CONNECTED NATION

A CONNECTED NATION WHITE PAPER

From Megabits to Milliseconds

Latency and America's AI Infrastructure Gap

Nathan Smith, Ph.D. · **Brent Legg** · **Jim Stewart**
Connected Nation, Inc. · July 2026

KEY TAKEAWAYS

What policymakers and broadband leaders should know

01 **The economic geography of AI will be set by deployment, not training.** Hosting a data center won't make a region the next Silicon Valley; the regions that thrive will be those where AI runs well for users.

02 **Latency, not bandwidth, is the new binding constraint.** Fiber and gigabit cable now reach most U.S. locations, but on a 1 Gbps link, network latency — not link speed — accounts for the great majority of perceived delay.

03 **Training depends on power. Inference depends on networks.** AI inference is governed by middle-mile fiber, peering density, and proximity to Internet Exchange Points (IXPs) — not by megawatts.

04 **Physical AI will make this a regional development issue.** Augmented reality, robotics, and agentic systems require ultra-low, ultra-reliable latency to nearby GPU capacity. Regions without it will see advanced applications arrive late, or not at all.

05 **A modest share of remaining BEAD funds (~$21B) could close the gap.** Targeted investments in middle-mile, IXPs, and regional GPU-as-a-Service capacity would protect the public's last-mile investment from being stranded by the next wave of demand.

INTRODUCTION

This Time Is Different

It is tempting to think that the places where artificial intelligence is "made" will be the places where its economic benefits accrue. That intuition has historical precedent. Detroit's rise as the center of automobile manufacturing reshaped the American Midwest for decades. Silicon Valley's role in the semiconductor and personal computing revolutions produced one of the most durable concentrations of wealth and innovation in modern history. As AI investment surges and ever-larger data centers are announced, it is natural for regions to hope that hosting AI infrastructure will make them the next great center of technological gravity.

But this time is different. Hosting a data center will not turn a region into Silicon Valley. Today, infrastructure matters in a different way.

AI is not a single technology but a stack, and different layers of that stack have different infrastructure needs. AI training, which produces new models, is increasingly constrained by power supply and transmission capacity. AI deployment—the inference phase where models are actually used—depends far less on power and far more on communication network performance. Especially important is latency: the "reflexes" of the internet, as distinct from bandwidth or throughput, which is often, somewhat misleadingly, called "speed."

This distinction will become more important as AI moves out of screens and into the physical world. As AI systems begin to augment human perception in real time, or guide machines through messy environments, milliseconds matter. Network performance requirements are poised to escalate, reshaping the geography of digital advantage and disadvantage and redefining what they mean. Some areas rich in last-mile fiber risk lagging behind because of a lack of local interconnection—though, fortunately, this problem is likely cheaper to solve than the tens of billions spent closing the last digital divide.

That earlier wave of investment has been remarkably successful. Decades of private and public effort have dramatically increased available bandwidth and closed most coverage gaps, even if some remote areas will remain dependent on satellite. Fiber deployment has accelerated to over ten million passings per year (see *The Broadband Competitive Landscape on the Eve of BEAD*). Even where fiber is absent, cable upgrades have made gigabit service widely accessible. Gigabit download speeds—far beyond the needs of most current applications—are now available to a majority of broadband service locations. The central constraint of the 2010s — insufficient throughput for streaming video — was not solved by accident. Carriers and ISPs competed their way past it, deploying capacity that now far exceeds what any mainstream application demands.

But success has changed the problem rather than solved it. The goalposts have moved.

As bandwidth has become abundant, other aspects of network performance have emerged as binding constraints. The most important of these is latency—the time it takes for data to travel across the network—and the interconnection structures that shape that journey. Latency is a "time tax" on every interaction that

gigabit bandwidth alone does not address. Indeed, as bandwidth increases, latency accounts for a larger share of total delay. While measurement remains imperfect, most places are likely already past the point where latency, not bandwidth, determines the effective "speed" of most internet use. As Jensen Huang highlighted in his recent keynote address at the GPU Technology Conference, agentic AI, which often does human-like web search tasks without being constrained by the speed of human thought processes, depends on ultra-low latency for high performance.

So far, this shift has been easy to overlook. As long as a human is in the loop, human perception imposes a natural ceiling: even mediocre latency feels instantaneous. But that constraint is disappearing. As AI systems begin to act in the world—guiding robots, aligning augmented reality overlays, coordinating machines—computer rather than human reflexes become the limiting factor. Latency becomes the binding constraint on syncing a local actuator with a remote "brain."

