
# AI-on-RAN: Enabling Monetizable Differentiated Connectivity for AI

AI-RAN Alliance —

Working Group 3 White Paper

ai-ran.org

AI-RAN™ ALLIANCE

# Contents

# 1. Executive Summary

AI is becoming a mobile-native experience: consumers increasingly expect always-available, context-aware AI assistants that listen, see, and respond in real-time. Enterprises are deploying AI into safety-critical workflows – such as autonomous vehicles or droids – where the network is no longer infrastructure but part of the value proposition.

Best-effort mobile networks, however, exhibit uplink variability, mobility-induced interruptions, and congestion-related latencies. For interactive and multimodal AI, this variance is more damaging than average throughput. An uplink micro-outage or problematic handover can degrade or terminate an AI session outright, even when average link speeds are high. For emerging AI services, predictability beats peak rate!

**This white paper argues that AI creates the technical imperative and the commercial opportunity for differentiated connectivity, particularly in the uplink and under mobility.**

The technical foundations include network slicing, deterministic latency mechanisms, uplink optimization, local breakout, analytics exposure, and intent-driven orchestration, among others. What has been missing is a compelling use case that justifies the investment, knowing there is a willingness to pay. AI changes that assumption by making network quality perceptible at the application layer, creating demand for premium connectivity that operators can provide and monetize.

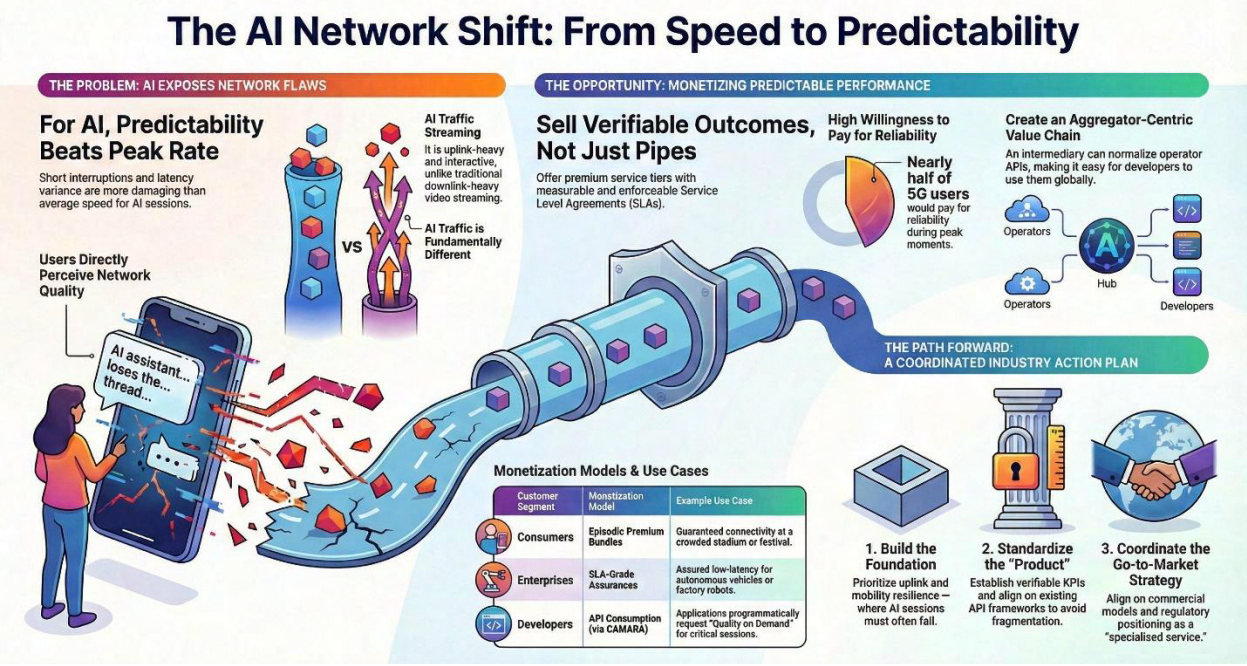
**Key findings.** AI applications make network quality perceptible at the application layer in ways previous services did not: users notice when an assistant loses context mid-sentence, and enterprises measure when connectivity failures halt autonomous operations. This perceptibility creates accountability and, with it, demand for premium connectivity as well as a commercial basis for monetizing it. The technical building blocks for differentiated connectivity already exist across 3GPP, ETSI, O-RAN, GSMA/CAMARA, and related initiatives; what is missing is alignment on common definitions, interoperable interfaces, and deployment sequencing. Crucially, "premium" is only commercially sustainable if service objectives are verifiable: enforceable KPIs with normative measurement methods connect technical capabilities to commercial viability. A phased deployment approach – beginning with uplink observability, progressing to enforceable service classes, and eventually to intent-driven orchestration – offers a pragmatic path that reduces operational risk while building toward full capability.

**The opportunity.** Operators can monetize differentiated connectivity through multiple channels: consumer tiers that guarantee reliability during high-demand moments, enterprise SLAs with per-application or per-flow granularity, API bundles that expose network capabilities to developers, hyperscaler partnerships that integrate connectivity into broader AI offerings, among others. An aggregator-centric value chain can normalize operator capabilities across markets,

enabling AI application providers to consume network services without bespoke integrations. Research indicates significant willingness to pay: nearly half of global 5G users experience connectivity challenges during peak hours, and about half of those would pay for reliability. For enterprises, the stakes are even higher: when connectivity failures ground autonomous operations, the cost of downtime dwarfs the cost of assurance. Operators who establish differentiated offerings early will secure positions in AI value chains that extend beyond commodity transport.

**Call to action.** Working Group 3 recommends that the industry act with urgency across five areas: first, prioritize uplink and mobility resilience as the foundational technical investment, addressing the scheduling, cell-edge, handover, etc., challenges where AI sessions actually fail; second, establish app-agnostic performance levels with verifiable KPIs that enable SLA enforcement and commercial accountability suitable to AI workloads; third, align telemetry and API exposure with existing industry frameworks rather than creating parallel interfaces; fourth, publish phased deployment guidance with security and privacy requirements embedded at each stage; and fifth, coordinate commercial constructs and regulatory positioning across the ecosystem to avoid fragmentation.

*In summary, AI applications represent compelling use cases that benefit from – rather than necessitate unique treatment beyond – the performance level approaches established by the industry. The building blocks exist; the demand is emerging. What remains is alignment and execution.*



# 2. The Rise of Mobile-Native AI Applications

Mobile networks have historically been optimized for consumers using downlink-heavy video streaming and some best-effort applications. AI-native services change that premise entirely.

## Consumer AI: From Contextual Search to Always-On Assistants

The use of **AI-native applications** on consumers' smartphones has increased sharply over the past few months.

The usage of AI-native applications on consumers' smartphones has increased drastically over recent months. Mobile-native generative AI adoption is accelerating, with active usage approaching 20% in leading European markets and expected to exceed 50% by 2030 if the current trajectory continues [2-1].

This growth is supported by rapid ecosystem expansion: mobile AI apps reached over 115 million global downloads in December 2024 alone, with more than 29,000 AI apps now available across major app stores [2-2]. While current user sessions remain relatively short (typically 1–2 minutes with limited interactions), more intensive multimodal use cases are emerging, particularly for visual and video generation [2-1].

Despite fast adoption, AI traffic still represents a very small share of overall mobile data volumes, reflecting lightweight overall usage patterns for most applications, compared to download-heavy video applications [2-2]. At the same time, AI traffic shows structurally different network characteristics, with higher uplink intensity compared to traditional mobile traffic – thus signaling a shift toward interactive, always-on assistant behaviors rather than passive content consumption [2-2][2-3].

Whilst some applications can use smaller, privacy-first on-device AI models, most high-end experiences require a hybrid or cloud-based approach. This creates a new and subtle expectation: whilst users will tolerate occasional buffering in video, they do not tolerate an assistant that "loses the thread" mid-interaction due to network deficiencies.

*Conversational continuity and responsiveness thus become primary measures of quality, and the network's variance becomes visible in ways it never was before.*

In addition to today's AI-native apps, two new consumer paradigms are emerging that will amplify this further: new AI-native devices and always-on personal AI agents.

New **AI-native devices**, such as AI-enabled glasses or other camera-enabled devices running multimodal assistants, introduce a distinct traffic profile, often generating request-driven bursts that trigger immediate responses. Indeed, glasses can act as streaming sensors, producing

persistent uplink flows of video snippets, a sequence of images, and/or audio. Here, even modest frame drops or short uplink micro-outages can disproportionately degrade the end-user experience or even break sessions.

Beyond latency, energy efficiency becomes a critical constraint: for glasses-class devices, the marginal power cost of uplink transmission – including radio on-time, retries, ramp-up, etc. – can be as significant as the compute cost of on-device inference [2-4]. Measurements further indicate that lightweight wearables often rely on intermediate proxy processing today, where raw sensor data is partially processed outside the device before being forwarded to applications or the cloud. While this reduces on-device computation, it introduces additional latency variability and dependency on external system states, highlighting the importance of predictable handling of the uplink [2-4].

**Always-on personal AI agents** used on smartphones, AI glasses, or other devices introduce yet another novel consumption behavior: short tasks mixed with longer multimodal sessions, and potentially micro-transaction or metering models tied to tokens or sessions. Moreover, token- or session-based pricing may indirectly shape user behavior, encouraging prompt fragmentation, iterative refinement, or even early termination – all of which amplifies burstiness. While token economics is primarily an application-layer concept and often executed in the cloud, it matters for networking because it correlates with bursty uplink prompts and interactive response loops.

## Enterprise AI: The Network Becomes Part of the Workflow

Enterprise and industrial AI use cases embed the network into operational workflows. This increases the value of assured performance and, with it, the cost of failure.

**Autonomous vehicles and V2X.** Autonomous driving today largely relies on in-vehicle AI powered by onboard sensors. However, an SLA-grade network enables unique capabilities: aggregating multimodal sensor data beyond a single vehicle, centralized or edge-AI processing to generate higher-level insights such as cooperative perception, and the possibility of remote assistance or remote operation where appropriate [2-5]. Cloud-processing via networks is not a replacement for in-vehicle autonomy; it is an augmentation layer that becomes valuable when connectivity is predictable and protected, particularly for uplink video/sensor sharing and low-latency coordination.

