POWERING SOVEREIGN AI AT SCALE

Contents

Executive Summary 1

The Sovereign AI Requirement 1

Delivering Sovereign AI with Purpose-Built Infrastructure 2

Telecom's Role in the National Sovereign AI Ecosystem 3

- Sovereign AI as a Business Opportunity 3
- Telenor 4
- SK Telecom 4

AI Factory Infrastructure Solutions 5

- AI at Scale with Supermicro AI Factory SuperClusters 6
- NVIDIA GPUs Optimized for AI Workloads 7
- NVIDIA RTX PRO™ 6000 Blackwell Server Edition GPU 8
- The NVIDIA H200 NVL GPU 8
- NVIDIA HGX Platform (HGX B300 / HGX B200) 8

More information 9

Executive Summary

Artificial intelligence is rapidly becoming a foundational capability for telecommunications operators, shaping network operations, service delivery, customer experience, and national digital competitiveness. As AI becomes central to economic growth and critical infrastructure, telecom operators must adopt AI in a way that is secure, compliant, trustworthy, and sovereign.

Sovereign AI ensures that data, models, and AI operations remain under operator and national control — and within defined legal and geographic boundaries — enabling responsible, high-performance AI adoption across the telecom domain. For telecom providers managing highly sensitive subscriber data, critical national infrastructure, and regulated services, Sovereign AI is not optional; it is a strategic prerequisite. Sovereign AI also positions telecom operators as national AI platforms, enabling them to support government, healthcare, finance, and enterprise AI workloads with trusted, domestic infrastructure.

Purpose-built AI Factories, powered by Supermicro systems and NVIDIA accelerated computing, allow operators to deploy secure, scalable, GPU-dense AI environments that support GenAI, RAG, LLM training, and real-time inference — while enabling new revenue opportunities such as Sovereign AI-as-a-Service.

The Sovereign AI Requirement

Telecom operators face a unique combination of security, regulatory, operational, and national-level pressures that make Sovereign AI mandatory. Beyond subscriber data and network telemetry, operators support critical infrastructure, emergency services, and lawful intercept obligations — all of which require AI systems that are trustworthy, explainable, and fully controlled.

© 2026 Copyright Super Micro Computer, Inc. All rights reserved

Security

Telecom environments handle highly confidential information, including subscriber data, network configurations, and operational intelligence. Sovereign AI ensures that subscriber identity, location data, signaling, and operational intelligence remain within operator-controlled, jurisdiction-approved boundaries, strengthening national cyber resilience and reducing exposure to foreign platforms.

Compliance

Regulatory compliance is a defining characteristic of the telecom industry. Operators must comply with national and regional regulations governing data protection, critical infrastructure, and lawful interception, as well as industry standards and internal policies. Sovereign AI enables compliance by ensuring transparency, auditability, and control over where data is stored, how it is processed, and which models are used. AI operations deployed within sovereign environments align with national digital sovereignty frameworks, enabling operators to demonstrate compliance with data residency, critical infrastructure protection, and sector-specific regulations.

Trust

AI models used in telecom operations must be reliable. Hallucinations or incorrect outputs can have disastrous consequences when applied to network operations, customer support, or service assurance. Trust in AI and the outcome of all queries is therefore critical.

One of the most effective ways to increase trust is to control which data is used to inform AI outputs. RAG (Retrieval-Augmented Generation) becomes essential for telecom AI. By grounding AI outputs in OSS/BSS, runbooks, engineering documentation, and authoritative operator data, RAG ensures accuracy, reduces hallucinations, and maintains operator control over knowledge sources. When deployed within a sovereign AI environment, RAG improves output quality without compromising data control, enabling operators to benefit from contextual intelligence while maintaining ownership of their data.

Together, security, governance, and trust define the Sovereign AI requirement for telecom—and set the stage for how AI must be built and deployed.

