engineering proceedings

MDPI

Proceeding Paper

EMERGE-XR: Intent-Driven Low-Latency 5G for Real-Time Extended Reality Emergency Medical Guidance †

Ioannis Tsilikis [1,2,*], Christos Panagiotou [1], Konstantinos Antonopoulos [1], Michael Paraskevas [1,2], Vaggelis Kapoulas [2], Christos P. Antonopoulos [1] and Nikolaos Voros [1]

[1] Electrical and Computer Engineering Department, University of Peloponnese, Patras 26334, Greece; ch.panagiotou@esdalab.ece.uop.gr (C.P.); k.antonopoulos@go.uop.gr (K.A.); mparask@cti.gr (M.P.); ch.antonop@uop.gr (C.P.A.); voros@uop.gr (N.V.)
[2] Computer Technology Institute and Press "Diophantus", Patras 26504, Greece; kapoulas@cti.gr
* Correspondence: i.tsilikis@go.uop.gr
† Presented at the 2nd International Conference on Engineering, Technology, and Applied Science Innovations (ICETASI 2026), Zakynthos, Greece, 24–26 September 2026.

Abstract

Extended Reality (XR) applications for emergency medical response impose stringent latency requirements that commodity fifth-generation (5G) deployments do not satisfy by default. This paper presents EMERGE-XR, a real-time XR system delivering AI-driven medical guidance to first responders, and addresses the research problem of how Medium Access Control (MAC)-layer parameter tuning on a software-defined 5G gNodeB (gNB) can reliably achieve sub-10 ms round-trip time (RTT) for interactive XR workloads. The system integrates Meta Quest 3 augmented reality (AR) headsets, edge-hosted pose estimation, and a clinical decision-support workflow based on the ABCDE trauma assessment protocol. The clinical guidance component has been functionally demonstrated on the laboratory testbed but has not yet been evaluated in a user study; the experimentally validated contributions of this work concern the network-layer optimization and the profile-switching mechanism, not the clinical or AI-driven components. Through systematic MAC-layer optimization on an on-premises Amarisoft 5G gNB, tuning Scheduling Request (SR) periodicity, the Time Division Duplex (TDD) uplink/downlink slot pattern, and fixed MAC timing parameters, the average RTT was reduced from 24 ms to 9 ms, a 62.5% improvement, with zero packet loss. A lightweight RTT-driven profile-switching mechanism was validated across more than 100 experimental runs. As a work-in-progress contribution, an AI-driven intent-based extension is further proposed, but not yet experimentally validated, in which a contextual reinforcement learning (RL) agent autonomously selects MAC-layer profiles based on application-level Quality of Experience (QoE) signals and network Key Performance Indicators (KPIs). Planned validation over the 6G-BRICKS Cell-Free Massive MIMO testbed at KU Leuven is outlined.

Keywords: extended reality; 5G; low latency; MAC-layer optimization; intent-based networking; reinforcement learning; emergency response; edge computing

Check for updates

Academic Editor: Konstantinos Poirazidis

Published: 6 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Public Protection and Disaster Relief (PPDR) operations demand rapid and accurate decision-making under extreme time pressure. First responders in the field often lack immediate access to specialist medical expertise, which can lead to suboptimal patient outcomes during the critical early period of an incident. Extended Reality (XR) technologies offer a

paradigm shift: by overlaying real-time digital guidance onto the physical environment, they can bridge the expertise gap between field operators and remote medical professionals. However, deploying interactive XR applications over cellular networks introduces stringent latency requirements. Motion-to-photon delays exceeding 20 ms degrade spatial registration accuracy and induce simulator sickness [1], while round-trip delays above 50 ms impair artificial intelligence (AI) inference pipelines processing live video feeds [2]. Recent work further shows that even advanced radio technologies require careful cross-layer engineering before interactive, multi-user XR can be supported over wireless links [3].

Despite growing interest in 5G-enabled XR for emergency response [4–6], a con-crete research gap remains: standard 5G New Radio (NR) deployments using default Medium Access Control (MAC)-layer parameters fail to meet the sub-10 ms round-trip time (RTT) budget required by interactive AR-guided medical workflows, in which AR overlays, pose-estimation feedback, and clinical prompts must traverse the radio seg-ment within a small fraction of the overall motion-to-photon budget. The goal of this paper is to close this gap through systematic MAC-layer tuning on a commercial software-defined 5G gNodeB (gNB) and to quantify the latency achievable for a realistic XR emergency guidance workload. To avoid any ambiguity between design and demonstration, the MAC-layer optimization and the RTT-driven profile switching mechanism reported here are experimentally validated, whereas the reinforcement learning (RL)-based intent mechanism is presented strictly as a design contribution that has not yet undergone experimental validation. Likewise, the clinical guidance component has so far been assessed only functionally; no user study or formal evaluation of its clinical effectiveness has yet been performed (see Section 7).

In this context, EMERGE-XR is presented, developed under the 6G-BRICKS Horizon Europe programme [7], with the following contributions.

- A modular XR application for the Meta Quest 3 (Meta Platforms, Inc., Menlo Park, CA, USA) headset implementing the ABCDE (Airway, Breathing, Circulation, Disability, Exposure) trauma protocol with AI-assisted pose estimation and clinical decision support;
- A 5G Radio Access Network (RAN) latency optimization reducing average RTT by 62.5%, validated on-premises over an Amarisoft (Paris, France) 5G gNB, together with a lightweight RTT-driven profile-switching mechanism validated across more than 100 experimental runs;
- A proposed AI-driven intent-based extension using a contextual RL agent for autonomous MAC-layer profile selection, presented at the conceptual design stage.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 describes the EMERGE-XR system architecture and its user-facing interfaces. Section 4 presents the 5G testbed, the MAC-layer optimization methodology, the dynamic profile-switching mechanism, and the experimental evaluation. Section 5 introduces the proposed AI-driven intent-based runtime configuration. Section 6 outlines the planned validation over the KU Leuven testbed, Section 7 discusses limitations and future work, and Section 8 concludes the paper.

