Edge AI: why telecoms operators should look again at edge computing

September 2026

Gorkem Yigit

As the focus of AI workloads shifts from training to inference, edge AI is emerging as the next layer of AI infrastructure. This is where telecoms operators could hold their strongest position in the AI compute stack, because it is where local compute and critical connectivity converge. However, it is also territory they have lost before. Analysys Mason's new strategy report argues that edge AI is a more credible proposition than multi-access edge computing (MEC) was, and sets out how operators can position themselves as the opportunity develops.

AI is reviving edge computing for operators with stronger foundations than MEC

MEC[1] was supposed to turn distributed network sites into a compute platform that enterprises and developers would pay for. The use cases were imagined in boardrooms with little feel for the markets they targeted, and built on what the infrastructure could theoretically enable rather than on what anyone was observably asking for. The conviction was that demand and ecosystem would follow the infrastructure, which did not materialise.

What is different now is that distributed AI infrastructure is becoming an architectural requirement rather than a speculative service. Our report identifies five dimensions on which edge AI differs from MEC.

- **Use case value.** MEC centred on narrow latency gains that were hard to monetise. Edge AI enhances the same use cases (for example, video analytics, public infrastructure and hyper-personalisation) with more intelligent capabilities that translate into measurable enterprise outcomes, and enables new services such as agentic applications and AI voice.
- **Demand**. MEC demand had to be manufactured. AI adoption across enterprise, government and regulated sectors now generates genuine demand for distributed, sovereign inference infrastructure.
- **Ecosystem.** MEC standards were poorly adopted, and operators relied on third parties that were often also competitors (for example, public cloud providers). The NVIDIA stack has since created a cross-vertical developer base on a unified platform and has a direct stake in making distributed inference work since inference now drives its growth. Other chipset manufacturers are also building competing ecosystems.

[1] An ETSI-defined system providing an IT service environment and cloud-computing capabilities at the edge of the access network, which contains one or more type of access technology, and is in close proximity to its users.

- **Operator differentiation.** In the MEC era, many players could deliver comparable capabilities. In edge AI, operators can bundle connectivity, compute and sovereign credentials into an offer that is difficult to replicate, particularly in regulated verticals.
- **Orchestration.** Control planes for distributed infrastructure (heterogeneous infrastructure management, workload placement and automation) were nascent during MEC. They are now maturing with GPUs in mind – with open frameworks and vendor investment – making edge AI viable at scale.

What should operators do differently with edge AI?

Better conditions are not the same as an operator advantage. What operators have that other edge players do not is a set of network assets and local credentials that can be bundled with AI compute, and the question is which use cases turn those into decisive advantages.

MEC use cases were designed mainly around lower latency. Edge AI use cases still value low latency, but they need much more: a performance envelope that holds up under mobility, dispersion, high concurrency and sovereignty. This is the kind of guarantee an operator can enforce at the network level.

- Take an emergency services body camera running live inference (for example, face or plate matching). The model is too large to sit on the camera, there is no fixed site to offload to in the field, the connection to the inference point has to perform consistently rather than most of the time, and the footage cannot leave the jurisdiction.

- A smart city's traffic intersections depend on aggregation: detection happens at the roadside; correlation cannot, as no single junction holds the picture. Running that correlation in a cloud region adds cost and delay to a loop that has to keep closing, so the aggregation has to happen between the endpoints and the cloud where an operator's network already sits.

