www.pipelinepub.com
Volume 22, Issue 3

The Quiet Backbone of the AI Economy

By: Rudy Hoebeke

Telecommunication has always been an industry defined by constant evolution and innovation. Over the past two decades, we have witnessed successive waves of technological advancement that have fundamentally reshaped how networks are built and operated. The next significant wave is already upon us, driven by the rapid ascent of AI, and its impact promises to be profound.

The echo of innovation, from video to AI

In the mid-2000s, the seemingly innocuous upload of the first video "Me At The Zoo" to YouTube heralded a complete paradigm shift in how video content was produced, distributed, and consumed. This marked the beginning of the video era.

The impact on network infrastructure has been tremendous.

IP video became the primary driver for network capacity growth for two decades, fundamentally altering network architectures. The need for more efficient video distribution led to distributed peering and the widespread implementation of Content Delivery Networks (CDNs), moving content closer to end-users at the network edge.

Now, a new wave is forming, triggered by the launch of ChatGPT in November 2022. This event propelled AI into mainstream consciousness, demonstrating its capabilities to a global audience.

The speed of adoption is staggering: while YouTube took three years to reach 100 million users, ChatGPT achieved this milestone in just two months (Figure 1). Adoption has continued at an extraordinary pace, reaching nearly 6 billion monthly visits by 2025. This suggests that AI's impact on networks will be felt much sooner and more intensely than that of IP video.

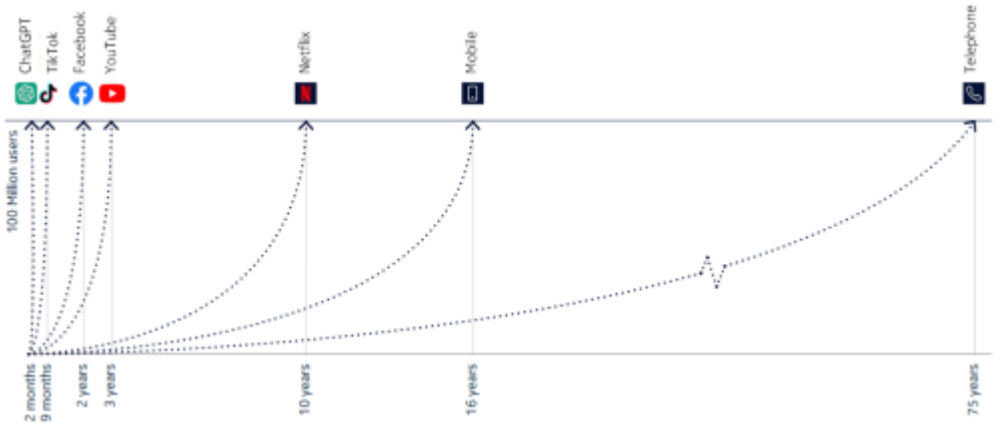

Figure 1. Time to reach 100 million users for various applications
click to enlarge

While today's industry conversation understandably gravitates toward massive GPU infrastructure buildouts, it's essential not to overlook the network's pivotal role. The cloud, in its very essence, exists because of the network. Its evolution, whether supporting IP video or the surge of AI workloads today, is inextricably linked to the network's own advancement. The network will ultimately serve as the gating factor for how far AI and cloud technologies can truly evolve.

AI's insatiable appetite for data

Just as the video wave dramatically reshaped network traffic patterns and architecture, the AI wave is poised to do the same.

According to Bell Labs Global Network Traffic report, wide area network (WAN) traffic is forecast to reach 3,386 exabytes per month by 2033. A substantial portion of this (1,088 exabytes) is attributed to AI traffic alone, which is projected to grow at a compound annual growth rate (CAGR) of 24 percent (Figure 2).