Delivering Sovereign AI with Purpose-Built Infrastructure

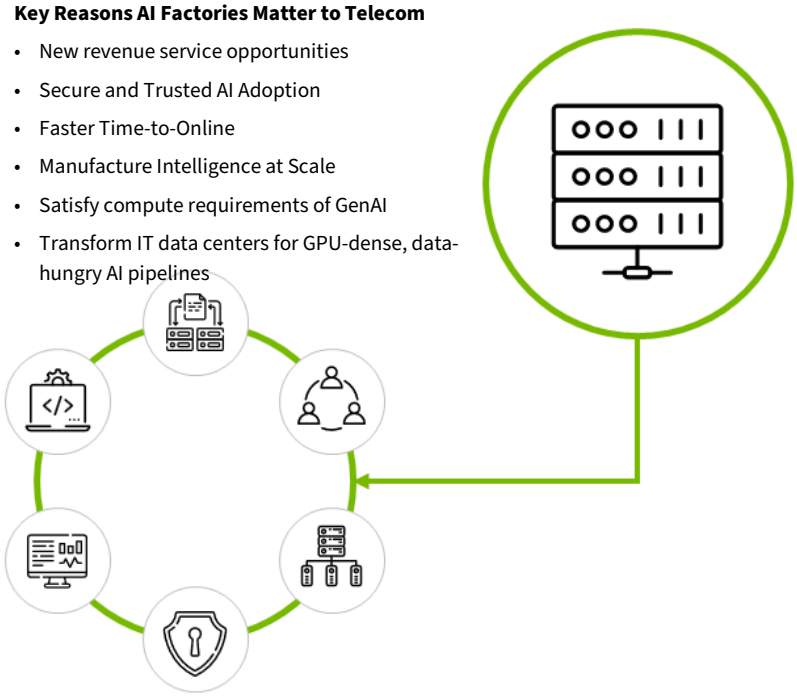
The rapid evolution of artificial intelligence is driving a strategic shift toward sovereign AI platforms that enable telecom operators to innovate while retaining full control over their data, models, and infrastructure. As AI becomes foundational to network operations and digital services, telcos must adopt AI architectures that not only accelerate innovation, but also address stringent requirements around security, regulatory compliance, and trust.

© 2026 Copyright Super Micro Computer, Inc. All rights reserved

AI Factories — sovereign, GPU-dense, modular AI data centers — are the execution engine of Sovereign AI. They serve as the sovereign AI control plane, the intelligence manufacturing center, and the foundation for AI-as-a-Service. AI Factories powered by Supermicro and NVIDIA enable operators to analyze network telemetry, logs, and performance data at scale, support GenAI and RAG workloads, and deliver secure AI services to enterprises and government agencies. In addition, AI factories can be used as the foundation for new service opportunities.

Key Reasons AI Factories Matter to Telecom

- New revenue service opportunities
- Secure and Trusted AI Adoption
- Faster Time-to-Online
- Manufacture Intelligence at Scale
- Satisfy compute requirements of GenAI
- Transform IT data centers for GPU-dense, data-hungry AI pipelines



Building these AI factories requires platforms that deliver true scalability, rapid time-to-market, and power-efficient performance tailored for telco demands. Purpose-built AI infrastructure must scale from small clusters to multi-rack systems with hundreds of GPUs, enabling operators to expand AI workloads without redesigning the core architecture. Integrated, turnkey solutions that combine compute, networking, storage, and software accelerate deployment and shorten time-to-online so telcos can launch services faster. Efficiency is equally critical: optimized hardware and cooling reduce power draw and maximize throughput in constrained environments, ensuring sustained AI performance with lower operational costs.

Telecom's Role in the National Sovereign AI Ecosystem

Telecom operators are not just AI users — they are national AI enablers. Because they operate national networks, edge infrastructure, data centers, and regulated operational environments, they are uniquely positioned to become trusted national AI platforms. Telecom-hosted Sovereign AI Factories can power government AI workloads, healthcare and finance AI, local language and cultural models, national AI research, and enterprise AI adoption across industries. This aligns telecom with broader national Sovereign AI priorities: economic development, digital competitiveness, cultural preservation, and strategic autonomy.

Sovereign AI as a Business Opportunity

Sovereign AI allows telecom operators to move up the value chain — from connectivity providers to trusted national AI platforms. Operators can offer GPU-as-a-Service (GPUaaS), Sovereign AI-as-a-Service, managed AI platforms, industry-specific copilots, secure inference services, local language model hosting, and RAG-based enterprise AI assistants. These services create recurring revenue streams while strengthening customer relationships built on trust, transparency, and regulatory alignment.

