ADDENDUM SE-1

LATENCY BUDGETS AND EDGE-AI PERFORMANCE REQUIREMENTS FOR PRIORITY BROADBAND PROJECTS*A*

Supplemental Engineering Narrative to:

Priority Broadband Projects in the AI Era: Statutory Mandates, Engineering Realities, and Long-Term Oversight Requirements Under the NTIA BEAD Program

DECEMBER 2025 EDITION

David J. Malfara, Sr.
Broadband Strategist/Principal Consultant – Big Bang Broadband LLC
Senior Member, Institute of Electrical and Electronics Engineers (IEEE)

FOREWORD

This Technical Addendum is issued as a companion to *Priority Broadband Projects in the AI Era* and expands on one of the most consequential engineering challenges facing BEAD-funded networks: the emergence of strict latency budgets as the governing performance constraint for real-time, AI-enabled applications. As artificial intelligence becomes embedded in industrial operations, public services, healthcare, logistics, agriculture, and a host of other mission-critical environments, the tolerance for latency, jitter, and upstream instability collapses. Networks that once served primarily as conduits for consumer traffic now function as components of timing loops, control systems, and distributed computational pipelines.

Congress anticipated the need for future-proof infrastructure when it defined the obligations of a Priority Broadband Project. What was not fully understood at the time—but is now undeniable—is that “evolving needs” and “advanced services,” as referenced in 47 U.S.C. § 1702(a)(2)(I), have materialized faster and with more demanding performance characteristics than traditional broadband paradigms ever contemplated. The AI Era is not a distant horizon; it is the operational reality during the BEAD construction period itself.

This addendum explains why deterministic latency, stable transport, and end-to-end architectural integrity must now be treated as statutory obligations, not optional enhancements. It provides a narrative framework for policymakers, state broadband offices, and PBP subgrantees to understand how AI Era performance requirements translate directly into compliance responsibilities during the ten-year performance period.

In publishing this addendum, Big Bang Broadband LLC seeks to ensure that BEAD investments yield durable, economically transformative infrastructure capable of supporting the emerging AI economy for the communities they serve.

CONTENTS

EXECUTIVE SUMMARY

The emergence of artificial intelligence (AI) as a real-time, latency-sensitive workload transforms how broadband networks must be engineered, evaluated, and governed. This Technical Addendum SE-1 provides the engineering foundation underlying the statutory interpretation and compliance recommendations presented in the main white paper, *Priority Broadband Projects in the AI Era*. It explains why Priority Broadband Projects (PBPs) must be treated as **end-to-end timing systems**, not simply as high-capacity access links, and why deterministic latency performance—not speed—is the defining requirement for supporting the "advanced services" contemplated in 47 U.S.C. § 1702(a)(2)(I).

AI workloads already impose strict timing constraints on networks, with allowable end-to-end delays frequently measured in tens of milliseconds and jitter tolerances measured in single-digit millisecond ranges [1]–[5]. These constraints are not future projections; they are current engineering realities. As inference workloads migrate toward the network edge and increase in density, tolerance for latency variance shrinks, and the ability of a broadband network to support real-time AI becomes dependent on the stability of the entire transport chain—from premises to local transport, aggregation, middle mile, and compute adjacency.

The BEAD statute anticipates this evolution. The PBP definition requires networks to (1) deliver specific performance attributes and (2) "easily scale" to meet evolving household and business needs, including the support of "other advanced services" throughout a ten-year performance period [6]. In today's technological context, these evolving needs include real-time AI workloads across manufacturing, healthcare, agriculture, logistics, public safety, utilities, education, and residential automation. A PBP that cannot maintain deterministic latency cannot support these services and therefore cannot satisfy its statutory obligations.

This addendum demonstrates that PBP compliance cannot be assessed through conventional speed tests, consumer feedback, or one-time activation checks. Compliance must be evaluated through **periodic, structured, engineering-grade assessment** of latency, jitter, tail latency excursions, route stability, queue dynamics, and compute-adjacent timing behavior. These measurements must be performed under realistic load because timing pathologies typically emerge during peak utilization, not idle periods.

This document is intended for state broadband offices, network architects, policymakers, and oversight personnel responsible for ensuring that BEAD-funded networks remain capable of supporting AI-driven advanced services throughout the ten-year performance period. The engineering standards and statutory interpretations presented here establish a robust, defensible basis for PBP oversight, ensuring that BEAD-funded networks become durable, future-ready infrastructure capable of supporting the distributed AI economy.

