OOKLA research™

Beyond Download Speed:

Benchmarking 5G Mobile Networks Against AI Workloads

Executive summary

AI workloads are rewriting the performance requirements that define mobile network quality. AI traffic is not one thing: text chat, conversational voice, multimodal and AR vision, generated video, and agentic background activity each load the network differently. The metrics that have historically benchmarked mobile networks, peak download speed and coverage, say little about whether a network can carry that mix. This report puts forward a framework for assessing 5G mobile network readiness against the performance requirements of AI workloads, across five dimensions: upload capacity, multi-server latency, loaded latency, cloud infrastructure latency, and jitter.

Most markets clear the baseline for today's text-based AI, but that readiness hides underlying gaps. Upload capacity is the widest, with operators allocating only a small share of throughput to the uplink and fewer than half meeting the bar for the more demanding use cases. The picture shifts again once networks are measured under load rather than in ideal conditions, and once the path beyond the network edge to the cloud, where inference runs, is brought into view. The report draws on Speedtest Intelligence data across 22 markets and 86 operators, covering the major developed and emerging 5G economies in North America, Europe, Asia Pacific, the Middle East, and Latin America. The 22 markets are listed on the following page.

Region	Markets Covered
North America	United States, Canada
Europe	United Kingdom, Germany, France, Ireland, Finland, Norway, Sweden, Italy, Spain
Asia Pacific	South Korea, Japan, India, Indonesia, Australia, Singapore, Malaysia, Thailand, Philippines
Middle East	United Arab Emirates
Latin America	Brazil

Key takeaways:

- **AI readiness ranks markets differently from headline 5G network performance.**
  The metrics that actually decide AI performance, upload capacity, latency under load, and the path to the cloud, produce a different leaderboard from download speed, and the gap widens as adoption shifts toward heavier use cases like conversational voice and multimodal AI.

- **Existing 5G infrastructure generally supports text-centric AI, yet falls short of the performance benchmarks required for emerging modalities.**
  While latency targets for text LLMs (under 50 ms) are achieved in 18 of 22 markets and conversational voice targets (under 40 ms) in 13, no market currently reaches the sub-10 ms requirement for AR.

- **Upload allocation is the widest and most persistent gap.**
  The typical operator devotes only around 10% of throughput to the uplink, fewer than half meet the 20 Mbps target for AR and multimodal AI, and upload share has declined or held flat in 12 of 22 markets since 2023.

- **Latency holds up under normal conditions but degrades sharply under load, and unevenly.**
  Degradation ratios run from 3.7x to 11.4x across markets, and the gap between operators inside a single market is as wide as the gap between markets.

- **The path from the network to the cloud is becoming its own constraint.**
  Cloud latency and worst-case jitter vary as much as the operator network does, and within one market the choice of cloud provider can swing latency by nearly 100 ms, enough to decide whether real-time AI is viable.

Contents

From background capability to network-defining workload

The transition of artificial intelligence from a supplementary background process to a primary foreground activity is occurring more rapidly than many infrastructure planning timelines had predicted. A growing share of consumers engage AI applications as a daily tool rather than an optional feature. Enterprises are integrating AI into operational workflows that depend on reliable connectivity. And the devices running AI will not be limited to the smartphone — they will include wearables, smart glasses, and connected industrial devices, each generating traffic that existing networks are not dimensioned to handle.

Data from Omdia, referenced in the **GSMA Mobile AI white paper**, indicates that global AI-driven network traffic is set to expand at a compound annual growth rate of 73% between 2025 and 2033. By 2031, AI traffic is expected to reach a critical inflection point where its total volume overtakes that of conventional traffic. The device categories generating the most demanding traffic profiles are still in the early stages of mass adoption. This is not a capacity problem. Operators have kept adding capacity. AI traffic is simply shaped differently from what networks were built for, heavier on upload, always on, and bursty, rather than download-led and session-based.

AI interaction is deepening, not just growing

The adoption of mobile AI has outpaced every other modern technology. According to **Sensor Tower's State of Mobile 2026 report**, total downloads reached 3.8 billion by 2025, a 30-fold increase in just three years. While these download figures are notable, session volume serves as a more vital metric for engagement. The report highlights that in 2025, sessions for generative AI apps surpassed one trillion, exceeding the growth rate of downloads. Network demand is intensifying not only due to a rapidly expanding user base but also because existing users are engaging in more frequent and longer sessions.

ChatGPT offers the clearest benchmark for the scale of that engagement. The app reached 800 million weekly active users by October 2025, processing more than 2.5 billion messages daily, and, in July 2025, became the fastest application in history to reach one billion global downloads across iOS and Google Play, according to Sensor Tower. On mobile specifically, ChatGPT alone reached 546 million active monthly users worldwide as of April 2025 and accounted for 60% of all measured AI app traffic on the operator network analyzed in **Ericsson's June 2025 Mobility Report**. The category is no longer a single-platform story: by the end of 2024, 16 generative AI apps had each exceeded $10 million in in-app purchase revenue, and 25 had surpassed 10 million downloads.

The trajectory is also shifting from session-based interaction to persistent agentic AI. **OpenClaw**, an open-source AI agent that went viral in early 2026, runs

autonomously — maintaining connections, calling APIs, and executing tasks without human input. Unlike session-based AI, agentic systems keep the network engaged around the clock rather than in short bursts during active use. Not all agentic workloads are equally demanding: messaging-based agents tolerate moderate delay, while coding and workflow automation agents require tighter responsiveness. But translating any of this into a network planning figure requires a more precise unit of measurement than downloads or sessions alone.

The more operationally relevant unit is the token — the basic segment of text, image, or audio that an AI model processes per interaction — because token volume is a direct proxy for the inference activity that generates data flows across mobile networks. **Google reported** processing 480 trillion monthly tokens as of April 2025, a figure it announced at its I/O developer conference in May. **Microsoft's Foundry APIs** processed over 500 trillion tokens in fiscal year 2025, up more than seven times year-over-year. What makes it more demanding for network infrastructure is that the data generated by each interaction is also growing.

Average **prompt length grew nearly fourfold** between early 2024 and late 2025 — from approximately 1,500 tokens to over 6,000 — as users shifted from single-sentence queries to submitting documents, images, and extended multi-part context. Completion lengths almost tripled over the same period as reasoning-capable models generate longer, more structured outputs. Each session now places a materially higher data demand on the network than it did twelve months ago. Volume and intensity are compounding simultaneously — and that combination is spreading into markets where networks are least equipped to absorb it.

Geographically, Asia expanded its lead as the top market for generative AI app downloads, **growing 80%** between H2 2024 and H1 2025 — well ahead of Europe at 51% and North America at 39%. North America's share of global generative AI app downloads fell from approximately 20% at the category's launch to 11% in H1 2025, reflecting global diffusion rather than domestic slowdown. AI-driven network demand is no longer concentrated in a small number of high-income, well-infrastructured markets. It is spreading across network environments with varying profiles, 5G penetration, and infrastructure maturity.

This geographic diffusion is being accelerated by hardware trends that are putting AI-capable devices into a much broader range of hands and, increasingly, onto a much broader range of form factors. A growing share of smartphones now ship with dedicated neural processing units (NPUs) that run AI models directly on the device rather than routing every request to the cloud. **Counterpoint Research** forecasts that generative AI-capable devices will account for more than one in three global shipments in 2025, with the mid-range tier already accounting for 38% of that total — a threefold increase from 2024. The extent to which these devices process AI workloads locally versus routing them to the cloud will shape how much additional network demand they generate, a question explored further in the metrics section. Beyond smartphones, AI-native wearables and smart glasses introduce a structurally more important variable. Omdia places global AI glasses shipments at **5.1 million units in 2025**, a 158% year-on-year increase, with the market projected to exceed 10 million units in 2026 and reach 35 million by 2030 at a compound annual growth rate of 47%.

From pilot to production, enterprise AI adds a different kind of load

Enterprise AI follows a similar adoption trajectory to consumer, but with a different network demand profile — lower in aggregate volume today, but more intensive in its infrastructure requirements. Consumer AI traffic skews toward mobile devices and consists of high volumes of short, user-initiated sessions. Enterprise AI runs predominantly on fixed networks — office Wi-Fi, ethernet, and dedicated connections — though a growing share of use cases depend on mobile connectivity. Field service operations, manufacturing computer vision, remote patient monitoring, and logistics fleet management all depend on devices continuously transmitting sensor data, imagery, and operational telemetry upstream to cloud or edge inference systems. Public 5G networks were dimensioned for the opposite traffic profile — downlink-dominant, session-based, human-paced. The structural mismatch between what these mobile enterprise AI workloads require and what public networks were designed to deliver is the central readiness challenge this report examines. Cisco projects that agentic AI could push enterprise network traffic to roughly 9x its current level by 2035, against about 2.5x without it.

Private 5G deployments reached 6,500 globally at the end of 2025 according to **Berg Insight**, concentrated in manufacturing, media, and logistics — sectors where connectivity requirements tend toward the uplink-intensive, low-latency profile this report examines. The willingness of these sectors to invest in dedicated infrastructure signals that the requirements are real, and public networks serving similar use cases will face comparable demands.

