Delay-Reliability Aware Downlink Scheduling for Multi-sensory Immersive Extended Reality in NextG

Moyukh Laha, *Member, IEEE,* Goutam Das and Gabriel-Miro Muntean *Fellow, IEEE*

***Abstract*—Multi-sensory Extended Reality (mXR) applications represent a cornerstone technology for immersive communications in the 6G and IMT-2030 frameworks, enabling synchronized stimulation across multiple human sensory modalities to create authentic immersive experiences. mXR applications demand simultaneous satisfaction of stringent data rate, delay, and reliability requirements across all sensory streams, necessitating an unprecedented convergence of enhanced Mobile Broadband (eMBB) and Ultra-Reliable Low-Latency Communications (URLLC) service paradigms. Existing 5G scheduling schemes, designed to optimize either eMBB or URLLC services independently and typically handling decoupled data streams, do not address the cross-modal synchronization and joint optimization challenges of mXR applications. This paper proposes *DREAM-X (Delay-Reliability Aware Scheduling for Multi-sensory XR Applications)*, a unified scheduling framework specifically designed for next generation (NextG) networks to support mXR services. This framework includes a novel Multi-modal Delay Tracking Queue architecture that organizes multi-sensory data units based on transmission deadlines, enabling precise cross-modal delay coordination. We formulate the scheduling problem as a constrained optimization and develop a Model Predictive Control (MPC) solution with rolling horizon optimization to ensure computational tractability for real-time deployment. The simulation results show that DREAM-X provides multi-fold gains in the number of satisfied mXR users compared to conventional and state-of-the-art XR scheduling approaches, while maintaining strict delay and reliability constraints.**

***Index Terms*—Multi-sensory Extended reality, Immersive Communication, 6G, Delay-reliability aware scheduling.**

I. Introduction

Imagine a future where individuals can reach across continents to feel the texture of ancient artifacts, inhale the aroma of distant spices, witness breathtaking landscapes, or savor exotic flavors, all while engaging in shared virtual environments. This is the promise of multi-sensory Extended Reality (mXR). While traditional XR applications have primarily concentrated on visual and auditory experiences, true telepresence requires the simultaneous stimulation of all five human senses: sight, sound, touch, smell, and taste. This shift from video-based XR to fully multi-sensory XR represents a significant paradigm change. It redefines immersive communications by enabling cyber-physical presence, an essential foundation for the Metaverse and a core objective of the ITU-R IMT-2030 vision [1]–[4]. The impact spans diverse critical domains, including students engaging in immersive education, engineers operating industrial equipment remotely, surgeons performing teleoperations with haptic feedback [2], [5]. These scenarios are not speculative projections; they are concrete enablers of the IMT-2030's vision for immersive communications [3].

However, delivering authentic mXR experiences over 6G and next generation (NextG) networks presents profound challenges that exceed conventional 5G paradigms and expose its critical limitations. Unlike traditional applications that primarily handle decoupled audiovisual streams, mXR involves the concurrent orchestration of five heterogeneous sensory modalities—visual, auditory, haptic, olfactory, and gustatory—each with distinct and stringent quality-of-service (QoS) requirements. According to the 3GPP TS 22.261 [6], immersive video requires high throughput (1–100 Mbps), sub-10 ms latency, and 99.9% reliability. Spatial audio demands data rates ranging from 5 to 512 kbps, with latency below 10 ms and similar reliability to preserve auditory coherence. Haptic feedback imposes even more stringent constraints, requiring 0.8–2 Mbps throughput, latency under 5 ms, and reliability levels between 99.9% and 99.999%. Although not yet formally specified by 3GPP, olfactory and gustatory streams—enabled through environmental and taste simulation technologies—are expected to require approximately 0.1 Mbps, latency below 100 ms, and tolerate packet loss rates up to 10^{-3} [5], representing more relaxed but still critical constraints for complete sensory immersion.

Recognizing these demands, 3GPP has classified mXR under *tactile and multi-modal communication services* [6], incorporating explicit cross-modal synchronization constraints to mitigate perceptual dissonance. This multi-sensory convergence disrupts the foundational 5G service triangle, traditionally composed of Enhanced Mobile Broadband (eMBB), Ultra-Reliable Low-Latency Communications (URLLC) and Massive Machine-Type Communications (mMTC), by necessitating a new service paradigm. Volumetric video aligns with the characteristics of eMBB, while haptic feedback exemplifies the URLLC demands. Their simultaneous satisfaction is beyond the scope of existing 5G service categories, prompting the emergence of *Mobile Broadband and Reliable Low-Latency Communications* (MBRLLC), a paradigm that encapsulates the joint high-throughput, ultra-low-latency, and high-reliability requirements identified in recent literature [7]. Ultimately, this convergence highlights a significant gap and calls for novel solutions in NextG networks.

