qvantel | NOKIA

5G AI SLICES:

HOW TELECOMMUNICATIONS PROVIDERS CAN UNLOCK NEW B2B REVENUES FROM THE NETWORK EDGE

UNLOCKING NEW REVENUES WITH AI-DRIVEN BSS AND AUTONOMOUS NETWORKS

Evolution towards autonomous networks, and AI driven BSS is set to increase efficiency and accelerate growth in the telecoms industry across service types. As a new element and service offering type for that landscape 5G AI Slices represent a growth and innovation opportunity.

5G AI Slices combine advanced 5G SA network slicing with GPU-powered AI compute at the edge, enabling new classes of enterprise services and consumer experiences that require both guaranteed connectivity and real-time AI inferencing close to where the experience or data is generated. This combination creates a fundamentally new commercial proposition. Telecommunications providers can offer enterprises not just a network connection, but a managed, SLA-backed platform for running AI compute needed by the services at the edge, without the latency penalties or data sovereignty risks of centralised cloud infrastructure.

For truly real-time, natural feeling experiences, such as voice interaction models, real-time rendered overlay information and guiding physical AI (robots or different moving devices) round-trip to a distant location for compute will often be too much (for first emerging examples, see Nvidia GTC sessions from 2026). The commercial and technical feasibility of this emerging proposition requires a high level of automation from sales and service assurance solutions, and evolution towards fully autonomous networks and AI driven BSS.

This opportunity, and evolution trajectory is being demonstrated through the TM Forum Moonshot Catalyst "Autonomy Accelerated: Connected Intelligence for Reliable Agentic Operations (Phase III)," developed by Qvantel, Nokia, BMC Helix, Infosys, and telecommunications provider Champions including Telenor, Deutsche Telekom, e&, Turk Telecom, VNPT, du and MTN.

THE MONETIZATION GAP

Despite significant investment in 5G, many telecommunications providers continue to sell mostly connectivity the same way they always have. Fixed data plans, monthly subscriptions, and volume-based pricing do not reflect the differentiated value that 5G SA and 5G Advanced networks can now provide.

Challenges in connectivity only approach call for new growth avenues. In most mature markets, ARPU has plateaued even as data consumption grows. B2B and vertical market use cases, including smart factories, remote healthcare, and autonomous logistics represent the most credible near-term 5G revenue opportunity. But capturing that value requires operational and commercial models that most telecommunications providers do not yet operate. A shift toward value-based monetization and establishing a new position in the AI computing and service delivery space can see telecommunications providers developing new AI based revenue streams.

AI-RAN: A NEW GENERATION OF NETWORKS WITH DISTRIBUTED CLOUD INFRASTRUCTURE BUILT IN

The Radio Access Network has historically been the most static component of mobile infrastructure. AI-RAN changes that entirely. It refers to the integration of artificial intelligence, including GPU-powered compute, directly into 5G cell sites and edge nodes, transforming them from passive signal relays into distributed, programmable intelligence and computing platforms. These platforms can process, predict, and adapt in real time, and act as distributed, cloud infrastructure for service applications that need the lowest possible latency and AI compute close to end-users.

AI-RAN transforms cell sites into dual-purpose assets, combining connectivity with monetizable edge compute. With traditional RAN running at roughly 30% utilisation on average, idle capacity becomes a commercial opportunity. Central to this is AI inferencing, the real-time execution of trained models at the network edge, where latency and data sovereignty matter most. Rather than routing requests and data to the centralised cloud for processing, inference runs locally, enabling faster decisions, lower bandwidth consumption, and greater control over sensitive data.

THE 5G AI SLICE OPPORTUNITY: GPU COMPUTE AT THE NETWORK EDGE

What Makes a 5G AI Slice Different

Network slicing is the ability to partition a single physical network into multiple logical networks, each with its own performance profile. A 5G AI Slice is more than a network performance guarantee. It is the combination of a dedicated, SLA-backed 5G network slice with GPU compute capacity at the network edge, enabling third-party service applications to run AI inferencing workloads close to where the experience and data is generated.

This is the key distinction from conventional network slicing. The value proposition is not connectivity alone but also managed AI compute delivered as a service over a guaranteed network path.

The latency advantage of edge-based AI inferencing over centralised cloud processing is substantial. A round trip to a central cloud facility may take hundreds of milliseconds. Placing GPU compute at the edge can reduce application response times to tens of milliseconds, an order-of-magnitude improvement for latency-sensitive workloads. This benefit comes primarily from the physical proximity of the compute to the application and the device, not from the 5G slice SLA itself.

The Commercial Opportunity This Creates

Three aspects of the 5G AI Slice model combine to create a distinctive B2B revenue opportunity for telecommunications providers.

AI edge compute as a service.