AI factories deployed in local data centers enable customers to train, fine-tune, and run AI models on trusted infrastructure that keeps data within national borders, meeting strict requirements for data residency, security, and compliance.

© 2026 Copyright Super Micro Computer, Inc. All rights reserved

Sovereign AI Factory Implementations

Sovereign AI is not a futuristic concept – it is being deployed and used today. Supermicro and NVIDIA collaborate with operators across the globe to design, build, and deploy regional and national AI clusters, used to accelerate AI adoption in businesses and governments. These deployments demonstrate that Sovereign AI for telecom is real, scalable, and economically viable.

Telenor

Telenor is a leading Nordic telecommunications provider with a long-standing reputation for operating trusted, secure, and resilient digital infrastructure across highly regulated markets. Building on this foundation, Telenor has launched Telenor AI Factory, Norway's first sovereign AI cloud platform, designed to accelerate AI adoption while ensuring full data sovereignty. The AI Factory provides enterprises, startups, and public sector organizations with access to high-performance, GPU-accelerated infrastructure hosted entirely within Norway, allowing customers to develop, train, and deploy AI models without their data ever leaving national borders. This positions Telenor not only as a telecom operator, but as a strategic AI platform provider.

The AI Factory delivers significant benefits through its combination of security, scalability, and sustainability. Built on open-source technologies and modern AI tooling, the platform offers flexible GPU clusters, predictable pricing, and seamless integration with existing customer environments. This lowers barriers to entry and shortens time-to-market for AI initiatives, even when working with sensitive or regulated data. Customers benefit from Telenor's operational expertise, regulatory compliance, and trusted infrastructure, while avoiding large upfront investments. By offering sovereign AI as a managed service, Telenor has created new revenue opportunities and a compelling blueprint for how telecom operators can expand their role in the AI value chain.

For more information, visit: https://www.telenoraifactory.no/

SK Telecom

SK Telecom, a leading telecommunications and AI company in Korea, has built the GPU cluster called Haein, a sovereign AI infrastructure platform designed to support large-scale AI workloads and drive national AI competitiveness. Named after Haeinsa Temple, the Haein cluster features Supermicro AI servers with over 1,000 NVIDIA Blackwell B200 GPUs, forming one of the country's highest-performance AI GPU clusters. Hosted at Gasan AI Data Center, the platform delivers GPU-as-a-Service (GPUaaS) for training, inference, and model development. Haein is currently being utilized for the government-led Sovereign AI Foundation Model Project, supporting AI-driven innovations in Korea.

Beyond raw compute, the Haein Cluster leverages SK Telecom's proprietary Petasus AI Cloud virtualization and AI Cloud Manager AIOps platform to rapidly provision and manage AI workloads, enhancing utilization and developer productivity. Strategic global partnerships with Penguin Solutions and Supermicro have accelerated deployment and integration, while collaboration with technology partners like Vast Data helps reduce provisioning times from weeks to minutes.

This sovereign AI infrastructure not only strengthens SKT's internal AI capabilities, but also enables enterprise and national stakeholders to build and deploy AI solutions securely within South Korea's borders—creating new service opportunities and reinforcing the company's strategic shift toward AI-driven offerings.

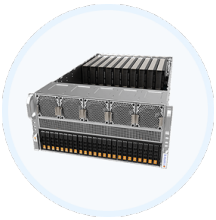
For more information, visit: https://news.sktelecom.com/en/2056

© 2026 Copyright Super Micro Computer, Inc. All rights reserved

AI Factory Infrastructure Solutions

AI factories from Supermicro and NVIDIA are complete, turnkey solutions simplifying the deployment of sovereign AI at scale for faster time-to-online and time-to-revenue, with full-stack solutions including compute, software, networking, and storage. Supermicro delivers AI infrastructure optimized for performance and efficiency, with fully-integrated solutions based on NVIDIA Enterprise Reference Architectures and NVIDIA-Certified Systems™ for guaranteed full-stack performance and compatibility. Supermicro's industry-leading rack-level testing, validation, and deployment services ensure quality and seamless plug-and-play deployment for complete AI confidence. These solutions support telco-grade requirements for uptime, compliance, multi-tenant isolation, and rapid time-to-online.