Although enabling mXR in NextG requires advances across

M. Laha, and G.-M. Muntean are with the School of Electronic Engineering, Dublin City University, Ireland and with the Lero Research Ireland Centre for Software. G. Das is with the Dept. of Electronics & Electrical Communication Engineering, Indian Institute of Technology Kharagpur, India.
Corresponding author: Moyukh Laha, E-mail: laha.moyukh@ieee.org.

This work was funded by Taighde Éireann – Research Ireland (Grant No. 13/RC/2094_2) and co-funded by the European Union (SyMeCo, Grant Agreement No. 101081459). The views expressed are solely those of the author(s). Neither the European Union nor the European Research Executive Agency can be held responsible for them.

multiple layers of the network stack and architectural redesigns, one of the most pressing challenges lies in developing novel scheduling mechanisms. Such mechanisms must satisfy stringent delay and reliability requirements across all sensory modalities—what we term *delay-reliability* constraints—necessitating fundamentally new schedulers. Despite extensive research on 5G scheduling strategies [8], [9], the current approaches are not suitable for mXR applications. Existing proposals primarily focus on optimizing for isolated service categories, such as eMBB or URLLC, or address objectives such as throughput maximization in visual XR domains [10]–[15]. However, they fail to address the compounded QoS demands required by multi-sensory streams. To the best of our knowledge, no existing work proposes a unified scheduler capable of concurrently handling all five sensory modalities while jointly ensuring their *delay-reliability* constraints. This work addresses this fundamental gap.

We introduce *DREAM-X*: Delay-Reliability Aware Scheduling for Multisensory XR Applications, a holistic scheduling framework specifically designed for the unique demands of mXR in 6G and NextG networks. The key contributions of this work are summarized as follows:

- *Multi-modal Delay Tracking Architecture*: A novel Multi-modal Delay Tracking Queue (mDTQ) architecture is proposed, which concurrently monitors Data Unit (DU) deadlines across all five sensory modalities while maintaining cross-modal synchronization, essential for immersive communications.
- *Joint Delay-Reliability Optimization*: A unified optimization problem is formulated employing this mDTQ framework to maximize the number of mXR users satisfying the QoS constraints across all five sensory modalities, with its NP-hard complexity established.
- *Tractable MPC Solution*: To address the computational intractability of the original optimization problem, the formulation is mapped into a Model Predictive Control (MPC) framework with a rolling horizon optimization, leveraging domain-specific insights to reformulate the problem and propose practical solutions suitable for real-time deployment.
- *Performance Evaluation*: Extensive simulations are conducted to analyze the impact of critical system parameters on the proposed solutions, demonstrating superior performance compared to state-of-the-art scheduling baselines.
- *Standardization Insights*: The implications of this work for 6G standardization are discussed, highlighting the changes required for the practical integration of multi-sensory XR into NextG network specifications.

The remainder of this paper is organized as follows: Section II describes related work. Section III presents the DREAM-X framework, including the multi-modal delay tracking mechanism, optimization problem formulation, and the Model Predictive Control solution methodology. Section IV provides a comprehensive performance evaluation through extensive simulations with various system parameters and discusses the standardization implications for 6G networks. Finally, Section V concludes the paper.

II. Related Works

Scheduling in 5G networks has been extensively studied, with extensive surveys [8], [9] categorizing approaches based on delay sensitivity, QoS guarantees, and adaptability. Classical schemes such as Round-Robin (RR) allocate resources uniformly but disregard channel dynamics, making them unsuitable for delay-critical services. Channel-aware schedulers like Maximum Rate (MR) and Proportional Fair (PF) balance throughput and fairness, with advanced implementations like GPU-accelerated PF scheduling [16] achieving microsecond-level decisions. However, these foundational approaches are not well suited for multi-sensory XR applications, that have strict delay-reliability requirements.

Machine learning has emerged as a powerful tool for enhancing scheduling adaptability. Reinforcement learning approaches such as DEMUX [17] enable dynamic allocation for eMBB–URLLC coexistence, while supervised learning techniques employ feedforward neural networks [18] for efficient decision-making. Integration of Time-Sensitive Networking (TSN) principles has spurred innovations like hierarchical particle swarm optimization with Double Q-learning [10] and meta-schedulers using cooperative learning across base stations [11]. Zhang et al. [12] propose constrained, risk-sensitive DRL frameworks for managing URLLC-induced puncturing overhead. While these ML approaches effectively address dual-service coexistence, they focus on traditional eMBB-URLLC scenarios rather than the simultaneous high-throughput and ultra-low-latency demands across five distinct sensory modalities.