AI traffic rewrites the rules of network design

AI workloads run across both fixed and mobile networks, but the challenges they pose differ between them. Fixed broadband operates in more controlled environments with higher average capacity and more stable latency characteristics. Mobile networks carry AI in the field, in motion, and on devices with no fixed connection. Although the underlying backhaul is frequently fiber-based, the radio interface remains the point where performance is least certain and infrastructure limitations are most apparent. This analysis focuses on the impact of AI traffic on the Radio Access Network (RAN), specifically within 5G mobile environments.

Most deployed 5G networks are configured around assumptions that users consume far more data than they produce, that sessions have natural start and end points, and that traffic is predictable enough to plan around. Those assumptions held for all users to date. AI challenges each of them — shifting the uplink/downlink balance, keeping connections active without human input, and introducing traffic patterns that do not follow traditional planning models. The infrastructural gaps resulting from increased AI traffic remain a subject of ongoing debate.

It is also important to acknowledge that not all AI traffic is the same. Each use case demands a distinct set of network requirements, exposes different infrastructure weaknesses, and requires different solutions. Understanding those differences is the starting point for any credible assessment of where current 5G networks stand in meeting the needs of emerging AI use cases.

Use Case 1:

Text-based large language model (LLM) interactions — the dominant AI traffic type on mobile networks today — generate a two-phase traffic pattern. The user's prompt travels upstream to a cloud inference server, then the response streams back token by token in irregular bursts determined by server-side compute load. That burst pattern conflicts directly with RAN scheduling algorithms tuned for smooth, sustained video flows. Individual users tolerate moderate delay — the interaction resembles messaging — but under peak load, many concurrent sessions competing for scheduling resources create queuing delays across the cell. A growing share of LLM inference runs locally on device neural processing units (NPUs), which reduces some of that cloud-bound traffic — but when a task exceeds local compute capacity or requires data not held on the device, the device reverts abruptly to cloud inference, generating an unscheduled uplink burst the network did not anticipate. The hybrid model does not reduce AI traffic so much as concentrate what remains into higher-intensity, less predictable bursts.

Use Case 2:

Conversational voice AI is far less forgiving of latency variation than text-based interaction. The entire processing chain — voice capture, transcription, LLM inference, and audio delivery — must stay within tight cumulative delay bounds. Jitter, the variation in delay between consecutive packets, determines whether a voice AI session feels natural or stilted. Networks with similar median cloud latency can exhibit very different jitter at the 90th percentile (worst case), large enough to produce audible gaps in conversational flow. Buffering can compensate to a point, but it adds latency, which in a conversational exchange makes the interaction feel unnatural. Voice AI is one of the clearest cases for 5G Standalone network slicing — the ability to isolate latency-sensitive traffic in a dedicated virtual channel, shielded from the congestion that degrades performance on shared infrastructure.

Use Case 3:

Multimodal and AR vision are the AI use cases that most directly challenge the downlink-dominated design of current mobile networks. AI glasses, AR headsets, and industrial computer vision systems all share the same basic architecture: continuous streams of camera, audio, and sensor data transmitted upstream for cloud-based inference, with contextual outputs returned in near-real time. A typical smartphone and network runs at a 90/10 downlink-to-uplink traffic split. AR glasses with cloud-offloaded spatial compute push that ratio toward **40% uplink or higher**, depending on application complexity. Ericsson's analysis of XR network requirements places the minimum viable connectivity threshold for these use cases at bitrates of tens of Mbps with latency below 20 milliseconds — a combination that requires 5G Standalone with compute placed close to the RAN, a configuration that remains the exception across most deployed networks as of early 2026.

Use Case 4:

AI-generated video is overwhelmingly downlink-driven and has the largest potential for traffic volume growth among the five modalities. OpenAI's Sora 2 and Google's Veo 3 are moving toward mass consumer deployment through 2026, and Ericsson identifies consumer AI video as the primary driver of future AI-related mobile traffic growth ahead of text-based applications. At 4K resolution, each stream requires 20 to 40 Mbps of downlink throughput — comparable to high-definition streaming — but generated on demand rather than served from Content Delivery Networks (CDNs). If adoption follows a trajectory similar to the video streaming transition of the 2010s, the capacity implications at scale could be comparable.

Use Case 5:

Agentic AI is different in kind from every AI use case above. Where the other four typically serve a defined task — whether triggered by a user or a sensor — agentic systems autonomously execute multi-step workflows, call external tools, coordinate with other agents, and maintain persistent connections without continuous human input. The traffic this generates is not a burst responding to a prompt — it is a steady background flow of API calls and data transfers that keeps connections active in-between human interactions. **Cisco** describes this as a shift from the spiky, session-based traffic of current generative AI to persistent, machine-paced activity that keeps connection levels consistently elevated. The scale of agentic traffic impact at the RAN level remains an active area of analysis, but the direction of the shift is undisputed.

These five use cases pressure the same layers at once — the downlink scheduler, uplink capacity, latency headroom, and the signaling layer that manages how devices connect and disconnect. Always-on devices, persistent AI sessions, and agentic background processes all prevent the network from reclaiming radio resources during natural pause periods — a load that conventional capacity planning, built around session-based traffic, was not designed to account for.

Moving beyond download speed as the primary benchmark

Mobile networks have been benchmarked, marketed, and invested in around a single primary metric for the better part of two decades: peak download speed. For the traffic mix that defined the 4G and early 5G era, dominated by streaming video, social media, and web browsing, it was a reasonable proxy for user experience. For AI workloads, it tells only part of the story.

When a user interacts with an AI application on a mobile device, the experience depends on two distinct systems working in sequence. The AI model that processes a voice command, generates a response, or analyzes an image runs on cloud infrastructure operated by technology companies, outside the mobile network. What the network controls is the speed, capacity, and consistency with which data travels between the user's smartphone or device and that cloud infrastructure. Operators cannot control inference processing time, model performance, or data flows inside data centers. They can control the performance of their own network and the quality of the path between their network edge and the cloud. It is on those variables that this report focuses.

Peak download speed measures the best a network can deliver to a single device under ideal conditions. AI workloads are shaped by a different question: how consistently does the network perform under real-world conditions, against the performance requirements of these workloads, across all users sharing the same cell at the moments they need it most? Answering that question requires a broader set of metrics. The definitions below establish the measurement framework used throughout this report.

The following parameters are utilized as directional research indicators for the scope of this study, rather than finalized proprietary benchmarking standards.

Multi-server latency	Measures a network's baseline responsiveness by averaging latency across multiple server connections during each Speedtest. It reflects the round-trip delay a user should expect under normal conditions when their device communicates with a variety of internet locations.
Loaded latency	Measures round-trip delay while the connection's bandwidth is fully saturated on both downlink and uplink during a Speedtest. It captures how much latency degrades when the network is under stress, revealing the gap between baseline performance and real-world worst-case conditions.
Degradation ratio	Divides 10th percentile loaded latency (representing the best-case connections under full utilization) by multi-server latency to quantify how much a network's responsiveness deteriorates from typical conditions to peak demand. Multi-server latency serves as the baseline because it captures the network's responsiveness under normal operating conditions. A ratio of 10x means the 10th percentile loaded latency is ten times the multi-server latency.
Cloud infrastructure latency	Measures round-trip latency from the device to cloud service providers such as AWS, Azure, Google Cloud (GCP), and Oracle (OCI) during Consumer Quality of Experience (CQoE) tests. It captures the full path between the user and cloud-hosted applications, including the operator's network, interconnection points, and the cloud provider's edge infrastructure.
Cloud infrastructure jitter	Measures the average difference between consecutive cloud infrastructure latency measurements. It indicates how stable or variable the connection to cloud services is over time. Low jitter means consistent performance. High jitter means unpredictable delay that can degrade real-time applications even when median latency is adequate.

Percentile conventions: The 10th percentile (P10) represents the best-performing measurements (lowest latency or jitter).
The 90th percentile (P90) represents the worst-performing measurements (highest latency or jitter).
The median (P50) represents the typical experience.

Not every metric applies equally to every AI modality. Text-based LLM interaction is constrained primarily by upload speed and latency. Conversational voice AI demands the tightest latency consistency. Multimodal AI and AR vision stress both upload capacity and latency simultaneously. The table below maps the primary, secondary, and minimal constraints for each modality.

The Relevance of Each Metric Varies by AI Modality, and that Variation Matters

Modality	Upload Speed	Latency	Cloud Latency & Jitter
Text LLM	▲	★	★
Conversational Voice AI	▲	★	★
Multimodal AI / AR Vision	★	★	★
AI-Generated Video	⬣	▲	★
Agentic AI	▲	★	★

★ **Primary constraint** – likely to directly determine whether the AI modality functions at acceptable quality.
▲ **Secondary constraint** – may materially affect experience but is not the primary determining variable.
⬣ **Minimal** – not expected to be a material constraint under typical conditions.