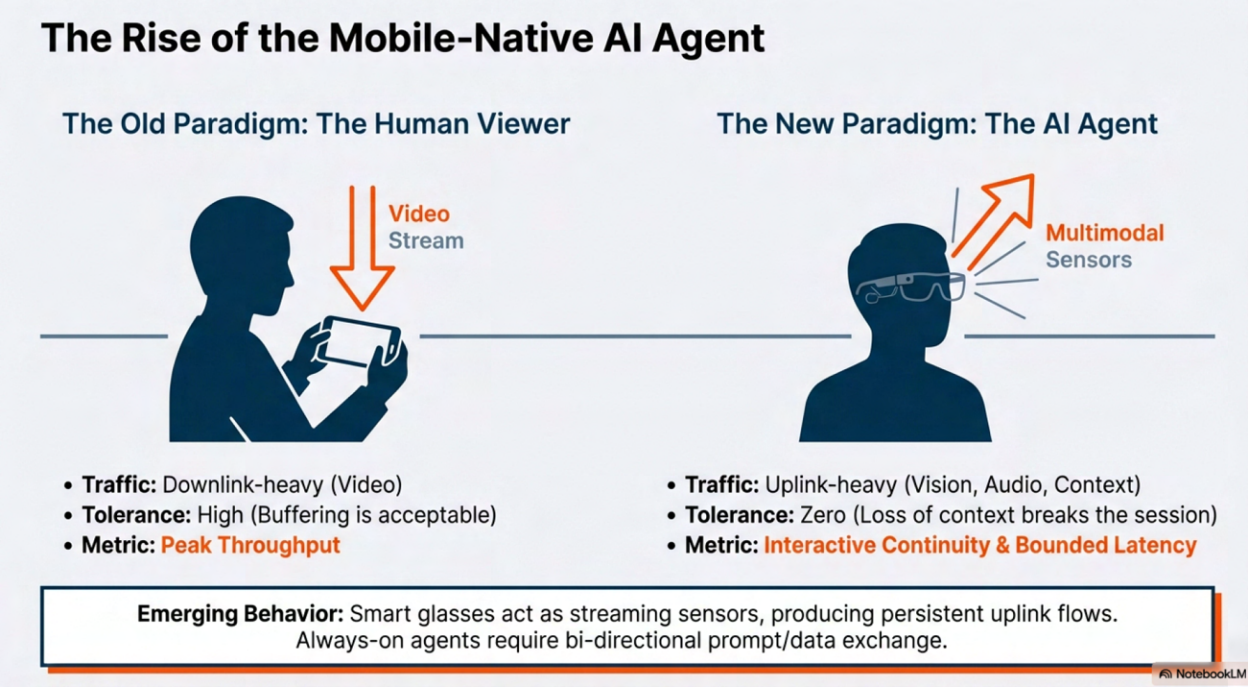
**Droids and closed-loop control.** Some emerging designs for mobile robots, drones, and other autonomous machines create a closed-loop dependence on communication. A perception stream flows uplink, drives inference, which drives a control action flowing downlink, always with stringent latency constraints. The emphasis here is on mobile intelligent robots rather than stationary automation, because mobility is precisely where wireless variability and handover behavior are most challenging.

**IoT and sensing at scale.** At scale, sensing becomes an AI data pipeline driven by many devices, heterogeneous uplink profiles, and the need to move data efficiently to inference and analytics. Even when individual sensors are low-rate, their aggregate uplink and the need for robust coverage will have a perceivable impact on networks.

## What AI Workloads Require from the Network

Throughout this white paper, AI workloads are assumed to follow an AI-on-RAN deployment model, where AI applications and models are hosted outside the RAN protocol stack. Still, they may be co-located with RAN infrastructure at the network edge. In this model, AI applications remain logically separated from AI-in-RAN functions by the user plane (UPF); this preserves architectural boundaries, security/trust domains, and operational responsibilities. Connectivity, therefore, becomes the critical coupling mechanism between user devices and edge-hosted AI execution environments. Differentiated connectivity here enables predictable interactions between AI workloads and mobile end devices via the mobile network, under strict isolation and policy control.

As a result of this execution model, AI workloads exhibit network interaction patterns that differ fundamentally from traditional mobile downlink-heavy applications. Interactive and multimodal AI applications increasingly require bi-directional prompt and data exchange, uplink-heavy multimodal streams, burstiness, and periodic transfer of long or multimodal context [2-6][2-7]. These flows are sensitive to short interruptions because they are tied to session states in LLMs not trained on intermittent networks.



For interactive and control-loop AI, bounded latency in the uplink and downlink matters more than average best-effort performance. Requirements include session continuity across mobility, reliable uplink scheduling, and resilience to micro-outages. Note that wearables and edge devices face thermal, battery, and other limits that can amplify variability.

*Networks that compensate through coverage robustness and proper scheduling can directly improve experience without requiring every application or device to over-engineer around worst-case conditions.*

AI also introduces new expectations around where data is processed, what is exposed, and how long it is retained. Hybrid and on-device models can reduce data movement, but they also introduce orchestration complexity and may increase the need for policy-aware network behavior. Compute placement is thus not only an infrastructure decision but also a service requirement that determines whether inference runs on-device, at the network edge, or centrally.

Beyond guaranteed uplink and bounded latency, mobile-native AI workloads introduce additional system-level challenges. For instance, efficient coordination across the device–edge pipeline becomes critical, as poorly aligned processing and data exchange can increase end-to-end latency and constrain scalability. Mobility further complicates execution by requiring temporal synchronization across multiple inference nodes, where handovers risk disrupting ongoing multimodal or control-loop sessions. At the same time, AI applications must operate across highly heterogeneous devices, often relying on non-standardized data formats that reduce efficiency, particularly in XR and IoT ecosystems. Finally, scaling AI workloads at the network edge exposes an energy–capacity trade-off: without optimized hardware support and intelligent orchestration, increased AI demand can translate directly into disproportionate energy consumption.

As multimodal, always-on AI sessions become more prevalent, the limitations of best-effort connectivity become increasingly visible, motivating the shift toward differentiated, verifiable performance levels presented in Section 3.

# 3. Why Differentiated Connectivity Matters

The need for differentiated connectivity emerges from distinct AI workload classes with fundamentally different traffic and latency characteristics. These include:

(i) AI-driven visual analysis and sensing workloads that generate uplink-heavy data streams;
(ii) AI-enhanced prediction workloads, such as XR and teleoperation, are highly sensitive to latency variance.

(iii) AI-enabled multi-agent coordination workloads, including autonomous vehicles and robotics, that require synchronized and reliable uplinks under mobility;
(iv) generative edge AI services, such as large language models and multimodal generation, where session continuity and time-to-first-token dominate user perception; and
(v) distributed learning workloads, including federated learning and split inference, which introduce frequent bi-directional exchanges between devices and the edge.

These workload classes provide a structured basis for defining performance levels and connectivity assurances.

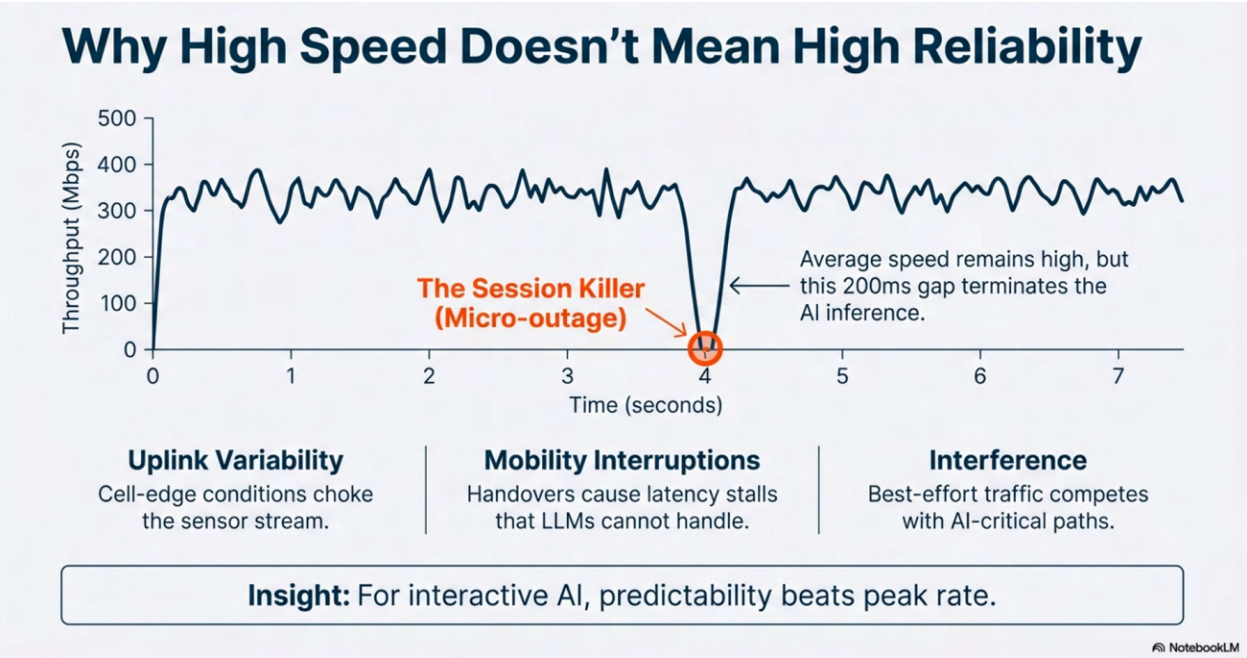
Across these workload classes, short interruptions, latency variance, and uplink contention directly degrade AI performance, making best-effort connectivity insufficient [3-1]. Therefore, delivering differentiated service quality is a strategic priority for operators. The cornerstone of this effort is protecting time-critical applications, including interactive AI applications, particularly in the uplink, where interactive workloads are most vulnerable.

## Challenges and Limitations

Indeed, interactive AI sessions are fragile under uplink data loss, mobility-induced micro-outages, interruptions to inference streams, broken multimodal alignment, or even session termination. Uplink capacity at the cell edge and latency stalls during handovers are particular challenges that require addressing through a modern communications stack.

In multimodal AI pipelines, latency variance can dominate perceived performance [3-2]. That is, a network that delivers “fast on average” but occasionally stalls can feel worse than a consistently moderate network. AI services increasingly involve multiple concurrent flows: sensor upload, control messages, token streams, background OS traffic, and application updates, among others.

Without per-flow prioritization and isolation, best-effort traffic will interfere with AI-critical paths, leading to a notable degradation in user experience.




## Building Blocks for Differentiated Connectivity

Differentiated connectivity begins with classic mechanisms: admission control, QoS enforcement, and slicing for isolation. It becomes commercially attractive when paired with programmable APIs and intents that allow applications and enterprises to request connectivity characteristics in a controlled way [3-3][3-4].

AI-application awareness should not be reserved for congestion moments only [3-5]. Indeed, the applications signal their performance requirements through standardized mechanisms (QoS flows, URSP), enabling the network to deliver appropriate treatment without application-layer inspection.