Supermicro



First-to-Market NVIDIA-Certified Systems

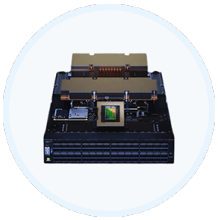
Rack-Scale Integration, Testing, and Validation before Shipping

Cluster-scale Deployment, Services, and Support

Storage and Networking Integration

NVIDIA

NVIDIA Accelerated Compute



NVIDIA Spectrum™-X Ethernet Networking Platform

NVIDIA Software Stack

NVIDIA AI Data Platform

© 2026 Copyright Super Micro Computer, Inc. All rights reserved

AI at Scale with Supermicro AI Factory SuperClusters

Supermicro's AI Factory SuperClusters are based on NVIDIA Enterprise Reference Architectures and provide enterprise customers with complete, rack-scale and cluster-scale solutions that ensure full-stack performance and compatibility, simplifying the deployment of complete AI factories. Supermicro's testing and validation process goes beyond industry standards, with complete testing of all nodes and cluster-level (L12) testing before shipment to ensure seamless plug-and-play deployment for customers of any size. Supermicro AI Factory solutions are endorsed by NVIDIA for Infrastructure Configuration, Spectrum-X networking, and Software Reference Stack and are based on the NVIDIA Enterprise Reference Architecture for RTX PRO™ 6000 Blackwell Server Edition and HGX B200. Supermicro AI Factory SuperClusters are ideal for telecom operators deploying multi-rack GPU clusters for GenAI, RAG, LLM training, and real-time inference.

NVIDIA Spectrum-X Networking
High-speed AI compute fabric tuned and validated across the full stack of NVIDIA hardware and software, creating an unmatched Ethernet solution for AI factories

NVIDIA GPUs
Choose from NVIDIA RTX PRO 6000 Blackwell Server Edition GPU and HGX B200 to handle a wide range of enterprise AI workloads

Supermicro NVIDIA-Certified Systems
Tested and validated for guaranteed compatibility and performance

Complete Rack-Level Integration and Validation
Built by Supermicro's expert teams, delivered ready to power on from day one

NVIDIA Software Stack
Run NVIDIA AI Enterprise, NVIDIA Omniverse, and NVIDIA Run:AI with guaranteed compatibility

Storage
Supermicro works with leading ISVs to create storage solutions that support the NVIDIA AI Data Platform and can be seamlessly connected to AI factories

Thermal Optimization
Architectures designed for maximum performance in air-cooled environments

© 2026 Copyright Super Micro Computer, Inc. All rights reserved

GPU	NVIDIA RTX PRO™ 6000 Blackwell Server Edition GPU	NVIDIA HGX B200
Maximum Cluster Size	Up to 32 nodes, 256 GPUs per scalable unit	Up to 32 nodes, 256 GPUs per scalable unit
Nodes per Rack (Typical)	4–8 per rack	4 per rack
GPU System Node SKU(s)	SYS-522GA-NRT SYS-422GL-NR AS -5126GS-TNRT2	SYS-A22GA-NBRT
Rack Power (4 Nodes)	33.3–36.6kW	53.6kW
Networking	NVIDIA Spectrum-X	NVIDIA Spectrum-X
NVIDIA Software Stack	NVIDIA AI Enterprise/NVIDIA Omniverse/NVIDIA Run:ai	NVIDIA AI Enterprise/NVIDIA Omniverse/NVIDIA Run:ai
NVIDIA Software Stack	AI inference / Retrieval Augmented Generation (RAG), HPC, and visual computing	Foundational AI model training, large-scale AI inference, and HPC workloads
	Learn More **Datasheet**	**Learn More** **Datasheet**

For more information on AI Factory SuperClusters, visit https://www.supermicro.com/en/accelerators/nvidia/ai-factory

NVIDIA GPUs Optimized for AI Workloads

The following NVIDIA GPUs are well-suited to accelerate AI workloads at scale. Each GPU—NVIDIA RTX PRO™ 6000 Blackwell Server Edition, NVIDIA H200 NVL, and NVIDIA HGX™ B200 and B300—offers unique combinations of memory, performance, and interconnect capabilities, enabling operators to match GPU resources to their AI workloads while maintaining low-latency, high-accuracy inference. Telecom operators can match GPU architectures to workload profiles:

© 2026 Copyright Super Micro Computer, Inc. All rights reserved

NVIDIA RTX PRO™ 6000 Blackwell Server Edition GPU

Built on the groundbreaking NVIDIA Blackwell architecture, the NVIDIA RTX PRO™ 6000 Blackwell Server Edition delivers a powerful combination of AI and visual computing capabilities to accelerate enterprise data center workloads. Suited for enterprises looking to run a range of enterprise workloads—in addition to generative AI—using small-model inferencing and fine-tuning (models less than 70B).