Download speed understates how much upload matters

Mobile networks were built around a historically accurate assumption: users consume far more data than they produce. The 90/10 downlink-to-uplink ratio that characterizes most deployed 5G networks reflects that reality. AI workloads challenge it. When a user submits a prompt to a large language model, the full context, potentially running to thousands of words, images, or documents, travels upstream before any response returns. According to the **Ericsson Mobility Report**, text-based LLM traffic already runs at approximately a 29/71 uplink-to-downlink split. Conversational voice AI and agentic AI are each estimated closer to 50/50, and AR and multimodal vision at around 40% uplink. Cisco reports the same shift in its own data: roughly 9% of AI inference flows are upstream-heavy, against about 0.5% for standard web traffic. Looking further out, the **GSMA** has modeled upload's share of total mobile traffic reaching between 25% and 35% by 2040 in its medium and high AI growth scenarios.

Upload speeds have risen globally, but the share of network capacity that operators allocate to the uplink has not kept pace, and in some markets has moved in the opposite direction, as **Ookla's recent analysis** of uplink capacity trends across major global operators shows. The upload-to-download ratio is a useful indicator of network design philosophy, reflecting how an operator balances capacity between the two directions, but it is not treated as a threshold criterion in this report. Absolute upload speed, the capacity available for upstream data transfer regardless of how much total throughput the network delivers, is the more direct measure of whether a network can support a given AI workload. A network with a low upload share but high total capacity may still deliver adequate absolute upload speeds. The analysis that follows benchmarks operators against absolute upload speed targets.

There are real uncertainties in how uplink demand will evolve. How much AI processing will ultimately run on the device itself, reducing the data that needs to cross the network? Apple's **Private Cloud Compute** model offers an early signal of the hybrid approach likely to emerge, keeping basic and privacy-sensitive AI workloads on-device while pushing more compute-intensive workloads to the cloud. Wealthier consumers will be able to afford devices with more on-device compute, potentially reducing their network dependence. Less affluent consumers on devices with less processing power will push more workloads into the cloud, over networks that may themselves be less capable, widening the gap. Advances in compression may ease the pressure from always-on vision applications. These questions remain open, and actual uplink demand from AI will depend heavily on how they are resolved.

Where latency becomes the binding constraint

Multi-server latency thresholds in this report are informed by available standards and industry research. For conversational voice AI, **ITU-T** sets the maximum acceptable one-way delay at 150 milliseconds, implying the network's round-trip contribution must remain well below that to leave headroom for processing. For text-based LLM interaction, **MLCommons MLPerf v5.0** establishes 450 milliseconds as the system-level ceiling for time-to-first-token. The network's share needs to stay small enough that inference has room to operate within that budget. For AR and multimodal vision, **3GPP** sets the motion-to-photon latency upper limit at 20 milliseconds to maintain stable visual rendering.

Loaded latency serves a different analytical purpose. It does not measure what a user experiences during a typical AI interaction. It measures what the connection delivers when fully utilized, a stress test that reveals how much performance degrades under peak demand. Loaded latency is not compared against application thresholds in this report because it captures a more extreme condition than typical AI sessions produce. Instead, the degradation ratio quantifies how much headroom each network loses under pressure. A network with low multi-server latency but a high degradation ratio may meet AI requirements under normal conditions but fall short when many users compete for the same cell simultaneously.

The path beyond the network edge

Cloud infrastructure latency and jitter extend the measurement beyond the operator's network to the full path between the user and the cloud platforms where AI inference runs. The cloud providers included in this analysis, AWS, Azure, Google Cloud (GCP), and Oracle Cloud Infrastructure (OCI), together host the large majority of commercial AI workloads globally, though other providers including regional and sovereign cloud platforms also play a growing role. Operators influence this segment of the path through their peering agreements, core network routing, and the proximity of their points of interconnection to cloud provider edge locations. They do not fully control it. No widely adopted standards yet define acceptable thresholds for cloud infrastructure latency or jitter in an AI context. Both are therefore treated in this report as observational rather than benchmarked, reported alongside network performance data to build a more complete picture of how the full path from device to AI model performs in practice. For real-time AI applications, consistency matters as much as speed. A connection with low median cloud latency but high jitter, meaning large differences between consecutive latency measurements, can be as disruptive as a consistently slow one.

Setting the thresholds

The table below sets out indicative minimum and target thresholds for upload speed and multi-server latency across the five AI modalities. Multi-server latency values reflect the network's contribution to delay under typical operating conditions, not the total time it takes for an AI application to respond. Cloud infrastructure latency and jitter are excluded from the threshold table because no agreed benchmarks exist for these metrics in an AI context. They appear as observational data in the analysis that follows.

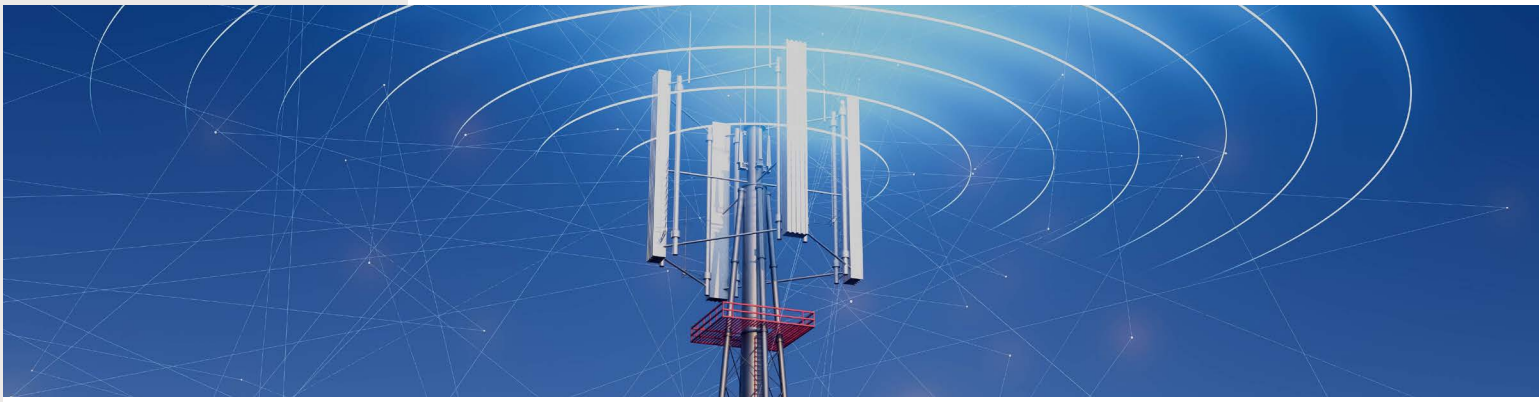
Indicative AI Workload Network Performance Thresholds

Modality	Median Upload Speed	Multi-server Latency
Text LLM	Min: 3 Mbps –> Target: 8 Mbps	Min: < 100 ms –> Target: < 50 ms
Conversational Voice AI	Min: 1 Mbps –> Target: 3 Mbps	Min: < 80 ms –> Target: < 40 ms
Multimodal AI / AR Vision	Min: 10 Mbps –> Target: 20 Mbps	Min: < 30 ms –> Target: < 10 ms
AI-Generated Video	Min: 5 Mbps –> Target: 10 Mbps	Min: < 150 ms –> Target: < 80 ms
Agentic AI	Min: 5 Mbps –> Target: 10 Mbps	Min: < 100 ms –> Target: < 50 ms

Sources: **ITU-T G.114** · **3GPP TR 26.928** · **3GPP TS 22.261** · **MLCommons MLPerf v5.0** · **Ericsson Mobility Report** · **GSMA GigaUplink White Paper** · **Cisco**

Where no direct standard exists, values should be read as directional. Minimum thresholds represent the level below which users are likely to notice degraded quality. Target thresholds represent where consistently good performance becomes achievable.

These are per-device requirements under typical conditions. They do not account for the compounding pressure when many AI-active users share the same cell simultaneously. The loaded latency analysis in the following sections captures that dimension.



Benchmarking 5G markets against the demands of AI workloads

To assess where 5G networks stand today against the requirements of AI workloads, this section draws on Speedtest Intelligence® data across 22 markets — covering the major developed and emerging 5G economies in North America, Europe, Asia Pacific, the Middle East, and Latin America — and the leading operators in each market.

The upload gap is wide and in most markets widening

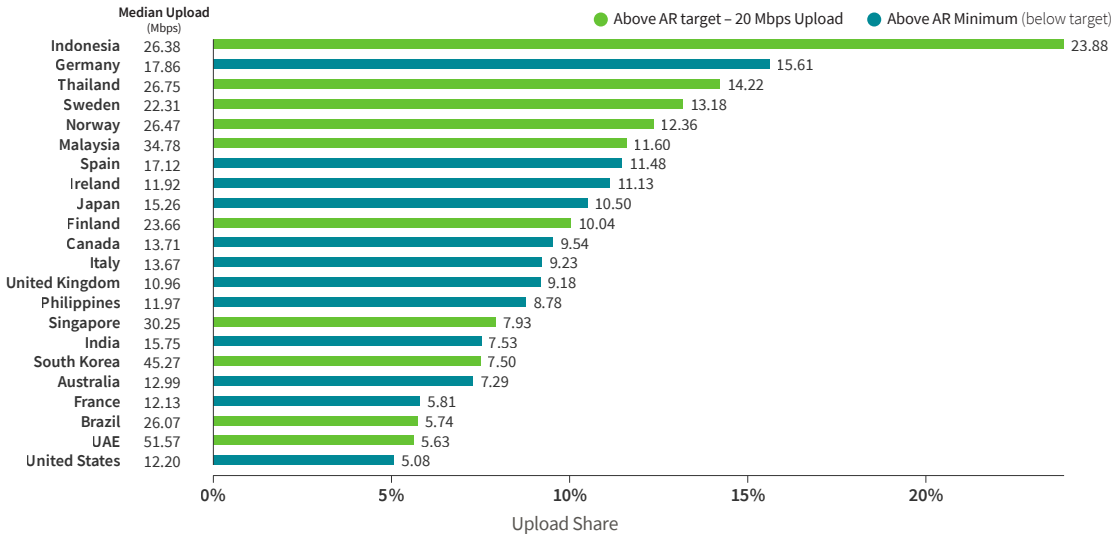
The most consistent finding across all 22 markets is that the proportion of network capacity allocated to the uplink is small and, in many markets, shrinking.