A persistent challenge is that the network needs a deterministic way to reconcile conflicting intents. A practical policy hierarchy for predictable operations is operator policy first, then enterprise SLA, then application and device intents [3-6].

Local breakout through, e.g., distributed UPF (dUPF)-based offload – complementing Wi-Fi offloading – can reduce end-to-end latency and keep traffic private, improving performance and dependability for verticals. This is of use in edge-sensitive scenarios, such as industrial control.

## Making Assurances Enforceable and Verifiable

As alluded to earlier, to enable a viable uplink boost with (near)deterministic latency commitments, premium applications must be protected via explicit admission control, flow isolation, policy enforcement, etc. To achieve this, differentiated service delivery requires observability.

Differentiated connectivity depends on more than high-level observability. Advanced AI workloads require fine-grained, real-time exposure of network metrics across temporal, spatial, and workload-specific dimensions. XR, teleoperation, and many other real-time systems rely on high-resolution temporal signals and per-cell / per-beam spatial indicators to maintain deterministic performance.

Applications, networks, and the orchestration layer thus need to share a closed loop: monitor performance, detect risk, and act, e.g., by adjusting scheduling, switching slices, moving compute, or changing delivery mode.

*To be monetizable at scale, differentiated connectivity thus needs KPIs that are measurable, enforceable, and ideally auditable.*

Such assurances rely on latency, throughput, dependability, and even location-relevant indicators. Furthermore, security assurance should be considered part of the KPI set since trust underpins premium services [3-7].

AI applications and vertical use cases increasingly demand fine-grained assurance metrics. Normalization and standardization of formats and interfaces, in combination with privacy-preserving mechanisms and strict access controls, should thus be treated as foundational [3-8].

Not all AI-driven content, however, is interactive. Repetitive payloads and forecastable peaks can be served more efficiently by shifting from one-to-one unicast to one-to-many multicast. Forecasting demand peaks enables proactive prefetching and multicast or broadcast offload, thereby freeing up valuable networking resources [3-9].

# 4. Deployment Models and Operator Strategies

This section translates requirements into actionable network capabilities, emphasizing an evolution path: near-term improvements that can be deployed widely, and longer-term architectures that make differentiated services scalable.

## Uplink Optimization

Uplink-heavy AI workloads strain scheduling to ensure proper interference management and robust mobility. Operators ought to prioritize enhancements to uplink scheduling, improved robustness at cell edges, and operational procedures that detect uplink saturation early. The uplink is ultimately limited by coverage and radio conditions, so software features and hardware modernization are practical levers for service differentiation [4-1].

Notably, uplink coverage improvements can be achieved by advanced multi-layer coordination, exploiting superior propagation characteristics to extend reach indoors and at the cell edge. Additional coverage gains are realized through FDD Massive MIMO deployments, high-performance passive antenna systems with improved beam efficiency and ultra-low passive intermodulation, and support for high-power UE classes, where applicable (e.g., not for wearables like smart glasses).

Uplink capacity improvements can be achieved through uplink carrier aggregation across FDD and TDD bands, uplink single-user MIMO, uplink multi-user MIMO with Massive MIMO and interference rejection combining, and layered FDD–TDD Massive MIMO designs that separate uplink reach from downlink throughput. These mechanisms increase usable uplink spectrum efficiency and enable short, high-bitrate AI traffic bursts without overloading individual carriers.

Uplink robustness under mobility can be improved through uplink-quality-aware cell and carrier selection, mobility algorithms that incorporate uplink performance metrics rather than downlink criteria alone, and future uplink–downlink decoupled operation that pairs the most suitable frequency bands for each direction. These approaches reduce uplink degradation for power-limited devices and improve session continuity for mobile AI workloads.

Uplink reliability and latency-tail control can be strengthened through uplink Coordinated Multi-Point Reception (uplink CoMP). This technique reduces worst-case latency events that disproportionately impact interactive and safety-relevant AI applications.

Uplink stability for time-critical AI traffic can also be supported through uplink configuration grants, service-specific QoS flows, and queue-aware rate-adaptation mechanisms such as L4S. Together, these features help bound uplink delay and prevent transient congestion from translating into visible AI session disruption.

## Impact of Latency

AI workloads differ fundamentally in whether their performance is constrained by computation or memory [4-12]. Compute-bound workloads, such as real-time vision inference or predictive control, require fast, deterministic execution, as network latency directly impacts end-to-end response time. Memory-bound workloads, such as large language models and generative foundation models, are dominated by model size and memory bandwidth; however, even in these cases, network-induced variance strongly affects perceived responsiveness. Today, this is prevalent during prompt submission, token streaming, or multimodal exchanges; in the near future, it will also arise during handover across foundation models, where the newly engaged LLM must recompute prior decoded tokens to preserve decoding context, etc.

*Differentiated connectivity is therefore relevant across both categories: reducing absolute latency for compute-bound workloads and minimizing variance and interruption for memory-bound workloads.*

For interactive AI, latency requirements are defined less by average delay and more by predictability. Short interruptions, jitter spikes, or mobility-induced stalls can break session continuity, disrupt token generation, or force re-synchronization between modalities. As a result, bounding latency and controlling its tails become more important than minimizing median latency alone.

Deterministic latency is generally not a single feature but rather a system upgrade that involves minimizing variance, ensuring predictable scheduling, reducing mobility-induced spikes, handling micro-outages, etc. The aforementioned mobility mechanisms and uplink tools can help stabilize sessions, which are vital for interactive AI sessions. For recurring uplink and downlink exchanges – such as token streams, multimodal uploads, or closed-loop control – scalable semi-persistent scheduling can reduce grant uncertainty and help bound latency, provided it is designed to scale without over-reserving resources.

It should be noted that, for large language models, inference latency is often dominated by compute-related factors, such as model size, key-value cache size, and memory bandwidth, among others, which define a lower bound on inter-token latency [4-2]. While mobile networks cannot reduce this intrinsic processing time, they play a decisive role in preventing additional delay from accumulating between successive exchanges. Stable and predictable connectivity is therefore essential to prevent network-induced latency from compounding already compute/memory-bound inference paths.

Furthermore, hybrid UE–edge inference models can introduce frequent round-trip times between the device and the network, which may significantly increase control-plane and uplink signaling. Operators ought to offer generic performance levels suitable for such bursty app traffic; without it, inference pipelines risk becoming chatty, reducing scalability and consuming latency budgets intended for user-facing interaction. These effects must be explicitly considered when assessing the feasibility and scalability of split-inference AI workloads.

Lastly, in industrial settings, local breakouts/dUPF enable low-latency, private traffic paths that reduce end-to-end delay by avoiding unnecessary core network traversals [4-3]. For AI-on-RAN in industrial settings, this matters both for latency and for governance, as it is easier to enforce policy and data locality when traffic stays local.

## SLA-Grade Assurances and Traffic Handling

Operators will likely need a toolkit that orchestrates static assurance (SLA-backed slices) and dynamic assurances (Quality on Demand), while treating service continuity across edge domains as a first-order requirement. For interactive AI sessions, seamless availability of session state and model context during mobility – particularly in industrial settings with edge-cloud deployments – is essential; without it, repeated session resets can undermine user experience and render SLA guarantees economically unsustainable [4-4].

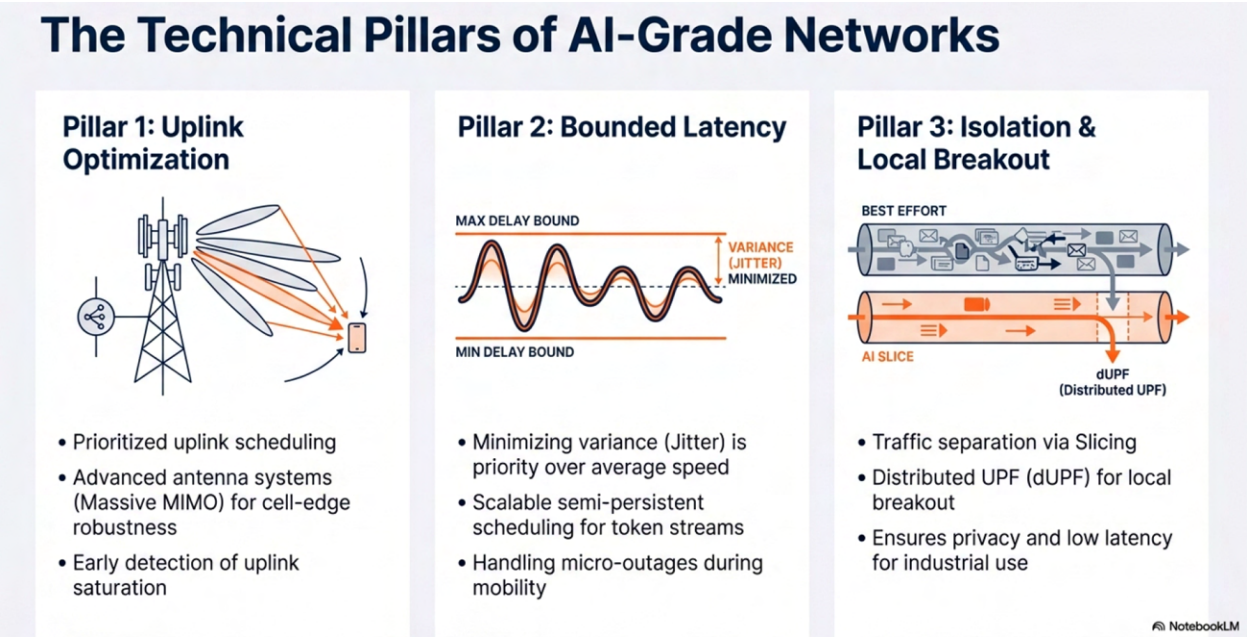
To this end, up-front verification will be important, since a service sold with (near) deterministic expectations will require testing and validation to de-risk deployments and reduce operational surprises. Such a testing framework could include full emulation, digital twin approaches, and/or limited rollouts.

Differentiated connectivity for AI workloads should not rely on overprovisioning or generic priority. Time-critical applications, such as interactive AI, are often composed of multiple concurrent flows, with relative flow importance changing over time. The network must thus evolve from coarse treatment to flow- and intent-aware handling, without compromising operational efficiency, nor privacy or neutrality principles.

In 5G, such QoS differentiation is realized through QoS Flows (QFI) within a PDU session, where QoS Flows are at the *finest* granularity within that session [4-5]. A PDU session is associated with a single S-NSSAI (slice selection) and a DNN, meaning dedicated QoS flows can be added/modified. However, they still inherit the session's slice context and cannot "hop" to another slice without steering to (or establishing) a different PDU session.