PART I — THE EMERGENCE OF LATENCY AS A DETERMINISTIC ENGINEERING CONSTRAINT

The rise of artificial intelligence (AI) as a real-time, inference-driven workload has transformed latency from a performance statistic into a **deterministic engineering constraint**. In traditional broadband contexts, latency was viewed as a secondary consideration—important for quality of experience but not a defining factor in whether a network could support a given category of applications. Today, this assumption no longer holds. Modern AI-enabled systems depend on **strict, end-to-end timing guarantees**, measured not simply in average latency but in the **predictability** of latency across an entire network path.

For clarity, **deterministic latency** refers to the ability of a network to maintain *stable, bounded, and low-variance* latency from one endpoint to another. It is not enough for average latency to be low. If jitter (latency variance) is high or if *tail latency*—the worst-case delay observed across a distribution—exceeds a narrow window, AI systems break down. Tail latency in particular is central to distributed inference: even occasional spikes of 80–150 ms can cause inference misalignment, sensor desynchronization, or full pipeline collapse in workloads that require end-to-end timing envelopes between 20 ms and 100 ms [1], [3], [4].

This transformation of latency into a deterministic requirement is driven by the increasing prevalence of **distributed inference chains**, in which data flows through a sequence of tightly-coupled processing stages:

- sensor capture,
- pre-processing,
- edge or near-edge inference,
- possible cloud escalation, and
- response or actuation.

Each stage consumes a portion of a finite, uncompromising **latency budget**. If the upstream network introduces delay or jitter that exceeds the allowable envelope, the entire chain collapses, regardless of how fast the access link is or how much bandwidth is available [1]–[5]. This is why prior broadband frameworks—which treated access-layer speed as the primary indicator of performance—are insufficient for AI ERA applications.

Crucially, the strictness of these latency budgets is not a speculative argument; it is repeatedly confirmed in peer-reviewed engineering studies. Machine-vision systems commonly require **50–100 ms** end-to-end timing windows and become unstable when jitter exceeds **10–20 ms** [3]. Teleoperation and remote-control systems require even tighter windows—often **<100 ms** end-to-end—and degrade precipitously when tail latency grows beyond **150–200 ms** [4]. Sensor-fusion environments, common in industrial and utility deployments, require sub-50-ms coherence between sensors; once timing drift exceeds this boundary, the system can no longer align or interpret data reliably [1], [2]. The literature is consistent: latency variance—not throughput—is the dominant factor in system viability.

This reality has immediate implications for the BEAD statute. Under **47 U.S.C. § 1702(a)(2)(I)**, Priority Broadband Projects must both **deliver defined performance characteristics** and **easily scale to meet evolving connectivity needs**, including support for “other advanced services” over a ten-year performance period [6]. In today's environment, advanced services are not hypothetical—they are AI-dependent systems with strict timing requirements. Thus, the ability to maintain deterministic latency must be treated as a **statutory obligation**, not a discretionary optimization.

Yet much of the upstream latency variance that breaks these AI workloads originates outside the access layer. Research consistently shows that **aggregation networks, local transport, and middle-mile paths** are the true sources of harmful jitter and tail latency [4], [5]. Misconfigured queues, oversubscribed upstream links, dynamic routing reconvergence, asymmetric paths, and traffic microbursts each consume portions of the latency budget. Even fiber networks—when dependent on shared distribution infrastructure—can exhibit destructive latency excursions under real-world load conditions if upstream segments are not engineered for deterministic performance.

This upstream sensitivity is particularly important because AI workloads are increasingly executed **near the edge**, where compute adjacency matters. *Compute adjacency* refers to the physical and logical proximity between end-user devices or sensors and the nearest inference-capable compute resource. When compute shifts toward regional or local facilities, the transport path between the user’s premises and the compute node becomes the critical determinant of whether AI workloads fall inside or outside their latency budget. This shift makes deterministic performance not merely a desirable trait but the **primary engineering requirement** for AI ERA broadband infrastructure.

For Priority Broadband Projects, this means that compliance must extend beyond initial installation. A PBP cannot be considered compliant simply because it delivers a high-speed access connection on day one. It must maintain deterministic performance **over time**, across seasons of utilization, evolving device density, new industrial applications, and shifts in traffic patterns. This is a profound shift in how broadband networks must be evaluated. AI makes clear that the relevant question is not, *“How fast is the network?”* but rather, *“Can it maintain stable timing under real-world, real-load conditions?”*

This shift reframes the statute’s scalability requirement. For decades, scalability referred primarily to increasing bandwidth. But with AI, **scalability now includes temporal scalability—**the ability to maintain deterministic latency even as inference frequency, workload concurrency, and data movement demands increase over the ten-year performance horizon. If a network loses timing stability as utilization grows, it is not scalable in the statutory sense, regardless of its initial throughput.