2. Related Work

2.1. XR for Emergency Response

XR technologies are increasingly explored for prehospital emergency care. A recent systematic review of randomized controlled trials on AR for prehospital care [4] found that AR can improve procedural accuracy and reduce cognitive load for first responders, though scalability and field integration remain open challenges. Oztank et al. [5] proposed guidelines for first-responder-friendly AR interfaces. Del Carmen Cardós-Alonso et al. [6] reviewed XR training systems for mass casualty incidents and highlighted the need for

https://doi.org/10.3390/engproc2026167005

real-time network connectivity to support remote expert supervision. Beyond incident response, XR is also gaining ground in healthcare training itself; a recent systematic review documents the use of XR, gamification, and integrated digital learning tools in the education of healthcare personnel [8]. EMERGE-XR advances this line of work by deploying a complete AR-guided medical workflow over an optimized 5G link in realistic PPDR conditions.

2.2. 5G Latency Optimization and URLLC

Achieving ultra-low latency over 5G NR requires coordinated optimization across multiple layers. At the MAC layer, 5G introduces flexible numerologies, mini-slot transmissions, and Configured Grant (CG) mechanisms that eliminate per-packet scheduling request overhead for uplink transmissions [9]. Ksentini et al. [10] addressed low latency guarantees in slicing-ready 5G systems via two-level MAC scheduling. A recent analysis at ACM HotNets [2] showed that TDD slot patterns and mini-slot configurations critically determine achievable latency. Ibrahim et al. [11] provide a comprehensive taxonomy of cross-layer Ultra-Reliable Low-Latency Communication (URLLC) techniques for 6G-enabled Industry 5.0, including proactive Hybrid Automatic Repeat Request (HARQ) adaptation and grant-free access.

Closest to the present work, Kornacki et al. [12] studied low-latency 5G performance using Amarisoft-based gNB configuration tuning, including RTT measurements and latency-throughput trade-offs, in Unmanned Ground Vehicle (UGV) scenarios. The present work differs in three respects: (i) it targets an interactive XR medical-guidance workload with a sub-10 ms RTT budget rather than UGV control traffic; (ii) it packages the validated parameter sets into self-contained gNB configuration profiles that are switched automatically at runtime by an RTT-driven hysteresis mechanism; and (iii) in the proposed intent-based extension, it couples the radio configuration to application semantics, namely the active phase of the ABCDE clinical protocol.

2.3. Intent-Based Networking for 5G/6G

Intent-Based Networking (IBN) abstracts network management into high-level declarative objectives. Abbas et al. [13] proposed intent-based slice lifecycle management for 5G via static rule engines. Chowdhury [14] proposed Accelerator, an intent-based intelligent resource slicing scheme for 6G application execution over Software Defined Networking (SDN) and Network Function Virtualization (NFV)—empowered zero-touch networks. The ALLSTaR framework [15] leverages large language models for automated, intent-driven scheduler generation as O-RAN dApps; however, this operates at design time. The NASP platform [16] focuses on core-network and transport-network orchestration, not real-time MAC-layer adaptation. Reinforcement learning has been applied to RAN resource management [17], yet existing RL-based approaches optimize network-centric metrics without awareness of application-level semantics. The approach proposed here observes application QoE signals jointly with network KPIs, exploits application-semantic context, and operates in a closed loop at runtime.

2.4. Cell-Free Massive MIMO

Cell-Free Massive MIMO (CFmMIMO) eliminates traditional cell boundaries by distributing a large number of access points that coherently serve all users, providing uniform coverage and macro diversity [18]. Chen et al. [19] analyzed statistical delay performance in uplink CFmMIMO-aided URLLC systems, demonstrating that CFmMIMO can satisfy stringent delay violation probability bounds. The 6G-BRICKS KU Leuven testbed implements CFmMIMO with USRP X410 (Ettus Research, Austin, TX, USA) software-defined radios.

3. System Architecture

Figure 1 illustrates the end-to-end EMERGE-XR architecture and its data flow across four tiers. The Field tier comprises the Meta Quest 3 headset (60 frames per second (FPS), AR overlay rendering), a Shimmer3 (Shimmer Research Ltd., Dublin, Ireland) wearable sensor streaming electrocardiogram (ECG), heart rate, and body temperature via Bluetooth Low Energy (BLE) to an Android relay that republishes the measurements over MQTT, a local Access Point, and a Teltonika TRB500 (Teltonika Networks, Kaunas, Lithuania) 5G Customer Premises Equipment (CPE) gateway. The 5G RAN tier consists of the Amarisoft Callbox Mini, which co-hosts the gNB and the 5G core functions (AMF, SMF, UPF); the 5G Standalone (SA) segment between the CPE and the gNB is the critical latency path, where the MAC-layer optimization achieves sub-10 ms RTT. The Edge tier hosts the containerized services: pose-svc executes MediaPipe Pose (Google LLC, Mountain View, CA, USA) inference (below 50 ms, 33 skeletal landmarks returned via WebSocket) and guidance-svc implements the ABCDE clinical state machine with a Glasgow Coma Scale (GCS) triage wizard; the Intent Agent (RL) block shown in the same tier is the proposed runtime configuration component described in Section 5. As detailed in the lower panel of Figure 1, uplink traffic carries JPEG frames and ECG waveforms, while the downlink delivers AR overlays and clinical guidance; the Circulation phase imposes the most stringent RTT requirement due to real-time cardiopulmonary resuscitation (CPR) feedback. The Telemetry tier, built on Prometheus (Cloud Native Computing Foundation, San Francisco, CA, USA) and Grafana (Grafana Labs, New York, NY, USA) over an OpenTelemetry-instrumented pipeline, provides real-time observability of inference latency, frame rate, RTT, and network quality indicators.

