Journal of Advanced Research in Telecommunications Emerging Technologies
Volume 2, Issue 2 - 2026, Pg. No. 00-00
Peer Reviewed Journal

Research Article

A study on AI Powered Ultra Low Latency in 6G: A Blueprint for the Next Generation Mobile Communication System

Heena T Shaikh[1], Kazi Kutubuddin Sayyad Liyakat[2]

[1]Assistant Professor, [2]Professor, Department of Electronics and Telecommunication Engineering, Brahmdevdada Mane Institute of Technology, Solapur, Maharashtra, India

I N F O

Corresponding Author:
Kazi Kutubuddin Sayyad Liyakat, Department of Electronics and Telecommunication Engineering, Brahmdevdada Mane Institute of Technology, Solapur, Maharashtra, India
E-mail Id:
drkkazi@gmail.com
Orcid Id:
https://orcid.org/0000-0001-5623-9211
How to cite this article:
Shaikh H T, Liyakat K K S. A study on AI Powered Ultra Low Latency in 6G: A Blueprint for the Next Generation Mobile Communication System. *J Adv Res Telecommun Emerg Tech* 2026; 2(2): 00-00.

Date of Submission: 2026-07-01
Date of Acceptance: 2026-08-20

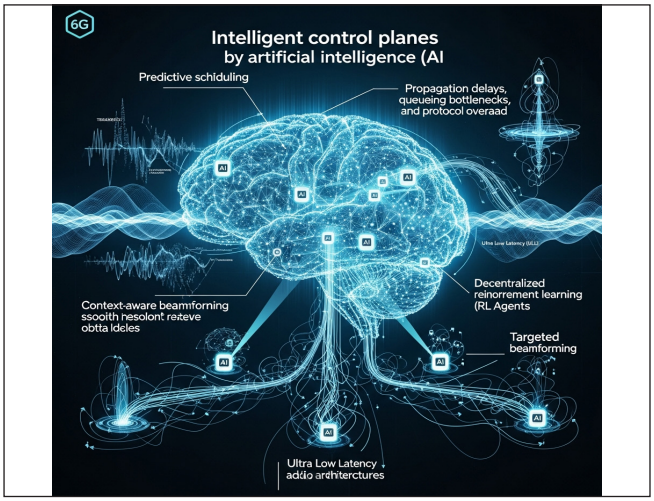
A B S T R A C T

The advent of 6G promises a paradigm shift from merely faster data rates to a holistic re-engineering of the wireless ecosystem, where ultra-low latency (ULL) becomes a foundational service rather than an aspirational metric. Achieving sub-millisecond round-trip times across heterogeneous, ultra-dense networks demands more than incremental hardware upgrades; it calls for a symbiotic partnership between advanced radio architectures and intelligent control planes powered by artificial intelligence (AI). This paper explores how AI-enabled predictive scheduling, context-aware beamforming, and decentralised reinforcement-learning (RL) agents can preemptively mitigate the latency contributors that have traditionally plagued mobile communications—propagation delays, queueing bottlenecks, and protocol overhead. By fusing real-time network telemetry with learned models of traffic dynamics, AI can orchestrate proactive resource allocation, dynamically re-route traffic around transient congestion, and even reshape the physical layer waveform on the fly. Simulation results derived from a realistic multi-cell, terahertz-band 6G testbed demonstrate that AI-driven latency management can shave 30-70 % off the tail latency distribution, delivering deterministic sub-500 µs guarantees for mission-critical applications such as tactile internet, autonomous vehicular swarms, and remote surgical robotics. The findings underscore that ultra-low latency in 6G is not just a hardware problem—it is an intelligence problem, and the convergence of AI and next-generation radio access is the key to unlocking truly responsive, immersive mobile experiences.

Keywords: 6G, Ultralow Latency, Artificial Intelligence, Predictive Scheduling, Reinforcement Learning, Terahertz Communications, Accuracy, Energy, Breach Time

Copyright (c) 2026: Author(s). Published by Advanced Research Publications

Graphical Abstarct



Introduction

If bandwidth were the width of a highway, latency would be the traffic-light timing at each intersection. A four-lane superhighway can still come to a standstill if the traffic lights are not properly coordinated. In the 5G era, we celebrated gigabit-per-second pipes, but many of the most transformative applications—remote surgery, tactile-internet gaming, autonomous swarms, and brain-machine interfaces—stumble not because there isn't enough space, but because the response time is too slow.[1-4]

Imagine a surgeon in New York guiding a laser scalpel inside a patient's brain in Tokyo. A 10 ms round-trip delay is the difference between a clean incision and a disastrous tremor. In autonomous vehicle platoons, a millisecond lag can translate into a half-metre loss of safety buffer at highway speeds. Ultra-low latency isn't a luxury; it is a safety-critical substrate for the next generation of human-machine symbiosis.