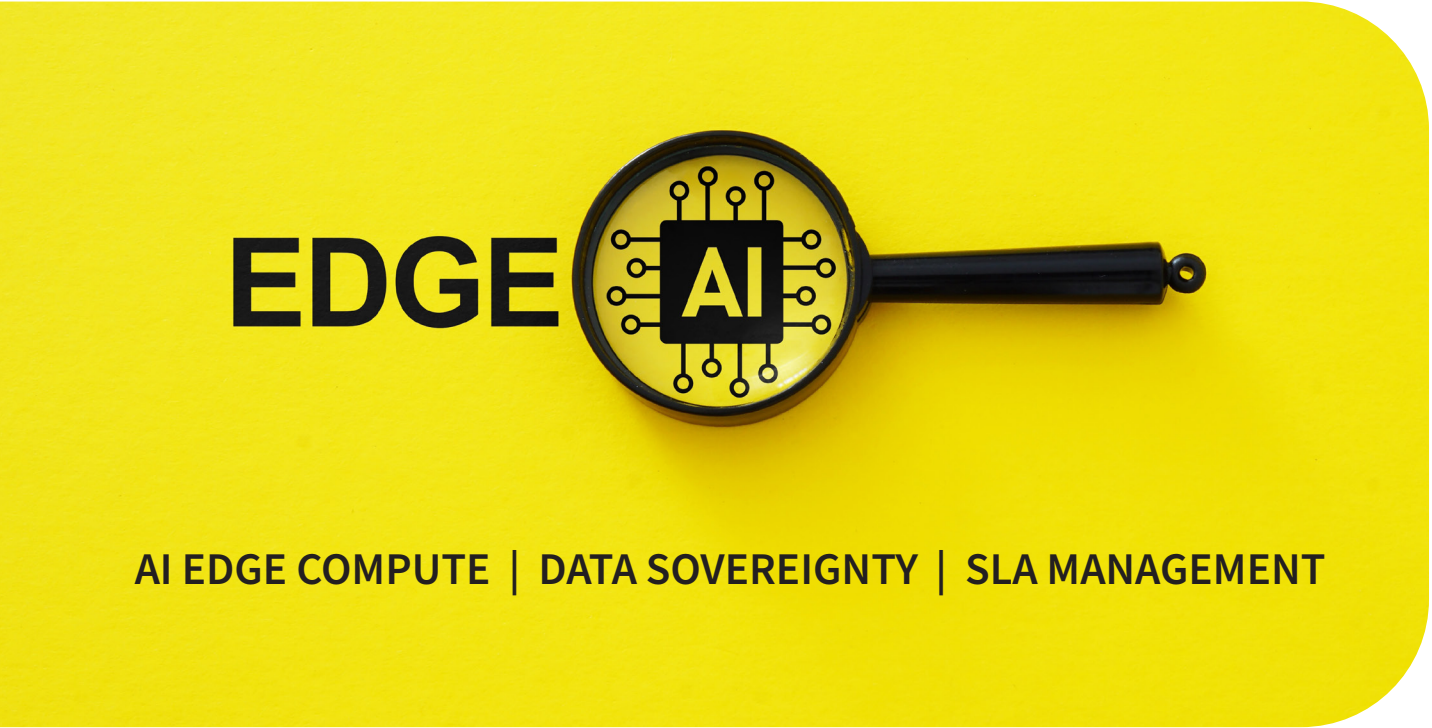
Telecommunications providers are no longer selling only connectivity. They are selling GPU compute capacity at the edge, AI inference capability, and the orchestration layer that integrates them, rated by tokens, API calls, or compute cycles, or even outcomes, rather than megabytes. Telecommunications providers are positioned as AI edge compute providers competing on superior experiences, proximity, sovereignty, and guaranteed performance rather than on raw compute cost alone.

Data sovereignty as a differentiator.

Telecommunications providers enjoy a high degree of enterprise trust and operate infrastructure that is geographically distributed and nationally anchored by design. For regulated industries where data cannot leave a jurisdiction, and for enterprises unwilling to route sensitive operational data through international hyperscale cloud infrastructure, or be reliant on non-domestic players, the telecommunications provider edge is a structurally advantaged compute location.

Guaranteed performance as a commercial product.

Because the 5G AI Slice combines a network SLA with a compute SLA, telecommunications providers can offer enterprises a contractual commitment covering both the connectivity and the compute required to run their AI application.



SAMPLE USE CASE FOR AI-SLICES

Supported by the TM Forum Moonshot Catalyst: Autonomy Accelerated: Connected Intelligence for Reliable Agentic Operations (Phase III)

This section explores a sample use case opportunity for 5G AI Slices enabled by evolution towards Autonomous Networks and AI-driven BSS, leading with a detailed example that is being actively demonstrated through the TM Forum Moonshot Catalyst programme. AI slicing use cases share a common characteristic: the AI inferencing required cannot be delivered from a centralised cloud facility within the latency and data sovereignty constraints the application demands. These are the scenarios where the 5G AI Slice is architecturally necessary, not merely convenient.

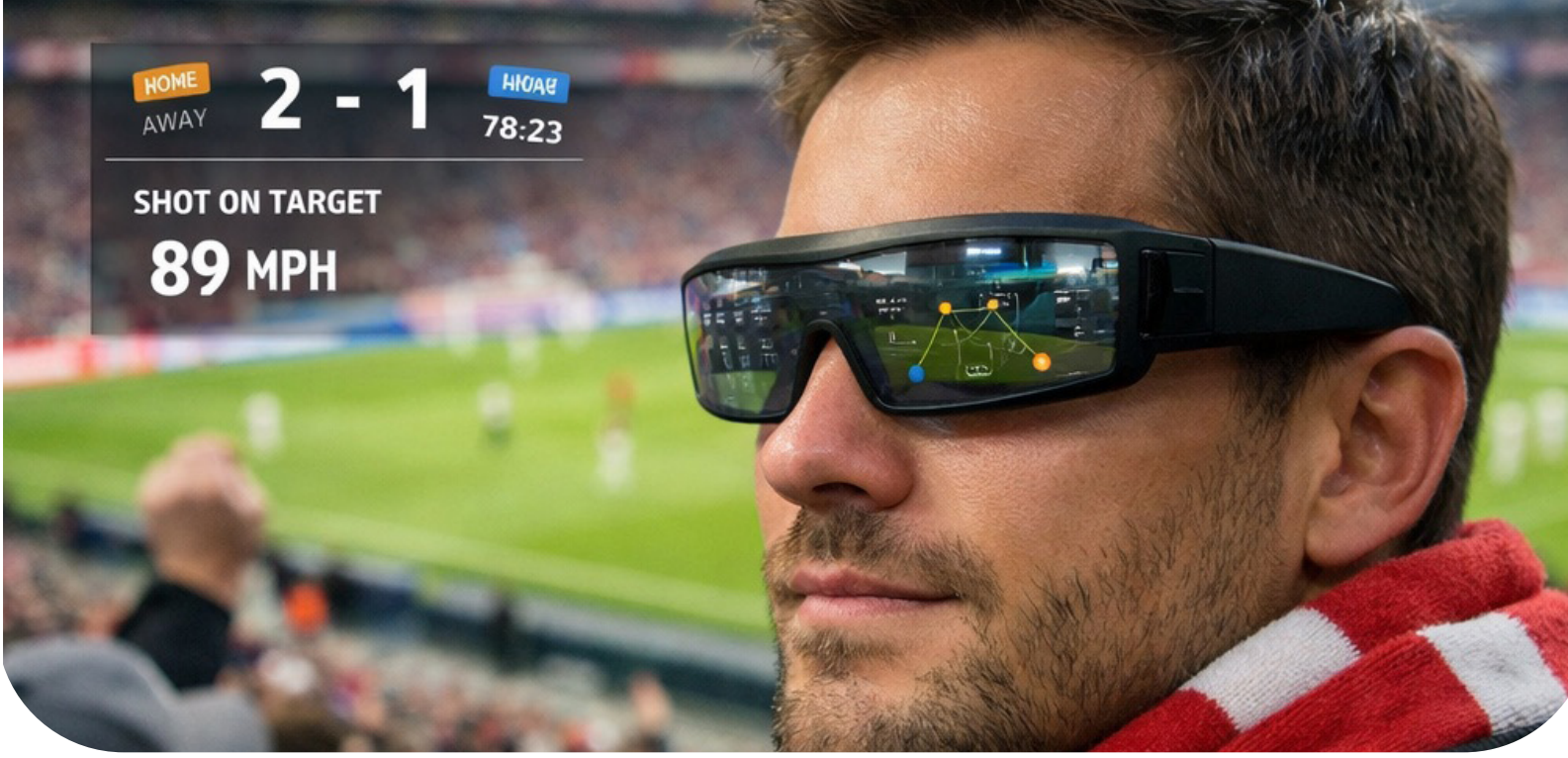
Augmented and Mixed Reality at Scale: AI Glasses at a Football Tournament