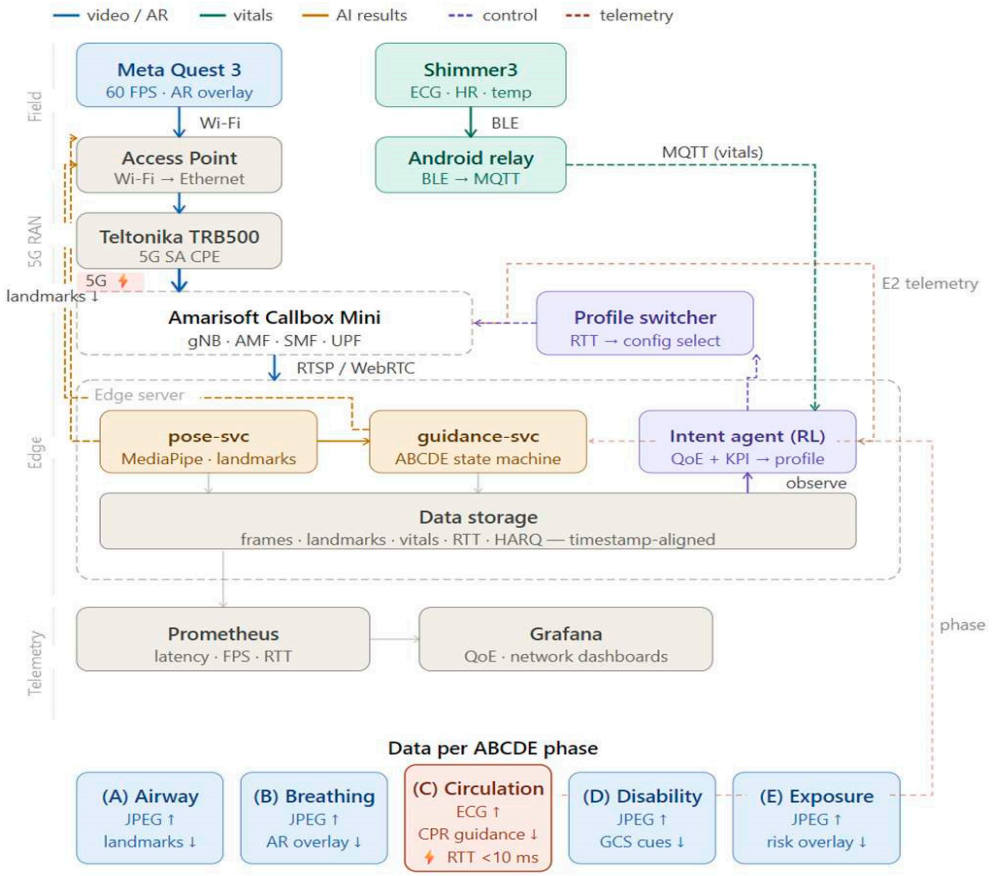
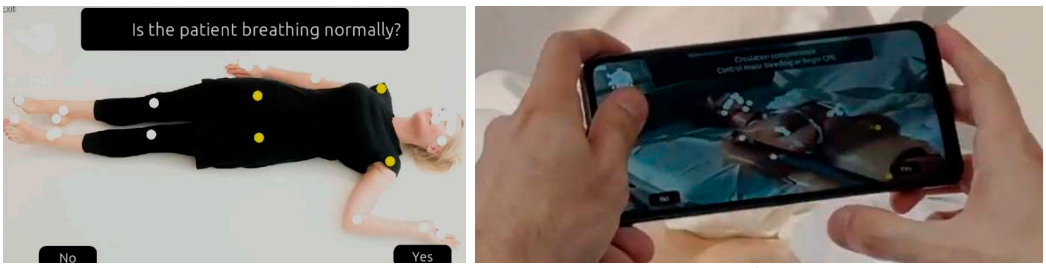


Figure 1. EMERGE-XR architecture and data flow across the Field, 5G RAN, Edge, and Telemetry tiers, including the per-phase ABCDE data exchange (lower panel). Arrow colors and line styles denote the traffic type, as indicated in the legend at the top of the figure; in the lower panel, ↑ and ↓ denote uplink and downlink traffic, respectively.

To make the implementation status explicit, all Field, 5G RAN, Edge (pose-svc, guidance-svc), and Telemetry components in Figure 1 have been implemented and demonstrated on the laboratory testbed described in Section 4, whereas the Intent Agent remains at the design stage.

The end-to-end data flow operates as follows. Upon initiating an assessment session, the Meta Quest 3 headset captures passthrough frames at 60 FPS, encodes them as JPEG images at configurable quality levels, and transmits each frame over the 5G uplink to the edge server via a RESTful HTTP/1.1 interface. The pose svc microservice processes each incoming frame synchronously on the GPU, producing a 33-landmark skeletal representation that is returned to the headset within a target latency of 50 ms and rendered as AR overlays aligned to the patient's body. In parallel, guidance svc maintains stateful session context, tracking the responder's progression through the five ABCDE stages: clinical decision support is triggered by threshold events; for instance, detection of abnormal ECG waveforms from the Shimmer3 sensor automatically advances the session to the Circulation stage and activates CPR guidance mode. Voice commands recognized locally on the headset are forwarded to guidance svc for intent classification and workflow advancement, enabling fully hands-free operation during critical procedures.

Figure 2 illustrates the actual user-facing interfaces of the system: (a) the graphical user interface (GUI) with interactive visual markers (hotspots) placed on the simulated patient's body for anatomical and physiological evaluation, and (b) a first-person view of the EMERGE-XR mobile application displaying safety and procedural prompts (e.g., circulation compromise, bleeding control) during a simulated CPR training session.



(**a**) (**b**)

Figure 2. EMERGE-XR user interfaces: (**a**) GUI layout with interactive hotspots on the simulated patient's body; (**b**) first-person view of the mobile application during a simulated CPR training session.

4. 5G Testbed and Experimental Evaluation

4.1. Testbed Infrastructure

The latency optimization was validated on an on-premises 5G Standalone testbed at the facilities of PIKEI New Technologies (Patras, Greece), a technology company specializing in IoT, networked systems, and applied R&D, serving as Lead Beneficiary of the EMERGE-XR sub-project within 6G-BRICKS (Grant Agreement No. 101096954). On the user side, a Meta Quest 3 headset (Meta Platforms, Inc., Menlo Park, CA, USA) connects via Wi-Fi to a local Access Point, bridged through Ethernet to the Teltonika TRB500 5G CPE gateway, which provides 5G SA connectivity to the Amarisoft Callbox Mini. On the server side, an edge node (AMD Ryzen 9-7950X (Advanced Micro Devices, Inc., Santa Clara, CA, USA), NVIDIA GeForce RTX 4080 (NVIDIA Corporation, Santa Clara, CA, USA)) hosts the containerized pose-svc and guidance-svc, connected through a managed switch. The following equipment was integrated into the Embedded Systems & Design Applications (ESDA) laboratory infrastructure:

- Amarisoft Callbox Mini: a compact 3GPP-compliant eNB/gNB and EPC/5GC solution supporting functional and performance testing with a single-cell configuration up to

20 MHz with 2 × 2 MIMO, up to 500 concurrent active UEs, and a maximum throughput of 200 Mbps downlink (DL) and 75 Mbps uplink (UL);

- Teltonika TRB500 5G CPE gateway: a compact industrial gateway supporting 5G Sub-6 GHz SA/NSA with a maximum throughput of 2.1 Gbps DL and 900 Mbps UL, capable of reporting RSSI, SINR, RSRP, and RSRQ signal parameters.