Upload share is calculated as median upload speed divided by the sum of median upload and download speeds, expressed as a percentage. It captures how much of the total throughput a network allocates to the uplink rather than the downlink.

AI workloads generate a materially higher share of upstream traffic — an estimated 29/71 uplink-to-downlink split for text LLM, closer to 50/50 for conversational voice AI and agentic AI. Traffic ratios do not translate directly into speed requirements, and operators with high total throughput can deliver adequate upload speeds despite low upload share. But fewer than half of operators meet the 20 Mbps target for AR and multimodal vision.

Most Markets Allocate Less than 15% of 5G Throughput to Upload

Speedtest Intelligence® | 2025



SPEEDTEST

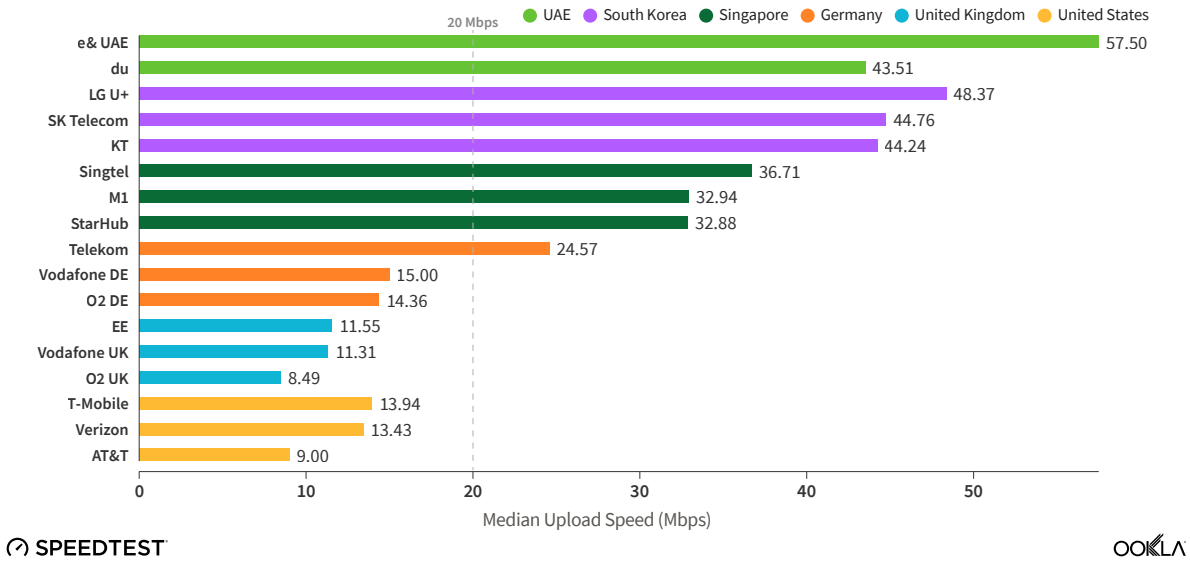
OOKLA

The range across markets is wide. Indonesia leads at 23.9%, followed by Germany at 15.6%, Thailand at 14.2%, and Sweden at 13.2%. The United States sits at 5.1% — the lowest in the dataset — alongside France at 5.8%, Brazil at 5.7%, and the UAE at 5.6%. Markets that lead on download speed do not necessarily lead on uplink capacity. The UAE and the US rank among the strongest on overall performance, and both still deliver competitive absolute upload speeds given their high total throughput, even though they devote a smaller share of capacity to the uplink. This is the first signal that the metrics shaping AI readiness produce a different ranking from the ones that have historically defined network quality.

Against AI workload performance thresholds, every market meets the minimum upload requirement for text LLM, voice AI, AI-generated video, and agentic AI at the median. The constraint appears at the AR and multimodal vision target of 20 Mbps. Only ten markets reach it.

Fewer Than Half of Operators Clear the Upload Target For AR And Multimodal AI

Speedtest Intelligence® | 2025



e& UAE leads the entire dataset at 57.50 Mbps median upload — more than four times any US operator. South Korea's three operators cluster between 44.24 and 48.37 Mbps, reflecting uniform C-band TDD deployment with comparable spectrum holdings. The contrast with Europe and North America reflects several overlapping factors. FDD spectrum deployed for 5G provides dedicated uplink capacity without the downlink trade-off inherent in TDD. In TDD, the uplink and downlink share the same frequency band and are separated by time slots, meaning any increase in uplink allocation comes at the direct expense of downlink capacity. FDD avoids this trade-off by using separate frequency bands for each direction. Operators with more FDD spectrum tend to perform better on upload. The maturity of 5G SA architecture matters because it enables uplink-specific scheduling and a broader range of carrier aggregation combinations. Spectrum composition plays a role too, since mid-band TDD delivers capacity but penetrates poorly indoors, leading to unpredictable upload speeds where users need them most.

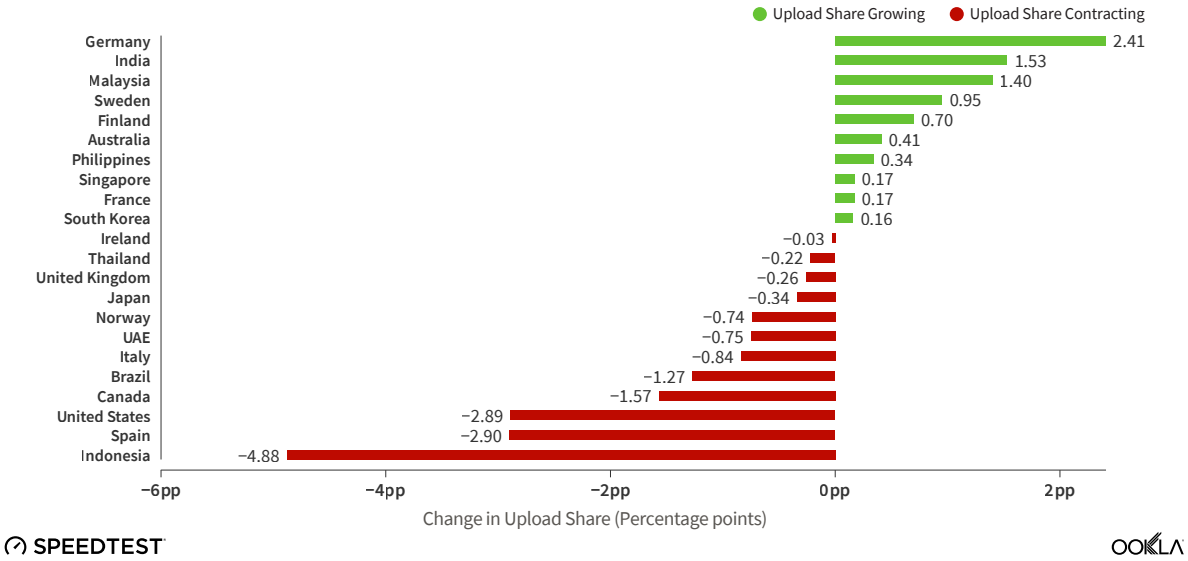
Markets that complement TDD mid-band with FDD low-band — including the Nordic markets, Germany, the UK, and Australia — tend to show more consistent uplink performance. South Korea illustrates the trade-off: relying almost entirely on C-band TDD, it achieves the second-highest market-level upload speed in the dataset at 45.27 Mbps, behind only the UAE, yet records one of the lowest uplink share figures at 7.5%.

In the US, T-Mobile leads at 13.94 Mbps, ahead of Verizon at 13.43 Mbps and AT&T at 9.00 Mbps. According to **Ookla's analysis**, all three carriers allocate roughly 20% of their midband TDD frame resources to uplink — a uniform configuration despite meaningful differences in their 5G SA maturity and carrier aggregation deployment. The result is a market where upload speeds vary more by operator architecture than by spectrum allocation.

Upload speeds have risen in absolute terms across most markets between 2023 and 2025. The proportion of capacity allocated to the uplink has not kept pace.