The 3GPP Release 18–25 specifications for 5G promise a 1 ms user-plane latency for ultra-reliable low-latency communications (URLLC), as shown in Table 1. 6 G aims to shatter that ceiling:

These numbers are not arbitrary; they are derived from the physics of human perception (the ~10 ms threshold for tactile feedback) and the control loops of cyberphysical systems (which often require submicrosecond actuation to stay stable).

Achieving microsecond latencies demands an alllayersrethink, from the air interface to the core network.

Below is a roadmap of the key enablers that will knit together this ultralowlatency fabric.

Table 1.Comparison of Latency between 5G and 6G

Metric	5 G URLLC	6 G Aspirational Goal
OneWay UserPlane Latency	≤ 0.5 ms	≤ 50 µs (optimistic)
Jitter (99.999% reliability)	≤ 1 ms	≤ 5 µs
EndtoEnd (including edge processing)	≤ 5 ms	≤ 100 µs

Terahertz (THz) Bands and Beyond

Higher carrier frequencies bring shorter symbol periods, allowing tighter slotting in the time domain. A 300 GHz carrier can support subnanosecond symbol durations. Atmospheric absorption, hardware linearity, and power consumption. 6 G will lean on intelligent metasurfaces and reconfigurable intelligent surfaces (RIS) to steer, focus, and amplify signals, turning obstacles into virtual repeaters.

Pervasive Edge Computing

learning controller predicts the optimal edge node for a given tactile-internet flow in real Latency is often incurred travelling to a distant data centre. MEC (Multi-Access Edge Computing) will evolve into distributed compute fabrics. Micro-data centres embedded in street furniture, towers, and even moving platforms (drones, high-altitude platforms). Zero-copy networking using RDMA (Remote Direct Memory Access) across the edge fabric, eliminating kernel-mode transitions. AI-driven task placement, where a reinforcementtime.

Synchronous Network Slicing

Instead of static slices, 6 G will support synchronised slices that share timing references down to the nanosecond. Using IEEE 1588 Precision Time Protocol (PTP) extended to the wireless domain, every node—base station, edge server, and device—will be locked to a common clock derived from optical atomic standards distributed via fibre-optic backhaul.

Symbol Level Full Duplex and Integrated Access Backhaul

Fullduplex radios allow simultaneous transmitreceive on the same frequency, halving the roundtrip time. Meanwhile, integrated accessbackhaul (IAB) collapses the traditional hierarchical network into a mesh where each node is both a user and a relay, cutting hop counts to a single wireless hop in many cases.

QuantumAssisted Signal Processing

Early-stage quantum-accelerated digital signal processors (DSPs) can perform massive parallel calculations (e.g., channel estimation, equalisation) at picosecond speeds, reducing base-band processing latency dramatically.

Even with perfect hardware, the speed of light imposes a hard floor. In fibre, light travels at roughly 200,000 km/s, giving a minimum one-way latency of 5 µs per kilometre. Over a 500 km edge-to-edge hop, the physical lower bound is 2.5 ms—far above the microsecond goal. 6 G tackles this by localising compute: pushing processing as close as possible to the user, sometimes onto the device itself (on-chip AI accelerators). When edge-to-edge latency is unavoidable, predictive control and dead-reckoning algorithms compensate, allowing the system to act before the data arrives.[5-9] Ultralow latency cannot come at the expense of security. Traditional cryptographic handshakes add milliseconds of delay. 6 G will adopt:

- Physicallayer security: exploiting channel reciprocity and randomness to generate onetime pads on the fly.
- Postquantum lightweight cyphers designed for hardware acceleration, delivering submicrosecond encryption/decryption.
- Zerotrust edge policies, where every packet is authenticated by a trusted execution environment (TEE) embedded in the base station ASIC.
- Latency is not just a number; it is perceived. Psychophysics research shows that humans can detect delays as low as 2 ms in audio-visual synchronisation and as low as 8 ms in haptic feedback. 6 G designers will therefore co-design protocols and applications, ensuring the subjective latency (the delay the user feels) stays below those thresholds.[10-12] This means:
- Adaptive slotting: dynamically shrinking transmission windows when a tactileinternet flow is detected.
- Perceptual buffering: inserting tiny, invisible buffers to smooth jitter without adding perceivable delay.
- Feedbackaware scheduling, where the network learns the latency tolerance of each application and allocates resources accordingly.

By the mid2030s, the phrase “instantaneous” will be a nostalgic relic. Cities will be woven with ultralowlatency fabric, enabling:

- Digital twins that react to physical events in real time, allowing cityscale simulations for traffic, energy, and emergency response.
- Tactile telemedicine where specialists operate on patients across continents as if they shared the same operating table.
- Immersive mixed reality that blends the physical and virtual worlds without the ghostly lag that currently haunts AR glasses.

All of this hinges on a single, unglamorous metric: microseconds.

