mistral AI for video analytics and language workloads, and a joint RE:AI–Cohesity service that lets enterprises and agencies search and reason over backup data while keeping it in sovereign environments. WEKA's NeuralMesh, a high-performance, software-defined storage system, is integrated with RE:AI to ensure GPUs are continuously supplied with data to optimise their productivity and performance. RE:AI also acts as a federation hub: Paragon connects to global and on-prem clouds so customers can keep regulated workloads on RE:AI while bursting non-sensitive jobs to hyperscalers or partner GPU clouds, adhering to regulatory requirements where applicable while still offering global scale and choice.

The Next Phase of Telco AI Infrastructure: AI Grid

Telcos are uniquely positioned to unlock value across all five layers of the AI stack—energy, chips, infrastructure, models, and applications. By combining their control of sites, power, connectivity, and data centers with NVIDIA's accelerated computing platforms and an ecosystem of model and application partners, they can turn AI factories into national platforms that activate the entire AI value chain, not just the infrastructure layer.

As they transform core networks and data centers into AI factories built on full-stack accelerated computing, the natural next step is to extend this intelligence across the distributed network edge—creating an "AI grid" that runs localized models and agents closer to where data is generated and decisions are made. This distributed AI posture lets telcos support latency-sensitive, AI-native applications in industries like manufacturing, healthcare, and mobility, while also applying the same capabilities to automate their own operations. By packaging this infrastructure as Compute-as-a-Service and Token-as-a-Service, operators can position themselves as high-value AI service providers powering the AI token economy.

Ready to Get Started?

Learn more about AI factories in telecom.

Contact us to get started.

NVIDIA