5G Upload Share Declined or Held Flat in More than Half of Markets Since 2023

Speedtest Intelligence® | 2023 – 2025

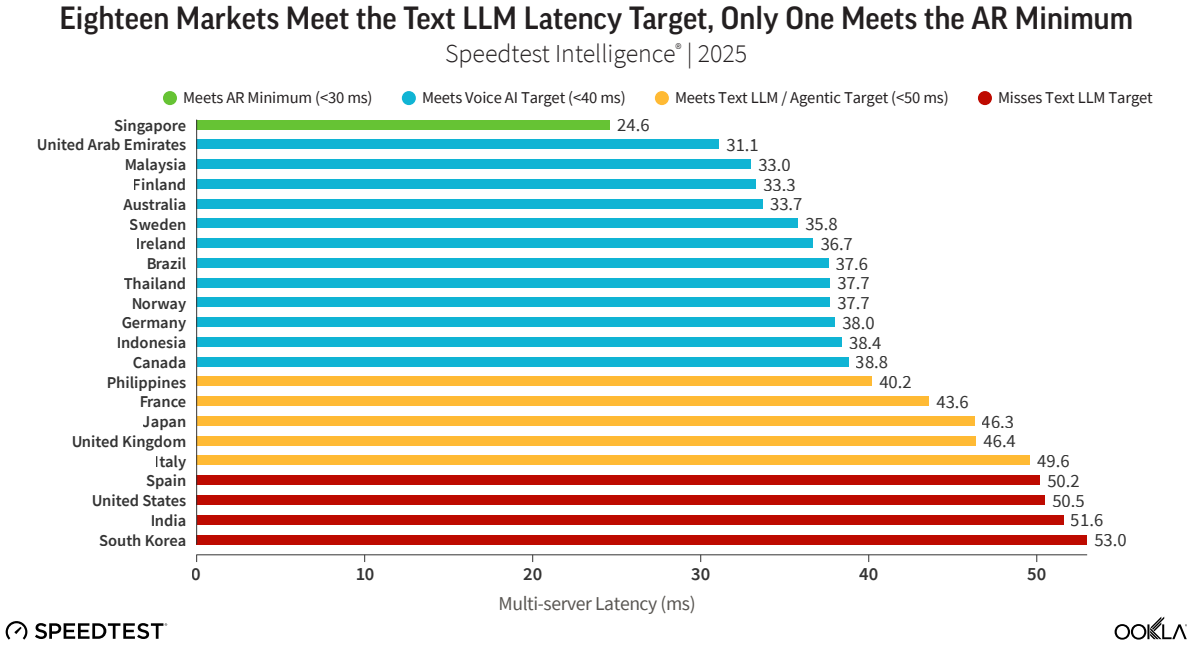


SPEEDTEST OOKLA

Twelve of the 22 markets recorded either no change or a decline in uplink share over the two-year period. Indonesia saw the largest drop, falling nearly 5 percentage points as downlink speeds grew faster than uplink speeds despite leading the dataset on upload share. The US and Spain each fell 2.9 percentage points, with the US declining from 8.0% to 5.1%, the lowest share of any market. Germany is the exception, rising 2.4 percentage points — the largest improvement — illustrating what targeted spectrum investment, 5G SA deployment, and carrier aggregation can deliver when applied together. Part of what makes this trend slow to reverse is a technical constraint: operators sharing the same TDD band in a geographic market must align their timing patterns to prevent cross-network interference, a coordination requirement that applies globally and means shifting uplink allocation requires industry-wide action rather than any single operator acting alone.

Under normal conditions, most markets meet the baseline for today's AI modalities

Multi-server latency, as defined in the metrics framework above, measures baseline network responsiveness under normal conditions. It is the metric this report benchmarks against the AI workload performance thresholds, because those thresholds describe what AI applications require under typical operating conditions, and multi-server latency is measured under those same conditions.



The results fall into three tiers when measured against the AI workload performance thresholds.

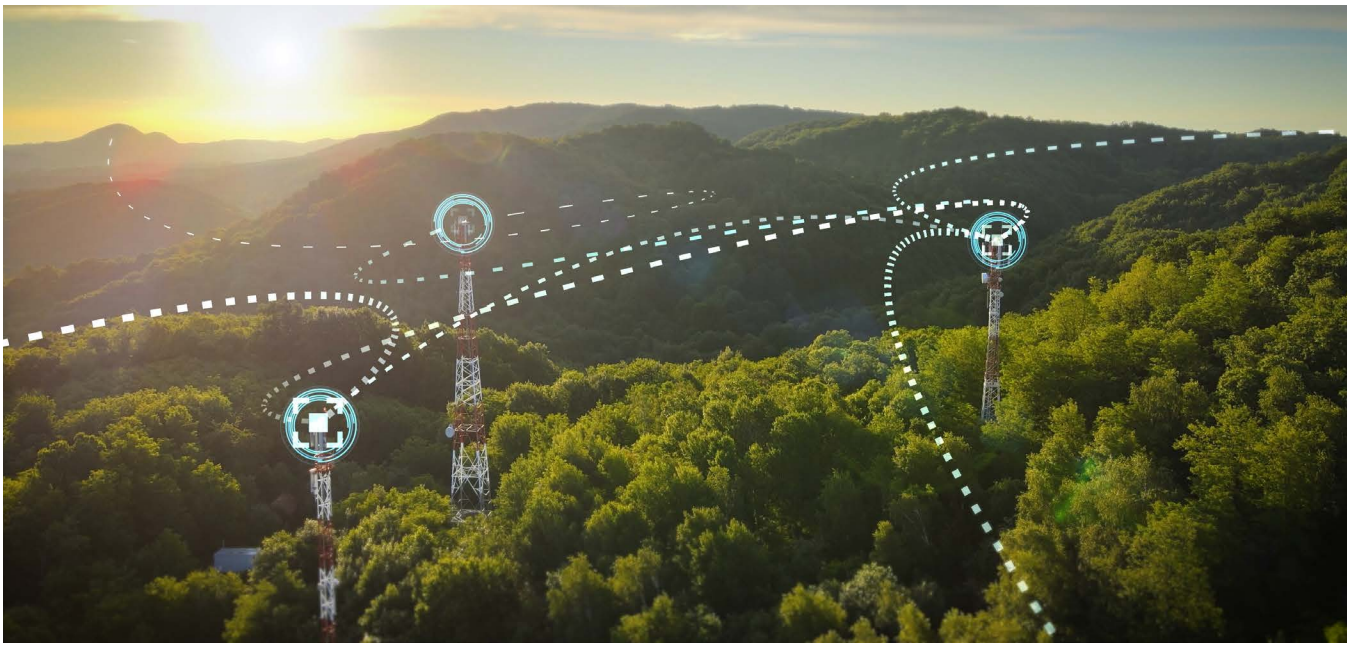
For the AI modalities that dominate mobile traffic today — text LLM interaction, AI-generated video, and agentic AI — readiness is near-universal. All 22 markets meet the minimum multi-server latency threshold of less than 100 ms. Eighteen of 22 meet the target of less than 50 ms. The four that miss are South Korea at 53.0 ms, India at 51.6 ms, the United States at 50.5 ms, and Spain at 50.2 ms. That the US and India appear here — despite ranking sixth and ninth respectively on download speed — reinforces the pattern from the upload analysis.

For conversational voice AI, the field narrows. Thirteen markets meet the target of less than 40 ms, led by Singapore at 24.6 ms and the UAE at 31.1 ms. The nine that miss the target but meet the minimum of less than 80 ms include Japan at 46.3 ms, the UK at 46.4 ms, and the US at 50.5 ms. Voice AI is functional in these markets under normal conditions, but with limited headroom. Once cloud transit, inference processing, and server queuing are added to the network's latency contribution, the cumulative delay in these markets approaches the boundary where users perceive degradation.

For AR and multimodal vision, no market meets the target of less than 10 ms. Only Singapore meets the minimum of less than 30 ms. The most demanding AI modality remains beyond the reach of current 5G networks globally, regardless of how advanced the deployment.

Singapore, the UAE, Malaysia, Finland, and Australia form the top tier on baseline latency — none of which leads the dataset on raw throughput. The US, Spain, and South Korea, which rank among the highest on download speed, sit in the bottom tier.

Several factors likely contribute to this gap. TDD mid-band spectrum delivers high throughput but adds round-trip overhead, as the device waits for its uplink transmission slot before responding. NSA architecture may also play a role, as the control plane routes through the 4G core rather than directly through a 5G core. South Korea, which deploys 5G almost entirely on C-band TDD and remains predominantly NSA, records the second-highest download speed in the dataset at 558.33 Mbps yet the highest multi-server latency at 53.0 ms. South Korea's interconnect environment, where sender-pays network peering policy may add routing overhead between mobile and fixed networks, is a further factor that could contribute to higher multi-server latency independent of radio architecture. The gap is narrowing. South Korea's market-level multi-server latency fell from 82.5 ms in 2024 to 53.0 ms in 2025, and further improvement is likely as 5G SA migration progresses.