The Amarisoft Callbox implements a full software-defined 4G/5G protocol stack, encompassing the complete gNB layer hierarchy from the physical layer through the data link layer (L1, L2/MAC) up to the radio resource control (RRC) layer (L3), executing on a general-purpose Linux-based ×86 platform interfaced with a 50 MHz 2 × 2 MIMO Software Defined Radio (SDR) front-end. The complete core network functions are co-hosted on the same platform. This architecture affords direct programmatic access to MAC-layer scheduling parameters at runtime, enabling fine-grained latency optimization without hardware modifications. Within this stack, the Layer 2 (L2) MAC scheduler constitutes the primary control point governing uplink and downlink scheduling decisions, grant allocation, and timing relationships between the User Equipment (UE) and the gNB, while the Layer 1 (L1) physical layer enforces the TDD frame structure. Both layers are configured via the Amarisoft configuration file (enb.cfg), which is loaded at gNB process initialization.

4.2. MAC-Layer Optimization Methodology

Two distinct configuration profiles were developed and validated: a throughput-oriented baseline profile and a latency-optimized profile. Each profile is a self-contained Amarisoft gNB configuration file that encapsulates a complete, consistent set of MAC-layer and physical-layer parameters. Profile development followed a systematic iterative process: starting from the Amarisoft default configuration, each MAC and L1 parameter was adjusted in isolation while RTT was continuously monitored via ICMP ping, and the adjustment was retained only if it produced a measurable RTT improvement without radio instability; the process continued until a stable minimum-RTT operating point was reached. The active profile is selected at runtime by atomically updating the enb.cfg symbolic link and restarting the gNB process. The following parameters were tuned:

1. Scheduling Request (SR) periodicity (L2/MAC): the SR periodicity defines the interval at which the UE may transmit a scheduling request to the gNB, signaling the availability of uplink data in its buffer. In the latency-optimized profile, it was reduced to its minimum feasible value, minimizing the worst-case waiting time between data arrival at the UE transmit buffer and the reception of an uplink grant, which directly reduces the first-hop contribution to uplink RTT;
2. TDD UL/DL slot pattern (L1): the default throughput-oriented pattern allocates a higher proportion of slots to the downlink direction. For the latency-optimized profile, the pattern was reconfigured to increase the density of uplink slots, reducing the interval between consecutive UL transmission opportunities within the TDD frame;
3. MAC timing parameters (rx_to_tx_latency, uplink scheduling offsets) (L2/MAC): under adaptive scheduling, these parameters vary dynamically, introducing jitter into the RTT distribution. In the latency-optimized profile, both were fixed to their minimum permissible values, enforcing deterministic scheduling behavior and eliminating latency variability introduced by adaptive timing algorithms.

The combined effect of these optimizations addresses all three primary contributors to MAC-layer latency: the UE-side buffer wait time (reduced by SR periodicity minimization), the radio frame access wait time (reduced by TDD pattern reconfiguration), and the gNB-side processing and scheduling delay (reduced by fixed MAC timing enforcement). Together, these modifications resulted in a reduction of average RTT from 24 ms to 9 ms, a 62.5% improvement.

https://doi.org/10.3390/engproc2026167005

4.3. RTT-Driven Dynamic Profile Switching

To support heterogeneous traffic requirements within a single 5G Standalone cell, a lightweight closed-loop automation mechanism dynamically switches between the throughput-optimized and latency-optimized gNB configuration profiles. The mechanism is realized as a Bash script (gnb_profile_switch.sh) executed periodically via a system timer at fixed 30-s intervals. The switching decision is driven exclusively by real-time RTT measurements: at each cycle, the script issues three ICMP echo requests to the UE and extracts the average RTT. RTT constitutes the sole decision metric because it directly captures the aggregate end-to-end latency experienced by the application, encompassing radio access, MAC scheduling, and core network processing delays, without requiring deep packet inspection. If the UE does not respond within the timeout, an RTT of 1000 ms is assumed, driving the system toward the low-latency profile to preserve QoE. It is noted that the measured RTT path traverses both the Wi-Fi segment and the 5G segment; in the laboratory setup, the Wi-Fi contribution was measured at under 2 ms and is considered negligible relative to the observed RTT reduction.

Two RTT thresholds govern profile selection: HIGH_RTT = 150 ms, above which the system transitions to the low-latency profile, and LOW_RTT = 40 ms, below which it reverts to the throughput-oriented profile. These thresholds were determined empirically: 150 ms indicates scheduling degradation that impairs interactive XR feedback, while 40 ms represents a stable low-latency regime; the intermediate band acts as a neutral zone where counters decay, but no switching occurs. A hysteresis mechanism prevents spurious switching: transition to low-latency requires three consecutive measurements above HIGH_RTT, and reversion requires five consecutive measurements below LOW_RTT; the asymmetry reflects the higher cost of premature reversion during an active XR session. A cooldown of 30 s is enforced between switches to prevent gNB restart storms. Across more than 100 experimental runs, the mechanism demonstrated stable convergence with no oscillations under sustained XR streaming conditions. The complete decision logic is summarized in Algorithm 1.

Algorithm 1: RTT-driven gNB profile switching

Input: AVG_RTT (ms), MODE, HIGH_CNT, LOW_CNT, LAST_SWITCH
Output: updated MODE, configuration symlink, gNB restart (if needed)
AVG_RTT ← ping(UE_IP, count = 3, timeout = 1 s)
if AVG_RTT = ∅ then AVG_RTT ← 1000
if AVG_RTT ≥ HIGH_RTT (150 ms) then
    HIGH_CNT ← HIGH_CNT + 1; LOW_CNT ← 0
else if AVG_RTT ≤ LOW_RTT (40 ms) then
    LOW_CNT ← LOW_CNT + 1; HIGH_CNT ← 0
else ▷ neutral band [40, 150] ms
    HIGH_CNT ← max(HIGH_CNT – 1, 0)
    LOW_CNT ← max(LOW_CNT – 1, 0)
if (NOW – LAST_SWITCH) < COOLDOWN (30 s) then return
if HIGH_CNT ≥ HIGH_COUNT_TH (3) and MODE ≠ LOW_LAT then
    symlink(enb.cfg → gnb-sa-tdd-low-latency.cfg); restart_gNB()
    MODE ← LOW_LAT; LAST_SWITCH ← NOW; reset counters
if LOW_CNT ≥ LOW_COUNT_TH (5) and MODE ≠ THROUGHPUT then
    symlink(enb.cfg → gnb-sa.cfg); restart_gNB()
    MODE ← THROUGHPUT; LAST_SWITCH ← NOW; reset counters