The most scalable approach to leveraging such capabilities is explicit signaling of performance requirements rather than deep packet inspection. Applications or trusted middleware should be able to indicate to the network the interactive AI application requirements and/or mark AI-critical flows to express a minimal set of requirements, including bounded latency, minimum uplink rate, resilience during mobility, etc. This enables predictable policy enforcement aligned with commercial expectations, and thereby simplifies enforcement because what was requested and what was delivered can be observed and compared.

The AI RAN Alliance opines that approaches that require payload inspection or inference on user content within the network should be avoided, as they introduce privacy risks and operational complexity.




## End-to-End Orchestration

Differentiated connectivity becomes monetizable when properly orchestrated, translating service intent into coordinated actions across the RAN, core, transport, and edge. A viable edge strategy requires orchestration that jointly manages network conditions and compute placement in near-real time, including proactive service migration, shared model catalogs, and mechanisms that prevent fragmentation across edge domains [4-6]. Orchestration further ensures operational stability and limits the security attack surface to what operators can safely govern.

Here, RAN automation platforms can drive valuable improvements in service assurance for time-critical applications, which are inherently end-to-end. The network segment where a problem is felt may not be the only segment that can resolve it. A robust AI-grade offering therefore benefits from a cross-domain orchestration approach that can coordinate RAN parameters and resource policy, slice admission and isolation controls, user-plane steering including local breakout, exposure of telemetry and predictions to authorized applications, service assurance verification and reporting, among others.

*The practical goal is not full autonomy everywhere, but rather repeatable, bounded end-to-end automation through orchestration that improves outcomes without introducing unpredictable side effects for AI-native applications.*

Orchestration expands in value, but also in risk. A policy decision layer is therefore suggested that enforces authorization and entitlement, implements a policy hierarchy for conflicting intents, provides safety checks and change control, applies rate limits and anti-abuse safeguards, supports safe rollback behavior, among others. This ensures security and operational stability

by preventing oscillatory automation in which closed loops conflict or amplify transient/non-ergodic operational conditions.

To scale beyond bespoke deployments, operators need a standardized way to expose connectivity as a product. A viable path forward is workload blueprints that define a validated mapping between workload types, target KPIs, and an operator-defined set of network and edge actions. For MEC-based AI workloads, these blueprints should also account for mobility across edge zones, including proactive service migration that transfers not only connectivity but also AI session state and model context to avoid disruptions, such as session restarts, in XR or robotic applications. By codifying these behaviors, blueprints reduce friction for developers and thus create a consistent basis for commercialization, while supporting interoperability across vendors and regions [4-7].

## Security by Design

Premium connectivity is ultimately a trust product [4-8]: if customers cannot trust that intentions are enforced safely, the offering will remain niche. Requirements thus ought to include strong encryption by default, strict tenant isolation and segmentation, clear incident detection and reporting processes, documented operational controls for premium offerings, etc.

A practical threat model for differentiated services suitable for interactive AI applications should address three categories: first, traffic risks such as privacy leakage, session hijacking, or flow manipulation. Second, programmability risks include API abuse, entitlement escalation, and denial-of-service attacks against premium mechanisms. Third, control and orchestration risks, such as policy bypass, automation instability, or compromise of control-plane components.

Slicing is an underpinning capability to enable security [4-9]. It should be complemented with continuous assurance for configuration validation, resource protection mechanisms that prevent (malicious) starvation, monitoring for cross-tenant leakage via telemetry, etc. Exposure APIs should enforce least-privilege access with time-bounded credentials, rate limits, anomaly detection, and audit trails that support operations and compliance.

In AI-on-RAN deployments, security extends beyond protecting traffic and APIs. AI models and execution environments must be isolated to prevent unauthorized access to model parameters, inference of training data, or cross-tenant leakage. Differentiated connectivity, therefore, operates alongside sandboxed AI execution environments in which user devices interact only with model inputs and outputs, while model internals remain protected. Connectivity APIs, telemetry exposure, and orchestration mechanisms must respect these isolation boundaries, ensuring that performance optimization does not compromise data privacy, model intellectual property, or regulatory obligations [4-10].

Furthermore, premium uplink boosts and reservation mechanisms must be protected against automated abuse, such as bot-driven exhaustion or malicious over-requesting. When AI models influence network control decisions, the integrity of the underlying models, pipelines, and configurations becomes a security and safety requirement, extending well beyond model optimization.

Lastly, conflicting intents could be a security signal! Operators should thus adopt a policy hierarchy and, ideally, a fine-grained telemetry that enables heightened security while being privacy-preserving, aggregated where appropriate, and limited by strict scope and retention policies.

## Deployment Strategy: A Phased Evolution

The industry should treat differentiated connectivity for interactive AI applications as an evolution. A pragmatic deployment path begins with improving uplink robustness and then observability; it enables measuring where AI sessions fail and identifying underlying uplink failure patterns. The next phase should introduce enforceable service classes and slicing with verification and admission control. Finally, intent-driven APIs and closed-loop orchestration can be added once telemetry and policy maturity can support safe automation. Such a phased approach reduces operational and commercial risks.

An important deployment strategy concerns edge compute, which is often touted as the default answer to AI latency. In reality, edge placement must be justified both per workload and per site, and is most likely to be commercially viable in industrial and enterprise contexts. A realistic strategy distinguishes between general-purpose edge footprints and use case-driven deployments, evaluating each location's power availability, thermal envelope, and backhaul conditions before scaling. Operators should evaluate how often edge placement truly improves experience (especially under mobility), whether the operational lifecycle is supportable, how utilization and monetization work outside peak windows, and how workload migration is handled without disrupting sessions [4-11].

As AI workloads grow, operators should expect higher uplink share and higher variance sensitivity, more performance disputes that require clear verification (particularly with premium customers), new support expectations from AI service providers (including fully automated agentic servicing), and the need for troubleshooting that connects app metrics with network metrics.

# 5. Monetization and Business Models

The emergent nature of GenAI, combined with its distinct network requirements, creates a unique monetization opportunity for operators. Unlike previous generations of mobile services, where connectivity was largely invisible infrastructure, time-critical applications make network quality perceptible to end users and even AI agents – thereby increasing the willingness to pay. This section outlines how operators can capture value across consumer and enterprise segments through connectivity bundles and new network-as-a-service constructs that serve AI workloads and other time-critical applications.

## Consumer Monetization

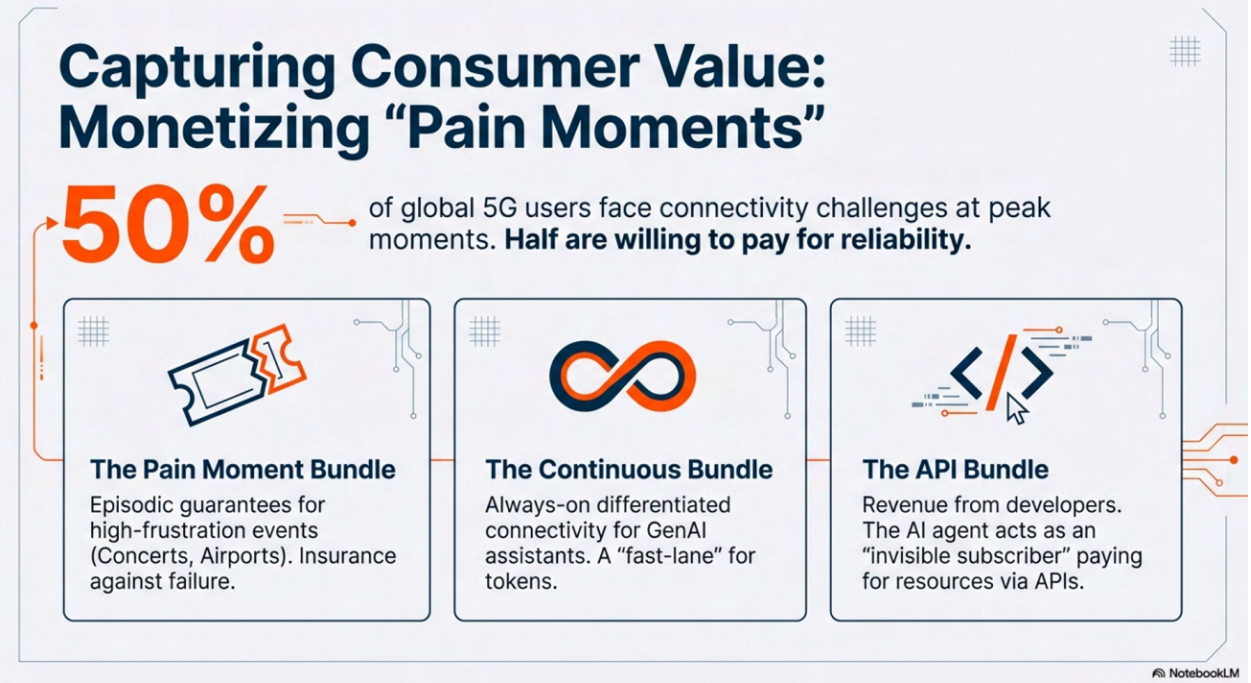
For consumers, the most natural monetization opportunities center on moments of highest frustration and highest willingness to pay. We outline three specific monetization bundles:

**The first bundle** focuses on monetizing critical connectivity moments, when users experience the highest frustration and are most willing to pay. These are situations like crowded stadiums, festivals, airports, or transport hubs where network load spikes unpredictably, and conventional best-effort connectivity breaks down. The action plan is straightforward: guarantee seamless and reliable connectivity exactly during these high-demand windows. Research shows [5-1] that nearly half of global 5G users face connectivity challenges in such moments, and about half of those users are willing to pay extra for reliability. By targeting these "pain moments," operators can generate one to two months of additional ARPU per year simply by offering assured connectivity during peak demand events. The bundle positions reliability as an episodic premium service that aligns with human behavior: people do not want to miss a goal, a live stream, a banking transaction, or a payment failure precisely when it matters most.

**The second bundle** addresses the broader transition toward differentiated connectivity, moving beyond episodic pain-moment guarantees toward a more continuous, tailored network experience. Here, the action plan introduces assurances for uplink speed, real-time responsiveness, location-aware optimization, and application-centered performance (using one of two slicing approaches outlined earlier). These capabilities resonate strongly with both AI-native applications and advanced consumer use cases, from real-time AR overlays to interactive GenAI sessions. Market studies indicate that differentiated connectivity can lift customer satisfaction by nearly 20% and brand equity by almost 50% [5-1]. That is proof that customers respond when operators make quality visible. This bundle positions the operator as a provider of predictable, personalized network performance that adapts to the user or application's behavior rather than offering a uniform best-effort pipe.