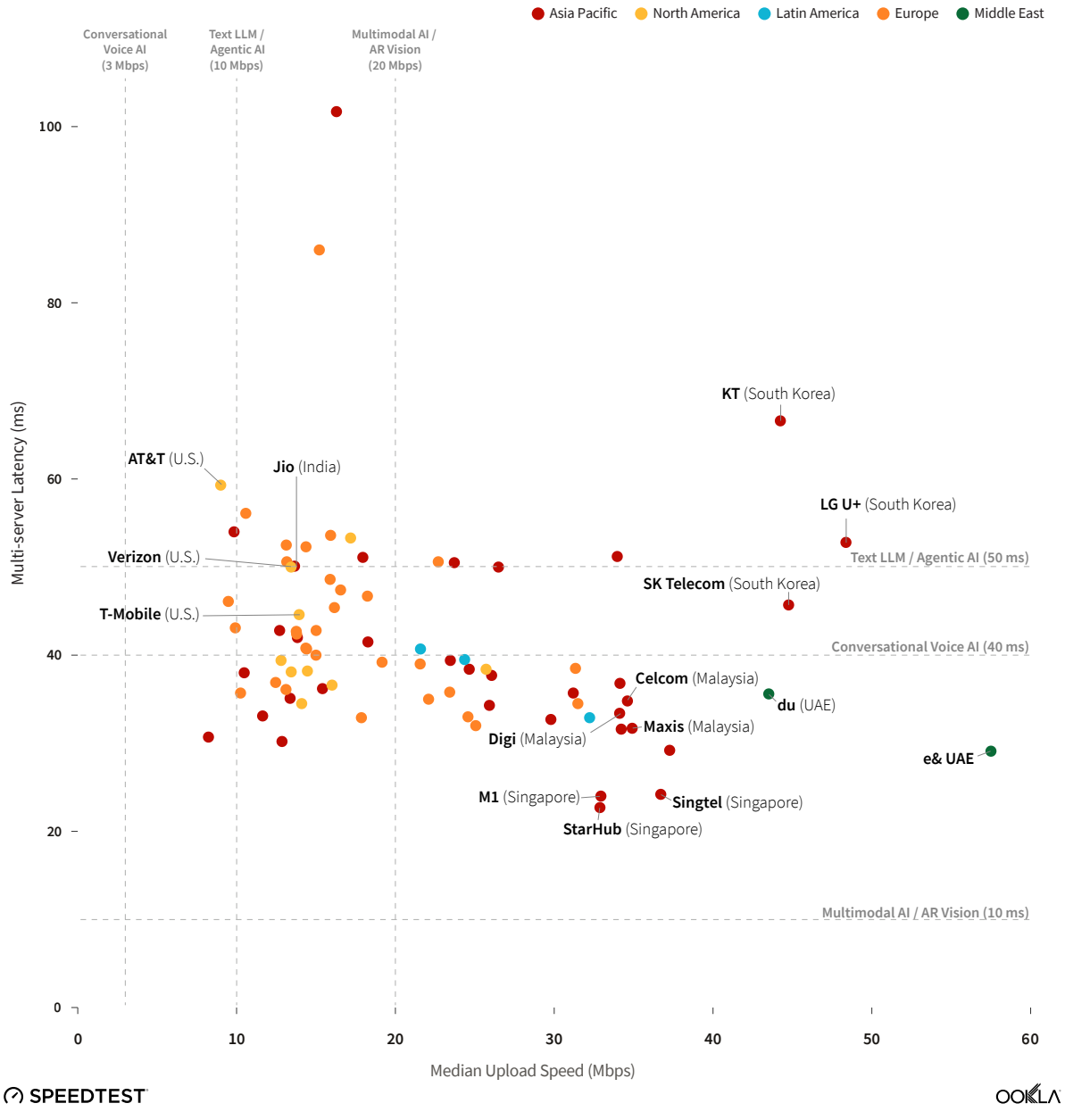


Few operators meet both upload and latency targets simultaneously

Each operator's users need both upload throughput and latency met simultaneously on the same network. The chart below plots all 86 operators on both axes, with the text LLM and voice AI target thresholds overlaid.

Select Global Operators 5G Upload Throughput and Latency Assessed Together Reveal Latency as the More Common Constraint

Speedtest Intelligence® | 2025



Of the 86 operators across all 22 markets, 65 meet both the upload and latency targets for text LLM and agentic AI (10 Mbps, 50 ms). That drops to 46 for conversational voice AI, where the 40 ms latency target eliminates nearly half the field. No operator meets the multimodal AI and AR vision combined target. The 10 ms latency threshold sits below every operator in the dataset.

Latency is the more common constraint. Of the 21 operators that miss the text LLM combined target, 15 have adequate upload speed but latency above 50 ms. Only three meet the latency target but fall short on upload. Most operators that fall short are constrained by responsiveness, not capacity.

Regional clustering is visible on the chart. The UAE's two operators occupy the lower-right corner, combining the highest upload speeds with latency well below every threshold except multimodal AI. Singapore's three major operators (Singtel, StarHub, M1) anchor the lowest-latency zone, meeting both the text LLM and voice AI targets. Malaysia's six operators cluster tightly in the mid-right, all meeting both the text LLM and voice AI targets. The South Korean operators deliver the highest upload speeds in the dataset outside the UAE, but their latency positions two of three above the text LLM threshold line.

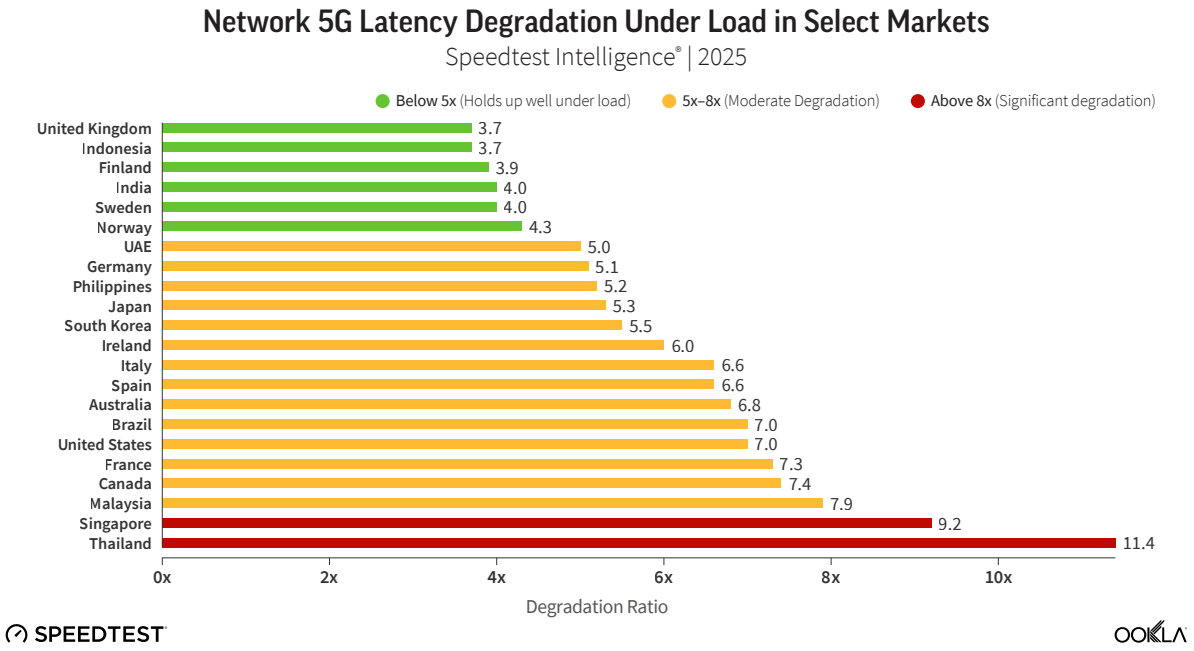
Deutsche Telekom (Germany) meets both the voice AI and text LLM targets at 24.57 Mbps and 33.0 ms, the strongest position among large European operators. Claro (Brazil) meets every threshold below multimodal AI at 32.23 Mbps and 32.9 ms, outperforming operators in markets with far more advanced 5G deployments. T-Mobile (US) meets the text LLM target but misses voice AI, sitting just above the 40 ms line. AT&T (US), at 9.0 Mbps and 59.3 ms, misses both the upload and latency targets for text LLM, a position it shares with only two other operators globally.

Across the full dataset, the operators closest to meeting every achievable threshold tend to be in geographically compact, densely peered markets.

Under stress, the gap between markets and operators widens

The analysis above captures operator readiness under normal conditions. **Loaded latency** captures what happens when the connection is fully utilized — a stress test that reveals how much headroom each network loses under peak demand. Loaded latency is not benchmarked against the AI workload performance thresholds, which describe typical conditions. Instead, it serves as a comparative metric to measure degradation across markets and expose operator-level gaps that market averages conceal.

The degradation ratio makes that comparison concrete. It divides the 10th percentile loaded latency — representing the best-case connections under full utilization — by the multi-server latency. Multi-server latency serves as the baseline because it captures the network's responsiveness under normal operating conditions, making it the natural reference point for measuring congestion-driven degradation. The 10th percentile was chosen as the numerator rather than the median because it isolates the best-case outcome under load. If even that connection shows substantial degradation, the experience for the median user will be substantially worse.