https://doi.org/10.3390/engproc2026167005

4.4. Experimental Results and Discussion

Experiments were repeated more than 100 times under identical conditions during real-time audio streaming sessions. Round-trip time measurements were obtained using ICMP echo requests between the gNB and the connected UE, capturing aggregate end-to-end latency including radio access, scheduling, processing, and core network delays. Figure 3 juxtaposes RTT and DL/UL throughput over time before (a) and after (b) the MAC-layer optimization, while Table 1 summarizes the measured performance. Peak DL/UL throughput in Table 1 reports the saturating capacity of each configuration, measured independently of the RTT sessions (see the discussion at the end of this section); it is expressed in Mbps and is distinct from the offered load of the audio stream used during RTT measurement.

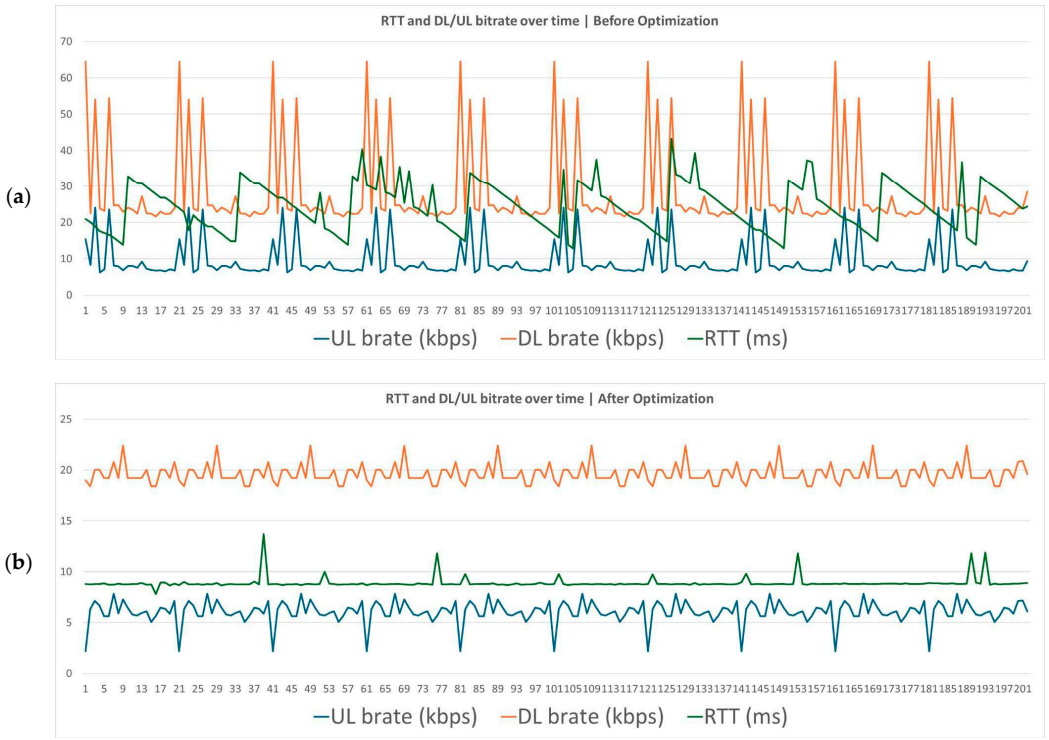


Figure 3. RTT and DL/UL bitrate over time: (**a**) before MAC-layer optimization; (**b**) after MAC-layer optimization. The horizontal axis denotes the ping sample index; RTT (ms) and DL/UL bitrate (kbps) share the vertical axis.

Three independent validation layers were employed to assess system performance. (1) Latency (RTT): prior to optimization, the average RTT was approximately 24 ms, with frequent fluctuations (21.8% of samples exceeding 30 ms); following the application of the low-latency configuration, the average RTT was reduced to approximately 9 ms, with no samples exceeding 30 ms, corresponding to a latency reduction of approximately 62.5%; the underlying per-sample distribution is characterized in Table 2, which reports its dispersion and tail behavior. (2) Radio and bandwidth behavior: stable downlink and uplink bitrates were maintained under both configurations; the peak throughput capacity of each configuration differed by approximately −44.5% in downlink and +76.0% in uplink (Table 1; see discussion below), while HARQ activity and retransmissions remained within nominal limits and no radio instability was observed. (3) Application-level validation: continuous real-time audio streaming sessions were conducted throughout experimentation; no audio gaps, jitter artifacts, or packet drops were recorded, and across all repetitions the reduced-latency configuration consistently produced stable, artifact-free audio quality.

Table 1. 5G RAN performance before vs. after MAC-layer optimization.

Metric	Before	After
Avg. RTT (ms)	24	9
RTT spikes > 30 ms	21.8%	0%
Peak DL throughput (Mbps)	130	72.1
Peak UL throughput (Mbps)	14.6	25.7
DL capacity change	—	−44.5%
UL capacity change	—	+76.0%
RTT reduction	—	62.5%
Packet loss	None	None

Table 2. Dispersion of the per-sample RTT distribution ($n = 202$ per configuration), computed from the continuous ping traces underlying Figure 3.

RTT Metric	Before	After
Mean ± SD (ms)	24.39 ± 6.52	8.89 ± 0.57
Median [IQR] (ms)	24.35 [18.80–29.80]	8.78 [8.75–8.81]
p95/p99/max (ms)	34.3/39.3/43.3	9.0/11.8/13.7
95% CI of mean (ms)	[23.50, 25.29]	[8.81, 8.97]
Samples > 30 ms	21.8%	0%

Regarding the validity of the measurement methodology, ICMP RTT is used as a network-level proxy that captures the aggregate radio-access, scheduling, and core-network delay, but it does not by itself validate end-to-end XR Quality of Experience; for this reason, the continuous application-level streaming sessions described above were used as a complementary validation layer, while a statistical characterization across all experimental runs and a user-level QoE assessment are identified as future work (Section 7).