This use case is the centrepiece of the TM Forum Moonshot Catalyst "Autonomy Accelerated: Connected Intelligence for Reliable Agentic Operations (Phase III)," developed by Qvantel, Nokia, BMC Helix, Infosys, and telecommunications provider Champion partners. The catalyst demonstrates the commercial and operational lifecycle of a 5G AI Slice, from AI-assisted offer creation, feasibility testing, sales through to autonomous service assurance. The offer demonstrated in this catalyst is an enhanced VIP customer experience using AI glasses, delivered over an eMBB 5G AI slice at a high-profile football match. The following describes both the fan experience being delivered and the underlying technical demonstration that brings it to life.

The Commercial Model

A telecommunications provider partners with a major football event organiser (B2B customer for telecommunications provider) who is planning to sell an offer aimed at VIP ticket holders who receive AI glasses and tablets as part of a hospitality package. The AI glasses connect to a 5G AI Slice, which provides QoS-tiered connectivity for VIP in-stadium performance even during peak crowd demand. As well as distributed edge compute nodes in close proximity to the stadium for best possible real-time experiences with the AI-glasses.

The slice also needs to be able to support the traffic from the hi-definition video cameras and tablets used in the venue during the event. The telecommunications provider provides the 5G AI slice. The event organiser provides the equipment (AI glasses, cameras, tablets), related software experience for commentary and video and any other content required as part of the VIP experience.



Use Cases for AI Glasses at the Match

1. Intelligent Match Commentary- Fans can select different commentary options by language, tactical depth, and team allegiance.

2. Player and Ball Data Overlay on the Pitch- Glasses can provide an easy experience to ask for player-specific statistics and profile information, with AI audio responses and supporting overlays displaying player information and statistics, as well as real-time heat map position and team formation context. All of this is rendered on edge GPU compute, with sub-20ms latency ensuring the overlay tracks live player movement without lag.

3. Multi-Camera and Tactical Feed Selection- The glasses display an edge-rendered secondary feed in peripheral vision, automatically selected by an AI director tracking where the significant action is.

4. Stadium Navigation- The glasses combine on-device positioning with edge-computed crowd density data from the venue to guide wearers in real time, to provide navigation and information for example, regarding less crowded service points.

5. Social and Shareable Moments- The AI glasses identify high-intensity match moments in real time, including goals, near-misses, and significant VAR decisions, and automatically clip a personalised highlight reel from the wearer's perspective. At full time, the fan receives a curated short video combining their in-glasses view with broadcast footage of the moments they witnessed. This is processed on edge GPU compute at the venue and delivered to the fan's device within minutes of the final whistle, creating a shareable personal artefact tied directly to the experience package.

Why Edge GPU Compute Is Essential

None of the above is achievable from a centralised cloud for superior, real-time experience. The overlay rendering, AI director feed selection, crowd density computation, and real-time highlight processing all require GPU inference running within proximity to the venue. Round-trip latency to a distant data centre would introduce visible lag on player overlays, break the responsiveness of navigation guidance, and make real-time clip generation impractical during the match itself. The 5G AI Slice provides both the dedicated, guaranteed network path and the edge GPU compute that make the experience reliable across many concurrent AI glasses users at the stadium.

How the Catalyst Demonstrates This End-to-End

The catalyst structures the demonstration around two complementary use cases that together cover the full intent lifecycle, from the first sales conversation through to service assurance and autonomous recovery from a live network fault.

This is summarised in figure 1 below:

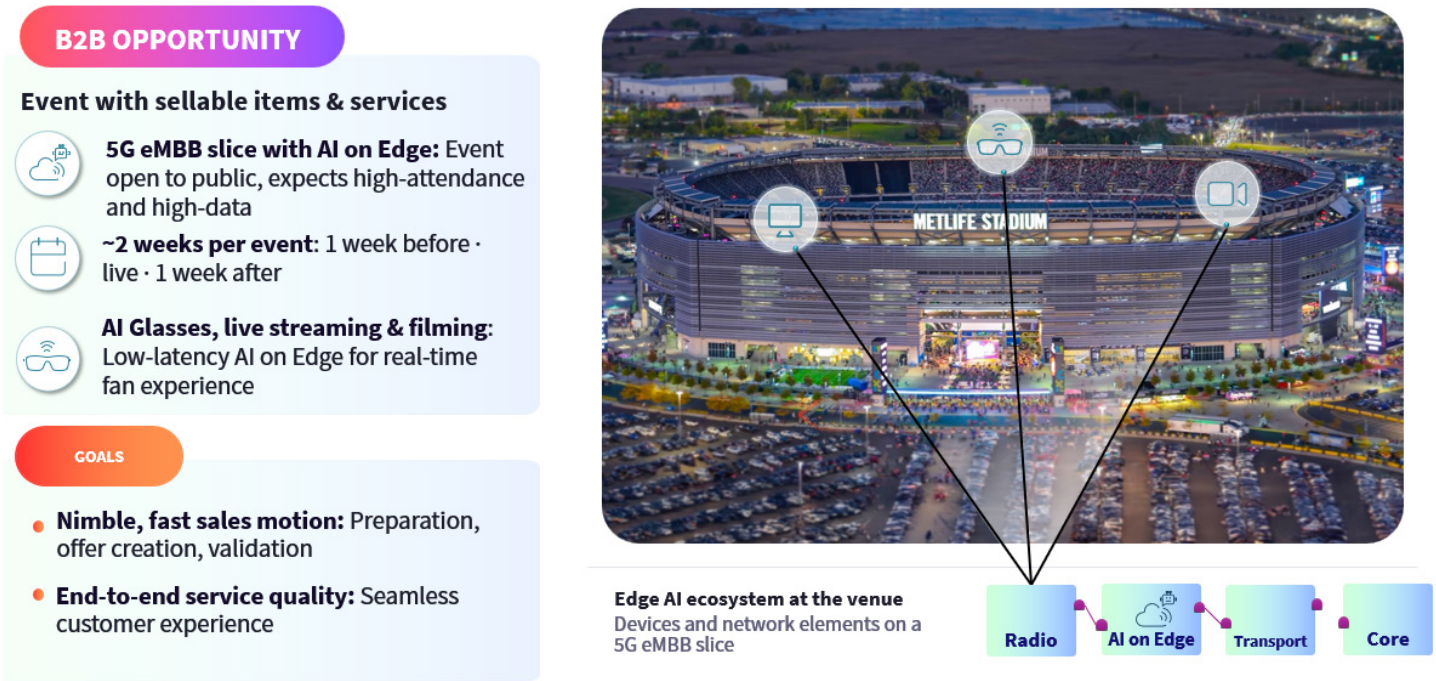


Figure 1: Summary of the moonshot catalyst - Autonomy accelerated

The high-level solution architecture is shown below in figure 2. This shows systems from Qvantel, Nokia, BMC Helix and Infosys. And Infosys integrated to provide the end-to-end solution for the Catalyst that enables telecommunications providers to create and monetize offers based on 5G AI slices.

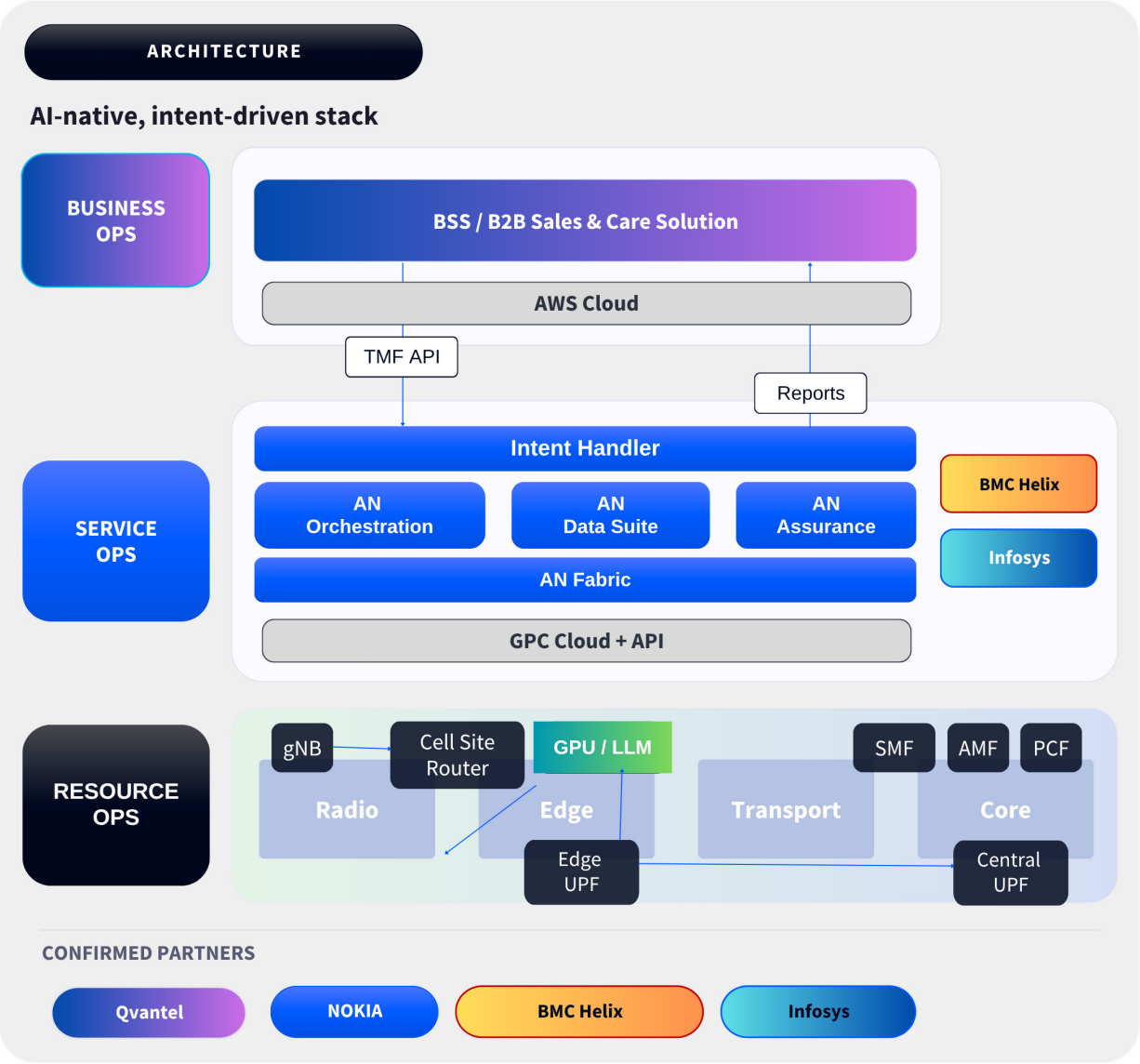


Figure 2: Solution architecture enabling creation, sales and assurance of 5G AI slices

Use Case 1: Offer and Sell the 5G AI Slice and Ensure it Will Work as Expected

The offering and selling process starts with the B2B sales representative of the telecommunications provider using AI driven BSS to capture the lead and start approaching the customer. The sales representative can use the B2B sales tool GUI, or alternatively MCP compatible native AI client (e.g. Claude Desktop tool) to create a lead, ask and compare other similar sales cases for ideas for preparing a meeting. After the sales meeting / call they can use the recorded meeting transcript to form a business intent on the customer business need.