Finally, because timing stability cannot be measured through consumer speed tests or spot-checking techniques, deterministic latency must be evaluated through **structured, periodic, and engineering-grade audits**. These audits must assess end-to-end latency under load, jitter distribution, tail-latency behavior, route stability, and the influence of upstream congestion—all factors repeatedly documented in the engineering literature as determinative for AI ERA functionality.

In sum, latency is no longer a behavioral characteristic of broadband networks. It is a **hard performance limit** dictated by AI, measurable through established methodologies, and enforceable through clearly stated statutory obligations. Priority Broadband Projects must therefore be designed, monitored, and maintained as deterministic timing systems from the moment they are constructed through the full ten-year performance period.

PART II — LATENCY BUDGETS AS A SYSTEM-LEVEL CONSTRAINT IN BEAD-FUNDED NETWORKS

The engineering literature establishes that latency budgets are not attributes of individual network segments but of the **entire service path** connecting premises to compute. This makes latency inherently a **system-level constraint**. Each component—access, local transport, aggregation, middle mile, and compute adjacency—consumes a portion of a finite timing envelope. If any segment introduces delay, jitter, or instability beyond its allowable allocation, the end-to-end workflow fails, even if every other segment performs well [1]–[5]. In the AI Era, this dependency is absolute.

To appreciate why, it is necessary to understand that timing constraints arise from the **structure of distributed inference pipelines**, not from any particular network technology. A typical AI-enabled workflow—such as machine vision, teleoperation, robotics, or real-time sensor fusion—relies on a tightly choreographed sequence of processing stages. Each stage is synchronized to the next, and the entire sequence must complete within a strict end-to-end latency budget. These budgets are unforgiving. When machine-vision pipelines require 50–100 ms round trip to maintain frame coherence, or when teleoperation systems require <100 ms latency with minimal jitter to ensure control stability, even occasional delay spikes compromise system integrity [3], [4]. This is why timing must be evaluated not as an average but as a **distribution**.

Tail latency—the worst-case delays occurring during congestion or routing instability—is the single most important metric for system reliability. Research shows that even when average latency meets requirements, tail latency excursions of 100–150 ms can desynchronize inference stages, disrupt control loops, or cause sensor misalignment [1], [3], [4]. These failures often occur under load, which means they are invisible to idle-hour measurements or traditional consumer speed tests. A network can appear fast while being incapable of sustaining deterministic performance during real-world peak conditions.

This has direct consequences for the statutory obligations of Priority Broadband Projects under **47 U.S.C. § 1702(a)(2)(I)**. The statute requires PBPs to meet defined performance attributes and to "easily scale" to support "evolving connectivity needs" and "other advanced services." In today's technological landscape, these advanced services are dominated by AI-enabled workloads with strict timing requirements. A PBP that cannot maintain deterministic timing across its full path cannot support these services and therefore does not satisfy the statutory definition of a PBP [6].

One of the most important insights from the engineering literature is that **most harmful latency originates upstream from the access link**. Fiber, fixed wireless, or hybrid access technologies can perform extremely well at the edge but still fail to support real-time inference if local transport or aggregation paths introduce jitter or queueing delay [4], [5]. Routing asymmetry, traffic microbursts, upstream oversubscription, and dynamic convergence events in the middle mile are the predominant sources of timing instability. This is why focusing on access-layer speed—as most broadband regulatory frameworks historically have—is insufficient. Deterministic timing must be validated across the entire transport chain.

Computing architecture magnifies this requirement. As AI inference increasingly shifts to **regional and near-edge compute nodes**, the distance and stability of the path from the customer premises to these compute resources becomes decisive. This path—often overlooked—can determine whether AI-enabled processes remain within their latency budgets. This concept, known as **compute adjacency**, emphasizes that real-time performance depends on the *network between the user and the compute resource*, not only on the local access connection. As inference becomes more decentralized, adjacency becomes more important, and deterministic performance becomes indispensable.

These system-level dependencies make clear that PBP compliance cannot be validated using consumer-grade testing methods. **Speed tests cannot measure system viability**. They cannot detect jitter envelopes, tail latency excursions, upstream queue buildup, or route instability. They only measure peak throughput on a lightly loaded path. For AI ERA workloads, this is essentially irrelevant. Compliance must instead rely on **periodic, engineering-grade measurements** that evaluate the timing behavior of the entire system under realistic load.

A proper compliance framework must therefore assess:

- end-to-end latency under peak or near-peak conditions,
- jitter distribution across representative load cycles,
- tail latency excursions during congestion events,
- aggregation and middle-mile queue dynamics,
- route stability, symmetry, and reconvergence timing, and
- compute-adjacent timing under load.

These are the phenomena that determine whether a network can support distributed AI workloads, and they are precisely the characteristics documented across the engineering literature [1]–[5]. A PBP that satisfies these timing requirements is scalable in the statutory sense. A network that satisfies only throughput targets is not.