"AI infrastructure" is beginning to loom in policymakers' thinking as a key driver of competitiveness. It is—but distinctions matter. The infrastructure constraints on AI *training* are primarily about power. But training capacity, while economically meaningful, is not the main determinant of broad regional advantage. AI *inference*—where AI is actually *used*—depends far more on communication networks: local peering, interconnection density, and the efficiency of data routing. The emerging domain of physical AI will depend especially on these factors, requiring ultra-low latency and highly reliable connectivity.

At some point, a physical AI "killer app" will emerge—something that people expect to work seamlessly, everywhere. When it does, some regions will find that their networks cannot support it—not because they lack fiber to the home, but because their data must travel too far, through too many hops, to stay within its latency requirements. The frontier is shifting from megabits to milliseconds—from scarcity of throughput to the barrier of reflex speed. The telecommunications infrastructure ecosystem must evolve.

> *"The frontier is shifting from megabits to milliseconds — from scarcity of throughput to the barrier of reflex speed."*

SECTION 01

The Arithmetic of Lag Time

Why bandwidth is no longer the bottleneck.

Broadband policy has long focused on bandwidth, for good reasons. In the 2010s, demand growth was driven by consumption, especially streaming video. In 2020, remote work elevated the importance of upload speeds as homes became workplaces. Federal programs such as the Rural Digital Opportunity Fund (RDOF) and the Broadband Equity, Access, and Deployment (BEAD) Program reflect these priorities, targeting the performance metrics most relevant to legacy internet uses.

With or without subsidies, fiber optic cable deployment to homes and businesses has accelerated to the point that a majority of U.S. locations now have access to fiber-based services and gigabit download speeds—a share that continues to grow. Fiber providers often describe their networks as "future-proof," and in one sense this is likely true. The bandwidth fiber networks can deliver far exceed the requirements of any internet use case that is widespread today or plausibly on the horizon, and fiber-based services are upgradable not by replacing the fiber line itself but rather the optical gear used to "light" it. And escalating bandwidths transfer the speed pain point to a different performance metric. This is partly a matter of simple arithmetic.

Total delay in an internet interaction consists of two components: transmission time and latency.

Transmission time depends on how fast a connection can push bits through a medium, such as a fiber optic cable. In contrast, latency depends on distance, routing efficiency, and physical limits—most fundamentally, the speed of light, although many factors drive practical latencies far above that theoretical limit.

On a low bandwidth (and in that sense "slow") connection, transmission time dominates. For example, a 1 MB payload takes roughly 320 milliseconds to transmit over a 25 Mbps connection. Add 80–120 milliseconds of latency—a realistic figure for underserved areas where traffic must reach a distant hub—and most of the delay is bandwidth-driven.

By contrast, on a faster 1 Gbps connection, that same payload transmits in about 8 milliseconds. With that bandwidth, even a modest 35ms of latency means that over 80 percent of the delay now comes from the network itself rather than the link speed. In underserved areas, like the communities targeted by the BEAD program, round-trip latency to a distant hub commonly runs 80–120 milliseconds, making the case even stronger.

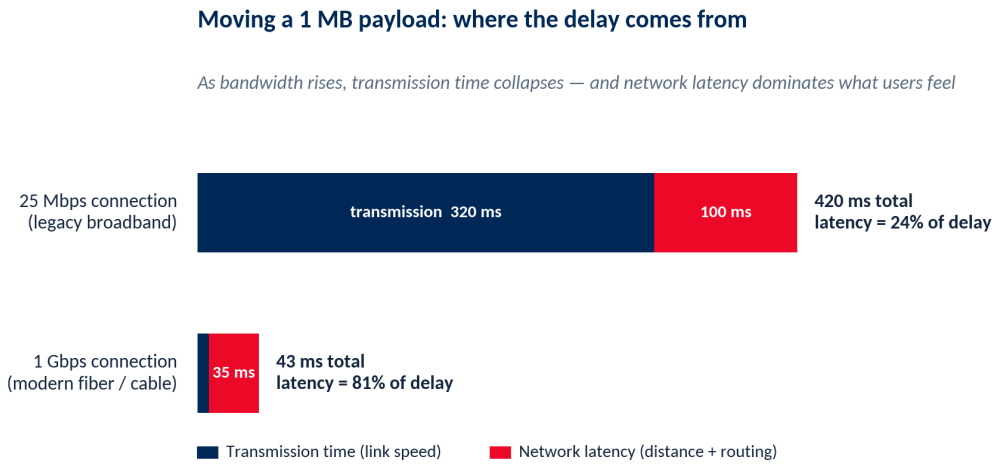

FIGURE 1 The arithmetic of lag time. On a 25 Mbps link, transmission time dominates. On a 1 Gbps link, the same 1 MB payload arrives in a fraction of the time — and latency becomes roughly 80 percent of the delay users feel.