In this case, the captured business intent from the customer included the following characteristics for the event that the customer is organizing and expected types and volumes of devices used in the event:

Expected device types: 100 pairs of AI glasses, 50 tablets and 50 cameras,

Location that the slice should cover: MetLife Stadium

Dates / times when the slice should be deployed: 00:00 on 1st August to 23:59 on 15th August.

Once the intent is captured, the Qvantel BSS initiates a feasibility probe via the TMF921 intent API. This feasibility probe is designed to validate the feasibility of the intent and get suggested services to be offered to the customer. Nokia's orchestration and agentic intent handling layer resolves the probe into constituent services, covering core, RAN, transport, and edge, using AI agents and LLMs to assess deliverability. In parallel, Infosys agents collect traffic forecasts from Nokia's Assurance Center and validate planned maintenance windows via BMC Helix to identify any delivery risks during the event window. The outputs are synthesised into a single recommendation report and Nokia produces TMF921 intent reports confirming feasibility and delivery confidence.

As part of the intent validation, Nokia's orchestration layer detected that the customer business intent included the use of AI Glasses. Using that information, the intent probe response suggests selling also edge computing service, for best possible AI experience delivery using edge located GPU services.

Once the B2B sales representative gets that suggestion, AI assisted offer creation and pricing is used to formalize the commercial offer to be sent to the customer. This includes adding token-based pricing model for the suggested edge computing service element.

Once the offer is accepted, the BSS provisions the service as required. This process flow is illustrated in figure 3 below.

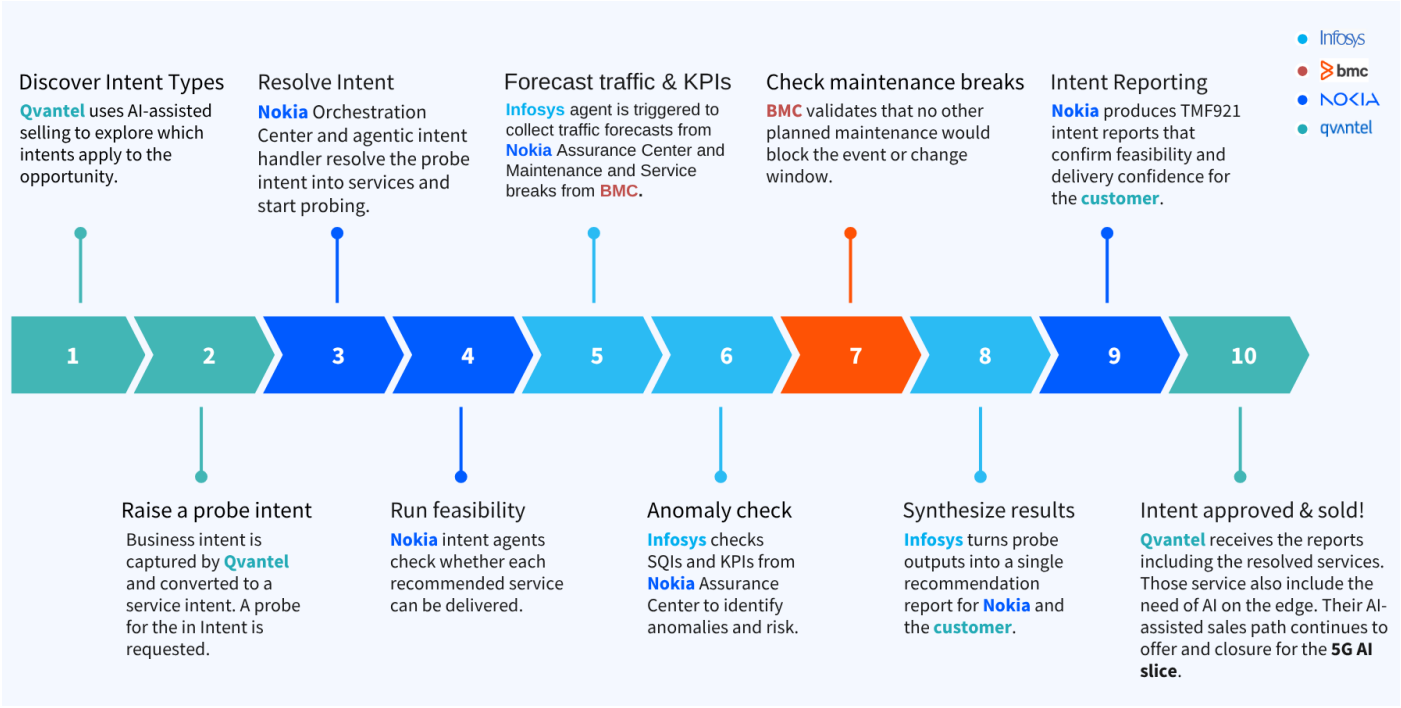


Figure 3: Process flow - from intent discovery to selling 5G AI slice

A demo of the catalyst solution to create, offer, monetize and assure 5G AI slices can be seen by playing the video below.

Use Case 2: Agentic Assurance for 5G AI Slice and Associated Services

Once the slice is live, the system continuously monitors subscriber experience on the AI glasses and connected devices. When a network event, such as a fibre cut, causes end-user experience to degrade, an incident is raised automatically, and root cause analysis is generated without human initiation.