Finally, because the BEAD statute introduces a **ten-year performance period**, determinism must be maintained over time. Timing stability is not static; research shows that even well-engineered networks drift under increasing load, evolving traffic patterns, and changing interconnection conditions [4], [5]. Therefore, system-level timing performance must be periodically evaluated to ensure that PBPs remain compliant throughout the full performance period—not merely at activation.

In sum, latency budgets constitute the engineering foundation upon which statutory compliance must be evaluated. A PBP is not a high-speed access link; it is an end-to-end timing system. And in the AI Era, system-level timing performance—not speed—determines whether the network is fit for its intended statutory purpose.

PART III — PERIODIC AUDITS AND TEMPORAL INTEGRITY ACROSS THE TEN-YEAR PERFORMANCE PERIOD

The statutory design of the BEAD program requires Priority Broadband Projects to remain compliant for a full decade following certification. This **ten-year performance period** distinguishes BEAD from all previous federal broadband programs and fundamentally alters the obligations of network operators. A PBP is not merely a construction project; it is a **long-term operational commitment** to sustain network performance in accordance with federal law.

This long horizon intersects directly with the deployment and mainstream adoption of real-time AI systems. As documented throughout the engineering literature, AI-enabled workloads are dominated by **strict timing requirements**—often with allowable end-to-end delays measured in tens of milliseconds and with jitter tolerances measured in single-digit millisecond ranges [1]–[5]. These workloads depend on **stable, predictable timing behavior** throughout the entire network path. Any drift in that behavior can break inference pipelines, disrupt sensor coherence, destabilize control loops, or render teleoperation or machine-vision systems inoperable [3], [4], [5].

To understand why periodic audits are essential, it is necessary to recognize that network behavior **changes over time**. Even high-quality networks experience natural timing drift as utilization patterns evolve. Research shows that:

- device density increases,
- inference workloads become more frequent,
- traffic microbursts appear as sensors proliferate,
- routing policies shift due to upstream changes,
- new services introduce unpredictable bursts of latency, and
- congestion events increase tail latency [4], [5].

A network that meets timing requirements at activation may fail to do so in year five or year eight. For example, a fiber-based PBP may deliver sub-20-ms jitter and stable 40-ms end-to-end inference loops at launch. But as additional sensing systems, municipal IoT devices, local robotics, or industrial automation come online, queueing behavior in aggregation layers may cause periodic tail latency spikes above 100 ms. These spikes—rarely visible in traditional speed tests—are catastrophic for AI-driven systems.

This reinforces the importance of understanding **tail latency**, which refers to the worst delays observed in a distribution. Policymakers often focus on "average latency," but the engineering literature consistently shows that AI ERA workloads are sensitive not to averages but to **the far-**

right tail of the distribution. Even small numbers of high-delay events can break real-time applications [1], [3], [4]. A network can be “fast” in consumer terms while being unusable for advanced services.

Understanding this distinction is essential for interpreting the statute. Under **47 U.S.C. § 1702(a)(2)(I)**, Priority Broadband Projects must not only deliver defined performance attributes but must also “easily scale” to meet **evolving needs** and support “other advanced services” over the full performance period [6]. In today’s environment, those evolving needs are dominated by real-time AI systems. If a network fails to maintain deterministic timing as those needs evolve, it is no longer scalable in the statutory sense.

This requirement persists even if **ownership of the network changes**, or if the operational environment evolves dramatically. The statute makes the obligation inherent to the network itself—not to a particular operator’s initial configuration. This means states must have compliance mechanisms that remain valid across operator changes, management transitions, mergers, or acquisitions during the ten-year window.

Because timing stability is dynamic, PBP oversight must rely on **periodic, structured, engineering-grade assessment**, not on one-time certification. Traditional broadband compliance methods—consumer surveys, spot speed tests, high-level reporting—cannot detect timing drift or tail latency behavior. A PBP could appear fully compliant while, in reality, being unable to support AI-enabled services during peak periods.

This is why periodic audits must include:

- full-path latency measurements under realistic load,
- jitter distribution analysis,
- tail-latency envelope evaluation,
- route stability and path symmetry examination,
- aggregation and middle-mile queue dynamics, and
- compute-adjacent timing assessment.

These audit elements are exactly the characteristics documented in the engineering literature as essential to AI ERA application viability [1]–[5]. They are the only way to ensure that timing stability remains intact across the ten-year period.

The Priority Broadband Project Operational Framework developed by Big Bang Broadband LLC directly addresses this requirement. By treating PBPs as timed systems rather than static access networks, the Framework enables states to verify compliance using structured, high-resolution timing assessments capable of detecting drift before AI-enabled services begin to fail. This operational discipline helps ensure that communities receive the long-term economic and service benefits Congress intended.

