Sovereign AI, autonomous networks and AI-RAN will drive future investment in telecoms AI infrastructure

November 2026

Gorkem Yigit

Analysys Mason's recent report *Telecoms AI cloud infrastructure: worldwide forecast 2025–2031* shows that telecoms operators' spending on AI cloud infrastructure will grow from USD7 billion in 2025 to USD17.6 billion in 2031, at a CAGR of 17%.

Growth will be driven by different factors over the forecast period. In the near term, much of the momentum will come from operators building AI infrastructure to serve the B2B/B2C market, particularly where they see an opportunity around sovereign AI.[1] Investment in infrastructure for operators' own AI use cases will build more gradually as AI adoption expands across network and operational processes and operators progress towards more autonomous networks. AI-RAN will add another source of spending on AI infrastructure as AI becomes more deeply embedded in the RAN.

Operators are betting on sovereign AI to support their B2B/B2C infrastructure expansion

Operators will cumulatively invest more than USD50 billion in AI infrastructure between 2025 and 2031 to support B2B/B2C services such as GPU as a service (GPUaaS), Model as a service (MaaS) and local language model development.

The Asia–Pacific region will account for the largest share of this investment. Chinese operators will be the biggest spenders as they sharpen their focus on AI services, while SoftBank and KDDI in Japan, SK Telecom and KT in South Korea, and Singtel in Singapore all have ambitious plans to expand their AI compute capacity. Investment is also accelerating in Western Europe, led by Deutsche Telekom and Iliad/Scaleway, while the EU AI Gigafactories initiative could create further opportunities for operators.

Sovereignty is central to many of these investments. Most operators cannot compete with hyperscalers and neoclouds on infrastructure scale or price alone, so they are targeting local opportunities where data residency[2], regulatory compliance and infrastructure operated by a trusted local provider can set them apart, particularly in highly regulated verticals. However,

---

1 Sovereign AI refers to the development and use of AI under defined sovereignty requirements, including where data and infrastructure are located and who controls the data, infrastructure, technology and operation of AI systems.

2 Data residency refers to the requirements that specify the country or geographic jurisdiction in which data must be stored and, where applicable, processed.

demand for sovereign AI is still developing. It remains unclear how much enterprise demand will specifically require sovereign infrastructure and how much of a premium, if any, enterprises will be willing to pay for it. These factors will ultimately determine the size of the addressable opportunity for operators. Analysys Mason will explore enterprise demand for sovereign AI, including sovereign infrastructure supplier choices and willingness to pay, in forthcoming research.

GPUaaS is the most visible way in which operators are monetising this capacity, and we forecast that worldwide telecoms operator GPUaaS revenue will reach USD15.4 billion by 2032.[3] Even so, operators accounted for only around 3% of the wider GPUaaS market in 2026, and the long-term economics of their investments remain uncertain. Sovereignty could give operators an advantage in selected markets, but it will not be enough on its own. Those that combine GPUaaS with AI platforms and applications, professional services, vertical expertise and their existing B2B/B2C propositions should be better placed to differentiate.

The current capacity build-out will test the strength of this opportunity. Many of the pioneering operators are pursuing 3-to-5-year expansion plans, which will frontload investment through 2028/2029. After that, the focus is likely to shift from large capacity investments to increasing utilisation and monetising that capacity. Whether another investment cycle follows will depend on the demand and returns that emerge from this initial build-out.

Autonomous network ambitions and AI-RAN will drive steadier growth in internal AI infrastructure spending

Operators will steadily increase spending on AI cloud infrastructure for internal operations through 2031. Network operations are expected to account for a growing share as operators deploy AI across operational processes and progress towards more autonomous networks.

The transition will take time. Our *Telco AI Readiness Index* found that more than half of operators plan to implement AI across multiple operational areas in the network, but the current technological and organisational readiness of most operators does not support these ambitions. Gaps in tooling, ROI clarity, skillsets and organisational alignment, alongside data that is not yet model-ready, are limiting operators' ability to move from isolated deployments to AI at scale.

As a result, the infrastructure build-out will be gradual. Early deployments in areas such as assurance, predictive maintenance and RAN energy management typically address only part of the operational workflow, so their compute requirements remain modest. Infrastructure demand should increase as operators scale successful use cases across operational areas and introduce more advanced autonomous and agentic operations. This transition will also require operators to establish a trust and validation layer to govern how AI models and agents operate, particularly as they are given greater autonomy over network operations.

[3] See Analysys Mason's *GPU as a service (GPUaaS): worldwide forecast 2025–2032*.

AI-RAN will create an additional source of infrastructure demand. AI/ML-assisted functions in the baseband, such as link adaptation and beamforming optimisation, are already being deployed, but fully inference-based, AI-native functions will take a growing share over the forecast period. This will increase demand for GPUs, custom ASICs[4] and other accelerators as more AI processing moves into the RAN.

The pace of AI infrastructure investment will hinge on operators proving ROI

AI is reshaping operators' data centre infrastructure strategies. After a decade in which many operators consolidated facilities and migrated workloads to public cloud, AI is giving the remaining estate a new role. Existing capacity is increasingly being repurposed to support inference for network, OSS/BSS and data platform workloads, while operators with established data centre and cloud businesses are expanding their infrastructure to address external AI demand.

This shift also creates an operational challenge. Few operators have the skills and tooling to build and operate large GPU-optimised infrastructure, and these capabilities differ materially from those developed for traditional IT and network cloud. Most operators will need to build them or rely on partners including IT vendors, public cloud providers, neoclouds and software vendors such as F5, Rafay and vCluster.

Whether investment continues at the projected pace will ultimately depend on ROI, which remains largely unproven. Operators will need to translate their infrastructure investments into measurable operational improvements or sustainable new revenue. The next few years will show whether these returns are sufficient to justify and determine the scale of the next round of AI infrastructure investment.

Analysys Mason's Cloud and AI Infrastructure research programme provides insight into how AI is reshaping infrastructure investment, architectures and competition across telecoms operators, vendors, hyperscalers and neoclouds. We help clients understand emerging opportunities in areas such as sovereign AI, GPUaaS and AI factories, and assess how the competitive positions of service providers and technology vendors are evolving.

Contact us to discuss how our subscription research and consulting services – from thought leadership and benchmarking to TCO and ROI analysis, surveys and competitive assessments – can support your business.

4 ASIC = application-specific integrated circuit