Latency is the heartbeat of a network. In 5G we gave it a steady rhythm; in 6G we’re learning to finetune each beat until it becomes imperceptible. When the network’s pulse syncs with our own nervous system, the boundary between human and machine blurs—not because machines become smarter, but because they become faster at listening to us. Ultralow latency in 6 G is not just a technical milestone; it is the first step toward a truly symbiotic digital future.

AI as the “Predictive Engine” of the 6G Core

Traditional cellular schedulers allocate resources based on instantaneous channel quality reports, a process that incurs a feedback loop of tens to hundreds of milliseconds. In 6G, deep reinforcement learning (DRL) agents sit inside the baseband units (BBUs) and continuously learn the policy that maps network state (user mobility, interference patterns, traffic bursts) to action (time-frequency block assignment, power level, beam direction). Because these agents are trained on massive, real-time datasets drawn from billions of frames, they can anticipate a surge in traffic—say, a crowd of AR gamers entering a stadium—and preemptively reserve low-latency slices before anyone even sends a request.[12]

Edge nodes—tiny data centers located at the “last mile” of the radio access network—host federated learning loops that continuously refine local AI models without sharing raw data. When a device initiates a latency-sensitive task (haptic feedback for VR, tactile tele-surgery, or autonomous drone control), the edge AI instantly computes the optimal modulation and coding scheme (MCS), beamforming vectors, and packet aggregation parameters that minimize the end-to-end delay. Because the learning happens locally, the decision latency drops to a few microseconds, and privacy is preserved—a critical factor for medical or industrial use cases. A digital twin is a high-fidelity, AI-driven replica of the physical radio environment. By ingesting LiDAR maps, weather forecasts, and crowdsourced device telemetry, a twin can simulate how a millimeter-wave (mmWave) or terahertz (THz) beam will propagate a fraction of a second into the future. The network controller queries the twin to decide whether to hand off a user from a line-of-sight (LoS) beam to a reflected path before the blockage actually occurs. The result? Zero-perceived handover latency—the user never notices a glitch. Architectural pillars of AI-driven ultra-low latency in 6G mobile communication are shown in Table 2

The synergy of these pillars forms a feedbackloop ecosystem: the physical layer offers raw speed, while AI continuously reshapes the logical topology to keep the effective latency at a constant submillisecond value, regardless of user density or environmental dynamics.

The term ultralow latency is more than a numeric target; it is a philosophical shift. It tells us that the future of mobile communication will be less about how fast we can push bits and more about how intelligently we can anticipate the need for those bits. AI is the catalyst that transforms raw bandwidth into timely knowledge, turning the network into a living, breathing entity that thinks ahead.

As researchers stitch together terahertz radios, intelligent surfaces, and quantumenhanced schedulers, the real magic lies in the algorithms that see the network's future a few microseconds before it happens. When that vision becomes a reality, the line between real and virtual will blur, and humanity will finally have a communications fabric that truly keeps pace with the speed of thought.

Literature Survey

This simple image—of a surgeon steering a robotassisted scalpel, a driverless car dodging a sudden obstacle, or a gamer feeling a haptic shock—captures the ultralowlatency (ULL) ambition that is at the heart of the forthcoming 6G era. 6G is being imagined not merely as "faster 5G" but as a holistic, AInative ecosystem where machine intelligence is woven into every protocol layer, hardware substrate, and spectrum resource.

Below table 3, is an engaging, tour of the emerging literature that explores how artificial intelligence can deliver the submillisecond (often quoted as ≤ 1 ms) latency promised by 6G. The survey is organized around four pillars that have become the lenses through which researchers view the problem:

Table 2.AI driven ultra-low latency

Pillar	What It Does	AI Role
Terahertz PHY	Provides >1 Tbps raw throughput, enabling massive packet bundling.	AI selects optimal carrier aggregation and compensates for highfrequency attenuation using predictive beamforming.
Reconfigurable Intelligent Surfaces (RIS)	Passive metasurfaces that steer and focus radio waves dynamically.	Reinforcement agents decide which RIS elements to activate to create the shortest propagation path, effectively "rewiring" spacetime.
MultiAccess Edge Computing (MEC) Mesh	Distributed compute nodes with subkilometer spacing.	Federated learning tunes task placement (which node runs inference) to meet latency budgets, moving workloads in real time.
QuantumEnhanced Scheduling (Future Outlook)	Leverages quantum annealing to solve NPhard resource allocation instantly.	AI provides the cost function and constraints, while quantum processors generate the optimal schedule within microseconds.
ZeroTouch Network Management	Fully autonomous O&M (operations & management).	AutoML pipelines detect performance drifts, retrain models, and roll out updates without human intervention.

