www.pipelinepub.com
Volume 21, Issue 11

Latency as the New Currency: Innovations Driving the Next Frontier in AI Connectivity

By: Ivo Ivanov

The meteoric impact of artificial intelligence over the past few years is difficult to overstate, and progress is moving so quickly that it is better measured in months rather than years. Some of the biggest tech events of the year, such as the Consumer Electronics Show (CES) and Mobile World Congress (MWC), were dominated by emerging AI use cases, from LLM-powered humanoid robots to "sight beyond sight" vehicle-to-cloud software capable of giving cars human-like senses. Businesses also have high expectations for the next generation of AI, testing and deploying everything from problem-solving AI agents to advanced data analytics and forecasting tools. We've been conditioned over the past decade or more to believe that all we need to apply AI and use it effectively is *data*. More data typically means broader applications and better results. In 2025, however, that singular approach is being brought into question.

For years, AI innovation was synonymous with the cloud. Training clusters and centralized hyperscale platforms, the brains behind AI, draw power and data into vast facilities designed to meet the demands of machine learning models. But this is just about the *training* of AI. With real-time demands and expectations of *inference* now resting heavily on its shoulders, AI is breaking free from these traditional gravitational centers. In the interests of speed and instant access, AI is now in action almost everywhere – embedded in devices, vehicles, retail environments, and digital agents that respond within milliseconds. So perhaps the question we should be asking isn't just how much compute power can be deployed or how much data can be gathered – but how intelligently it can all be connected.

This direction of travel has brought one vital characteristic to the fore: *latency*. No matter how massive the model or how sophisticated the silicon, high latency is kryptonite to AI. It's the obstacle in the corridor, the roadworks on the highway – slowing everything to a crawl.

It interrupts the flow of information between people, devices, and AI systems, introducing delays that can degrade user experiences, limit real-time decision-making, and even compromise safety in critical applications such as autonomous vehicles or remote health monitoring, where split-second responses can make all the difference. Make no mistake, in this post-AI world, *latency* is the new currency, and the next frontier of AI will be won or lost on the network layer.

The hidden enabler of AI innovation

Latency has always been a technical consideration in network design, but post-AI, it has become a business-critical variable. Every additional millisecond can distort outcomes in systems that rely on real-time processing, learning, and adaptation. AI agents designed to predict, recommend, or act autonomously are only as good as the data pipelines that feed them. If information arrives too late, decisions are made on stale insights, rendering even the most powerful models ineffective. Consider predictive maintenance in industrial manufacturing. Here, AI agents continuously monitor sensor data from machinery to flag anomalies before they escalate into failures. But if that data is delayed by even a fraction of a second, the insight may arrive too late to prevent damage or downtime. The same logic applies to AI in fraud detection, where instantaneous analysis of transaction patterns can mean the difference between blocking fraud and letting it through. In both cases, latency isn't just a technical hurdle – it directly affects business continuity and customer trust.

Technically, the challenge lies in the sheer volume and velocity of data that AI applications must handle. Unlike traditional AI deployments, modern AI workloads are not simply about processing large datasets and delivering results; they require high-frequency data exchange between edge devices, sensors, data centers, cloud platforms, and end-users. Each hop across a network introduces potential delays, whether due to physical distance, network congestion, or inefficient routing. Minimizing latency, therefore, is not a matter of optimizing a single link; it demands a holistic rethinking of how digital infrastructure is architected, interconnected, and managed. Moving forward, we need to consider latency reduction as a foundational design principle rather than an optional variable – even with legacy networks, it can still be achieved.

The rise of AI hubs and engineering near- zero latency

Traditionally, Internet Exchanges (IXs) were designed to facilitate efficient data exchange between networks, improving performance and reducing transit costs. But as AI workloads migrate to the edge, the role of IXs is evolving. Rather than serving solely as aggregation points for global, largely content- based Internet traffic, IXs are becoming localized interconnection "hubs" for AI within their city areas – dense ecosystems where enterprises, cloud providers, AI developers, and edge networks converge. Here, we see the emergence of the interconnection triangle for AI inference, enabling stable and low- latency data exchange between AI agents and devices through high-powered transmission technologies.

The logic is simple: the closer AI models and services are to the people and devices they serve, the faster and more reliable their performance will be. Direct, local or regional interconnection minimizes

the number of hops, reducing latency and jitter that would otherwise cripple real-time AI operations. The evolution of IXs into local and regional AI hubs is not simply a scaling

exercise; it is a fundamental re-architecture of digital infrastructure to meet the speed, proximity, and redundancy demands of intelligent, real-time systems – preparing for the age of Zero – or at least near-zero – Latency.

This shift also requires a new kind of governance and service model. AI hubs will increasingly need to support dynamic provisioning, policy-based routing, and workload-specific traffic prioritization. Instead of passively exchanging traffic, these IXs must evolve into intelligent coordination points that allocate network resources in real-time to meet the latency sensitivity of each AI application. This is where the fusion of interconnection and orchestration will define AI-ready network infrastructure.

Global integration to solve the challenges of AI

While proximity through metro interconnection is critical, it alone cannot solve the latency challenge for AI at scale. Long-haul connectivity still underpins the broader AI ecosystem, linking regional hubs, data lakes, power-intensive training clusters, and cloud platforms across vast distances. To keep pace with the innovations of AI, network rollout is necessary across the board. In parallel with the build-out of AI data center infrastructure, we see the continued development of fiber and optical transport technologies, the work going into developing the 6G mobile standard, and the accelerating rollout of LEO satellite constellations to provide a redundant backbone in space.

Each of these network technologies, on their own, is simply one piece of the gigantic AI puzzle. What binds them together is peering, or the direct interconnection of networks via an Internet, Cloud, or AI Exchange. For a regionally or globally operating company, it is necessary to go beyond the support for ultra-low latency requirements at the local level. Because the hundreds of local markets where AI products and services are being deployed need to be harmonized and synchronized as far as possible. This requires the integration of hyper-local AI hubs within regional and pan-regional interconnection ecosystems, and even within global ecosystems as far as practicable and sensible. Only in this way can a globally acting enterprise roll out their products and services and ensure they serve their customers efficiently and with the same level of seamless performance everywhere.

Latency is the new currency

Connectivity is not a destination; it's a moving target. It isn't something that can be statically "achieved" – it needs to be resilient, adaptable, and – ideally – autonomous. The networks we build and upgrade today will shape the limits of what AI can do for years to come, paving the way for some of the exciting new use-cases demonstrated at events like MWC to move from concept to reality. It all boils down to latency. What used to be a technical metric buried in a service level agreement, causing the odd degree of frustration, is now the biggest constraint on our ability to shape and develop the biggest technological development in a generation.

Much like electricity enabled the industrial age, and broadband catalyzed the digital one, low-latency interconnection will be the backbone of AI's coming era. It will determine not only how fast we can process information, but also how quickly we can understand, act, and adapt in a world driven by intelligent systems. The breakthrough is coming, but it won't come from faster processors or larger sets of data – it will come from the invisible architecture of connection, engineered to move at the speed of thought itself.