**The third bundle** looks ahead to revenue unlocked through API-driven networks. Instead of selling connectivity only to end users, operators expose programmable interfaces that let third-

party developers tap into network capabilities such as on-demand QoS, enhanced uplink, mobility insights, or location verification. These APIs become foundational for the emerging class of AI agents and agentic AI systems that increasingly act as autonomous clients on the network [5-2]. Such AI entities need stable, reliable connectivity guarantees to function properly, thus effectively becoming "invisible subscribers" whose traffic must be managed with predictable quality. By enabling this through APIs, operators create a new revenue stream that scales with developer adoption rather than subscriber count. At the same time, this approach fosters innovation by giving app builders direct levers to optimize how their AI applications behave over mobile networks, turning the telecom network into an active platform rather than a passive transport layer.




## Enterprise Monetization

Enterprise customers require a different value proposition, one built around contractual assurances and integrated with their operational workflows.

**Per-device, per-application, and even per-flow SLAs** allow enterprises to specify which workloads require premium treatment. A logistics company might purchase assured connectivity only for its autonomous fleet management AI, while a manufacturer might require deterministic latency exclusively for robotic control systems. This granularity prevents enterprises from overpaying for blanket coverage while ensuring critical applications receive protected treatment.

**API consumption pricing** introduces a usage-based dimension. Enterprises pay for what they consume: slice reservations, QoD (Quality on Demand) activations, telemetry subscriptions,

intent-based requests, etc. This model aligns operator revenue with enterprise value creation and scales naturally as AI adoption grows. Workload-aligned pricing goes further by packaging connectivity characteristics around specific use cases with pricing that reflects the operational value (!) at stake rather than raw bandwidth consumed.

**Workload-aligned and value-based monetization** in the enterprise is most effective when connectivity is priced based on the business value of the workload it supports [5-3], rather than raw bandwidth or uniform service tiers. Different enterprise applications carry vastly different operational risks, performance sensitivities, and revenue impact. By aligning monetization to specific workloads – such as autonomous fleet management, real-time robotics, or AI inference pipelines – operators can charge in proportion to the outcomes being protected. This approach ensures enterprises pay only for the connectivity assurances that truly matter, while operators capture value commensurate with the criticality of the services they deliver.

## Aggregator-Centric Value Chain for Differentiated Connectivity

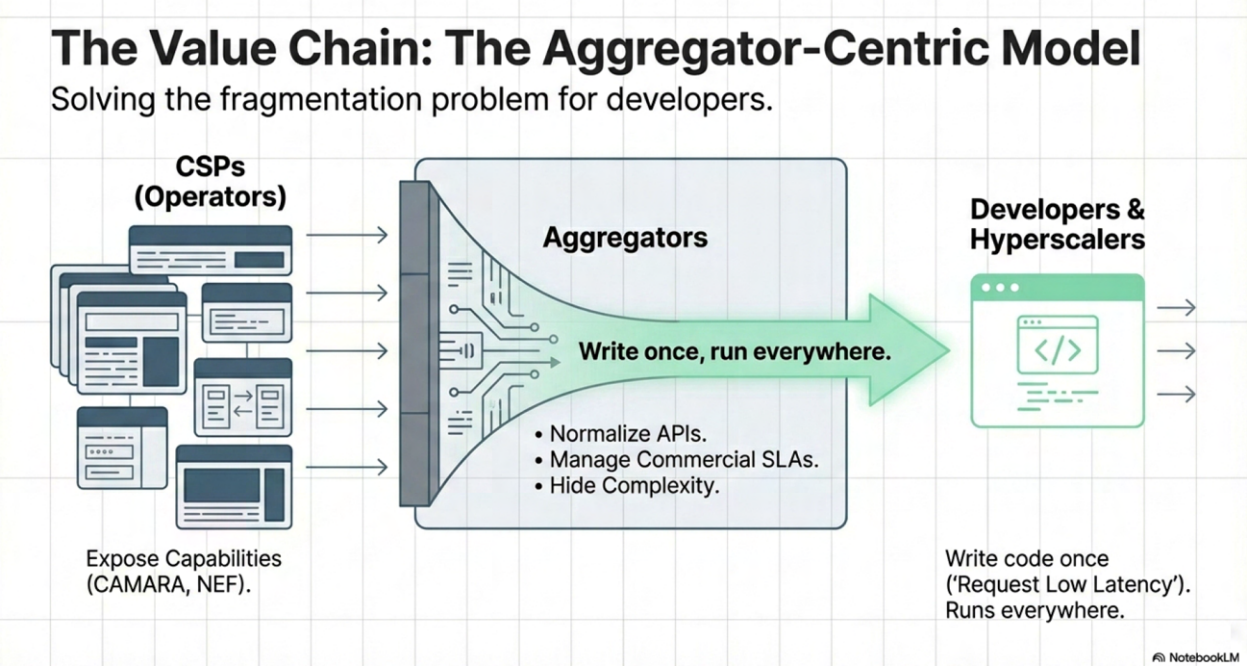
AI-grade connectivity will not scale if every application provider must integrate with each operator separately. An aggregator-centric value chain addresses this by inserting a specialized intermediary between communications service providers (CSPs) and the developer ecosystem [5-4].

In this model, an aggregator connects to multiple CSPs' differentiated offerings (for example, slices, Quality on Demand, and exposure APIs), normalizes their capabilities and KPIs into globally consumable APIs, and manages commercial relationships both upstream with CSPs and downstream with developer platforms and application providers. This hides CSP-specific implementation details (such as S-NSSAI values, 5QI mappings, or URSP templates) and mitigates cross-market fragmentation for application and AI developers.

The resulting value chain coexists with today's direct mobile broadband model. CSPs deploy and expose AI-grade services via network APIs. Aggregators purchase these capabilities on a pay-per-use basis (for example, per unit time, per data volume, or per QoD activation) and map them to a normalized capability model. Developer platforms, hyperscalers, and AI platforms consume these normalized APIs and embed them into SDKs and toolchains, often bundled with cloud compute and AI models [5-5]. Application and AI service providers then use these SDKs to request AI-grade connectivity for specific sessions or flows and embed the connectivity cost into their own products, such as consumer subscriptions, in-app passes, or enterprise SLAs. End users and enterprises ultimately experience AI applications that are consistently connected across operators and regions.

Commercially, this model enables usage-aligned pricing [5-6][5-7][5-8]: CSPs monetize differentiated connectivity per use; aggregators add value through aggregation, normalization,

and SLA management; and application providers monetize indirectly via application-centric business models rather than selling raw bandwidth. For AI-native services, the aggregator layer also provides global consistency (a single semantic, such as “low-latency uplink for multimodal interaction,” can work across markets), allows CSPs to evolve internal implementations without breaking applications, and aligns with emerging AI ecosystems where agents can programmatically request, adapt to, and verify AI-grade connectivity.




## Monetization Enablers

Rather than selling raw network primitives, operators can package differentiated connectivity into bundles that reduce complexity for buyers and create repeatable commercial products applicable to consumers and enterprises alike.

**Per-application slicing and composable bundles.** Network slicing enables operators to offer “slice-as-a-feature” bundles where specific applications receive dedicated treatment. A GenAI assistant application might be bundled with a slice configuration optimized for bounded-latency uplink and consistent token delivery. An enterprise AR training platform might receive a slice tuned for high uplink throughput and mobility resilience. The key is composability, where enterprises and developers can mix and match slice characteristics to fit their workload profiles without requiring bespoke network engineering for each deployment.

**Quality on Demand during pain points.** Not every situation requires persistent premium connectivity. QoD mechanisms allow real-time interventions for detected pain points (e.g., handover micro-outage risk, congestion onset, coverage degradation at the cell edge, etc.) without requiring always-on slice reservations. This creates a more efficient commercial model:

users or applications pay for protection when it matters, and the network allocates premium resources dynamically rather than statically. For AI workloads, QoD can be triggered automatically when session telemetry indicates elevated risk, creating a closed loop between application needs and network response.

**API bundles and the CAMARA ecosystem.** The GSMA-led CAMARA initiative [5-9] and similar efforts are creating standardized APIs that expose network capabilities to developers and enterprises. Operators can bundle these APIs into commercial packages: a "developer starter" tier might include basic QoD and location verification; an "enterprise" tier might add slice booking, predictive connectivity insights, and enhanced telemetry. API bundles transform the telecom network from passive transport into an active platform, creating revenue streams that scale with developer adoption rather than subscriber count. For AI applications specifically, these APIs serve as essential infrastructure, enabling agents and agentic systems to request and verify connectivity guarantees programmatically.

## GenAI Connectivity Tiers as Network-as-a-Service

A connectivity-as-a-service model purpose-built for GenAI becomes viable when operators can offer standardized and commercially repeatable products.

**Operator-hyperscaler bundles.** Partnerships between operators and hyperscalers can create integrated offerings that combine connectivity, compute, and model access into unified products. An enterprise purchasing AI capabilities from a hyperscaler could simultaneously provision the network connectivity required to support those capabilities, and all that with SLAs that span the full stack. This reduces integration friction, clarifies accountability, and allows operators to participate in AI value chains rather than being relegated to commodity transport. The commercial arrangement might involve revenue sharing, wholesale connectivity agreements, or even a joint go-to-market for vertical solutions.

**Vertical-specific performance-level templates.** Different verticals have distinct AI workload profiles, and operators can create pre-configured performance level templates with slice configurations tailored to specific AI application workload patterns. An XR slice might prioritize symmetric bandwidth and ultra-low latency for immersive experiences. An industrial slice might emphasize deterministic latency bounds and high reliability for closed-loop control. An automotive slice might focus on mobility resilience and the integration of edge computing for V2X applications. Multiple slices may serve different applications within a vertical. These vertical slices become repeatable products that can be sold across multiple customers within a segment, reducing the need for bespoke configuration while ensuring workload-appropriate treatment.

**Premium analytics and telemetry for AI performance tuning.** AI developers and enterprises increasingly need visibility into how network conditions affect their applications. Operators can

monetize this need by offering premium telemetry services, including real-time performance dashboards, historical analytics, predictive degradation alerts, detailed session-level metrics, and more. This telemetry enables AI application providers to tune their systems by adjusting encoding, buffering, and fallback behaviors based on network conditions, etc., whilst giving operators a new revenue stream from data they already collect. The key is packaging telemetry in actionable, privacy-preserving formats that developers can integrate into their optimization loops.