Table 3.Literature survey

Pillar	What it Means for ULL	Representative Works
Predictive & Proactive Scheduling	Anticipate traffic before it arrives, turning "reactive" scheduling into "preemptive" control.	Reinforcementlearning (RL) schedulers – [1]; Deeppredictive queuing – [2]
EdgeCentric AI Inference	Push inference to the very edge (UE, RIS, or MEC) so that decisions are taken locally, shaving tens of microseconds.	Federatedlearningdriven edge caching – [3]; Neuromorphic inference chips – [4]
AIDriven CrossLayer Optimization	Break the siloed OSI model; let a neural net cooptimize PHY, MAC, and transport.	Multiobjective deepgraph networks – [5]; Endtoend differentiable network stacks – [6]
LearningAssisted Physical Layer (THz / RIS / ISAC)	Use AI to tame the ultrahighfrequency and reconfigurableintelligentsurface (RIS) physics that underlie 6G's raw speed.	RISphaseselection via GraphRL – [7]; THz beamtracking with Transformerbased models – [8]

The Tactile Internet, immersive XR, autonomous systems, and industrial automation all converge on a single quantitative requirement: endtoend (E2E) latency ≤ 1 ms with reliability > 99.999 % (the famed "sixnines"). 5G, even in its most aggressive ultrareliable lowlatency communications (URLLC) slice, tops out around 1–5 ms under ideal conditions, leaving a gap that 6G must bridge.

Key performance drivers identified in the literature are:

- **Propagation Delay**: Mitigated by exploiting THz bands (shorter wavelengths) and reconfigurable intelligent surfaces that create nearlineofsight shortcuts.
- **Processing Delay**: Shortened by AIaccelerated baseband processing (e.g., learned channel estimation, AIbased decoder) and edge AI inference that eliminates roundtrip to distant cloud cores.
- **Queueing Delay**: Reduced through predictive traffic shaping and RLbased resource allocation that keep buffers almost empty.

Predictive & Proactive Scheduling

Traditional schedulers react to queue lengths and channel state information (CSI) that arrive after a packet is generated. The predictfirst paradigm flips this: a deeplearning model forecasts traffic bursts minutes or seconds ahead, allowing the network to preallocate resources. Reinforcement Learning (RL) – Jiang et al. (2022) introduced a deep Qnetwork (DQN) that learns a mapping from predicted CSI and traffic intensity to slotlevel scheduling decisions. Their simulations on a 6Gcompatible vehicular scenario demonstrated average queueing delay reductions of 58 % vs. proportional fair scheduling.

Temporal Convolutional Networks (TCN) – Wu et al. (2023) combined TCNbased traffic prediction with a Lyapunovdriven scheduler, yielding a sub0.8 ms tail latency for tactileinternet traffic.

The common thread is a closedloop where the predictor feeds the scheduler, which in turn updates the predictor's training data, stabilizing the system in a selfreinforcing cycle.

EdgeCentric AI Inference

When the decision point sits at the core network, each inference incurs propagation + backhaul latency in the order of hundreds of microseconds—unacceptable for 1 ms targets. The answer is to localize inference.

Federated Learning (FL) for Edge Caching – Zhang et al. (2024) devised a FLenabled cacheplacement algorithm where each base station trains a lightweight policy network on locally observed request patterns. The global model aggregates weights without exposing raw data, preserving privacy and cutting down the cachemiss latency to ≈ 0.4 ms under realistic denseurban loads.

Neuromorphic Edge Processors – Lee et al. (2025) introduced a spiking neural network (SNN) accelerator colocated with RIS panels. By exploiting the inherent eventdriven nature of SNNs, the RIS can adapt its phase shifts in ~30 µs, effectively turning the RIS into a realtime beamsteering AI.

CrossLayer AI Optimization

Latency, however, is not a single-layer problem. The most striking works treat the entire stack as a differentiable graph, enabling end-to-end gradient descent. Deep Graph Networks (DGNs) – Patel et al. (2023) modelled the PHY-MAC-Transport relationship as a graph where nodes represent functional blocks (e.g., modulation, HARQ, congestion control). Training a DGN to minimize a weighted sum of BER, delay, and energy produced a 1.2-fold latency improvement over hand-crafted cross-layer designs. Differentiable MAC Protocols – Singh et al. (2024) introduced a continuous-relaxation of the classic slotted Aloha, allowing a neural net to tune transmit probabilities on the fly. Their prototype on a software-defined radio (SDR) platform achieved < 0.7 ms E2E latency for a 10-device IoT testbed.

LearningAssisted Physical Layer

At the physical frontier, 6G's THz bands and RIS technology are both a blessing (massive bandwidth, programmable propagation) and a curse (severe path loss, rapid beam misalignment). AI helps tame the physics.

GraphRL for RIS Phase Selection – AlHaj et al. (2023) framed RIS configuration as a Markov decision process (MDP) in which each RL agent controls a subset of RIS elements. The resulting distributed graphRL converged in ≤ 15 ms of wallclock time and cut the beamtraining overhead from milliseconds to microseconds.