Across the dataset, the ratio ranges from 3.7x in the UK and Indonesia to 11.4x in Thailand. Three patterns stand out:

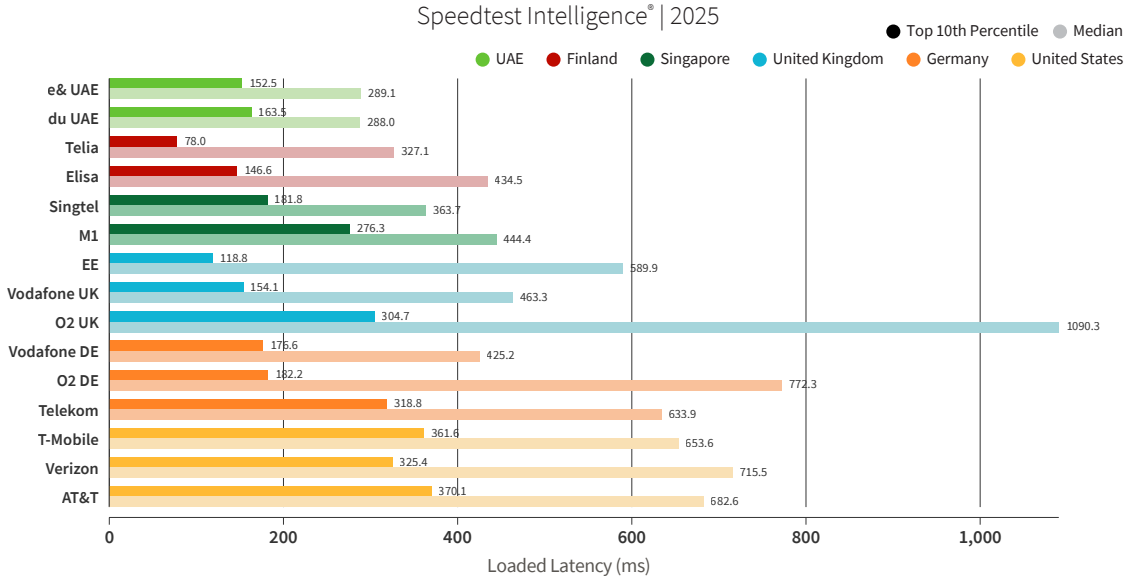
1. Absolute loaded latency matters as much as the ratio. The UAE records a degradation ratio of 5.0x and the lowest median loaded latency of any market, at 288.4 ms. For AI applications running on congested cells, the absolute value determines whether the experience holds. The UAE's result reflects coordinated investment in 5G Advanced capabilities across both e& and du — carrier aggregation, enhanced MIMO, and active 5G SA deployment — which sustain responsiveness under load rather than just improving peak speeds.

2 5G SA deployment does not automatically translate into strong loaded latency performance. The US records a 7.0x ratio and a median loaded latency of 684.6 ms, despite T-Mobile operating **one of the first nationwide** commercial 5G SA networks since 2020 and substantial C-band investment across all three operators. T-Mobile records 653.6 ms, AT&T 682.6 ms, and Verizon 715.5 ms. Indonesia records a 3.7x ratio despite limited progress on the factors that typically produce low degradation — it lacks deep fiber backhaul, advanced 5G SA deployment, and the spectrum depth seen in markets like the UK and the UAE. The most likely explanation is spectrum composition: Indonesia's 5G is deployed primarily on sub-3 GHz lower-mid bands rather than on C-band, providing better indoor penetration and more consistent coverage under load. The result should be read with caution, as Indonesia's 5G deployment is still early-stage and the user base remains small relative to overall mobile subscribers, meaning congestion patterns may differ materially from more mature 5G markets.

3 Markets with the lowest multi-server latency do not necessarily maintain that advantage under stress. Singapore records the lowest baseline in the dataset at 24.6 ms, but its degradation ratio of 9.2x is among the highest — the result of dense urban demand competing for cell resources during peak periods. Thailand records a competitive multi-server latency of 37.7 ms but the highest degradation ratio at 11.4x, with a median loaded latency of 960.3 ms, pointing to backhaul and cell-load management constraints rather than coverage or spectrum gaps. For AI workloads, Thailand presents the widest gap between baseline capability and loaded performance of any market in the dataset.

Within-market Operator Gaps on Loaded Latency Are as Wide as Cross-market Differences

Speedtest Intelligence® | 2025



SPEEDTEST OOKLA

At the operator level, the chart reveals that loaded latency performance varies as much within markets as between them. In the UK, EE records a 10th percentile (best-case) of 119 ms while O2 records 305 ms — a 2.6x gap between operators sharing the same market. That dimension of variation is one that market-level averages do not capture.

Cloud infrastructure latency adds a dimension that network benchmarks alone do not capture

Upload capacity and loaded latency measure what the operator network delivers. Cloud infrastructure latency and jitter measure the next segment of the AI inference path, from the operator's network edge to the cloud endpoint where inference runs. Operators influence this path through peering agreements, core network routing, and the proximity of their interconnection points to cloud provider edge locations, but do not fully control it. In markets where cloud latency runs high, it becomes a bottleneck regardless of how well the network itself performs.

Brazil stands apart from every other market in the dataset, with median cloud latency ranging from 149.7 ms to 163.6 ms across all four providers in 2025. Cloud infrastructure in Brazil is concentrated in São Paulo.

The **fragmented ISP market**, where most providers lack direct peering with hyperscalers, causes traffic to pass through multiple intermediary hops before reaching any cloud endpoint.

Europe leads the dataset. Based on 2025 Speedtest Intelligence data, Germany reaches AWS at 42.2 ms, the UK at 44.0 ms, France at 45.9 ms, and Ireland at 46.3 ms. Finland, Norway, and Sweden lead on Azure at 49.1 ms, 47.8 ms, and 50.7 ms. Users in these markets have the most latency headroom available for inference — the cloud path consumes a smaller share of the end-to-end delay, leaving more room for model processing before the cumulative delay crosses the threshold where users perceive degradation.

Latency to Cloud Infrastructure Varies Sharply Across Markets and Cloud Providers

Speedtest Intelligence® | 2025

		AWS	Azure	Google Cloud	Oracle Cloud Infrastructure
East Asia	South Korea	40	39	48	43
	Japan	46	46	62	54
Europe	Germany	42	43	43	45
	France	46	47	47	48
	United Kingdom	44	48	49	48
	Ireland	46	51	57	55
	Finland	53	49	55	56
	Norway	56	48	57	55
	Sweden	53	51	62	63
	Italy	61	62	60	59
	Spain	66	70	67	65
Southeast Asia	Singapore	74	75	73	96
	Indonesia	63	85	70	105
	Malaysia	77	86	86	111
	Thailand	77	88	85	113
	Philippines	104	106	88	133
Americas	Canada	55	59	63	80
	United States	68	67	68	77
	Brazil	157	162	164	150
Oceania	Australia	69	132	147	166
South Asia/ Middle East	UAE	125	125	139	135
	India	114	109	121	158

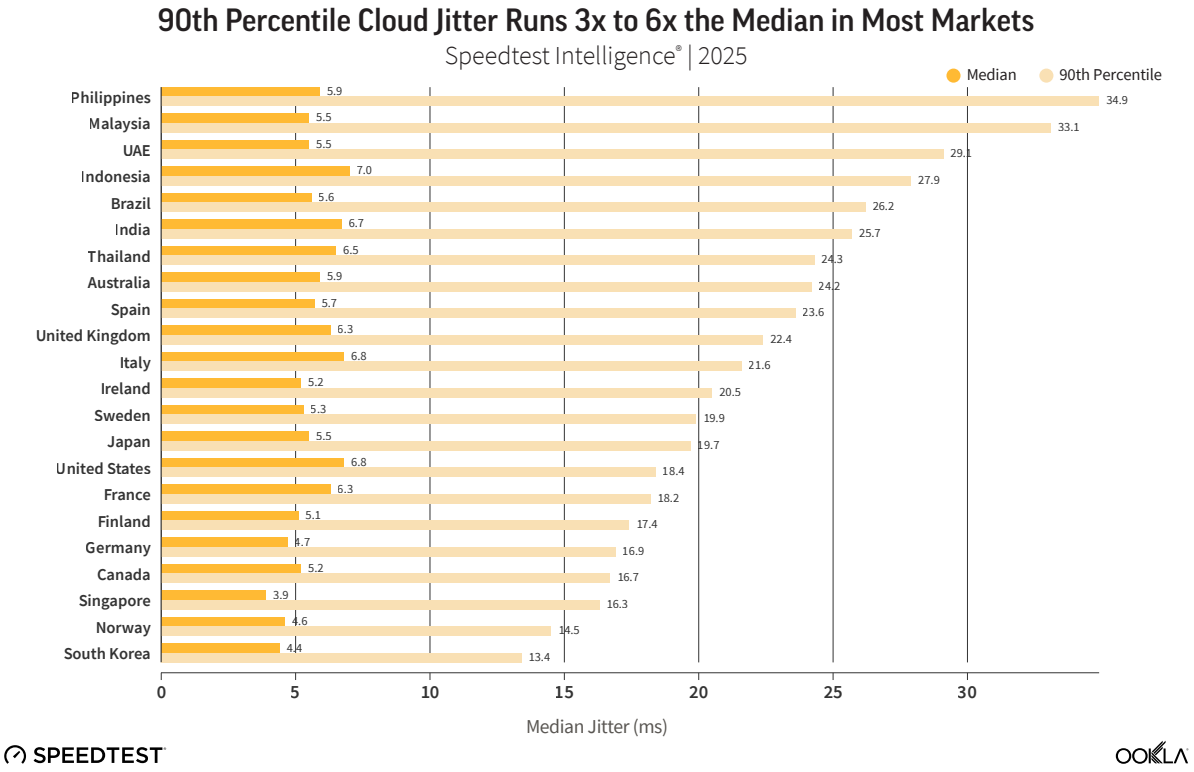
Cloud Latency (ms)

40 60 80 100 120 140 160

SPEEDTEST OOKLA

Across much of Asia Pacific, the most consequential infrastructure decision for AI deployment is not which mobile operator to use, but which cloud provider to use. Australia records 69.3 ms to AWS, compared to 165.9 ms to OCI — a gap of 96.6 ms within a single market that directly determines what AI modalities are viable. An enterprise using OCI has 96.6 ms more cloud latency than one using AWS, enough to push voice AI and agentic

applications past the point of perceptible degradation regardless of network quality. The pattern holds across Indonesia, Malaysia, Thailand, and the Philippines, where OCI consistently sits furthest from the other three providers. In Europe, the gap between providers is much narrower — Germany records a spread of just 2.7 ms between its fastest and slowest provider. The difference matters because **OCI is expanding its role as AI training and inference infrastructure**, making cloud provider selection an increasingly direct factor in AI experience quality for enterprises in affected markets.