Agentic operations drive an agent-to-agent workflow between BMC Helix's service management tooling and Nokia's autonomous network. Trouble tickets are updated across service and physical layers, the best corrective action is selected, and a human-in-the-loop step allows the operations team to approve the proposed network change before it is applied, ensuring governance and oversight. Service monitoring then confirms the intent has been restored, tickets are closed, and an updated intent state report is sent to the Qvantel BSS. The outcome is intent restored, the AI glasses experience is recovered and the SLA maintained. This is illustrated in figure 4 below.

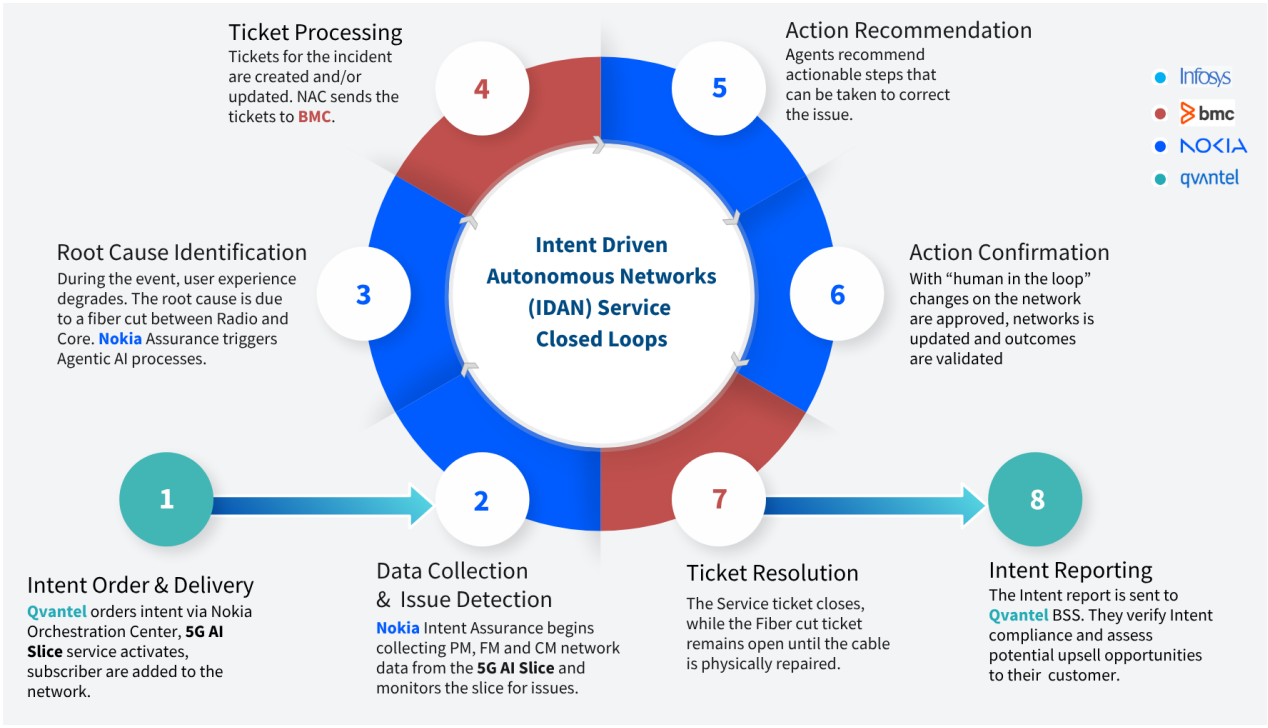


Figure 4: Closed loop ensuring service assurance of 5G AI slice

AI-Native, Intent-Driven Architecture

The catalyst is built on an AI-native, intent-driven architecture spanning three operational layers. At the business operations layer, the Qvantel BSS handles B2B sales, care, and commercial management, including AI-assisted offer creation, intent capture, and real-time SLA reporting. At the service operations layer, Nokia's intent handler, autonomous network orchestration, data suite, and assurance center manage service and resource intents and the closed-loop automation that keeps the AI slice performing to its contractual commitments.

The resource operations layer encompasses the physical and virtual network elements, with Cloud and AI providing infrastructure for edge compute and AI inferencing. This architecture directly instantiates a closed-loop model: a service is requested, the BSS validates commercial terms against network feasibility, the autonomous network creates the slice and activates edge compute, AI monitors SLA compliance continuously, and the BSS charges and reports in real time.

What the Catalyst Validates

The catalyst validates three core claims of the 5G AI Slice proposition.

First, that a telecommunications provider can sell a 5G AI slice with confidence, using digital feasibility checking and AI-assisted sales in the BSS to ensure the offer will perform as committed before the contract is signed.

Second, that the slice can be created and activated end-to-end using intent-driven orchestration, with all constituent services. Including the eMBB slice, and edge compute, from a single business intent.

Third, that the service can be continuously assured using agentic operations, with AI-driven root cause analysis and agent-to-agent workflows between the BSS and the autonomous network restoring intent automatically when degradation occurs, with human oversight at the point of change approval.

OTHER USE CASES FOR 5G AI SLICES

The football tournament use case illustrates the broader pattern. Each of the following scenarios shares the same underlying logic: AI inferencing is required at the network edge because latency or data sovereignty constraints make centralised cloud processing unworkable, or the device or environment has no independent local compute.

Computer Vision and Video Analytics at the Edge

Running AI inference on live camera feeds for public safety, crowd behaviour monitoring, traffic management, and environmental surveillance generates data volumes that make centralised cloud processing impractical. A deployment of high-resolution cameras across a city, a port, or a major venue produces continuous streams that must be analysed locally to deliver real-time detection and alerting. Routing that data to a distant cloud introduces latency and bandwidth costs that undermine the operational value. By running inference at the network edge close to where cameras are deployed, telecommunications providers can offer enterprises and city operators a managed vision AI service with guaranteed response times and local data sovereignty.

Real-Time Voice AI and Conversational Agents

Conversational AI services are acutely sensitive to latency. Responses exceeding 500ms feel unnatural to users. Placing GPU inference capacity at the edge reduces both network round-trip time and queuing latency, making human-grade voice AI viable at scale. For telecommunications providers with AI-RAN infrastructure, the same GPU capacity managing radio resources can host voice inference workloads turning idle edge compute into a monetizable service without additional capital expenditure.