**Developer-friendly consumption models.** To foster ecosystem adoption, operators should offer consumption models that lower barriers to entry. Freemium tiers allow developers to experiment with network APIs at no cost, graduating to paid tiers as their applications scale. Usage-based pricing aligns costs with value, where developers pay for QoD activations, slice-hours, or API calls consumed rather than committing to fixed capacity. *Such an approach mirrors successful patterns in cloud computing and encourages experimentation that drives innovation, ultimately expanding the market for premium connectivity services.*

**Assured connectivity as a premium tier.** Differentiated connectivity with performance levels suitable for interactive AI-native applications is an "assured" tier that combines all the elements discussed: enforceable KPIs, verifiable performance, security guarantees, auditability, etc. This tier is not merely "faster" or "higher priority" – it is designed for an AI era that is measurable, auditable, and trustworthy. Customers purchasing assured connectivity receive contractual commitments backed by real-time monitoring, incident reporting, and SLA enforcement. For AI workloads where reliability directly affects business outcomes, this tier commands premium pricing because the cost of failure and the value of assurance are high.

The above suggestions will invoke an ecosystem shift from "selling pipes" to "selling outcomes". AI applications make network quality perceivable at the application layer, creating accountability that previous generations of services lacked. *Operators that can deliver on that accountability will capture disproportionate value as AI becomes an important driver of mobile network demand.*

# 6. Business and Regulatory Implications

Differentiated connectivity sits at the intersection of performance, innovation, and policy. As operators introduce premium tiers and programmable APIs, they must navigate regulatory frameworks designed for an earlier era of telecommunications; all while anticipating how those frameworks will evolve. Poorly designed offerings risk regulatory pushback or even public mistrust; well-designed offerings can demonstrate that differentiation and fairness coexist, potentially shaping regulatory thinking in constructive directions [6-1].

## Compliance with Neutrality While Enabling Specialized Services

Net neutrality frameworks in most jurisdictions distinguish between general internet access, which must treat traffic without discrimination, and specialized services that operate alongside best-effort internet. Performance-based connectivity should be positioned squarely in the latter category [6-2].

This is defensible when offerings meet established criteria: they serve applications with requirements that best-effort delivery cannot meet, they operate with dedicated resources that do not cannibalize general-purpose capacity, and they do not degrade internet access quality for other users—premium slices for closed-loop robotic control or (quasi) deterministic connectivity for autonomous vehicles.

For consumers, the tiered products should ensure the base tier remains genuinely usable; i.e., differentiation should enhance the experience for those who pay, not artificially degrade it for those who do not.

Operators should proactively document how their offerings satisfy specialized service criteria and engage regulators before commercial launch rather than after.

## Privacy by Design for AI Data and API Exposure

Differentiated connectivity with performance levels suitable for interactive AI-native applications involves data flows with significant privacy implications, including session telemetry, location information, application metadata, and fine-grained performance metrics. Here, privacy considerations apply not only to user and application data but also to fine-grained network telemetry, which can reveal sensitive behavioral or contextual patterns if exposed without proper safeguards. APIs that surface this data to third parties amplify the risk!

Operators must therefore embed privacy-by-design principles throughout, with attention to GDPR and analogous frameworks [6-3]. Data minimization means collecting only what is necessary and retaining it only as long as required. Purpose limitation means data collected for network optimization should not be repurposed for advertising or profiling without explicit consent. Technical safeguards include aggregation thresholds, anonymization techniques, differential privacy where applicable, and strong access controls, among others.

Exposure APIs require particular care! Design should enforce privacy by default, using aggregated metrics rather than individual session data; implement rate limits to prevent bulk extraction; require explicit justification for fine-grained access; and maintain audit trails of all queries, etc.

Privacy is a foundation of trust: users and enterprises who believe their data is protected will engage more with premium offerings. Operators who lead on privacy and security will gain a competitive advantage [6-4].

## Safety and Resilience for AI-in-the-Loop Operations

When AI participates in network control affecting safety-critical applications, the stakes extend to public safety; examples include autonomous vehicles, remote medical procedures, and industrial robotics. An optimization that inadvertently degrades connectivity for such safety-critical applications could have consequences far beyond commercial disappointment.

Operators should thus adopt safety frameworks commensurate with risk: AI systems should enhance human decision-making rather than replace it for high-stakes scenarios; automated actions should be bounded and reversible. Fail-safe defaults should ensure graceful degradation; critical applications should have fallback paths that work even if AI is not functioning correctly; and systems should be stress-tested under adversarial conditions.

For applications where failures could cause physical harm, additional safeguards may be warranted, e.g., dedicated capacity that AI cannot reallocate, or human-in-the-loop approval for changes affecting safety-critical slices. Regulators in automotive, aviation, healthcare, and industrial safety are increasingly attentive to connectivity dependencies. *Operators with robust safety frameworks will be better positioned to serve these sectors.*

## Unified Industry Voice on Policy Direction

The regulatory landscape for differentiated performance-based connectivity is still forming. Decisions in the next several years will shape what operators can offer and on what terms. A fragmented industry voice risks fragmented outcomes: inconsistent national rules, overly restrictive requirements, or too permissive approaches that invite backlash.

The industry should prioritize consensus on key questions: What constitutes a specialized service requiring differentiated connectivity for interactive AI applications? What transparency requirements suit AI-driven control? How should privacy frameworks apply to telemetry exposure? What safety standards govern AI-in-the-loop operations? Operators, vendors, hyperscalers, and AI providers all benefit from coherent answers.

Proactive regulatory engagement is more effective than reactive defense. Operators who participate in drafting processes will have more influence than those who wait and object. The goal is to ensure frameworks that reflect technical/commercial realities whilst enabling beneficial innovation and addressing genuine risks. When industry and regulators share a common understanding, both can move forward with confidence, ultimately benefiting the end-user.

# 7. Cross-Industry Collaboration and Standardization

## Alignment Across Industry Organizations and AI Protocols

3GPP – network requirements for AI applications. 3GPP TS 22.261 defines three AI operation modes—split inference, task-specific model downloads, and federated learning [7-1]. To support these modes, the 5G system must expose resource utilization and predicted changes in bitrate, latency, and reliability, notify authorized applications of upcoming QoS changes, and allow aggregated QoS adjustments for groups of UEs.

O-RAN – RAIE enabling third-party network connectivity requests. O-RAN's RAN Information Exposure (RAIE) framework [7-2] provides third-party applications — including AI services — with access to live RAN status, enabling them to request connectivity tailored to their needs. By exposing metrics and control points through the Service Management and Orchestration (SMO) platform and the R1 interface, RAIE allows these applications to optimize performance or request network slices and configuration updates based on current conditions [7-2][7-3].

ETSI – AI phone use case and slice management. ETSI's ENI report envisions "AI phones" whose intelligence is provisioned on demand [7-4]. Users express intents in natural language; the device's agent validates multi-modal inputs and sends them to an intent manager. An AI orchestrator decomposes the request and issues a derived intent to the 5G core domain manager—for example, to create a slice with latency < 50 ms and bandwidth > 100 Mbps—and the domain manager configures resources accordingly.

## Common Taxonomy, KPIs, and Workload Profiles

Industry should align on app-agnostic performance-level specifications and map AI workload types (interactive generative models, real-time inference, federated learning, and robotics) to these standardized levels based on their throughput, latency, and reliability needs. O-RAN's cross-domain AI report stresses that AI enablers (data collection, algorithms/,models, and computing resources) must be coordinated across domains with secure data access, distributed compute, and governance [7-5]. Standardizing KPIs would let applications articulate their requirements; CAMARA's slice booking API already accepts throughput and latency parameters when reserving slices [7-6].

Standardized interfaces for edge compute (covering telemetry, service migration, model catalogs, privacy-preserving execution, among others) are required to prevent bespoke integrations and ensure that edge assets remain interoperable and economically viable. Additionally, standardization efforts should extend to ML state transfer, context migration,

mobility-aware orchestration interfaces, and related capabilities to ensure that edge nodes can reliably clone or transfer AI sessions. Without such frameworks, inter-edge mobility becomes a source of fragmentation, thereby increasing operational risk and diminishing consistent performance.

## Normalized Interfaces for Telemetry Exposure

Fine-grained telemetry is essential for time-critical service quality. 3GPP Network Data Analytics Function (NWDAF) collects network data and exposes analytics via the Network Exposure Function (NEF), enabling applications to subscribe to predicted mobility or slice SLA assurance insights. O-RAN's RAIE offers secure APIs to expose RAN data. CAMARA's Connectivity Insights provides current throughput, latency, and jitter [7-7], while Predictive Connectivity Data estimates future conditions using radio coverage and antenna catalogues [7-8]. Together, these efforts point toward a common telemetry schema with unified metrics, time-stamped measurements, and privacy-preserving aggregation, which are essential to preventing fragmentation, reducing integration overhead, and ensuring consistent interpretation by AI systems across heterogeneous RAN and core implementations.

## Blueprint of Best Practices

Best practices should combine technical and commercial considerations. Networks must ensure deterministic latency, reliable uplink performance, and isolation for time-critical applications, including interactive AI applications, through slicing, coverage optimization, and reliable handovers. Integrating RAIE, NWDAF, and NEF helps maintain time-critical workload performance. O-RAN urges adoption of its specifications as national standards to avoid fragmentation and ensure global alignment, and 5G Americas emphasizes collaboration and zero-trust architectures [7-9]. On the commercial side, operators can monetize differentiated connectivity via subscription tiers or usage-based pricing as discussed in earlier sections.

## Open SDKs and Reference Implementations

An ecosystem of open tools supports programmable networks with differentiated connectivity. CAMARA publishes open-source reference implementations for its APIs [7-6], fostering experimentation without vendor lock-in. Model Context Protocol (MCP) [7-10] provides language-agnostic SDKs and, as an open standard, encourages the integration of telecom-grade network APIs into AI agents. Ericsson's Intelligent Automation Platform offers a non-real-time RAN Intelligent Controller and SDK for developing rApps across multi-vendor networks [7-3].

In summary, the initiatives mentioned in this section demonstrate how industry bodies are aligning around a common framework for connectivity with predictable performance levels: 3GPP defines the network vocabulary and operational modes, ETSI illustrates agents

dynamically requesting slices with specific latency and bandwidth, and O-RAN's RAIE framework provides open interfaces for third-party connectivity requests. Open APIs and protocols, such as those defined by CAMARA, fill the remaining gaps, creating the opportunity for applications – including AI – to specify connectivity requirements.

# 8. Concluding Remarks

AI is transforming connectivity from invisible infrastructure into an experiential (consumers) and operational (enterprise) dependency. For the next wave of mobile-native AI services, the "good network" will be defined less by peak downlink speed and more by **predictable uplink, bounded latency, resilient mobility, secure programmability, and verifiable outcomes**. This is, in fact, an immediate requirement for AI applications being deployed today.

Differentiated connectivity represents both a technical necessity and a commercial inflection point. Technically, it enables AI workloads that are otherwise fragile under real mobile conditions by protecting interactive sessions and multimodal streams. Commercially, it creates a pathway for operators to monetize assured performance through repeatable products such as consumer tiers, enterprise SLAs, API bundles, telemetry services, hyperscaler partnerships, etc.