In short, **temporal integrity**—the sustained ability of a network to deliver deterministic latency across time—is the defining characteristic of a Priority Broadband Project. Periodic audits are not ancillary or optional; they are the operational tool by which statutory obligations are verified and maintained. Without them, neither states nor subgrantees can ensure that BEAD-funded networks remain capable of supporting the advanced, latency-sensitive services that will shape regional economic competitiveness for the next decade.

PART IV — EVOLVING WORKLOAD INTENSITY AND THE SHRINKING MARGIN FOR LATENCY VARIANCE

As artificial intelligence becomes more deeply embedded in real-world operations, the margin for latency variance steadily decreases. This phenomenon is counterintuitive for many policymakers and even for experienced network operators, because traditional broadband engineering has long assumed that networks become more efficient as utilization patterns stabilize. In the AI Era, the opposite is true: **the more widely AI is adopted, the stricter its timing requirements become**, and the less tolerance networks have for latency drift, jitter, or tail-latency excursions [1]–[5].

This shift is driven by the fact that AI ERA applications are not monolithic. Instead, they consist of **multiple concurrent, real-time inference chains** that often operate in parallel:

- machine-vision frames processed dozens of times per second;
- sensor-fusion pipelines synchronizing data from distributed IoT devices;
- robotics control loops executing continuously;
- teleoperation sessions with human-in-the-loop dynamics;
- municipal and environmental sensors generating microbursts;
- industrial or utility systems performing edge inference at increasing frequency.

Each pipeline consumes part of the available end-to-end latency budget. As more pipelines operate concurrently—which is precisely what happens during AI adoption—the effective allowable variance shrinks, because real-time workflows become more interdependent. This is sometimes referred to as **latency stacking**, where multiple inference processes amplify each other's sensitivity to jitter and tail latency [4], [5]. A delay that might be tolerable in isolation becomes catastrophic in a multi-pipeline environment.

The research supporting this tightening effect is consistent across engineering domains. Machine-vision workloads operating at 20 frames per second may tolerate 50–80 ms latency variance at deployment, but as inference frequency scales to 40 or 60 frames per second, the allowable jitter window may shrink to **10–20 ms** [3]. In distributed robotics, control-loop timing that initially tolerates <100 ms end-to-end latency may tolerate only half that under more sophisticated or concurrent workloads [5]. In IoT inference systems, increased device density naturally produces traffic microbursts that inflate queueing delay and tail latency even when average utilization remains low [2]. These effects are not theoretical—they are documented empirical phenomena.

This dynamic has powerful implications for Priority Broadband Projects. Under **47 U.S.C. § 1702(a)(2)(I)**, PBPs must "easily scale" to support evolving household and business needs and must remain capable of supporting "other advanced services" over the ten-year performance period [6]. But “scaling” in the AI Era does not simply mean supporting more devices or more bandwidth. It means supporting **increasingly strict timing requirements** as AI adoption intensifies. A PBP that maintains deterministic timing at activation but loses timing stability as load evolves is not scalable in the statutory sense.

This reality also underscores why deterministic performance must be engineered into PBPs **at the outset**. Timing stability cannot be retrofitted easily once a system is built. Aggregation and middle-mile infrastructure may require architectural changes if they cannot maintain bounded jitter or if they are prone to tail-latency events during moderate or peak loads. Routing policies may need to be constrained to ensure path symmetry and stability. Oversubscription ratios must be engineered with margin for inference growth—not simply for throughput. Without proactive design, networks will drift outside their latency budgets as AI workloads proliferate.

For state broadband offices, this means that PBP oversight must anticipate that **timing requirements tighten naturally over time**. Compliance cannot be measured against the conditions that prevail at the moment of network activation. Those conditions are transient and typically represent the least challenging period in a network’s operational life. The statute’s ten-year performance window is designed to ensure that networks remain viable not under initial conditions, but under **future conditions**—precisely when timing constraints become most strict.

Further, the engineering literature makes it clear that these tightening dynamics can produce **binary failure modes**. Real-time AI systems often do not degrade gracefully. When timing slips beyond the acceptable window, systems do not perform “somewhat worse”—they fail outright [3], [5]. Machine-vision pipelines become incoherent. Teleoperation becomes unstable. Distributed robotic systems misalign. IoT inference loops lose synchronization entirely. In policy language, this means that networks either meet timing requirements or they do not. There is no partial compliance.

This binary nature reinforces the statutory argument: A network that cannot maintain deterministic latency at increasingly narrow tolerances is not capable of supporting advanced services for the full performance period and therefore ceases to qualify as a Priority Broadband Project. Periodic auditing is the only mechanism by which drift can be detected early enough to prevent catastrophic failure of AI-enabled services.