As fiber becomes widespread, this inversion becomes universal. Latency, which has improved only modestly over time even as bandwidths have surged, is probably becoming the dominant source of lag time in most interactions with the internet, although this claim faces both statistical availability and definitional barriers to being firmly established.

Critically, the enormous investments and important innovations in last-mile connectivity do little to address this fact. Latency depends primarily on the length and efficiency of the data journey: middle-mile infrastructure, routing policies, and proximity to Internet exchange points (IXPs). Most milliseconds of latency lie outside the last-mile access network. Throughout this article, "IXP" is used to refer to a spectrum of interconnection facilities—from neutral traffic exchange points where networks peer directly, to carrier hotels, regional colocation facilities, and edge compute sites. These differ in ownership model and function, but share a common characteristic: they are the places where data changes hands between networks, and where proximity and peering density determine how efficiently traffic moves. Being near such places is key to achieving low latency.

Fiber expansion is on track to solving the bandwidth problem permanently in most places. But as deployment expands and bandwidth constraints recede, other aspects of network performance—especially latency—emerge as the new binding constraint.

SECTION 02

The Zoom Choir Problem

Why some applications break when milliseconds slip.

Have you ever tried to sing "Happy Birthday" over Zoom? It's awkward. While conversation works fine and one can almost forget the physical separation, singing together is another matter. It turns into cacophony. The main reason for that is latency.

Videoconferencing has become remarkably effective. Signals travel thousands of miles through fiber, routers, and data centers, yet conversation feels nearly instantaneous. During the COVID-19 pandemic, this capability preserved much of the organizational backbone of the economy. Because conversation is largely turn-based and asynchronous, delays of even a hundred milliseconds are barely noticeable.

Choral music is different. Its beauty depends on tight synchronization. Singers adjust continuously based on what they hear from others. Delays of just a few tens of milliseconds break that feedback loop. Many disappointed choirs during the pandemic, starved for shared singing, discovered fast and painfully how harmony falls apart in the face of transmission delay.

The "Zoom choir problem" is one of a class of problems where some functionality is dependent on tight synchronization and intolerant of latency. Similar problems occur in gaming and augmented reality. Expect them to proliferate as technologists seek to leverage AI to do more work in the physical world. Their

emergence will often be counterintuitive, because we don't usually analyze daily tasks, functions, and situations in terms of milliseconds of delay tolerance between events. Intuitively, if Zoom works fine for conversation, it should work for choral singing. But the hidden differences in delay tolerance make or break different applications.

Human perception and reflexes impose certain thresholds. Responses under roughly 100 milliseconds generally feel instantaneous. Delays of several hundred milliseconds are noticeable but tolerable. Multi-second delays disrupt engagement. But for tightly coupled feedback systems—music, gaming, driving, or robotic manipulation—even tens of milliseconds matter. And technologists could find uses for even tighter remote synchronization if it were discernibly and reliably available, especially in the substitution of remote compute for on-device compute.