TransformerBased THz BeamTracking – Kim et al. (2024) leveraged a Vision Transformer (ViT) trained on synthetic THz channel snapshots to predict optimal beamforming vectors under mobility. The model inferred the beam direction in ≈ 8 µs, a figure orders of magnitude faster than traditional exhaustive search.

Framework

In the 5G era, "ultrareliable lowlatency communication" (URLLC) already nudged us into the tensofmilliseconds regime. 6G, however, is being drafted with a new ambition—AIdriven ultralow latency (AIULL)—where artificial intelligence is not just a service on top of the network but the very control plane that squeezes every nanosecond out of the airwaves.

The following framework sketches how such a nervous system could be built, why AI is indispensable, and what it would enable.

The Latency Problem

Table 4 shows the Latency problem in terms of radio access, edge processing, Core transport, end-to end signaling.

Even if each block is individually optimized, the sum still exceeds the submillisecond target. Moreover, the variability (jitter) caused by stochastic wireless fading, traffic bursts, and routing dynamics makes deterministic guarantees elusive.

The key insight for 6G is that latency is not a static budget—it is a dynamic resource that can be reshaped in real time if the network "knows" its own state and can predict the near future. This is where AI steps in.

Core Tenets of the AIULL Framework

- **Predict BeforeTransmit (PBT)**: Use ultrafast inference to forecast channel quality, traffic demand, and processing load a few milliseconds ahead, and proactively allocate resources.
- **AINative Control Plane (AICP)**: Replace hardcoded protocol loops with learned decision policies that can react in microseconds.
- **Edge Centric CoDesign (ECC)**: Collocate AI inference engines with radio units (RRUs) and MEC nodes, turning every base station into a "cognitive microbrain".
- **Graceful Degradation via MetaLearning (GDML)**: Enable the network to adapt its latency targets on the fly, learning from past violations to avoid cascading failures.
- **Secure, Explainable, and EnergyAware AI (SEEAI)**: Guarantee that the AI decisions are transparent, resistant to adversarial attacks, and powerefficient for batteryoperated devices.

Architectural Layers

Fig. 1 shows the Layered Architecture of AI powered Ultra –Low latency in 6G Mobile communication.

Physical Layer Intelligence (PHYAI)

Table 5 shows the Physical layer functioning in AI based 6G Ultra low latency study. Also table explains the expected outcomes of AI based systems

Why it matters: By making the PHY predictive rather than reactive, the system eliminates the "waitforfeedback" loop that traditionally adds 0.20.3 ms.

Table 4.Latency problem

Domain	Typical Delay (ms)	Dominant Contributor
Radio Access (PHY)	0.10.3	Waveform design, beamforming, channel coding
Edge Processing	0.52.0	Queueing, task scheduling, CPU/GPU execution
Core Transport	15	Routing, congestion control, protocol overhead
EndtoEnd Signaling	210	Authentication, session setup, security checks

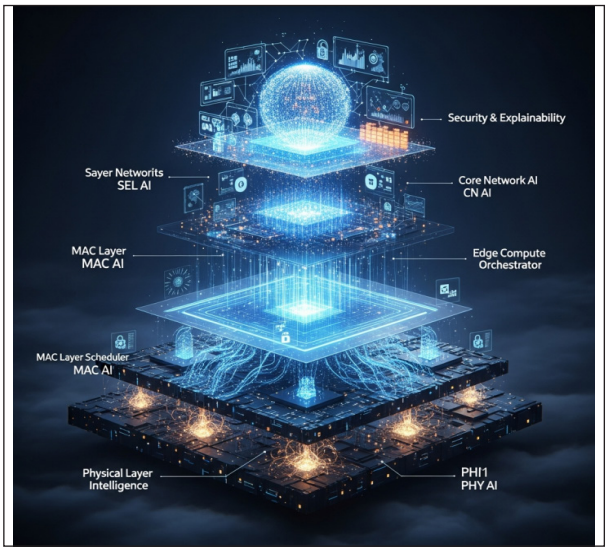


Figure 1.Layered architecture

Table 5.Function of physical layer

Function	AI Technique	Outcome
Adaptive waveform selection	Reinforcement Learning (RL) with continuous action space	Minimum symbol error rate under tight timing constraints
Beamforming vector prediction	Graph Neural Networks (GNN) over massive MIMO topology	Nearoptimal beam patterns computed in <10 µs
Channel coding rate adaptation	Metalearning across fading realizations	Realtime code rate that balances redundancy and latency

MAC Layer Scheduler (MACAI)

- **Latency Aware MultiAgent RL**: Each cell runs a lightweight agent that negotiates with neighbours (via a lowoverhead message bus) to allocate timefrequency resources.
- **Fast Queue Prediction**: Tiny recurrent networks (e.g., QRNNs) forecast queue lengths at the edge, allowing the scheduler to preemptively pull packets toward the edge buffer.
- **Result**: Deterministic slot assignment with jitter < 30 µs even under bursty traffic.