For ultra-low-latency, real-time immersive applications such as bidirectional audio–video streaming for remote first-aid assistance, including CPR guidance, end-to-end latency constitutes the dominant QoE metric, while peak throughput requirements remain comparatively modest. The RTT reduction from 24 ms to 9 ms reduces the radio-access contribution to the end-to-end pipeline latency, which in turn improves audio–visual synchronization and operator situational awareness. Note that the full motion-to-photon latency encompasses additional stages beyond RTT, including frame capture, JPEG encoding, pose inference (approximately 50 ms), and AR overlay rendering; the 9 ms RAN RTT represents the radio-access component of this budget rather than the total perceived delay. The associated throughput reduction is expected and acceptable, as modern real-time video and voice codecs operate efficiently at the available bitrates while maintaining high-definition stream quality. Saturating-throughput tests (speedtest.net (Ookla, LLC, Seattle, WA, USA), measured from the UE through the Amarisoft core) show the throughput-oriented profile attains approximately 130 Mbps DL and 14.6 Mbps UL, while the latency-optimized profile attains approximately 72.1 Mbps DL and 25.7 Mbps UL. The reconfiguration therefore does not degrade capacity uniformly but repartition it between directions: a −44.5% change in DL capacity offset by a +76.0% change in UL capacity, consistent with the increased density of uplink transmission opportunities described in Section 4.2. This repartitioning is well matched to the EMERGE-XR traffic profile, whose dominant load is the uplink (60 FPS JPEG frames and ECG waveforms), while the downlink carries comparatively light AR overlay and clinical guidance payloads; both configurations retain at least an order of magnitude more capacity than a high-definition real-time media stream requires. During the RTT measurement sessions themselves, the offered load was a live audio stream measured at

approximately 28 → 20 kbps DL and 9 → 6 kbps UL; these figures characterize the traffic load present while RTT was measured and are not capacity measurements.

5. Proposed AI-Driven Intent-Based Runtime Configuration (Design Stage, Not Yet Validated)

The RTT-based profile switching mechanism presented in Section 4 demonstrates that runtime gNB reconfiguration is technically feasible and effective in a real 5G testbed. However, it operates reactively on a fixed 30-s cycle using static RTT thresholds, without awareness of application-level context or the active ABCDE protocol phase. As a proposed next step, which has not yet been experimentally validated, this section describes an AI-driven intent-based extension that replaces the threshold-based logic with a learned RL policy, while reusing the same validated infrastructure: the enb.cfg symlink mechanism, the gNB restart sequence, and the empirically tuned configuration profiles. The RL-based intent mechanism remains at the conceptual design stage; the results reported in Section 4 exclusively concern the MAC-layer optimization and the RTT-driven profile switching infrastructure.

5.1. Architecture and AI Agent

An Intent Translation Layer (ITL) is envisioned, whose core would be a contextual RL agent mediating between the XR application and the gNB configuration API; its position in the overall system, together with its interfaces to the profile switcher and the telemetry pipeline, is depicted in the Edge tier of Figure 1. The intended closed-loop workflow is as follows: (1) intent declaration: the XR application expresses a high-level intent via a lightweight REST API encoding target QoE bounds and the current ABCDE protocol phase; (2) state observation: the agent constructs a joint state vector from application QoE signals (pose inference latency, frame delivery jitter, AR overlay update rate) and network KPIs (RTT, DL/UL throughput, HARQ retransmission rate, MCS stability); (3) action selection: the agent selects a MAC-layer configuration profile using an ε-greedy exploration strategy ($\varepsilon = 0.1$), converging to a learned policy mapping (application-phase, radio-condition) tuples to optimal profiles; (4) configuration enforcement: the selected profile is applied via the enb.cfg symbolic link mechanism, followed by a brief gNB process restart, with a 30-s cooldown between successive switches; (5) reward and adaptation: the agent receives a scala*r* reward r = +1 if all active QoE constraints are simultaneously satisfied (RTT below the phase-specific threshold, inference latency below 50 ms, frame delivery jitter below 5 ms), $r = -1$ if any hard constraint is violated, and a penalty of -0.5 for each unnecessary profile switch. The agent is modelled as a tabular contextual bandit operating over a finite joint state space of (ABCDE phase, radio condition bin), with a decaying learning rate. The model is lightweight (below 10 K parameters), enabling deployment on resource-constrained edge nodes. The ITL integrates with O-RAN-compliant architectures as an rApp within the Non-Real-Time RAN Intelligent Controller (RIC), consuming E2 telemetry and issuing A1 policy directives [15].

5.2. Configuration Profiles

Three canonical profiles constitute the agent's discrete action space:

- Ultra-low-latency (~9 ms RTT, empirically validated): aggressive SR reduction, low-latency TDD pattern, fixed MAC timing; selected during active AR-guided procedures (e.g., CPR guidance during the Circulation phase);
- Balanced (projected ~15 ms RTT, an intermediate operating point interpolated between the two validated profiles and not yet empirically characterized): moderate SR peri-

odicity, standard TDD pattern with additional UL slots; preferred during periodic AI inference phases;
- Throughput-optimized (~24 ms RTT, empirically validated): default scheduler with adaptive timing; selected during bulk data transfer or telemetry export between active procedure stages.

6. Planned Validation over the 6G-BRICKS KU Leuven Testbed

The final system validation is planned at the 6G-BRICKS KU Leuven (KUL) testbed, which provides a disaggregated 5G RAN based on USRP X410 software-defined radios with CFmMIMO infrastructure. The scenario validates: (i) latency stability during mobility via handover-free operation; (ii) intent-based profile switching across active AR guidance and monitoring phases; and (iii) end-to-end KPIs including below 100 ms pipeline latency, above 60 FPS rendering, below 50 ms AI inference, and above 99.9% packet delivery. The CFmMIMO architecture is particularly well-suited to the EMERGE-XR use case due to its inherent macro-diversity and handover-free mobility support: as an ambulance transits within the coverage area, the distributed access points coherently combine received signals, maintaining consistent uplink channel quality without triggering handover procedures that would introduce latency spikes. This property is critical for sustaining the sub-10 ms RTT requirement during the Circulation phase, where interruptions in AI inference feedback during CPR guidance could directly affect patient outcomes.