Finally, the accelerating pace of AI adoption ensures that timing pressures will increase over the BEAD performance period. Manufacturers, hospitals, utilities, logistics operators, and households are deploying AI-enabled systems at an unprecedented rate. The BEAD deployment timeline coincides with the period in which these systems will become standard practice across sectors. Therefore, deterministic timing is not a forward-looking aspiration—it is an immediate engineering requirement central to BEAD compliance.

In short, as AI workloads intensify, the acceptable margin for latency variance **shrinks**, not expands. This makes deterministic timing performance a moving target that must be monitored and preserved over time. Without intentional engineering and structured oversight, networks will naturally drift outside the timing boundaries required by AI workloads. Because BEAD’s statutory definition of a PBP requires long-term support for advanced services, the ability to maintain deterministic timing must be treated as a core statutory obligation. Compliance depends on it.

PART V — LATENCY BUDGETS AS THE ENFORCEMENT BACKBONE OF PRIORITY BROADBAND PROJECT COMPLIANCE

Latency budgets do more than define the engineering parameters of AI ERA networks—they provide the **enforcement mechanism** by which Priority Broadband Project compliance can be credibly maintained throughout the ten-year performance period. A PBP is not defined by its construction attributes alone. It is defined by its sustained ability to support advanced, time-sensitive services over a decade, regardless of changes in utilization, ownership, or traffic conditions. Latency budgets give states and federal agencies the measurable tools required to verify, enforce, and maintain that obligation.

Traditional broadband oversight relies on one-time certification, periodic speed tests, consumer-complaint systems, or high-level reporting. None of these mechanisms are capable of validating the central property upon which AI ERA services depend: **deterministic, low-variance end-to-end latency**. AI-driven systems fail not when throughput is insufficient, but when timing becomes unpredictable, unstable, or drift-prone. This mismatch between legacy oversight practices and modern engineering requirements is precisely why latency budgets must now anchor the enforcement framework for PBPs.

The engineering literature establishes unequivocally that timing pathologies reveal themselves only through **longitudinal observation**: tail-latency excursions, jitter spikes, routing instabilities, queue buildup, and inference desynchronization all emerge not at idle but under real-world load [1]–[5]. A network that appears fully compliant during a speed test may nevertheless experience unpredictable timing spikes that break inference chains or destabilize real-time control loops. These conditions can remain invisible for months or years unless deliberately monitored.

This is the crux of the statutory issue. Under **47 U.S.C. § 1702(a)(2)(I)**, Priority Broadband Projects must:

i. deliver defined performance attributes, and

ii. “easily scale” to meet evolving household and business needs, including support for “other advanced services,”

for the full duration of the ten-year performance period [6]. AI-enabled advanced services cannot function without deterministic timing. Therefore, networks that fail to maintain deterministic latency cannot satisfy statutory PBP requirements—even if they continue to meet nominal speed benchmarks.

Latency budgets make this enforceable. They provide:

- **measurable timing thresholds** tied directly to AI-system viability;
- **methodological rigor** aligned with IEEE and systems-engineering best practices;
- **repeatable evaluation frameworks** that can be applied consistently through time;
- **objective criteria** that do not depend on consumer perception or provider attestation.

By adopting latency budgets as a core oversight tool, states can shift from **reactive** oversight to **proactive statutory compliance management**. Instead of discovering PBP failure after a decade of underperformance—when clawbacks or corrections are infeasible—states gain the ability to identify drift early and require corrective action long before service quality collapses.

The Priority Broadband Project Operational Framework developed by Big Bang Broadband LLC enables this shift. It operationalizes deterministic latency requirements through structured measurement practices that evaluate full-path timing stability, jitter distribution, tail-latency envelopes, aggregation-layer behavior, and compute-adjacent timing under real-world load. By benchmarking performance against engineering-derived latency budgets, the Framework allows states to verify compliance in a rigorous, defensible manner.

This enforcement capability also reduces **legal and financial risk**. Without periodic audits, states risk funding networks that degrade into noncompliance well before the end of the performance window. Subgrantees risk losing PBP status, facing clawback exposure, or being found noncompliant with federal law. By integrating latency budgets into ongoing oversight, states can demonstrate due diligence, subgrantees can demonstrate continuous compliance, and communities can rely on networks that will remain capable of supporting time-sensitive services throughout the performance period.

Ultimately, latency budgets transform the PBP definition from a forward-looking aspiration into an **enforceable technical standard**. They reveal whether a network is scalable in the statutory sense—not merely in bandwidth, but in its ability to preserve the timing stability required for next-decade applications. This makes latency determinism the central criterion for PBP viability. Through structured measurement, it becomes the central mechanism for PBP enforcement.