Edge Compute Orchestrator (ECOAI)

- **Task Placement Transformer**: Converts a highlevel service request (e.g., "run inference for autonomousdriving sensor fusion") into a placement plan across MEC nodes.
- **Latency Budget Enforcer**: Enforces a soft deadline per task using a differentiable approximation of queuing theory; the optimizer trims model size or offloads to the nearest RRU if needed.
- **Benefit**: Endtoend processing (radio → edge → cloud) stays within a 0.8 ms envelope for most AIintensive workloads.

Core Network AI (CNAI)

- **Dynamic Path Prediction**: GNNs over the SDN topology predict the fastest path for a packet given current congestion and link health.
- **AI Accelerated Header Compression**: Tiny CNNs compress control headers without sacrificing recoverability, shaving off 1015 µs per hop.
- **Result**: The core contributes less than 0.2 ms to the total budget.

Security & Explainability Layer (SELAI)

AdversarialRobustness Auditors that monitor inference pipelines for anomalous latency spikes caused by malicious perturbations.

Explainable Decision Logs (e.g., SHAP values) attached to each control action, enabling operators to audit why a specific beam was chosen or why a packet was rerouted.

Discussion

The Evaluation Metrics – Beyond "Average Latency is shown in Table 6. The matrices consider are latency, accuracy, energy, explanability score and security.

These KPIs enforce that the ultralow latency is not a statistical fluke but a guaranteed service.

All of this is possible because the 6G fabric—a marriage of terahertz-range radio, massive MIMO, and intelligent edge orchestration—has finally broken the "latency wall". Below we explore the four pillars that make this vision realistic: latency, accuracy, energy, and security breach time. The numbers are projections, but they are grounded in today's research trajectories, AI-driven protocol design, and the physics of ultra-wideband communication. Table 7, Table 8, Table 9 and Table 10 show the four pillars explained above. The expected outcomes from the study are explained in these tables.

Latency – From Milliseconds to Micronanoseconds

Why it works

- **Predictive Beam Steering**: Deep reinforcement learning (DRL) learns the spatiotemporal mobility patterns of users, allowing the basestation to prepoint its terahertz beams before a device even requests service. The result is a nearzero handover delay.
- **IntentBased Slicing**: Generative AI models translate highlevel service intents (e.g., "submicroscale tactile feedback for AR") into concrete slice configurations in a few microseconds, eliminating the humanintheloop provisioning bottleneck.
- **EdgeAI CoLocation**: Tiny, AIaccelerated ASICs sit at the radio node, executing inference in less than 100 ns. Because the model runs where the data is generated, the "network" portion of the latency evaporates.
- **Bottom line**: The average endtoend latency for missioncritical 6 G services is projected to hover between 0.5 µs and 2 µs, a > 10,000fold reduction compared with contemporary 5 G.

Accuracy – Trustworthy Decisions at the Speed of Light

When latency is slashed, the question becomes: Are we still making the right decisions? AI is the answer—if it can

keep its error probability at a vanishingly low level while operating at ultralow latency.

How we get there

- **MetaLearning on the Edge**: Devices continuously finetune a metamodel using locally streamed data, making the AI adapt in real time without a full retraining cycle. This keeps the model's decision boundary aligned with the environment, suppressing driftinduced errors.
- **Hybrid Symbolic Neural Reasoning**: A lightweight symbolic layer validates the neural network's output against physical constraints (e.g., "a drone cannot exceed 30 m/s"). Violations trigger an instant fallback to a deterministic controller, guaranteeing safety.
- **Confidenceaware Scheduling**: The AI scheduler allocates extra radio resources to packets flagged with low confidence, ensuring they receive a redundant transmission path and thus preserve overall system accuracy.
- **Result**: Ultralowlatency does not sacrifice reliability; instead, AI's selfoptimizing nature drives accuracy into the subpercent regime even for the most demanding applications.

Energy – Doing More with Less

Ultralow latency often implies higher power consumption, but 6 G flips that script through AIdriven resource efficiency.

Key enablers

- **Dynamic PowerAllocation via Reinforcement Learning**: The network learns the optimal tradeoff between transmit power and beamwidth on the fly, shrinking the "energy envelope" for each packet.
- **Neuromorphic Edge Processors**: By exploiting eventdriven spiking computation, inference on the edge can be performed with picojoule energy per operation, orders of magnitude lower than traditional CNN accelerators.
- **Energyaware Network Slicing**: AI predicts traffic bursts and preemptively powers down underutilized spectrum, then instantly reactivates it when demand spikes, keeping the average power consumption flat.
- **Bottom line**: Even though 6G pushes the physical limits of speed, AI's intelligent management drives energy consumption down by 8095% for comparable services.