Emergency Services and Public Safety

Where first responders operate under a priority slice sharing infrastructure with the public, the SLA carries a safety obligation. AI inference at the edge also powers real-time video analytics for connected cameras and sensors, enabling faster detection and response while keeping sensitive data local.

Clinical Telepresence and Remote Diagnostics

Specialist consultations, real-time diagnostic collaboration, and AI-assisted patient assessment over mobile networks are rapidly expanding realities. A video feed that stutters at the wrong moment can directly affect clinical judgement. Community clinics, paramedic units, and remote diagnostic facilities typically have none or very limited local compute infrastructure. The GPU compute at the base station is therefore the only point in the network where AI-assisted monitoring and remediation can operate within the required response window. Data sovereignty reinforces the case, as sensitive patient data processed at the edge stays within jurisdictional boundaries by design.

Sovereign and Regulated AI Workloads

Government services, financial institutions, and regulated healthcare organisations require strict control over where data is processed. Routing inference through international cloud infrastructure creates compliance exposure that many enterprises cannot accept. GPU compute at the telco network edge keeps processing within national or regional boundaries by design, offering regulated enterprises a path to AI-native services that is compliant by default.

Hyper-Personalised Consumer and Enterprise Services

Real-time content personalisation, AI copilots, and dynamic media insertion require high concurrency within tight cost budgets. Distributed edge inference addresses the economics as much as the latency. Benchmarks from early NVIDIA AI Grid deployments, including data published by Comcast, demonstrate cost-per-token reductions of up to 76% compared to centralised architectures for real-time voice, vision, and media workloads (source: NVIDIA GTC 2026, AI Grid reference design benchmarks). For telecommunications providers, the distributed network footprint that has historically been a cost burden becomes a structural advantage in a world where inference economics depend on proximity to the user.

NEED FOR BSS TO EVOLVE FOR 5G AI SLICE MONETIZATION

The network capabilities described above are increasingly deployable. The typical legacy BSS infrastructure to monetize them is not keeping pace. Traditional BSS platforms were designed for standardised products, monthly billing cycles, and a small number of distinct service types. They cannot support products defined by SLA parameters, pricing that adjusts in real time, charging triggered by slice lifecycle events, API-based service requests from enterprise developers, or multi-party revenue sharing across ecosystem partners.

Modern, agile BSS changes that. It enables composable, SLA-driven product models where a single offer might bundle a latency guarantee, a reliability tier, an edge GPU compute allocation, and an API access right, priced dynamically based on what the network is delivering at any given moment. No/low-code configuration means product and commercial teams can define new offers, modify pricing rules, and launch bundles without IT development cycles, moving from concept to market in hours rather than months.

DIMENSION	LEGACY BSS (SERVICE-HEAVY)	MODERN BSS (NO/LOW-CODE)
Change model	Vendor-led customisation	Telco-led configuration
Time to launch	Weeks to months	Hours to days
Who makes changes	IT + vendor professional services	Business users + product teams
Pricing flexibility	Rigid, hard-coded logic	Dynamic, configurable rules
Experimentation	Limited, high cost to test	Rapid A/B testing and iteration
Cost structure	High OPEX (change requests, integrations)	Lower OPEX (self-service changes)
Response to market	Slow, reactive	Fast, proactive and AI-driven
Support for AI monetization	Limited integration	Native support for real-time and AI-driven pricing

AI embedded in the BSS layer enables predictive offer management, dynamic pricing based on real-time network conditions, and proactive churn intervention. As telecommunications providers expose network capabilities through APIs, the BSS must also handle multi-party revenue sharing natively, settling revenue flows across the telecommunications providers, edge compute providers, and software partners automatically and at scale.

The most powerful evolution is a fully integrated closed loop. A service is requested, the BSS defines commercial terms and validates them against a network digital twin, the autonomous network allocates the appropriate slice and edge GPU compute, AI monitors SLA compliance continuously, and the BSS charges in real time. Credits are applied where performance fell short and upsell interventions are triggered where demand is growing. This is what transforms AI-driven network capabilities from a cost efficiency tool into a scalable revenue engine.

NEW PRICING MODELS FOR A VALUE ECONOMY

As telecommunications providers are moving from selling undifferentiated connectivity based on volume to selling services that can be priced based on value we are seeing several new pricing models emerge. A sample of these are represented at a high level in figure 5.

QoS-based pricing ties charges to contractually defined performance levels rather than data volumes. Dynamic pricing adjusts terms in real time based on demand, location, and time of day. Usage-based and elastic pricing aligns more closely to cloud infrastructure models, which is particularly suited to variable IoT and enterprise workloads.

Outcome-based pricing is the most ambitious model. It ties charges to measurable business outcomes. A broadcaster could pay per successfully delivered event, while a healthcare provider could pay per completed remote consultation at the guaranteed quality level. API monetization opens a new high-margin revenue stream by allowing developers and enterprises to consume specific network and compute capabilities on a per-call or subscription basis.