- Passive heatsink taps chassis airflow so **450–600 W** cards pack densely, boosting rack efficiency and lowering PUE.
- Most powerful PCIe Gen5 upgrade for L40/L40S/A40 slots—built for air-cooled racks, & up to **8-GPU scale-out servers**.
- Powers a range of enterprise workloads, including enterprise inference, model fine-tuning, HPC, virtual desktops, and container-scale distributed graphics.

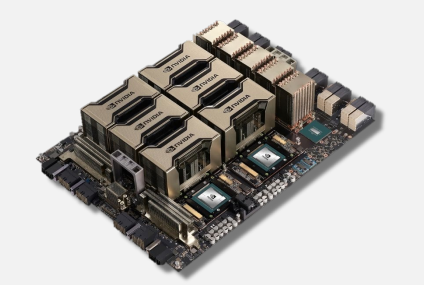
The NVIDIA H200 NVL GPU

The NVIDIA H200 Tensor Core GPU supercharges generative AI and high-performance computing (HPC) workloads with game-changing performance and memory capabilities. Ideal for customers training foundational models or using large models for inferencing (models greater than 70B).

- 141 GB HBM3e at 4.8 TB/s per GPU and 900 GB/s NVLink pair form a 282 GB logical device for huge models.
- Full-scale LLM training, high-batch GenAI inference, large-graph recommenders, and memory-bound HPC (CFD/weather).
- Maximizes per-node footprint to avoid sharding—cutting latency and energy per token while boosting throughput.

NVIDIA HGX Platform (HGX B300 / HGX B200)

As a premier accelerated scale-up platform with up to 30x more AI Factory output than the previous generation, NVIDIA Blackwell Ultra-based HGX systems are designed for the most demanding generative AI, data analytics, and HPC workloads. Designed for enterprises with intensive AI training and inference workloads, this is the most powerful NVIDIA GPU available for training and inference.



- Use when you need maximum per-node performance and 1.8 TB/s NVLink across 8 GPUs for large-model training and high-throughput inference.
- Full LLM training, AI reasoning, large-batch GenAI, and FP64/HPC that benefit from fast multi-GPU collectives.
- Facility ready for higher TDP (≈700–1,200 W/GPU, often liquid-cooled) to sustain clocks and performance density.

© 2026 Copyright Super Micro Computer, Inc. All rights reserved

More information

Telecom operators are at the center of national digital transformation. By deploying Sovereign AI Factories powered by Supermicro and NVIDIA, operators can secure their networks, accelerate innovation, and become trusted AI platforms for enterprises and government.

Build your Sovereign AI future. Start with the right foundation.

For more information about sovereign AI and AI factories for telecom, visit:

https://www.supermicro.com/telco-ai

SUPERMICRO

As a global leader in high-performance, high-efficiency server technology and innovation, we develop and provide end-to-end green computing solutions to the data center, cloud computing, enterprise IT, big data, HPC, and embedded markets. Our Building Block Solutions® approach allows us to provide a broad range of SKUs and enables us to build and deliver application-optimized solutions based upon your requirements. See www.supermicro.com.

NVIDIA

NVIDIA accelerated computing platforms power the new era of computing, performing exponentially more work in less time with much lower energy consumption than traditional CPU-based computing. Accelerated computing revolutionizes energy efficiency across industries by harnessing NVIDIA GPUs, CPUs, and networking, all optimized through NVIDIA enterprise software solutions.
More information at https://nvidianews.nvidia.com.

https://www.supermicro.com/en/accelerators/nvidia/ai-factory

© 2026 Copyright Super Micro Computer, Inc. All rights reserved