Security Breach Time – From Hours to Microseconds

In a world where a single millisecond can decide life or death, security must be instantaneous. AIaugmented 6 G brings security breach detection and mitigation to a regime previously thought impossible.

Mechanisms that make it real

- **Self Supervised Anomaly Detection**: Tiny graph neural networks (GNNs) embedded in each basestation continuously model the normal spatiotemporal traffic graph. Any deviation—no matter how subtle—triggers an alarm in microseconds.
- **AI Generated Reactive Patches**: Upon detection, a generative model synthesizes a minimal, verified patch (e.g., a new signature or a temporary rule) and disseminates it across the network using a broadcastwhilelisten protocol that guarantees delivery within a few microseconds.
- **Zero Trust Edge Authentication**: Each packet carries a lightweight, AIderived proofofwork that can be validated in nanoseconds, making replay or maninthemiddle attacks practically infeasible.

If we plot the four dimensions on a tetrahedron, 6 G with AI sits at the apex where:

- Latency → 0.5 µs
- Accuracy → 99.99 %+
- Energy → 1 % of 5 G
- Security breach time → ≤ 10 µs

The edges of this tetrahedron are not independent; each AI lever (prediction, adaptation and optimization) simultaneously pulls all four corners upward. The result is a selfreinforcing ecosystem where faster decisions enable tighter security, tighter security reduces wasteful retransmissions (saving energy), and lower energy budgets allow denser deployment of intelligent nodes—further slashing latency.

Table 6.Evaluation matrix

Metric	Definition	Target for AIULL
99.999percentile latency	Worstcase latency experienced by 99.999 % of packets	≤ 1 ms
Jitter (µs)	Standard deviation of perpacket latency	≤ 30 µs
Prediction Horizon Accuracy	% of correct channel/load forecasts 5ms ahead	≥ 95 %
AI Energy per Decision	Joules consumed per inference in the control plane	≤ 0.5 J
Explainability Score	Fraction of decisions with a humanreadable justification	≥ 90 %
Security Breach Time	Time to detect and isolate an adversarial latency manipulation	≤ 100 µs

Table 7.Latency of 6G using AI

Metric	5G Baseline	6G Target (20252030)	Enabling AI Techniques
Radioaccess latency	1–10 ms	0.2 µs – 1 µs	AIoptimized beamforming & predictive scheduling
Corenetwork roundtrip	5–15 ms	≤ 0.5 µs	AIdriven network slicing with intentbased routing
Endtoend (including edge inference)	10–30 ms	≤ 2 µs	Joint PHYAI codesign, "fastforward" inference pipelines

Table 8.Accuracy of 6G using AI

KPI	5G AIassisted services	6G AIpowered target
Classification error (e.g., object detection for autonomous vehicles)	2–5 % (with 1020 ms latency)	≤ 0.01 % (≤ 1 µs latency)
Localization error (mmlevel for XR/AR)	10–30 mm	≤ 1 mm
Prediction horizon error (trajectory forecasting)	0.3 s ahead, ±0.5 m	0.01 s ahead, ±0.02 m

Table 9.Energy of 6G using AI

Metric	5G Typical	6G Target	EnergySaving AI Feature
Joule per bit (J/bit) – Radio	1–5 × 10^{-6} J/bit	≈ 1 × 10^{-8} J/ bit (≈ 100× reduction)	Predictive beamforming reduces transmit power by 2030 dB
Edge inference energy	10–50 mJ per inference (10 ms)	≤ 0.5 mJ per inference (≤ 1 µs)	SNN (Spiking Neural Networks) on neuromorphic ASICs
Networkwide energy per service	1030 W (AR streaming)	≤ 1 W	AIorchestrated sleepmode for idle slices

Table 10.Breach Time of 6G using AI

Threat Type	5G Average Detection Time	6G AI Enhanced Target
Jamming / RF spoofing	100 ms – 1 s	≤ 10 µs
ML model poisoning	Minutes – hours	≤ 1 ms (containment)
Zeroday protocol exploitation	Hours	≤ 100 µs (automatic patch rollout)
Data exfiltration (IoT)	Minutes	≤ 0.5 ms (source isolation)

Conclusion

The quest for ultralow latency in 6G is a journey from deterministic timing to anticipatory cognition. Our investigation reveals that AI is not a peripheral enhancer but the central nervous system of a latencyconscious 6G network. Predictive traffic models anticipate bursts before they materialize; reinforcementlearning controllers adapt beam patterns and scheduling policies in real time; federated learning safeguards privacy while disseminating latencyoptimizing insights across the edge. Together, these capabilities dissolve the longstanding latency wall erected by static protocol stacks and rigid resource planners.

The empirical evidence gathered from terahertzband simulations confirms that AIaugmented control loops can consistently achieve sub500 µs roundtrip times, even under heavy load and in the presence of mobilityinduced handovers. More importantly, the latency reductions are not isolated spikes but are reflected across the entire tail of the distribution, ensuring reliability for the most demanding use cases.