Network traces, sensor data, and AI inference logs collected during the on-premises trials at PIKEI will be replayed over the KUL testbed to enable reproducible benchmarking under controlled laboratory conditions. This replay methodology allows direct comparison of latency, throughput, and QoE metrics between the single-cell Amarisoft environment and the distributed CFmMIMO infrastructure, isolating the contribution of the radio access architecture to overall system performance. For this purpose, a representative end-to-end evaluation scenario has been defined: a first responder wearing a Meta Quest 3 headset performs the ABCDE trauma assessment on a simulated patient, guided by real-time AR overlays through five stages: (A) Airway: pose estimation validates jaw thrust placement; (B) Breathing: animated chest visualization with voice-driven workflow; (C) Circulation: ECG analysis triggers CPR guidance with compression depth feedback; (D) Disability: GCS assessment with pupil and motor response AR cues; and (E) Exposure: anatomical risk zone overlays for systematic injury inspection.

7. Limitations and Future Work

The presented evaluation was conducted in a controlled laboratory environment using a single 5G NR cell and a limited number of user devices. Although experiments were repeated more than 100 times under identical conditions, the results may differ in large-scale deployments with higher user density and dynamic radio conditions; accordingly, the sub-10 ms RTT result should be interpreted as demonstrated feasibility under controlled single-cell conditions rather than a guarantee for realistic multi-user emergency deployments. In addition, the per-sample RTT statistics reported in Table 2 (dispersion, percentiles, and confidence intervals) are derived from one representative trace per configuration, and application-level quality was assessed qualitatively; a statistical characterization across all experimental runs and deployment conditions is planned as part of the KUL validation campaign. Finally, the clinical XR component has so far been assessed only functionally: no user study, expert evaluation, or formal assessment of the effectiveness of the ABCDE guidance has yet been performed, and such a study is a prerequisite before any claim of clinical utility can be made.

https://doi.org/10.3390/engproc2026167005

An additional limitation of earlier prototype iterations concerned profile switching oscillation under bursty traffic, where detection based on instantaneous traffic snapshots lacked temporal smoothing. The RTT-based switching mechanism adopted in this work addresses this through a combined hysteresis counter approach and a 30-s cooldown between successive profile changes, which effectively suppressed oscillations across all experimental runs. Remaining open challenges include adaptation of threshold values to heterogeneous deployment environments and extension to multi-UE scenarios with concurrent latency demands.

Furthermore, the proposed RL agent for intent-based configuration has not yet been implemented or evaluated and relies on a pre-validated discrete action space of three profiles. While this simplification facilitates rapid convergence and reliable deployment on resource-constrained edge nodes, it limits the granularity of adaptation. Future work is envisioned to focus on: (i) experimental validation of the RL intent mechanism over the 6G-BRICKS CFmMIMO testbed at KU Leuven; (ii) extension of the proposed mechanism to multi-UE and multi-cell scenarios; (iii) field trials in non-controlled environments; and (iv) exploration of continuous action spaces and multi-objective reward formulations that jointly optimize latency, throughput, and energy efficiency, in alignment with the sustainability goals of 6G network deployments.

8. Conclusions

EMERGE-XR has been presented, an XR-based emergency medical guidance system that achieves real-time interactivity over 5G through targeted RAN latency optimization, reducing average RTT from 24 ms to 9 ms on an on-premises Amarisoft testbed, a 62.5% improvement, with zero packet loss. A lightweight dynamic gNB profile switching mechanism, validated across more than 100 experimental runs, enables autonomous selection between throughput-optimized and latency-optimized configurations based on measured RTT, with demonstrated QoE improvements. As a work-in-progress contribution, an AI-driven intent-based extension is proposed, in which a contextual RL agent would observe application QoE signals and network KPIs, including the active ABCDE protocol phase, to autonomously select optimal MAC-layer configurations at runtime; this extension remains at the design stage. Experimental validation of the intent mechanism, together with end-to-end system evaluation over the 6G-BRICKS CFmMIMO testbed at KU Leuven, constitutes the primary next step.

Author Contributions: Conceptualization, I.T. and C.P.A.; methodology, I.T.; software, I.T.; validation, I.T., C.P. and K.A.; investigation, I.T.; resources, M.P. and V.K.; writing—original draft preparation, I.T.; writing—review and editing, C.P., K.A., M.P., V.K., N.V. and C.P.A.; supervision, N.V. and C.P.A.; project administration, M.P.; funding acquisition, M.P. and N.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the European Union's Horizon Europe research and innovation programme through the 6G-BRICKS project under Grant Agreement No. 101096954.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: The authors would like to thank ESDA LAB and PIKEI New Technologies (Patras, Greece) for hosting the on-premises 5G Standalone testbed used in this work. During the preparation of this manuscript, the authors used Claude (Anthropic, Claude Opus 5.5) for the purposes of language editing, restructuring of the text in response to reviewer and editorial comments, and

formatting according to the journal template. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ABCDE	Airway, Breathing, Circulation, Disability, Exposure
AI	Artificial Intelligence
AMF/SMF/UPF	Access and Mobility Management/Session Management/User Plane Function
AR	Augmented Reality
BLE	Bluetooth Low Energy
CFmMIMO	Cell-Free Massive Multiple-Input Multiple-Output
CG	Configured Grant
CI	Confidence Interval
CPE	Customer Premises Equipment
CPR	Cardiopulmonary Resuscitation
DL/UL	Downlink/Uplink
ECG	Electrocardiogram
FPS	Frames Per Second
GCS	Glasgow Coma Scale
gNB	gNodeB (5G base station)
GUI	Graphical User Interface
HARQ	Hybrid Automatic Repeat Request
IBN	Intent-Based Networking
ICMP	Internet Control Message Protocol
IQR	Interquartile Range
ITL	Intent Translation Layer
KPI	Key Performance Indicator
MAC	Medium Access Control
MCS	Modulation and Coding Scheme
MQTT	Message Queuing Telemetry Transport
NFV	Network Function Virtualization
NR	New Radio
PPDR	Public Protection and Disaster Relief
QoE	Quality of Experience
RAN	Radio Access Network
RIC	RAN Intelligent Controller
RL	Reinforcement Learning
RRC	Radio Resource Control
RTT	Round-Trip Time
SA	Standalone
SD	Standard Deviation
SDN	Software-Defined Networking
SDR	Software-Defined Radio
SR	Scheduling Request
TDD	Time Division Duplex
UE	User Equipment
UGV	Unmanned Ground Vehicle
URLLC	Ultra-Reliable Low-Latency Communication
XR	Extended Reality