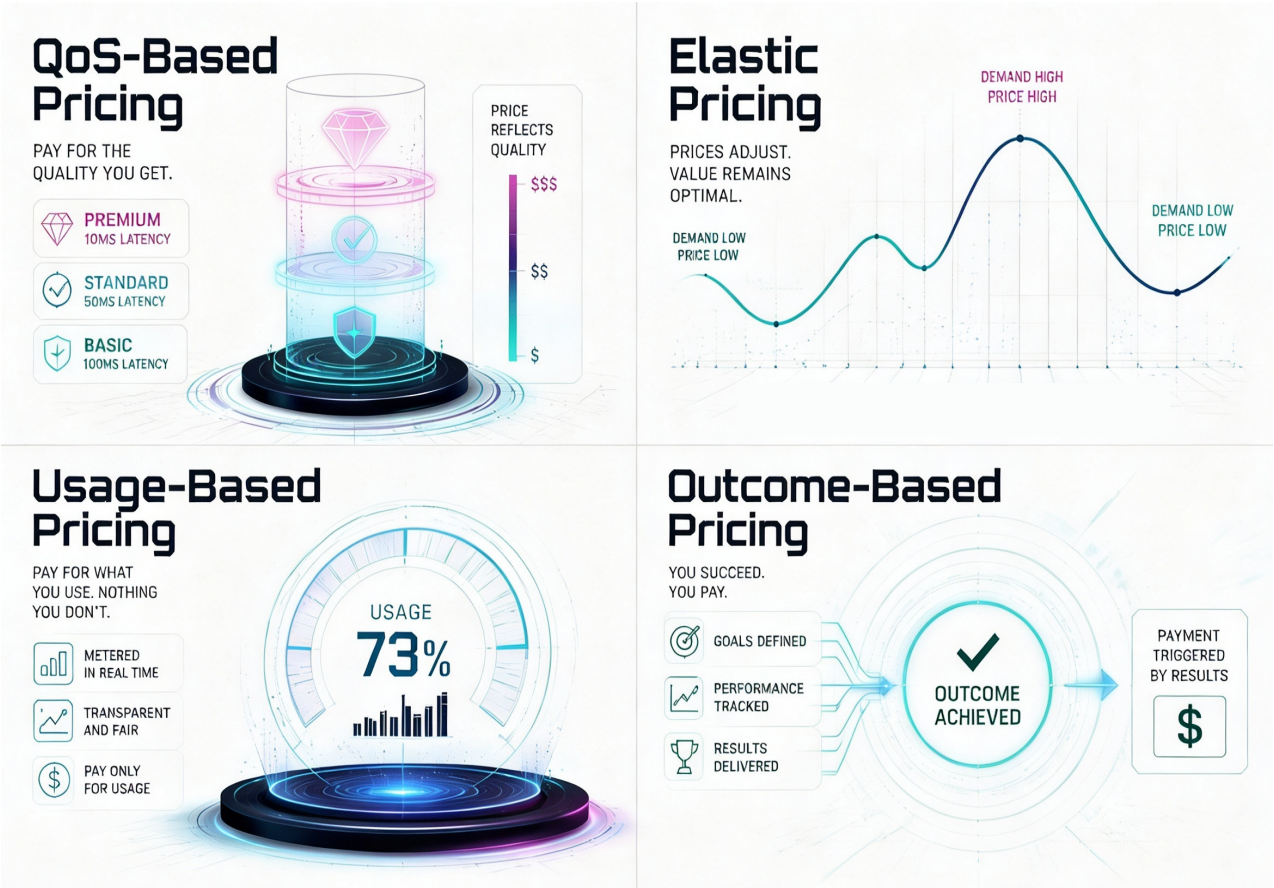


Figure 5: High level view of a sample of pricing models

STRATEGIC PRIORITIES

Three priorities stand out for telecommunications providers seeking to capture the 5G AI Slice revenue opportunity.

Lead with GPU inferencing use cases in enterprise verticals and real-time end-user experiences.

The specific combination of guaranteed connectivity and edge GPU compute enables AI applications to run where centralised cloud cannot reach. Differentiated experiences in events, broadcasting, healthcare, public safety, and smart city operations are where this proposition is most immediately credible and where enterprise willingness to pay for guaranteed AI inferencing performance is clearest.

Invest in modern, agile BSS.

Monetizing 5G AI Slices requires BSS capable of billing for GPU compute cycles, tokens, and API calls alongside network SLA parameters, handling dynamic pricing, and managing multi-party revenue sharing across the ecosystem. This is not an IT infrastructure upgrade. It is the commercial foundation of a fundamentally different kind of business.

Build toward the closed loop.

The long-term prize is a fully automated, real-time connection between network and compute intelligence and commercial monetization, converting every network and inference event into a potential revenue signal. Telecommunications providers who build toward this architecture now will have a structural commercial advantage as the market matures.

SUMMARY

5G AI Slices, combining guaranteed 5G connectivity with GPU-powered AI compute at the network edge, represent the foundations of a fundamentally different and new commercial opportunity. Here telecommunications providers can sell outcomes, guarantee performance, and participate in digital ecosystems as edge AI platform providers rather than just suppliers of connectivity.

The TM Forum Moonshot Catalyst "Autonomy Accelerated: Connected Intelligence for Reliable Agentic Operations (Phase III)," developed by Qvantel, Nokia, BMC Helix, Infosys, and telecommunications provider champions, demonstrates how telecommunications providers can deliver AI slices and open new revenue streams and play a bigger role in the AI value chain. Using a high-attendance football match as the real-world scenario, the catalyst shows the complete lifecycle in action: AI-assisted offer creation and intent probing before the sale to test and prove offer feasibility, automated slice creation and autonomous agentic operations maintaining the SLA once the service is live. This is just one use case. The others discussed in this paper show there are many new opportunities to create value using AI slices.

What determines which telecommunications providers capture this value is the ability to monetize it at scale. That requires BSS that is as intelligent, agile, and event-driven as the networks it supports and is capable of billing for GPU compute cycles and API calls, handling dynamic SLA-based pricing, and managing multi-party revenue sharing across the ecosystem in real time and also being able to support outcome based pricing. The catalyst demonstrates this integration with Qvantel's BSS driving AI-assisted sales and intent capture, Nokia's autonomous network delivering and assuring the slice, and BMC Helix and Infosys providing the agentic operations layer that closes the loop between the commercial and network domains.

The central question for telecommunications providers is not whether they can build smarter networks. The evidence is clear that they can. The question is whether they can build the commercial systems to turn network and AI edge compute capability into revenue, consistently, at scale, and continuously.

ABOUT QVANTEL

Qvantel, established in 1995, is a leading provider of digital BSS and monetization software for the telecommunications and digital services industry. Qvantel Flex is an AI-driven BSS and monetization suite enabling operators to launch new offers rapidly using low/no-code configuration and automation. In 2025, Qvantel expanded its operations and portfolio with the acquisition of charging and BSS supplier Optiva. Qvantel supports 70+ operators in over 40 countries, powered by 1,000+ specialists across all global regions.

For more information, visit www.qvantel.com.

qvantel

ABOUT NOKIA

Nokia is a global leader in connectivity for the AI era. With expertise across fixed, mobile, and transport networks, we're advancing connectivity to secure a brighter world.

NOKIA