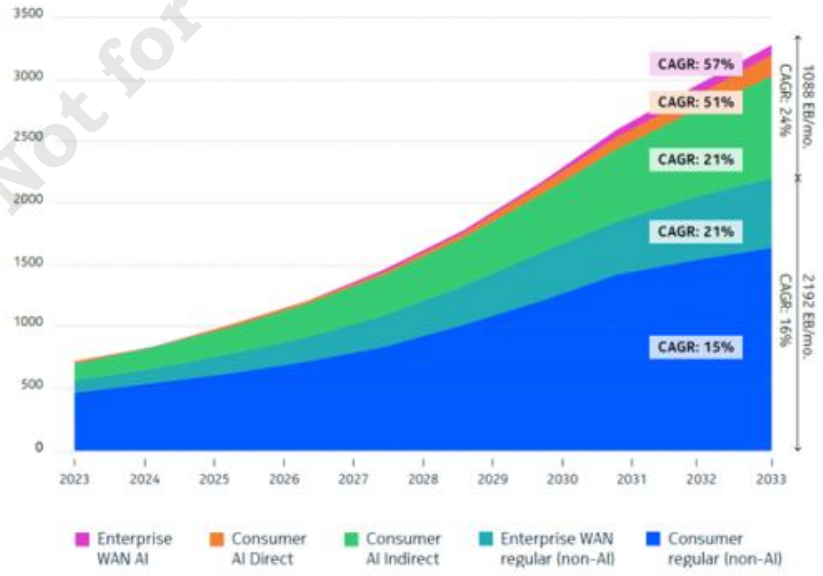

Figure 2. Global WAN AI traffic projections, EB per month
click to enlarge

This massive growth reflects not only the rising role of AI in everyday life but also a shift from today's largely text-based interactions to far more visually intelligent systems. As AI moves into multimodal perception, seeing, interpreting, and understanding the world, network capacity demands will climb even faster.

And this is only part of the story. A substantial share of tomorrow's traffic won't come from people at all, but from machines. As AI agents, autonomous robots, and intelligent systems operate at scale—gathering context, reasoning, collaborating with other systems, making decisions, and acting without human intervention, they will generate an immense surge in machine-to-machine communication.

Distributing intelligence

The impact of AI extends beyond mere traffic volume; it also necessitates a fundamental architectural transformation.

Today, the majority of AI processing is concentrated in the training of large language models (LLMs) housed within massive, centralized "AI factories." These hyperscale facilities consume extraordinary amounts of compute and storage to create ever more capable models.

However, this dynamic is set to shift dramatically. McKinsey & Company estimates that by 2030, 60 to 70 percent of all AI workloads will be dedicated to real-time AI inferencing. As AI adoption accelerates, the center of gravity will move from training models to serving predictions and answering questions, responding instantly to humans, machines, and other AI agents.

This shift will push AI workloads outward, closer to where the requests originate. In many ways, it echoes the transformation brought by CDNs in the video era. Just as video delivery moved to the edge to improve performance and scale, AI inference will increasingly be distributed across networks to meet the latency, cost, and capacity demands of real-time intelligence.

The network cloud continuum

Several powerful forces are driving the need to distribute AI workloads.

For business-critical and mission-critical applications, real-time processing is non-negotiable. Achieving this requires high-speed connectivity and ultra-low latency. Bringing inference closer to the user or device not only improves responsiveness but also reduces the volume of traffic traversing the WAN, cutting bandwidth consumption and improving overall efficiency.

Security and privacy concerns add another compelling reason. Localization of sensitive data and minimizing its movement across networks reduces exposure and aligns with increasingly stringent regulatory requirements.

Operational considerations matter as well. Distributing AI workloads allows operators to place compute resources where power and cooling resources are available and cost structures are most favorable. And in environments with limited or expensive connectivity, localized inference delivers far better performance and reliability.

All of this is pushing the industry toward a network cloud continuum, a world where AI processing happens everywhere. Some inference will run directly on devices such as smartphones, laptops, and wearables, albeit limited to tiny or very small models due to constraints in compute, memory, and power. A much larger share will be executed across a

spectrum of cloud environments: on-premises enterprise clouds, metro and edge clouds, and centralized hyperscale regions, each interconnected by high-performance networks (Figure 3).

This continuum is already emerging, and it will become the defining architecture of the AI era.

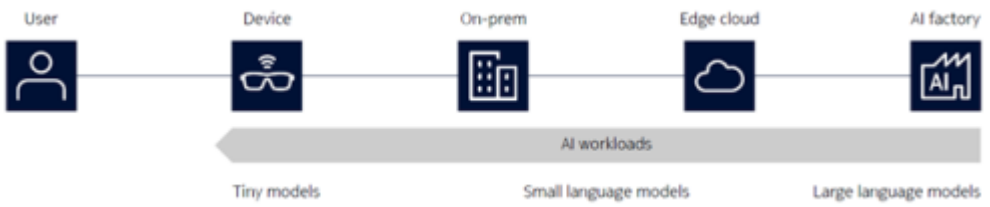

Figure 3. The network-cloud continuum
click to enlarge

The shift from centralized data centers to a highly distributed architecture has profound implications for network design. It's no longer just about adding capacity to support new use cases and AI-driven applications. The architecture of the network itself must be reimagined.