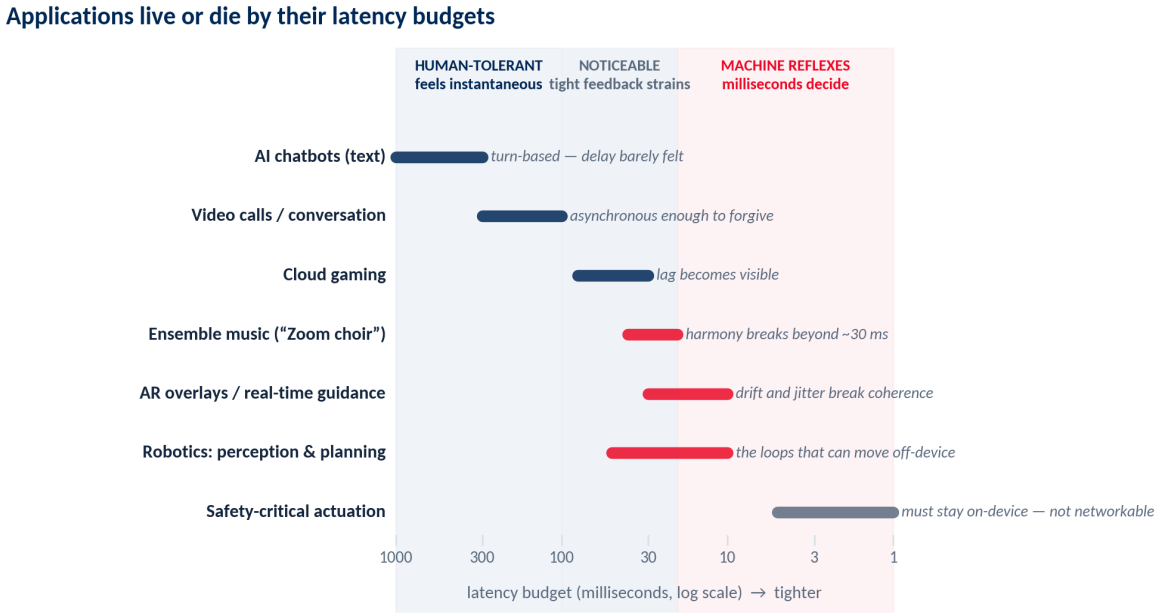

FIGURE 2 Applications live or die by their latency budgets. Turn-based uses forgive delay; tightly coupled feedback systems do not. Ranges are illustrative thresholds drawn from the discussion in this paper.

So far, most economically important AI applications live comfortably on the tolerant side of the divide.

Chatbots are not latency-sensitive. When they assist with writing and analysis, a response that arrives in a few hundred milliseconds feels instantaneous. Compute latency inside the model often dominates total response time anyway, for now. This balance could shift, as hardware and software improvements continue to reduce inference compute times, but the chatbot format, like conversation over Zoom, is inherently turn-based and asynchronous, and the human in the loop will continue to be satisfied with the tens of milliseconds that are typical today. Chatbot interactions, especially when they exchange only text, are typically even less dependent on bandwidth than latency, so they underscore the shift in what connectivity metric matters most at the margin. But they're not a driver of demand for better network performance.

Other LLM applications are more demanding. Systems that require real-time coordination—whether human–AI collaboration or machine control—are much less latency-tolerant than AI chatbots. Meanwhile, improving compute latency on the inference side will make *network* latency become a larger share of total delay.

It's important to keep in mind that none of this has anything to do with the business of AI training. Creating models and putting them to work are separate tasks.

SECTION 03

Networks, Not Just Power, Will Decide AI Deployment

Training needs megawatts. Inference needs milliseconds.

If training is where models are made, inference is where AI enters daily life.

A trained model is a large but movable digital artifact—hundreds of gigabytes to a few terabytes. Once trained, it can be replicated and deployed wherever sufficient compute and connectivity exist. From a user's perspective, it does not matter where the model was trained. What matters is where it runs.

Electricity still matters, but power availability is rarely the binding constraint. What matters most is network performance—especially latency and reliability. Compute is therefore placed in metropolitan data centers, carrier hotels, and cloud regions close to population centers, often in facilities with dense fiber connectivity and direct peering. Internet Exchange Points (IXPs), peering density, and middle-mile fiber are the key determinants of how low latency can get. IXPs allow networks to exchange traffic locally, reducing distance, cost, and jitter.

Regions without strong interconnection infrastructure may require traffic to traverse hundreds of miles to reach major hubs, imposing a small but persistent time tax on every interaction. While this may be individually trivial, aggregated across billions of interactions, it is quite meaningful.

Network infrastructure that enables low latency is thus becoming the general-purpose input to AI deployment, much as electricity is to LLM training. The network geography that determines latency performance for other traffic can differ systematically from the network geography that determines latency performance for AI inference, since the GPUs that power AI inference are not available at every Point of Presence on the network edge.

	AI TRAINING	AI INFERENCE
What it produces	New AI models	AI put to work for users
Binding constraint	Power supply and transmission capacity	Network performance — latency and reliability
Location logic	Location-flexible; follows megawatts	Follows users, workflows, and latency budgets
What it decides regionally	Economically meaningful, but not broad regional advantage	Where AI runs well — the basis of regional advantage

TABLE 1 Two tasks, two constraints. Creating models and putting them to work are separate businesses with different infrastructure needs.