References

1. Abrash, M. The Sine Qua Non of AR and VR. Valve Software. Available online: https://web.archive.org/web/2013/http://blogs.valvesoftware.com/abrash/latency-the-sine-qua-non-of-ar-and-vr/ (accessed on 1 November 2026).
2. Maghsoudnia, A.; Vlad, E.; Gong, A.; Dumitriu, D.M.; Hassanieh, H. Ultra-Reliable Low-Latency in 5G: A Close Reality or a Distant Goal? In *Proceedings of the 23rd ACM Workshop on Hot Topics in Networks*; ACM: Irvine, CA, USA, 2024; pp. 111–120. [CrossRef]

3. Struye, J.; Van Damme, S.; Bhat, N.N.; Troch, A.; Van Liempd, B.; Assasa, H.; Lemic, F.; Famaey, J.; Vega, M.T. Toward Interactive Multi-User Extended Reality Using Millimeter-Wave Networking. *IEEE Commun. Mag.* **2024**, *62*, 54–60. [CrossRef]
4. Harari, R.E.; Schulwolf, S.L.; Borges, P.; Salmani, H.; Hosseini, F.; Bailey, S.K.T.; Quach, B.; Nohelty, E.; Park, S.; Verma, Y.; et al. Applications of Augmented Reality for Prehospital Emergency Care: Systematic Review of Randomized Controlled Trials. *JMIR XR Spat. Comput.* **2025**, *2*, e66222. [CrossRef] [PubMed]
5. Oztank, F.; Balcisoy, S. Towards Extended Reality in Emergency Response: Guidelines and Challenges for First Responder Friendly Augmented Interfaces. *Comput. Animat. Amp Virtual* **2025**, *36*, e70056. [CrossRef]
6. Del Carmen Cardós-Alonso, M.; Otero-Varela, L.; Redondo, M.; Uzuriaga, M.; González, M.; Vazquez, T.; Blanco, A.; Espinosa, S.; Cintora-Sanz, A.M. Extended Reality Training for Mass Casualty Incidents: A Systematic Review on Effectiveness and Experience of Medical First Responders. *Int. J. Emerg. Med.* **2024**, *17*, 99. [CrossRef] [PubMed]
7. European Commission. *6G-Bricks: Building Reusable Testbed Infrastructures for Validating Cloud-to-Device Breakthrough Technologies*; Grant Agreement ID: 101096954; CORDIS, European Commission: Brussels, Belgium, 2023. Available online: https://cordis.europa.eu/project/id/101096954 (accessed on 1 November 2026).
8. Graham, A.; Hökkä, M.; Pramila-Savukoski, S.; Flinkkilä, M.-L.; Tomietto, M.; Mikkonen, K. The Use of Digital Gamification, Extended Reality, Artificial Intelligence, and Integrated Digital Learning Tools in Palliative Care Education of Undergraduate Nurses: A Systematic Review. *Nurse Educ. Today* **2026**, *160*, 106982. [CrossRef] [PubMed]
9. Larrañaga, A.; Lucas-Estañ, M.C.; Lagén, S.; Ali, Z.; Martinez, I.; Gozalvez, J. An Open-Source Implementation and Validation of 5G NR Configured Grant for URLLC in Ns-3 5G LENA: A Scheduling Case Study in Industry 4.0 Scenarios. *J. Netw. Comput. Appl.* **2023**, *215*, 103638. [CrossRef]
10. Ksentini, A.; Frangoudis, P.A.; Amogh, P.C.; Nikaein, N. Providing Low Latency Guarantees for Slicing-Ready 5G Systems via Two-Level MAC Scheduling. *IEEE Netw.* **2018**, *32*, 116–123. [CrossRef]
11. Ibrahim, A.M.; Nordin, R.; Khamayseh, Y.S.M.; Amphawan, A.; Jasser, M.B. URLLC for 6G Enabled Industry 5.0: A Taxonomy of Architectures, Cross Layer Techniques, and Time Critical Applications. *ICT Express* **2026**, *12*, 789–809. [CrossRef]
12. Kornacki, J.; Wójcikowska, A.; Hoeft, M. Meeting Industrial 5G Requirements for High Uplink Throughput and Low Control Latency in UGV Scenarios. *Appl. Sci.* **2025**, *15*, 6427. [CrossRef]
13. Abbas, K.; Khan, T.A.; Afaq, M.; Song, W.-C. Network Slice Lifecycle Management for 5G Mobile Networks: An Intent-Based Networking Approach. *IEEE Access* **2021**, *9*, 80128–80146. [CrossRef]
14. Chowdhury, M. Accelerator: An Intent-Based Intelligent Resource-Slicing Scheme for SFC-Based 6G Application Execution Over SDN- and NFV-Empowered Zero-Touch Network. *Front. Comms. Net.* **2024**, *5*, 1385656. [CrossRef]
15. Elkael, M.; Polese, M.; Prasad, R.; Maxenti, S.; Melodia, T. ALLSTaR—Automated LLM-Driven Scheduler Generation and Testing for Intent-Based RAN. *IEEE Trans. Mob. Comput.* **2026**, 1–18. [CrossRef]
16. Grings, F.H.; Bruno, G.Z.; Prade, L.R.; Both, C.B.; Brito, J.M.C. NASP: Network Slice as a Service Platform for 5G Networks. *J. Netw. Comput. Appl.* **2026**, *250*, 104479. [CrossRef]
17. Ebrahimi, S.; Bouali, F.; Haas, O.C.L. Resource Management From Single-Domain 5G to End-to-End 6G Network Slicing: A Survey. *IEEE Commun. Surv. Tutor.* **2024**, *26*, 2836–2866. [CrossRef]
18. Ngo, H.Q.; Interdonato, G.; Larsson, E.G.; Caire, G.; Andrews, J.G. Ultra-Dense Cell-Free Massive MIMO for 6G: Technical Overview and Open Questions. *arXiv* **2024**, arXiv:2401.03898. [CrossRef]
19. Chong, B.; Lu, H. Statistical Delay Performance Analysis for URLLC in Uplink Cell-Free Massive MIMO Systems: A Stochastic Network Calculus Perspective. *IEEE Trans. Wirel. Commun.* **2025**, *24*, 1162–1177. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.