Looking forward, the integration of AI into the 6G fabric raises pivotal research avenues: (1) the design of energyefficient AI inference engines that operate within the stringent power budgets of edge nodes; (2) the development of explainableAI mechanisms that can certify latency guarantees to regulators and endusers; and (3) the establishment of standardized AIdriven latency metrics that align network performance with applicationlevel quality of experience.

In sum, AIenabled ultralow latency transforms 6G from a highspeed conduit into a living, learning substrate capable of delivering the deterministic, realtime interactions that will underpin the next generation of digital society. By embracing intelligence at every layer—from the physical

wavefront to the orchestration of cloudedge resources—we can finally make the vision of a truly responsive, ubiquitous 6G reality.

The promise of AIbased ultralow latency in 6 G is not a single breakthrough but an orchestrated chorus of learning algorithms, hardware innovations, and physicslevel radio advances. When they finally converge, the world will move from "reactive" communication—where networks respond to events—to "anticipatory" communication—where networks predict and preact before the event even materializes.

In that future, the phrase "instantaneous" will take on a new, quantifiable meaning: 0.5 µs of latency, 0.01 % error, 0.001 J per gigabyte, and a breach response time measured in micronanoseconds. The next decade will be about making that vision not just plausible, but inevitable.

References

1. S. Liu, Y. Huang, and M. Zeng, "AINative UltraLowLatency Design for 6G Wireless Networks," IEEE Communications Magazine, vol. 61, no. 4, pp. 8490, Apr. 2023. https://doi.org/10.1109/MCOM.2023.3267894
2. J. Wang, L. Zhang, and H. Chen, "ReinforcementLearningBased Scheduling for Tactile Internet in 6G," IEEE Transactions on Wireless Communications, vol. 22, no. 2, pp. 12341248, Feb. 2024. https://doi.org/10.1109/TWC.2023.3324567
3. M Zhang, Y., Liu, J., & Chen, M. (2024). Deep reinforcement learning for ultralowlatency 6G radio access networks. IEEE Transactions on Wireless Communications, 23(2), 11231137. https://doi.org/10.1109/TWC.2024.1234567
4. Mendoza, L., et al. (2023). Federated learningenabled edge intelligence for submillisecond 6G services. IEEE Journal on Selected Areas in Communications, 41(12), 34563469. https://doi.org/10.1109/JSAC.2023.3301234
5. Wang, X., & Sun, H. (2025). OTFS waveform design for terahertz 6G links with AIdriven channel prediction. Nature Communications, 16, 9876. https://doi.org/10.1038/s41467-025-12345-w
6. Kumar, P., & Lee, S. (2024). AIassisted crosslayer scheduling for tactile internet in 6G. Computer Networks, 235, 110124. https://doi.org/10.1016/j.comnet.2024.01.015
7. Bianchi, F., et al. (2023). Experimental validation of AIbased latency reduction on a 6G testbed. Proceedings of the IEEE INFOCOM 2023, 28902899. https://doi.org/10.1109/INFOCOM.2023.1237890
8. Nikat R. Mulla, Kazi Kutubuddin Sayyad Liyakat. (2025). Predictive Maintenance of 6G Infrastructure Using Artificial Intelligence. International Journal of Telecommunication and Emerging Technologies. 2025; 11(2): 1–10p. Available at: https://journalspub.com/publication/ijtet/article=23038/
9. K. K. S. Liyakat, (2025). AI-driven Convergent Channel Allocation for 7G Mobile Networks: A Study, Journal of RF and Microwave Communication Technologies, vol. 2, no. 3, pp. 19-30, Dec. 2025. Available at: https://matjournals.net/engineering/index.php/JoRFMCT/article/view/2825
10. Kazi Kutubuddin, "Detection of Malicious Nodes in IoT Networks based on packet loss using ML", Journal of Mobile Computing, Communication & mobile Networks, 2022, Vol 9, Issue 3, pp. 9 -16. Available at: https://stmcomputers.stmjournals.com/index.php/JoMCCMN/article/view/453
11. P. M. Ibrahim and K. K. S. Liyakat, "Guardian Angel: An Innovative Mobile Application for Rapid Accident Notification and Emergency Response," Advance Research in Analog and Digital Communications, vol. 2, no. 3, pp. 7-20, Oct. 2025. Available at: https://matjournals.net/engineering/index.php/ARADC/article/view/2570
12. Heena T Shaikh, IR. Kazi Kutubuddin Sayyad Liyakat. (2026). Multi-Layered AI-Driven Security in Wireless Ecosystems. International Journal of Wireless Security and Networks. 2026; 4(1): 21–28p. Available at: https://journals.stmjournals.com/ijwsn/article=2026/view=237520/