Jitter tells a different story from latency. At the median, markets look alike, with median cloud jitter between 3.9 ms and 7.0 ms across all 22 markets. The separation appears at the 90th percentile, where worst-case jitter ranges from 13.4 ms in South Korea to 34.9 ms in the Philippines, a spread of more than 2.5x. South Korea, Norway, Singapore, Canada, and Germany hold the steadiest connections to the cloud, while the Philippines and Malaysia show the widest worst-case swings. The ranking only partly overlaps with the other metrics. South Korea, for instance, posts the lowest worst-case jitter in the dataset despite recording the highest multi-server latency of any market. Speed and stability are separate properties of the path to the cloud.

For real-time AI, consistency matters more than raw speed. A market can reach the cloud quickly on average yet still deliver unstable timing, and that instability degrades a voice or agentic AI session as much as a slow connection does. The gap is widest where headline numbers hide it: worst-case jitter runs three to six times the median in most markets. This is one more case where the markets that lead on speed are not the ones that lead on AI readiness. The markets best placed for real-time AI are not the fastest ones, but those that hold their timing steady when the path to the cloud is busy.

Network investment priorities for the mobile AI era

The data in this report identifies four areas where network investment can directly improve AI workload performance. Each maps to a specific dimension of the benchmarking analysis, and each responds to deliberate operator action.

Network investment priorities for AI workload readiness

Four areas where operator action can directly improve AI application performance

Speedtest Intelligence® | 2025

Priority	What the data shows	Why it matters	Operator action
Upload rebalancing	Fewer than half of operators meet the 20 Mbps upload target. The median allocates just 10.1% of through-put to uplink.	AI traffic runs at 29/71 to 50/50 uplink-to-downlink. Upload capacity constrains most AI modalities beyond text LLM.	Deploy 5G SA. Activate uplink carrier aggregation and NR-DC. Coordinate TDD frame alignment.
Latency and jitter	65/86 operators meet text LLM targets. Only 46/86 meet voice AI. Cloud infrastructure jitter diverges more between operators than median latency.	Latency is the constraint most operators fail on. Jitter determines whether voice AI and physical AI sessions stay consistent.	Add loaded latency and cloud infrastructure jitter to KPI frameworks. Invest in fiber backhaul and 5G SA.
Cloud peering	Cloud latency rose in 18/22 markets over the past year. Within-market provider gaps reach 96.6 ms.	The cloud path is often the binding constraint on AI performance, regardless of network quality.	Establish direct peering with hyperscalers. Optimize routing to reduce the gap between best-case and median cloud latency.
Next-wave readiness	0/86 operators meet multimodal AI targets. The 10 ms latency threshold is beyond every network globally.	The AI modality mix is shifting from text to multimodal and voice, with sustained uplink and tighter latency demands.	Deploy network slicing for latency-sensitive traffic. Evaluate AI-RAN for edge inference. Position for physical AI.

Based on 86 operators across 22 markets.

SPEEDTEST

OOKLA

Rebalancing the link toward upload

Upload capacity is the most direct lever operators can pull. Fewer than half of operators in this study meet the 20 Mbps target for the most demanding AI modalities today.

The next wave of uplink gains depends on capabilities that require 5G SA architecture. 5G SA enables uplink-specific scheduling and dynamic switching between supplementary uplink and TDD configurations. Uplink carrier aggregation across FDD and TDD bands delivers immediate throughput gains without new spectrum. **T-Mobile launched 5G-Advanced uplink Tx switching** in 2025, enabling devices to dynamically switch transmit antennas between frequency bands to maximize uplink carrier aggregation performance.

Closing the gap on latency

Most networks in this study meet baseline latency requirements for the AI modalities that dominate mobile traffic today. The gap appears under stress. Loaded latency reveals how much each network degrades when demand peaks. The operators with the lowest degradation ratios tend to share deep fiber backhaul, active 5G SA deployment, and favorable spectrum composition.

Cloud infrastructure jitter adds a dimension that latency metrics alone do not capture. Worst-case cloud jitter varies as widely across markets as cloud latency does, and the two do not track each other: a network can post competitive median cloud latency yet still deliver the unstable timing that degrades real-time AI. For AI applications that require consistent timing between consecutive exchanges, such as conversational voice AI and **physical AI**, reducing jitter is as important as reducing absolute latency.

Operators planning network investment should factor loaded latency and cloud infrastructure jitter into their KPI framework alongside throughput and coverage. These metrics capture how the network performs under the conditions where AI application quality degrades first.

Treating cloud peering as network infrastructure

between the operator's network edge and the hyperscaler endpoint where the model runs. Data from the selected markets shows that this dimension varies as widely across operators as upload speed and loaded latency, and it responds to deliberate action from both operators, through peering and routing optimization, and cloud providers, through data center placement and expansion.

The operators with the lowest cloud latency tend to serve markets with hyperscaler data center presence nearby, and those that also maintain direct peering arrangements with cloud providers show the narrowest routing gaps. **AWS Local Zones**, Azure peering arrangements, and direct GCP interconnects all reduce the number of network hops between the user and the inference endpoint. The effect is visible in the routing gap — the difference between the 10th percentile (best-case) and the median — which measures how consistently an operator routes traffic along the most efficient available path.

For operators building network propositions around AI-dependent services, cloud peering is not an IT procurement decision. It is a network performance decision with a direct impact on the AI experience operators can deliver. Across Asia Pacific alone, hyperscale cloud providers have committed tens of billions of dollars in data center investment during 2024 and 2025, with comparable commitments across Europe and the Americas. Operators that align their peering and routing strategy with this infrastructure buildout position themselves to capture the latency advantage as cloud capacity scales.

Preparing the network for the next wave of AI modalities

The AI workloads dominating mobile networks today — text LLM interactions and lightweight agentic workflows — sit below the most demanding AI workload performance thresholds. That will not last.

The shift underway is toward multimodal interaction, in which a single session combines voice, image, and video inputs. These applications generate sustained uplink traffic rather than the short bursts that characterize text queries, and they split inference between on-device models and cloud endpoints depending on task complexity. On-device capability is improving, but when a task exceeds local capacity, the fallback to cloud inference creates the kind of unpredictable uplink demand the benchmarking analysis identified.

Conversational voice AI and real-time multimodal applications are also among the clearest use cases for 5G SA network slicing — the ability to isolate latency-sensitive traffic in a dedicated virtual channel, shielded from the congestion that degrades performance on shared infrastructure. As these workloads move from early adoption to mainstream use, the ability to guarantee consistent latency for specific traffic classes will become a competitive differentiator for operators, and a practical requirement for enterprise deployments where service-level commitments apply.

A more fundamental architectural question is emerging alongside these incremental gains. The **AI-RAN Alliance**, an industry group with more than 100 members, is developing reference architectures for converged AI and RAN compute, where base station hardware processes radio signals and hosts edge AI inference on the same platform. Whether that compute runs on GPUs, CPUs with built-in AI acceleration, or a hybrid remains an open debate, and the commercial model is unproven. The investment signals, however, are substantial. **NVIDIA committed $1 billion in Nokia** in October 2025 to develop GPU-based RAN platforms. **T-Mobile is working with Nokia and NVIDIA** to integrate AI-RAN technologies into its network development. Ericsson is pursuing AI-assisted RAN optimization through its own silicon and software platforms.

Further out, physical AI — robotics, autonomous systems, industrial automation — represents the most network-demanding wave. These workloads require sustained upload for sensor telemetry, low latency for real-time control, and low cloud infrastructure jitter for coordination between autonomous systems. For applications where milliseconds determine outcomes, cloud-based inference may not be fast enough even on well-peered networks. Edge AI inference — whether from dedicated facilities or converged RAN infrastructure — closes that gap by placing compute as close to the device as the network allows.

The operators investing in upload rebalancing, latency reduction, and cloud peering optimization today are laying the foundation for these workloads. The performance gaps this report identifies are already wide enough to shape competitive positioning — and they will widen as the AI modality mix shifts.