SECTION 04

What "Near the Edge" Really Means

From content close to users to compute close to users.

Inference does not happen at training data centers for the same reason retail does not happen at factories: proximity matters. If training is heavy industry, inference is a distributed service business. It follows users, workflows, and devices—and increasingly it follows *latency budgets*. But just as not every retail store carries every item, not every "edge" location offers the same compute capabilities. An IXP can be nearby and still not have the right kind of infrastructure—or the right commercial offerings—to support modern AI workloads.

> *"If training is heavy industry, inference is a distributed service business. It follows users, workflows, and devices — and increasingly it follows latency budgets."*

This is where regional GPU deployments enter the picture.

For the last decade, the main "edge" story was the rise of content delivery networks (CDNs). CDNs brought content closer to users by caching popular files and serving them from geographically distributed points of

presence. This model economized on long-haul transport and, as a byproduct, reduced latency. It was especially well-suited to video streaming: lots of people repeatedly consume the same limited catalog of large files, so it's inefficient to haul those bits across long distances every time. Streaming isn't particularly latency-sensitive, but CDNs also support latency-sensitive web applications by keeping traffic local and reducing congestion and routing complexity.

Historically, that edge model optimized delivery and CPU-light compute. AI is revolutionizing the demand that the network edge must meet. Inference is often interactive, increasingly chained across multiple services, and frequently benefits from being close to the user or device. At the same time, the workloads that matter most—modern vision models, multimodal systems, voice, real-time translation, robotics/AR workloads, and high-volume API inference—are increasingly accelerator-bound. They run best (and often only economically) on GPUs or other specialized chips, not on general-purpose CPUs.

As a result, the edge is evolving from "content close to users" to "compute close to users," and CDNs and cloud providers are rapidly expanding GPU-backed inference footprints. Much inference remains centralized today, but edge inference is growing—because the constraint is shifting from *can you do it?* to *can you do it cheaply, predictably, and everywhere?* In places without nearby GPU-backed inference capacity, it can still be true that the latency to fetch cached video is lower than the latency to reach an inference endpoint. Text chat can tolerate that. But the next wave of applications—real-time collaboration, voice, AR overlays that must remain aligned, multi-agent systems that make frequent API calls, and eventually many physical-AI "control-adjacent" loops—will feel the penalty.

GPUs deployed regionally are the mechanism for closing that gap. It extends the CDN logic from caching bits to hosting accelerated compute: GPUs deployed in edge facilities so models can run nearer to end users, with lower round-trip latency and less reliance on distant hubs. This is the logic driving the rapid growth of what the industry is beginning to call "neoclouds"—providers purpose-built for GPU-dense, AI-optimized infrastructure, in contrast to the general-purpose hyperscalers that dominated the prior era. When it works well, local GPU deployment doesn't just make responses faster; it makes performance more consistent by reducing the number of long-haul hops and the exposure to congestion, jitter, and fragile routing paths.

But this is not a trivial upgrade from CPUs to GPUs. GPUs are optimized for massively parallel workloads—the core operation of modern neural networks—and inference increasingly depends on accelerators rather than general-purpose processors. That implies higher power density, more demanding cooling, different rack designs, and more careful siting. GPU-dense racks commonly draw 20–40 kilowatts, compared to 5–8 kilowatts for traditional edge infrastructure—a difference that many existing colocation facilities cannot support without significant electrical and cooling upgrades. This is not a counterargument to regional GPU deployment; it is precisely the reason to begin planning and investing now, before demand arrives and options narrow. In other words, regional GPU deployments require a real evolution of "edge" infrastructure, not just a software rollout.

> *"Edge is not only a geographic concept. It is an interconnection concept."*

And crucially, "edge" is not only a geographic concept. It is an interconnection concept. A facility can be physically close and still behave like it is far away if traffic must trombone through distant hubs before reaching major networks or cloud services. By contrast, a well-connected site—often one colocated with, or tightly integrated into, an IXP and regional carrier ecosystem—can deliver materially better latency and reliability even if it is not the closest building in miles. Dense peering and rich middle-mile options make the difference between "an edge site" and "the edge" in the practical sense that matters for AI.