The industry stands at an important inflection point: AI workloads are growing rapidly, and the applications being built today will shape user expectations and commercial relationships for years to come. Operators who move early to deliver differentiated performance levels will establish positions in value chains that extend far beyond transport. Those who wait risk being relegated to commodity pipe as hyperscalers, device makers, and AI platforms capture the value that differentiated connectivity enables.

## Recommended Focus Areas

Working Group 3 recommends that the ecosystem prioritize five areas with urgency and coordination:

**Prioritize uplink and mobility resilience as the technical foundation:** We identified uplink micro-outages, mobility-induced interruptions, and uplink variance as the primary technical barriers to reliable AI connectivity. The AI RAN Alliance should thus be an industry ambassador for focus on uplink enhancements, such as improved uplink scheduling, robustness at cell edges, lightweight mobility mechanisms that stabilize sessions during handovers, uplink interference techniques, spectrum needs, etc. These foundational improvements are prerequisites for more sophisticated intent-driven orchestration and can be deployed incrementally without waiting for end-to-end architectural transformation.

**Establish app-agnostic performance levels with verifiable KPIs:** The commercial viability of differentiated connectivity depends on measurability: "premium" is only sustainable if service objectives can be enforced and audited. The industry must converge on a set of standardized performance levels that are not workload-specific configurations but rather app-agnostic network specifications defined by bounded latency, uplink throughput floors, and reliability targets. Critically, these performance levels must include normative measurement methods that enable SLA enforcement, incident reporting, etc. AI workload types can then be mapped to appropriate performance levels based on their requirements, creating a common vocabulary for the telco and adjacent industries.

**Align telemetry and API exposure with existing industry frameworks:** The building blocks for programmable connectivity already exist across 3GPP (NEF, NWDAF), O-RAN (RAIE, R1 interface), and GSMA/CAMARA. Rather than creating parallel interfaces, the AI RAN Alliance should map AI-on-RAN requirements to these existing frameworks and identify specific gaps requiring extension, such as uplink-specific telemetry and mobility prediction exposure through e.g., localization. Proposed extensions should be contributed through established standards bodies to ensure interoperability and avoid fragmentation. Open-source SDKs and reference implementations should build upon established industry interfaces, enabling developers to integrate with network capabilities without bespoke engineering or vendor lock-in.

**Publish phased deployment guidance with security & privacy embedded throughout:** Differentiated connectivity for AI should be treated as an evolution, not a single transformation. A pragmatic deployment path begins with improving uplink observability and detecting where AI sessions fail; it then introduces enforceable service classes with slicing, verification, and admission control; finally, it adds intent-driven APIs and closed-loop orchestration once telemetry and policy maturity can support safe automation. The AI RAN Alliance should produce deployment guidance that sequences these capabilities with clear milestones. Security requirements — including tenant isolation, API access controls, automation bounds, and audit trails — must be specified at each phase rather than treated as a separate workstream. Premium connectivity is a trusted product; security is foundational, not an afterthought.

**Coordinate commercial and regulatory positioning across the ecosystem:** Fragmented commercial approaches will undermine the adoption of differentiated AI connectivity and weaken the telco industry's position. OTT providers. The AI RAN Alliance should facilitate alignment on commercial constructs – particularly the aggregator model that enables cross-operator consistency and allows AI application providers to consume network capabilities through normalized APIs rather than bespoke integrations with each operator. Simultaneously, the industry requires a unified position on regulatory compliance: differentiated connectivity for AI applications should be positioned as a specialized service that meets established net

neutrality criteria, with proactive documentation demonstrating that premium offerings do not degrade general internet access. A fragmented industry voice risks fragmented regulatory outcomes.

## Final Message

The building blocks for performance-based connectivity exist today across 3GPP, O-RAN, GSMA/CAMARA, TM Forum, and all commercial deployments worldwide. What is missing is alignment: common definitions, interoperable interfaces, coordinated commercialization.

The cost of fragmentation is high: inconsistent implementations or fragmented developer ecosystems cannot scale. The reward for alignment, however, is substantial: a new category of monetizable services, strengthened operator positioning in AI value chains, and – importantly – AI applications that work reliably for users who increasingly depend on them.

The industry does not need to wait for 6G. The moment for enabling differentiated connectivity with performance levels suitable for interactive AI-native applications is now. The operators and ecosystem players who act decisively will shape the next decade of mobile services and, with it, capture disproportionate value as AI becomes the defining workload of the new era.

# References

[2-1] Vodafone, proprietary measurements in live networks in 2025.

[2-2] Ericsson Mobility Report, "GenAI's impact on network data traffic today", June 2025; GenAI data traffic today – Ericsson Mobility Report.

[2-3] Antonio Montieri, Alfredo Nascita, Antonio Pescapè, "From Prompts to Packets: A View from the Network on ChatGPT, Copilot, and Gemini," submitted 13 Oct 2025; https://arxiv.org/abs/2510.11269.

[2-4] Paper by Northeastern on the measurements.

[2-5] Khan MJ, Khan MA, Malik S, Kulkarni P, Alkaabi N, Ullah O, El-Sayed H, Ahmed A, Turaev S. Advancing C-V2X for Level 5 Autonomous Driving from the Perspective of 3GPP Standards. Sensors (Basel). 2023 Feb 17;23(4):2261. doi: 10.3390/s23042261. PMID: 36850858; PMCID: PMC9967342.

[2-6] Ericsson Mobility Report, "AI, cloud and mobile set to drive significant growth in uplink traffic," https://www.ericsson.com/en/reports-and-papers/mobility-report/articles/ai-cloud-mobile-drive-uplink-growth.

[2-7] Nokia, "From voice to video to AI-shaped traffic: Why network architecture must evolve at software speed," https://www.nokia.com/blog/from-voice-to-video-to-ai-shaped-traffic-why-network-architecture-must-evolve-at-software-speed

[3-1] GSMA Intelligence – Telco AI: State of the Market, Q3 2025 (overview and strategic insights) https://www.gsmaintelligence.com/research/telco-ai-state-of-the-market-q3-2025

[3-2] Amey Agrawal, Anmol Agarwal, Nitin Kedia, Jayashree Mohan, Souvik Kundu, Nipun Kwatra, Ramachandran Ramjee, Alexey Tumanov. Etalon: Holistic Performance Evaluation Framework for LLM Inference Systems. 2024. https://arxiv.org/abs/2407.07000

[3-3] Ericsson. "The network platform is redefining telecom – here's how." 2023. https://www.ericsson.com/en/blog/2023/9/network-platform-redefine-telecom

[3-4] GSMA. "What is Open Gateway?" https://www.gsma.com/solutions-and-impact/gsma-open-gateway/what-is-gsma-open-gateway

[3-5] Microsoft. "https://www.fierce-network.com/sponsored/using-modern-apis-deliver-network-aware-applications" https://www.fierce-network.com/sponsored/using-modern-apis-deliver-network-aware-applications

[3-6] Ericsson. “Why managing conflict between rApps is vital for RAN automation at scale,” https://www.ericsson.com/en/blog/2024/6/the-benefits-of-rapps-for-ran-automation

[3-7] Keith Dyer. “Why Service Assurance matters for 5G Standalone,” https://the-mobile-network.com/2021/06/why-service-assurance-matters-for-5g-standalone

[3-8] ITU-T. “Recommendation ITU-T Y.3176: Machine learning marketplace integration in future networks including IMT-2020.” 2020. https://www.itu.int/rec/dologin_pub.asp?id=T-REC-Y.3176-202009-I%21%21PDF-E&lang=e&type=items

[3-9] The 5G Broadcast Collective. “Introduction to 5G Broadcast.” 2024. https://gsacom.com/paper/introduction-to-5g-broadcast

[4-1] F. Kronestedt, T. Chen, A. Kaur, A. Furuskär. Enhancing 5G uplink performance to enable differentiated services. Ericsson Technology Review, Nov 12 2025. https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/achieving-networks-with-high-performing-uplinks

[4-2] Michael Davies, Neal Crago, Karthikeyan Sankaralingam, Christos Kozyrakis. “Efficient LLM Inference: Bandwidth, Compute, Synchronization, and Capacity Are All You Need.” arXiv 2025. https://arxiv.org/abs/2507.14397v1

[4-3] NVIDIA. Accelerated and Distributed UPF for the Era of Agentic AI and 6G. 2024. https://developer.nvidia.com/blog/accelerated-and-distributed-upf-for-the-era-of-agentic-ai-and-6g

[4-4] Ericsson. “Enabling time-critical applications over 5G with rate adaptation.” 2023/2024 white paper. https://www.ericsson.com/en/reports-and-papers/white-papers/enabling-time-critical-applications-over-5g-with-rate-adaptation

[4-5] 3GPP. “TS 23.501 V18.7.0 — System Architecture for the 5G System (5GS); Stage 2.” 2025.

[4-6] Ericsson. “Driving 5G monetization through intent-based network operations.” 2025. https://www.ericsson.com/en/reports-and-papers/white-papers/driving-5g-monetization-through-intent-based-network-operations

[4-7] TM Forum. “AI-Native Blueprint Project,” https://www.tmforum.org/ai-native-blueprint-project.