This leads to an unavoidable conclusion:

If a network cannot maintain deterministic latency, it is not a Priority Broadband Project—regardless of throughput, technology, or initial certification.

This principle ensures that BEAD-funded networks remain capable of supporting the AI-driven systems that will define economic participation and competitiveness over the next decade. It aligns statutory intent with engineering reality and equips states with the tools needed to guarantee that Priority Broadband Projects deliver the long-term benefits Congress mandated.

The engineering realities documented in Parts I–V, and the statutory obligations they inform, are not merely academic observations. They are echoed with increasing urgency across the broadband, wireless, and cloud industries, where operators, analysts, and vendors are independently recognizing that real-time AI workloads impose strict timing constraints that legacy network architectures cannot satisfy without deterministic performance. This growing alignment between research and industry underscores the need to examine how commercial leaders themselves are framing these challenges.

INDUSTRY PERSPECTIVES ON AI TIMING, EDGE STRESS, AND NETWORK DETERMINISM

While the preceding sections rely on peer-reviewed engineering and systems research, it is equally important to recognize that the same timing pressures, latency constraints, and architectural considerations are now being acknowledged across the broadband, wireless, and cloud industries. Industry analysts, network operators, and major infrastructure vendors are independently arriving at the same conclusion documented in the academic literature: **AI Era workloads fundamentally change the performance requirements of broadband networks.**

Industry publications warn that AI-generated traffic will place unprecedented stress on edge infrastructure, magnifying the importance of deterministic timing and exposing weaknesses in networks that were never designed for real-time inference. Analysts have begun using terms such as **"AI traffic armageddon"** to describe the growing mismatch between legacy broadband architectures and the demands of distributed inference systems. These warnings mirror the system-level behaviors observed in the research community: timing instability, jitter bursts, and tail-latency excursions will undermine AI pipelines long before raw bandwidth becomes insufficient.
Source: K. Wollenberg, "Edge networks face AI traffic armageddon," *Fierce Network*, 2024. [Online]. Available: https://www.fierce-network.com/broadband/opinion-edge-networks-face-ai-traffic-armageddon

Equally significant is the accelerating shift of AI inference toward **edge and near-edge compute nodes**, a trend highlighted frequently in cloud and telecom reporting. As inference reduces its physical distance to data sources, the stability of the path between end users and compute resources becomes the decisive factor for application viability. Industry observers note that this shift increases the importance of predictable, low-jitter connectivity—directly validating the engineering principle that deterministic timing is the governing requirement for modern networks.
Source: M. Dano, "AI opportunity heads to the edge," *Fierce Cloud*, 2024. [Online]. Available: https://www.fierce-network.com/cloud/ai-opportunity-heads-edge?itm_source=parsely-api

Major telecommunications vendors are also now framing the problem explicitly in terms of **"physical AI"**—a recognition that AI is constrained by real-world network physics, including latency, timing synchronization, and jitter tolerance. Nokia's CTO has emphasized that AI performance will increasingly depend on predictable, measurable timing characteristics across the entire network, not merely on higher throughput or faster access speeds. This industry positioning closely aligns with the conclusions of Parts I–V: deterministic timing is not optional—it is foundational.
Source: M. Takahashi, "Nokia CTO wades into 'physical AI'," *Fierce Wireless*, 2024. [Online]. Available: https://www.fierce-network.com/wireless/nokia-cto-wades-physical-ai

Finally, industry engineering commentary has begun articulating the concept of **latency budgets** directly. Articles targeted at network architects and infrastructure leaders explain that edge AI pipelines must operate within strict end-to-end timing windows and that network unpredictability—especially in the form of jitter and tail latency—introduces cascading failures in real-world distributed inference. These industry perspectives reinforce the observation that AI ERA networks require system-level timing discipline, not merely access-layer bandwidth.
Source: R. Hensley, "Latency Budgets for the Real World: Designing Edge AI Pipelines," *AutomationInside*, 2024. [Online]. Available: https://www.automationinside.com/article/latency-budgets-for-the-real-world-designing-edge-ai-pipelines

Taken together, these sources demonstrate a widening consensus: **deterministic timing is emerging as the core requirement for AI ERA networks**, recognized not only by researchers but also by the industries responsible for deploying and operating next-generation infrastructure. This consistency between academic and industry perspectives underscores the need for BEAD-funded Priority Broadband Projects to be evaluated and governed as end-to-end timing systems. It further validates the necessity of periodic, engineering-grade audits to ensure that BEAD-funded networks remain compliant as AI adoption intensifies throughout the ten-year performance period.