This matters even more because AI systems are increasingly composed. A single user interaction may trigger calls to a vector database, a search index, a tool-using model, a safety service, and a specialized model—sometimes multiple times. These machine-to-machine chains compound latency and magnify the cost of poor interconnection. If every hop pays an avoidable distance tax because local peering is thin, performance degrades and the economics of deploying advanced services locally get worse.

That's where the policy gap bites. Broadband policy has historically focused on last-mile throughput—crucial work, and largely successful—but has paid much less attention to the interconnection layer that determines whether "nearby compute" is actually reachable as nearby compute. Regional GPU deployments can serve effectively where there is dense peering, robust middle-mile, and clear demand aggregation. But in places where interconnection is thin or absent, regional GPU deployment becomes harder to justify, and the regions that most need modern "edge compute" can be the last to get it.

Regional GPU deployments, including GPU-as-a-Service ("GPUaaS"), at well-interconnected edge facilities is therefore becoming critical infrastructure—not for training, but for deployment, and increasingly for physical AI. The emerging risk is not that AI won't exist in those places at all, but that the most advanced, latency-sensitive, and reliability-sensitive applications will arrive unevenly: first where GPU-backed inference and interconnection are already dense, and later—possibly much later—where the last-mile story succeeded but the peering story never began.

SECTION 05

From Screens to Bodies: The Rise of Physical AI

Why embodied AI raises the stakes for latency.

The humanoid robot has become a visual shorthand for AI's future. The imagery is misleading, insofar as it gives the impression that humanoid robots are everywhere when in fact they remain niche, with few or no economic applications. Publications use humanoid robots to symbolize AI because there's nothing visually impressive enough to draw clicks about a person using a chatbot via a laptop or a smartphone. But the

imagery accidentally carries an important truth: that to really live up to its promise, AI will need to come out of the screen and get physical.

In other words, AI must move from digital cognition to embodied action if it is to meet a wider range of human needs. Humans cannot eat text and images. Economic value increasingly requires machines that can cook, drive, see, fix or build things, move around, manipulate objects in the physical world, or something analogous. It's hard to forecast what the "killer app" will be that will mark the breakthrough of physical AI from nifty demonstration projects into mass economic impact. The history of technology forecasting is full of embarrassments. But some kind of transition from thoughts to things is surely in the cards. One killer app will change this dynamic forever.

Physical systems operate through nested control loops. The fastest loops must remain on-device for safety. Safety-critical actuation loops—the sub-5-millisecond signals that tell a motor to stop or a joint to hold—cannot tolerate network round-trips and are not candidates for remote compute. That layer will always live on the device. But higher-level perception and planning loops—operating at tens to hundreds of milliseconds—can be offloaded to remote compute, if connectivity is sufficiently reliable and fast. It is this planning layer—scene understanding, path selection, task coordination—where network latency becomes the binding constraint, and where regional GPU infrastructure makes the difference between a system that works and one that cannot.

This hybrid architecture is already common in robotics. Low-level control remains local; higher-level intelligence migrates into the network. The feasibility of that migration depends critically on ultra-reliable, low-latency connectivity—and increasingly on access to nearby compute. And increasingly that must be GPU.

SECTION 06

From AR to Robotics

AR as the proving ground — and forcing function — for low-latency GPU infrastructure.

Augmented reality (AR) provides a benign proving ground for embodied AI. It could be very impactful in its own right, by layering virtual knowledge and skills onto human sensorimotor capabilities. The future might, and likely will, feature jobs in which human workers serve as the "hands" for machine "eyes" and "brains." Or AR may matter primarily as a stepping stone, generating a new kind of data that can be used to train future robots in the skills of moving and manipulating objects like humans do. Either way, meeting the infrastructure needs of AR is important—not only for AR itself, but for what comes next.

In AR systems, a human provides dexterity and judgment while AI provides perception and guidance. An AR headset captures the scene and overlays instructions in real time. For this to function, overlays must remain spatially and temporally aligned. Even modest latency spikes cause drift and jitter, breaking perceptual coherence. Users disengage quickly. AR thus outsources perception to the network: if the network is slow or inconsistent, the system fails—not gracefully, but fundamentally.