[4-8] GSMA. “Securing the 5G Era.” 2024. https://www.gsma.com/solutions-and-impact/technologies/security/securing-the-5g-era

[4-9] Stan Wong, Bin Han, Hans D. Schotten. “5G Network Slice Isolation.” Network 2022, 2(1):153–167. https://doi.org/10.3390/network2010011

[4-10] Vonage. Network APIs | Secure Network-Powered Solutions. 2025. https://www.vonage.com/network-apis

[4-11] Vodafone Business. Dedicated Multi-Access Edge Computing for business. Vodafone UK. 2025. https://www.vodafone.co.uk/business/cloud-solutions/multi-access-edge-computing/dedicated-edge-computing

[4-12] Nvidia. Mastering LLM Techniques: Inference Optimization. 17 November 2023. https://developer.nvidia.com/blog/mastering-llm-techniques-inference-optimization/

[5-1] Ericsson ConsumerLab. Winning in the market with differentiated connectivity offerings. 2026. https://www.ericsson.com/en/reports-and-papers/consumerlab/reports/5g-next-wave

[5-2] TM Forum & GSMA. “Network API Monetization,” https://www.tmforum.org/oda/solutions/network-api-monetization

[5-3] GSMA Intelligence. The State of 5G 2024. 2024. https://gsmaintelligence.com/research/research-file-download?file=210224-The-State-of-5G-2024.pdf&id=79791087

[5-4] RCR Wireless. “Aduna: The cornerstone of an API-enabled telco future.” 17 Jan 2025. https://www.rcrwireless.com/20250117/analyst-angle/aduna-api-enabled-telco-future-analyst-angle

[5-5] GSMA. What is GSMA Open Gateway? 2025. https://www.gsma.com/solutions-and-impact/gsma-open-gateway/what-is-gsma-open-gateway

[5-6] Vodafone Developer Marketplace – Quality-On-Demand API. https://developer.vodafone.com/api-catalogue/quality-demand/overview

[5-7] Omdia (Informa Tech). “Deutsche Telekom and Ericsson form alliance to monetize 5G with global API platform.” 2025. https://omdia.tech.informa.com/om119640/deutsche-telekom-and-ericsson-form-alliance-to-monetize-5g-with-global-api-platform

[5-8] Verizon. Wireless Internet for the AI Era: Network Slicing Enables SLA-Backed Service for Verizon Business. 2024. https://www.verizon.com/about/news/wireless-internet-ai-era-network-slicing-enables-sla-backed-service-verizon-business

[5-9] The Linux Foundation. CAMARA, the Global Telco API Alliance, Delivers First Major Release with Innovative APIs for Seamless Access to Network Functions. 2024. https://www.linuxfoundation.org/press/camara-the-global-telco-api-alliance-delivers-first-major-release-with-innovative-apis-for-seamless-access-to-network-functions

[6-1] GSMA. Regulatory Environment – Public Policy. 2025. https://www.gsma.com/solutions-and-impact/connectivity-for-good/public-policy/regulatory-environment

[6-2] Ofcom. Network neutrality (open internet) – guidance and compliance. Updated guidance on interpretation and application of UK net neutrality rules (Open Internet Access Regulation). https://www.ofcom.org.uk/internet-based-services/network-neutrality

[6-3] GDPR-Info. Privacy by Design & Default (Article 25 GDPR). 2024. https://gdpr-info.eu/issues/privacy-by-design/

[6-4] PwC. TMT Customer Transformation POV. 2024. https://www.pwc.com/gx/en/industries/tmt/emea-customer-transformation-tmt-pov.html

[7-1] 3GPP TS 22.261 §6.40: AI/ML model transfer and split inference [link]

[7-2] O-RAN RAIE framework: enabling third-party applications to access live RAN status and optimize their applications or request connectivity adjustments [link]

[7-3] O-RAN R1 interface: rApps can subscribe to network performance data and request configuration updates via the SMO/Non-RT RIC [link]

[7-4] ETSI ENI 055 [link]

[7-5] O-RAN cross-domain AI report: data, algorithm/model, and computing elements [link]

[7-6] CAMARA Network Slice Booking API: slice reservation parameters [link]

[7-7] CAMARA Connectivity Insights API: real-time throughput, latency, and jitter [link]

[7-8] CAMARA Predictive Connectivity Data API: future connectivity forecasts [link]

[7-9] O-RAN & 5G Americas: collaboration and zero-trust advocacy [link]

[7-10] MCP specification: tools, resources, and prompts over JSON-RPC [link]

# Appendix A: AI Workload Profiles

This appendix exemplifies repeatable workload profiles that map AI application types to connectivity requirements and standardized performance levels.

## Profile 1: Intermittent Interactive Generative AI Assistant

**Typical scenario:** Conversational AI agent, contextual search, short compositional tasks, and similar tasks.

**Critical metrics:** During the session, bounded latency matters more than the lowest *average* latency. Session continuity is critical; micro-outages that require re-sends or resets are highly disruptive. Throughput is moderate, with uplink bursts for prompts and downlink response streams.

**Network actions:** Priority handling for the agent flow via per-app marking and per-flow treatment; optional QoD boost during detected pain points such as handover risk or congestion onset; real-time telemetry exposure, including latency, etc, as well as risk flags.

## Profile 2: Always-On Multimodal Assistant and Agentic AI

**Typical scenario:** AI-enabled glasses, agentic companion devices, multimodal mobile always-on or on-demand assistant with video, image sequences, audio uploads, etc.

**Critical metrics:** Uplink robustness to minimize stalls and grant uncertainty; consistent uplink scheduling to avoid session destabilization; fast recovery from mobility events. Even modest frame drops or short uplink micro-outages can disproportionately degrade experience.

**Network actions:** Similar to Profile 1, plus: uplink scheduling enhancements for continuity plus coverage optimization; admission control and isolation if offered as a premium tier; optionally local breakout or dUPF for latency control when edge inference applies.

## Profile 3: Physical AI / Closed-Loop Control

**Typical scenario:** Perception stream flows uplink, drives inference, which drives control action flowing downlink. Includes autonomous vehicles, droids, mobile robots, drones, etc. Often safety-relevant.

**Critical metrics:** Bounded latency for control-loop stability; high reliability since loss or interruptions can cause unsafe behavior; strong security posture. Mobility is precisely where wireless variability and handover behavior are most challenging.

**Network actions:** Dedicated QoS or slice with strict admission control; deterministic latency posture to protect control flow; strict policy hierarchy for intent plus verified entitlements; logs for auditability.

## Profile 4: Remote Inspection and Field Support

**Typical scenario:** Technician streams uplink video, AI provides deep, industry-specific analysis, and a remote expert joins for collaboration.

**Critical metrics:** Sustained uplink throughput and stability; resilience across coverage variations; insight into whether the network or the device is the bottleneck.

**Network actions:** Uplink-focused QoS tier; optionally slice reservation for known sites and times; real-time telemetry exposure to application for adaptive encoding and behavior.

## Profile 5: Batch Inference and Broadcast Distribution

**Typical scenario:** Predictable content updates, model or prompt pack distribution, AI-generated content pushed to many users.

**Critical metrics:** Efficiency through steering to multicast or broadcast where eligible; timing to align delivery windows with off-peak periods or ahead of events.

**Network actions:** Forecast-based delivery mode selection using prefetch and multicast where applicable; analytics-driven triggers with usage-based settlement hooks.

# Appendix B: Glossary

This glossary defines acronyms and specialized terminology used throughout the white paper.

| Term | Definition |
|---|---|
| 3GPP | 3rd Generation Partnership Project. The standards body responsible for 5G and mobile network specifications. |
| 5G SA | 5G Standalone. A 5G deployment architecture with its own 5G core network, not reliant on 4G infrastructure. |
| 5QI | 5G QoS Identifier. A scalar value referencing specific QoS characteristics (latency, reliability, priority) in 3GPP TS 23.501. |
| API | Application Programming Interface. A standardized interface enabling software components to communicate. |
| AR | Augmented Reality. Technology overlays digital information onto the real-world environment. |
| ARPU | Average Revenue Per User. A key telecommunications metric for revenue performance. |
| CAMARA | A GSMA-led open-source project creating harmonized network APIs for exposing operator capabilities to developers. |
| DAPS | Dual Active Protocol Stack. A 3GPP Rel-16 handover mechanism enabling near-zero interruption time. |
| DDoS | Distributed Denial of Service. A cyberattack attempting to overwhelm systems with traffic. |
| dUPF | Distributed User Plane Function. A UPF deployed closer to the edge to reduce latency and enable local breakout. |
| eMBB | Enhanced Mobile Broadband. A 5G use case category focused on high data rates. |
| GBR | Guaranteed Bit Rate. A QoS resource type where dedicated network resources are allocated for a flow. |
| GDPR | General Data Protection Regulation. European Union regulation on data protection and privacy. |
| GSMA | GSM Association. An industry organization representing mobile network operators worldwide. |
| Intent | A request describing desired service outcomes (e.g., bounded jitter, reserved resources) rather than low-level network configuration. |
| IoT | Internet of Things. Network of connected devices exchanging data. |

| Term | Definition |
|---|---|
| KPI | Key Performance Indicator. A measurable value demonstrating how effectively objectives are achieved. |
| LLM | Large Language Model. AI models are trained on large text datasets for natural language tasks. |
| Local breakout | Routing user traffic to local networks or the internet at the edge rather than through the central core. |
| MCP | Model Context Protocol. An open protocol for AI agent interoperability, contributed to the Agentic AI Foundation. |
| MEC | Multi-access Edge Computing. Computing capability at the network edge to reduce latency. |
| Micro-outage | A brief interruption in connectivity (typically 100ms to several seconds) that can disrupt interactive sessions. |
| MIMO | Multiple Input Multiple Output. Antenna technology using multiple transmitters and receivers. |
| mMTC | Massive Machine Type Communications. A 5G use case for connecting large numbers of IoT devices. |
| NEF | Network Exposure Function. A 3GPP 5G core function exposing network capabilities to external applications. |
| Non-RT RIC | Non-Real Time RAN Intelligent Controller. An O-RAN component for RAN optimization on longer timescales. |
| NWDAF | Network Data Analytics Function. A 3GPP 5G core function providing network analytics and predictions. |
| O-RAN | Open RAN: an industry initiative promoting open, interoperable RAN interfaces and architectures. |
| PDB | Packet Delay Budget. The upper bound on acceptable packet delay between the UE and the UPF is specified in the 3GPP specifications. |
| PER | Packet Error Rate. The rate of unsuccessfully delivered packets is used to define reliability targets. |
| Performance Level | An app-agnostic specification of network connection characteristics (latency, throughput, reliability) suitable for a class of applications. |
| QoD | Quality on Demand. Dynamic, contextual quality boosts or protections, typically time-bounded and entitlement-controlled. |
| QoS | Quality of Service. Mechanisms for differentiating traffic treatment based on requirements. |

| Term | Definition |
|---|---|
| RAIE | RAN Information Exposure. An O-RAN framework exposing RAN data to third-party applications via open interfaces. |
| rApp | RAN Application. Applications running on the Non-RT RIC in the O-RAN architecture. |
| SDK | Software Development Kit. Tools and libraries enabling developers to build applications. |
| SLA | Service Level Agreement. A contract defining expected service performance and remedies. |
| SMO | Service Management and Orchestration. The O-RAN framework for managing and orchestrating RAN functions. |
| S-NSSAI | Single Network Slice Selection Assistance Information. Identifiers for network slices in 3GPP. |
| TM Forum | TeleManagement Forum. An industry association focused on digital business transformation. |
| UE | User Equipment. The mobile device or terminal in 3GPP terminology. |
| UPF | User Plane Function. The 3GPP 5G core function handling user data forwarding. |
| URLLC | Ultra-Reliable Low-Latency Communication. A 5G use case category targeting < 1ms latency and 99.999% reliability. |
| URSP | UE Route Selection Policy. Rules enabling UEs to select appropriate network slices for applications. |
| V2X | Vehicle-to-Everything. Communication between vehicles and other entities (infrastructure, pedestrians, network). |
| XR | Extended Reality. An umbrella term covering AR, VR, and mixed reality technologies. |