Together, the engineering literature and industry analyses reinforce a single, unavoidable conclusion: deterministic timing performance is now the governing condition for broadband networks expected to support AI-enabled "advanced services" throughout BEAD's ten-year performance period. The following references provide the evidentiary basis for this Technical Addendum and form the authoritative foundation for its statutory and engineering conclusions.

REFERENCES

[1] Z. Chen, Y. Wang, J. Li, "An Empirical Study of Latency in an Emerging Class of Edge Applications," Carnegie Mellon University, technical report, 2017. [Online]. Available: https://www.cs.cmu.edu/~zhuoc/papers/latency2017.pdf

[2] A. Bourechak, D. Montin, F. El Hamzaoui, "At the Confluence of Artificial Intelligence and Edge Computing: Use Cases, Architectures and Emerging Trends," *Sensors*, vol. 23, no. 12, pp. 6103–6132, Jun. 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/articles/PMC9920982/

[3] S. Konakanchi, "Edge Computing for Latency-Sensitive AI Applications," *International Journal of Computer Engineering & Technology (IJCET)*, vol. 15, no. 6, pp. 2036–2045, 2024. [Online]. Available: https://iaeme.com/MasterAdmin/Journal_uploads/IJCET/VOLUME_15_ISSUE_6/IJCET_15_06_174.pdf

[4] D. P. Mtowe, H. H. Chan, "Edge-Computing-Enabled Low-Latency Control for Wireless Networked Control Systems," *Electronics*, vol. 12, no. 14, p. 3181, 2023. [Online]. Available: https://www.mdpi.com/2079-9292/12/14/3181

[5] A. Abouaomar, I. Mohammed, K. Alam, "Resource Provisioning in Edge Computing for Latency-Sensitive Applications," arXiv:2201.11837 [cs.NI], Jan. 2022. [Online]. Available: https://arxiv.org/pdf/2201.11837

[6] U.S. Congress, "47 U.S.C. § 1702(a)(2)(I) – Priority Broadband Projects: Definitions," *United States Code*, 2022. [Online]. Available: https://www.govinfo.gov/content/pkg/USCODE-2022-title47/pdf/USCODE-2022-title47-chap16-subchapI-sec1702.pdf

Note on Citation Methodology:
The references cited in this Technical Addendum conform to the IEEE citation format, which—consistent with standard engineering practice—does not require pinpoint (page-specific) citations for technical, regulatory, or systems-engineering documents unless a specific sentence or quotation is used verbatim. All sources listed in the References section are cited for the purpose of supporting the technical principles, system behaviors, or statutory frameworks discussed in the text. Because the concepts presented here draw upon integrated findings from entire documents—standards, performance analyses, regulatory filings, and engineering studies—page-specific references would be both impractical and misleading, implying a narrow textual dependency that does not reflect how engineering knowledge is synthesized or applied in practice.

This approach is fully consistent with IEEE professional publications, systems-engineering documentation, and regulatory technical briefs, all of which rely on whole-document citations when the referenced material informs broad analytical conclusions rather than discrete textual assertions.

ABOUT THE AUTHOR

David J. Malfara, Sr. is the Founder and Principal Consultant of **Big Bang Broadband LLC**, specializing in broadband network design, grant program compliance, and long-term infrastructure strategy. A **Senior Member of the Institute of Electrical and Electronics Engineers (IEEE)**, Mr. Malfara brings more than 45 years of executive-level experience in telecommunications, broadband engineering, and the operational launch of multiple network operating companies he helped build from inception. He has served **multiple times as a testifying subject matter expert in federal court and in formal proceedings before state public utility commissions**, bringing litigation-grade analytical rigor to broadband engineering, operational statutory interpretation, and long-term infrastructure compliance.

He is recognized nationally for his expertise in emerging BEAD **Priority Broadband Projects (PBPs)** and the statutory requirements embedded in **47 U.S.C. § 1702**, including scalability mandates, evolving performance obligations, and multi-year compliance standards. Mr. Malfara developed the **Priority Broadband Project Operational Framework**, one of the first practical mechanisms for auditing PBP performance and verifying compliance over the full ten-year statutory performance period. His framework is under active discussion with major broadband software platforms and infrastructure partners seeking to operationalize BEAD's long-term oversight requirements.

In addition to his engineering and policy work, Mr. Malfara serves on the **Local Technology Planning Team (LTPT)** for Marion County, Florida, advising county leadership on broadband deployment, infrastructure modernization, and the integration of **AI Era** requirements into public-sector planning. His research and analysis on AI Era networking demands, fiber scalability, and edge-compute readiness have informed discussions across the broadband industry, including fiber manufacturers, construction firms, data-center operators, and national broadband associations.

Mr. Malfara consults with public- and private-sector organizations seeking to build resilient, future-proof broadband networks aligned with federal law, engineering best practices, and the emerging performance requirements of the **AI Era**.