Because AR is latency-sensitive but not usually safety-critical, it offers a testbed for pushing compute outward into GPUaaS edge environments. It also generates rich structured data about physical environments, feeding future AI models. As AR matures, it may act as a forcing function for low-latency GPU infrastructure near population centers. If AR breaks through into mass adoption, it will turn latency-to-GPUaaS into a regional economic development pain point. If it matters mainly as a data source for robot training, technologists will still need islands of reliable ultra-low latency to deploy inventions at scale.

The same latency constraints appear even more forcefully in robotics and mobility systems. Robots need not be humanoid to perform human-like work. The practical robots of today—mostly in warehouses and factories—do not look like people, but they perform tasks humans would otherwise have to do. What matters is not the humanoid form but humanlike flexibility in adapting to messy environments.

In these systems, a separation of "thinking" from "acting" is already emerging. Onboard systems handle immediate safety, while higher-level perception and coordination increasingly depend on shared compute resources. Autonomous vehicles benefit from low-latency connections to share hazard information and coordinate fleets. Drones, constrained by weight and power limits, may rely even more heavily on edge compute. But across applications, a pattern persists: robots' Achilles heel is rigidity. They perform well in controlled environments but struggle to adapt when conditions change.

That limitation points toward a hybrid architecture. LLMs, or neural networks applied to spatial data, may help bridge the gap between rigid automation and flexible intelligence. Combining robot "hands" with an LLM "brain" is difficult, but it is a plausible path toward machines that can operate in unstructured environments. What matters for present purposes is not the exact form these systems will take, but the infrastructure they require: nearby GPUaaS capable of delivering reliable ultra-low latency to insert flexible intelligence into perception-control loops.

The boundary between local and remote intelligence shifts as latency falls. Tasks vary along dimensions of time sensitivity and compute intensity, and the hardest tasks are those that are both time-sensitive and compute-intensive. Compute is generally cheaper and more powerful in centralized environments, but only if latency is low enough to preserve acceptable "reflexes." As latency constraints relax, more of the "thinking" can migrate off-device.

This is the common thread linking AR and robotics. AR reveals the problem in a tolerable form; robotics amplifies it into a hard constraint. In both cases, the practical deployment of embodied AI depends on whether networks can deliver compute quickly, reliably, and locally enough to close perception-action loops. GPUaaS at the network edge is therefore not just an optimization—it is a prerequisite for moving AI from screens into the physical world.

SECTION 07

The Next Digital Divide Is a Latency Divide

Where the gap will show up next — and why.

The next digital divide will not be about bandwidth. Even in the few places where the slowness of streaming video is still a pain point, many cutting-edge applications may, like AI chatbots, be data-light. The fiber buildout will continue apace and keep producing benefits, but a public policy focus on last-mile fiber deployment alone will not suffice to stay at the cutting edge, nor promote the US's pursuit of global AI dominance. Latency is the space where public policy and private sector investment must now concentrate.

Latency remains poorly measured and poorly mapped. The FCC has dramatically improved public broadband mapping for throughput, but no comparable public-facing maps of interconnection quality or effective latency exist. That makes latency reduction difficult for policymakers to target. Latency mapping needs to improve, but infrastructure planning can't afford to wait for better data. Investments must get moving.

Some states have fiber everywhere but lack IXPs. In such places, much traffic has to follow "trombone" routes—traveling hundreds of miles out of state before returning. A packet originating in Salt Lake City, for example, historically traveled to Los Angeles or Denver to reach a major exchange point before returning to its in-state destination—adding 40 or more milliseconds to what should have been a local transaction. Such tromboning has been happening for a long time, and the extra milliseconds of delay usually haven't been particularly noticeable, but the goalposts keep moving, and what was good enough soon won't be.

> *"Failures will not appear as outages. They will appear as unreliability — systems that work elsewhere but not locally."*

Failures will not appear as outages. They will appear as unreliability, inconsistency, systems that work elsewhere but not locally. There will be a reputational cost of being the place where the latest AI-enabled system could not function properly while broadband leaders worked through the process of getting an IXP stood up. It's worth getting ahead of this.

SECTION 08

Infrastructure Strategy for Fast Reflexes

From last-mile to middle-mile, from passings to peering.

Last-mile fiber is probably future-proof for bandwidth, as its champions like to say, but it is not a complete solution. It does little to improve latency. Until recently, latency has been a secondary concern, even as reducing it quietly motivated major technological advances—from low-Earth orbit (LEO) satellite

constellations to content delivery networks. Now, the abundance of bandwidth is elevating latency from a background issue to the primary connectivity constraint. That shift transforms the infrastructure agenda. The question is no longer simply how much data can move, but how quickly systems can respond. Broadband leaders should begin to pivot accordingly—from last-mile to middle-mile, from passings to peering, and from pole attachments to network interconnection.