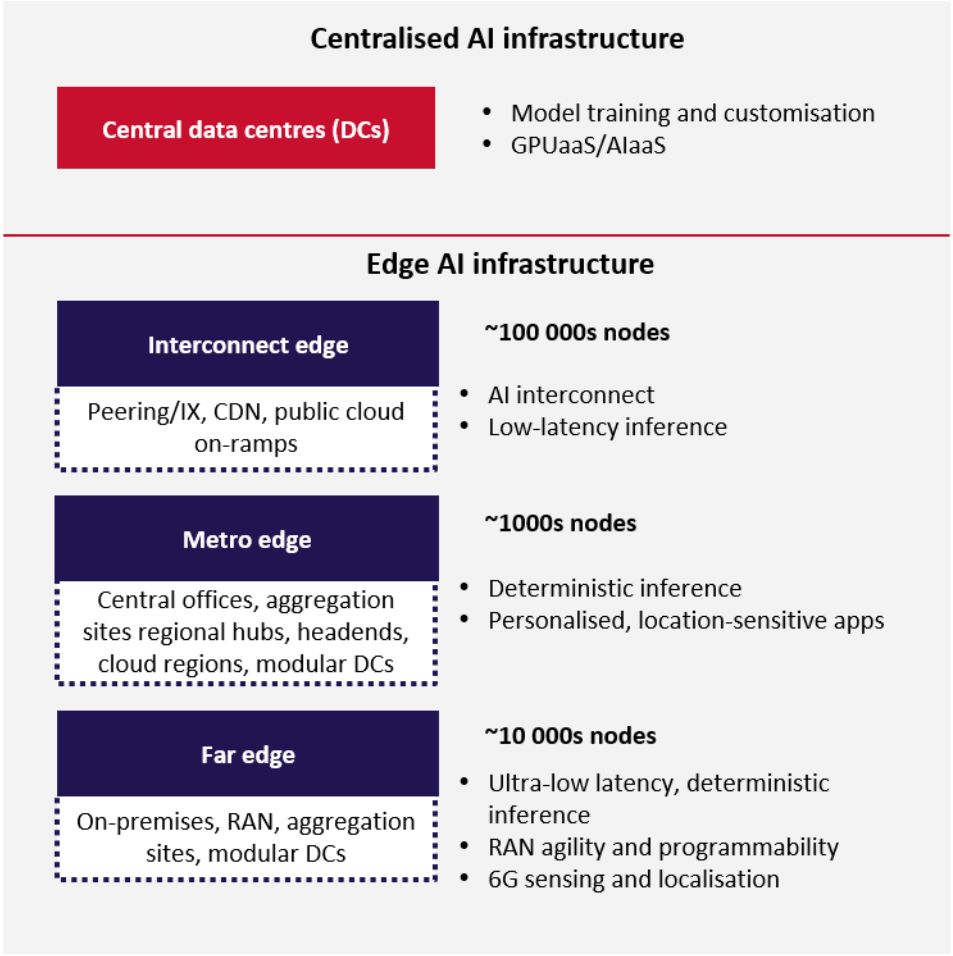
- Real-time ad insertion into a live stream by a cable operator adds concurrency; it needs strong performance guarantees to hold across thousands of simultaneous sessions, each personalised, in a very short window.

Operators have two edge infrastructure options: metro edge and far edge.[2] The spotlight is on the far edge, driven by AI-RAN with a cell site focus, but metro edge (central offices, aggregation sites and regional hubs) is better suited to the dispersed, aggregation-heavy workloads above. It sits high enough in the network to pool data across many sites and cells, while staying close enough to preserve the performance guarantees these use cases demand. Metro sites are also

[2] The interconnect edge is dominated by IX/co-location providers, content delivery networks (CDNs) and hyperscalers, leaving limited room for operators.

more concentrated than far edge and likely to have more favourable conditions (power, space, fibre and security), all of which make metro edge easier to scale and more suitable for operators' investment strategies.

Figure 1: Overview of AI grid with edge AI infrastructure



Source: Analysys Mason

Building out metro edge AI will still be a significant investment (approximately USD2–8 million or more per site), so it is realistically an opportunity for the operators that have a scalable metro footprint, enterprise connectivity relationships in target verticals and the appetite to invest. Not all will build alone: operators can pursue financial partnerships to develop, co-invest in or lease co-location sites in repurposed properties to scale infrastructure more capital-efficiently. That said, for a smaller set of operators with the assets, capabilities and market conditions to pursue AI RAN beyond internal RAN efficiencies, the far edge could become viable over the mid-to-long term, particularly as AI RAN matures and metro and far-edge capabilities combine. For most, however, metro will be the pragmatic starting point.

The direction of core network evolution also reinforces the metro edge. Connectivity and compute can increasingly converge at the same metro locations as operators disaggregate the 5G core and distribute functions such as the User Plane Function (UPF) towards the network edge,

and in future 6G brings in-network computing that places inference directly in the data path. This also eases the edge AI business case through shared build-out costs. In addition, edge AI could increase the demand for network slicing and advanced network APIs such as quality on demand, which have so far been lacking. These are the tools for delivering and monetising the deterministic connectivity that these use cases require.

Edge AI gives operators a second chance at edge computing. The opportunity is still nascent, but the demand, ecosystem and assets are aligning as they never did for MEC, and it will reward operators disciplined about where and what they build.

Analysys Mason's Cloud and AI Infrastructure research programme covers the convergence of cloud, AI and network technologies and its strategic implications. We have been at the forefront of the edge computing market, supporting operators, vendors and financial institutions with in-depth research and insights. Our recently published report, From MEC to edge AI: capturing the distributed AI infrastructure opportunity, examines the key pillars of a successful edge AI strategy for operators and sets out practical recommendations.