AI introduces requirements that stretch from the heart of the data center to the most remote points of end-user access. Delivering real-time intelligence at scale demands an end-to-end infrastructure built for low latency, massive bandwidth, and seamless coordination across every layer of the network.

Networks within data centers

Within AI factories, where the most intensive AI training and inference occurs, networks must operate at extreme speed, with exceptional reliability and near-lossless performance. Training large models requires enormous volumes of data to move in parallel across multiple lanes, tightly synchronized between GPUs. A single dropped packet or stalled flow can force the system to roll back to its last checkpoint, resetting hours or even days of work. Congestion can be just as damaging, leaving costly GPU resources idle while they wait for data or for other GPUs in the cluster to finish their part of the job.

In short, the data center network is becoming just as critical to AI performance as the GPUs themselves.

This demanding environment calls for an evolution in network architecture and platforms to deliver an essentially deterministic and lossless environment. Sophisticated telemetry and congestion-management techniques have become essential to maintain smooth data movement and maximize GPU utilization.

Ethernet, the world's dominant networking technology, is stepping up to meet these requirements and is becoming the preferred technology for scaling-out AI infrastructure worldwide. Link speeds are accelerating from 800 Gbps toward 1.6 Tbps and beyond. New mechanisms for load balancing and congestion control have emerged to handle the massive "elephant flows" characteristic of AI factories. And Ultra Ethernet, defined by the Ultra Ethernet Consortium (UEC), is reimagining Remote Direct Memory Access (RDMA) into an open and interoperable communications stack purpose-built for AI and High-Performance Computing (HPC) at scale.

The industry's shift toward Ethernet is no accident. It offers a broad, mature ecosystem, rapid innovation in speeds and protocols, global operational familiarity, seamless scalability, and true

multivendor flexibility. These strengths make Ethernet the most practical, future-proof foundation for AI networks.

Connecting the AI ecosystem beyond data centers

AI training workloads increasingly span distributed infrastructures, with GPU clusters spread across multiple data centers to address space, energy, and operational constraints. In this context, optimized network architectures enabling low-latency, high-bandwidth 'scale-across' interconnects are critical, as the network becomes the unifying fabric that transforms isolated facilities into a cohesive AI system.

Once models are trained, they must be efficiently delivered to inference locations, while large datasets must flow back to AI factories for training or fine-tuning. This demands robust, reliable cloud connectivity and seamless access networks that ensure AI workloads can be fed and consumed without bottlenecks.

Architecting for increased scale is paramount. AI applications can trigger a cascade of data requests and responses, leading to rapid traffic bursts that can overwhelm traditional networks. Without robust last-mile delivery, even the most advanced AI capabilities remain inaccessible.

To cope with these demanding requirements, several high-capacity connectivity options are available at both the IP and optical layers. The choice of technology depends on various factors, including distance, bandwidth requirements, latency, security considerations, and cost. Advanced network technologies, such as coherent optical engines, enable long-distance links with high speed, reliability, and energy efficiency, providing the backbone for this distributed AI ecosystem.

Securing and automating the AI network

In the AI era, network responsiveness is a key differentiator. Intelligent, reliable network automation allows networks to adapt dynamically to evolving demands of distributed AI workloads, optimizing performance and resource allocation in real time.

Equally important, AI is increasingly being embedded into network automation. Modern AIOps platforms are transitioning from concept to operational practice, providing capabilities such as conversational and context-aware interaction with the network, intelligent alarm correlation, accelerated root-cause analysis, and automated recommendations for remediation while maintaining operator oversight. By integrating AI directly into automation, network management becomes more proactive, efficient, and resilient.

Accompanying the expansion of data centers and the proliferation of AI workloads is the growing prevalence, volume, and sophistication of global cybersecurity threats. Protecting sensitive data is paramount. Technologies such as quantum-safe encryption and advanced DDoS detection and mitigation are essential, and many of these security functions must be embedded directly into the network. By integrating security at the infrastructure level, organizations can maintain high-speed, efficient operations without overburdening costly compute resources.

The network as the foundation for AI's future

Much like how the video revolution reshaped network architecture two decades ago, AI is now driving a profound evolution in how we conceive the cloud and the network.

Meeting AI's demands requires a forward-looking approach to network design and automation, both within individual data centers and across distributed infrastructures. By proactively transforming and evolving the network, organizations can establish the foundational infrastructure for a seamless and efficient cloud continuum—one that is resilient, adaptable, and ready to respond to whatever innovations the age of AI brings next.

This transformation is the essential foundation for unlocking AI's full potential and enabling the next wave of digital innovation.