This is not an abstract optimization. It is about enabling a new class of systems with near-instantaneous "reflexes": machines that perceive, decide, and act in tight coordination with humans and with each other. Augmented tools that guide skilled work in real time. Vehicles and drones that coordinate safely at scale. Robots that can adapt to unstructured environments rather than fail outside narrow routines. These capabilities will not emerge everywhere at once. They will emerge first where networks can reliably deliver compute within tight latency budgets—and they will bypass regions where they cannot.

That reality should inform federal and state decisions about the disposition of remaining BEAD funds—the roughly $21 billion left after last-mile commitments. Having invested heavily to close the coverage gap, it would be a costly mistake to stop just short of usability for the next generation of applications. Regions with excellent fiber access could still find themselves at a disadvantage if weak interconnection forces data onto long, inefficient routes. The result will not be outages, but something more insidious: systems that work elsewhere but not locally. A relatively modest share of BEAD surplus funds could instead catalyze high-leverage investments in middle-mile routes, Internet exchange points, and regional interconnection hubs—unlocking the conditions needed for low-latency, GPU-enabled services to take root.

States that wish to remain competitive in the age of physical AI must think beyond megabits. They must invest in the infrastructure that allows intelligence to reside near users: dense peering, IXPs, resilient high-capacity transport, and regional GPU capacity at the network edge. The AI economy will not announce itself first through massive training clusters. It will show up in systems that feel immediate, responsive, and reliable—systems that extend human capability into the physical world. And where those systems flourish will depend not on where AI models were built, but on where networks can give them fast, dependable reflexes.

ABOUT THE AUTHORS

Who Wrote This Paper

Nathan Smith, Ph.D. — Director of Economics & Policy at Connected Nation, Inc.

Brent Legg — Executive Vice President & IXP.US Group Lead at Connected Nation, Inc.

Jim Stewart — Strategic advisor at Connected Nation; recently retired as Chief Technology Officer of the Utah Education and Telehealth Network (UETN).

REFERENCE

A Quick Glossary

Bandwidth / Throughput. How many bits a connection can move per second. Often called "speed," somewhat misleadingly.

Latency. The time it takes for a single packet to make the round trip across the network. Governed by distance, routing efficiency, and ultimately the speed of light.

BEAD. Broadband Equity, Access, and Deployment Program — the $42.5B federal program funding last-mile deployment, with roughly $21B remaining after state allocations.

RDOF. Rural Digital Opportunity Fund — earlier FCC program targeting rural last-mile coverage.

IXP (Internet Exchange Point). A neutral facility where networks meet to exchange traffic directly, avoiding longer routes through distant hubs. In this piece, used broadly to include carrier hotels, regional colocation, and edge compute sites — places where data changes hands between networks.

Peering. Direct network-to-network traffic exchange, typically at an IXP, that bypasses the public internet's longer paths.

Middle Mile. The long-haul fiber routes connecting last-mile networks to internet backbones and exchange points.

CDN (Content Delivery Network). Distributed caches that store popular content close to users to reduce long-haul transport.

Edge Compute. Compute resources placed close to end users, reducing round-trip time for latency-sensitive applications.

Inference vs. Training. Training produces an AI model (power-intensive, location-flexible). Inference runs the model for users (latency-sensitive, locally constrained).

GPU / GPUaaS. Graphics Processing Units — specialized chips that run modern AI workloads. GPU-as-a-Service rents capacity remotely instead of requiring on-premises hardware.

Neocloud. New-generation cloud providers purpose-built for GPU-dense, AI-optimized workloads, in contrast to general-purpose hyperscalers.

Trombone Routing. When local traffic loops out of region to reach an exchange point before returning — e.g., Salt Lake City traffic historically routing through Denver or LA, adding 40+ ms.

Connected Nation co-owns Connected Nation Internet Exchange Points, LLC (d/b/a, IXP.US), a 50/50 joint venture with carrier hotel pioneer Hunter Newby to develop neutral Internet Exchange Point (IXP) facilities across the United States. This initiative is an expression of Connected Nation's mission to ensure that no community is left behind and that the internet works for everyone today and into the future.

connectednation.org · **ixp.us**