Research targeting XR scheduling remains limited and primarily focuses on video-centric applications. Energy-efficient strategies exploit Discontinuous Reception (DRX) mechanisms [13], while theoretical efforts model XR scheduling as periodic Markov Decision Processes [19], [20] approximated via nonlinear knapsack formulations. Architectural approaches include heuristic bin-covering schedulers that prioritize XR PDU-sets based on delay budgets [14], Weighted Round-Robin scheduling for inter-cell interference mitigation [21], and frame-level integration schemes [15]. AI-driven service provisioning techniques enhance QoE [22] but lack comprehensive delay-reliability scheduling strategies. These existing XR schedulers assume simplified arrival processes, periodic traffic patterns, or fixed packet sizes—conditions rarely met in practice—and primarily address single-modality (video) optimization. Our work eliminates these constraints by handling realistic traffic conditions while jointly optimizing delay-reliability across all five human sensory modalities.

Complementary research in traffic modeling and metaverse architectures has explored empirical datasets for video XR offloading [23], temporal prediction of VR traffic bursts [24], and semi-Markov models for metaverse resource allocation [25]. XR loopback control mechanisms [26] enable fine-grained traffic shaping through adaptive packet sizes and frame rates. Recent metaverse-focused work models streaming as massive ultra-reliable low-latency communications (mURLLC) with Neyman-Pearson criterion-driven NFV/SDN architectures [27], while blockchain-based collaborative VNF man-

agement schemes optimize end-to-end network slices [28]. Although these frameworks provide valuable traffic characterization and architectural insights, they focus on adaptability and resource provisioning without addressing the underlying scheduling policies required to enforce per-modality delay and reliability constraints across heterogeneous sensory streams.

In summary, while prior work spans diverse scheduling methodologies, ML-driven optimizations, and architectural innovations in 5G networks, the literature lacks a unified framework capable of managing the coupled delay-reliability requirements of multi-sensory XR applications. Existing approaches either address traditional service categories (eMBB, URLLC) independently, focus on single-modality XR optimization, or provide architectural solutions without scheduling mechanisms. This paper addresses this critical gap by introducing DREAM-X, a comprehensive delay-reliability-aware scheduling solution specifically tailored for the unique characteristics of multi-sensory XR applications in NextG networks.

III. DREAM-X: Delay-Reliability Aware Scheduling for Multi-sensory XR Applications

Current 5G networks handle multi-modal data streams as independent QoS flows, where MAC schedulers apply scheduling algorithms (e.g., Proportional Fair) based on individual channel quality and QoS requirements (5QI values) without awareness of the relationships between streams. Synchronization relies entirely on higher-layer protocols using application-level buffering and timestamp coordination [29], [30]. This stream-agnostic approach proves fundamentally inadequate for mXR applications, where tight temporal coupling between sensory modalities requires stringent end-to-end synchronization. Independent scheduling leads to differential delays that break immersive experiences, and inefficient resource allocation.

We propose *DREAM-X*, a novel downlink scheduling framework that maximizes mXR users satisfying *delay-reliability* requirements across all sensory modalities simultaneously. Each modality operates under specific delay bounds (D^{σ}) within which corresponding *Data Units* (DUs)—a generic term used for packets, frames, or Protocol Data Units (PDUs)—must be successfully delivered. Since immersive experience depends on joint delivery across all modalities, any single-modal violation significantly impairs overall experience. Each user-modality pair requires a minimum percentage (e.g., 99%) of DUs meeting delay constraints for acceptable QoE, creating the core challenge of jointly satisfying these heterogeneous delay-reliability constraints while efficiently utilizing network bandwidth (Resource Blocks).

DREAM-X adopts 3GPP TS-22.261's *simultaneous strict delivery* strategy where all modalities individually satisfy delay bounds (10ms for video/audio, 5ms for haptics), extending it to include emerging olfactory and gustatory streams for complete five-sense immersion [5], [6]. The framework implements two core components: (1) *Multi-sensory Delay Tracking Mechanism* that continuously monitors DU delay status in real time, providing precise time-to-deadline tracking across all user-modality pairs, and (2) *Delay-Reliability Aware Scheduler* that leverages these insights to prioritize DU transmission for maximal delay-reliability satisfaction across the user base.

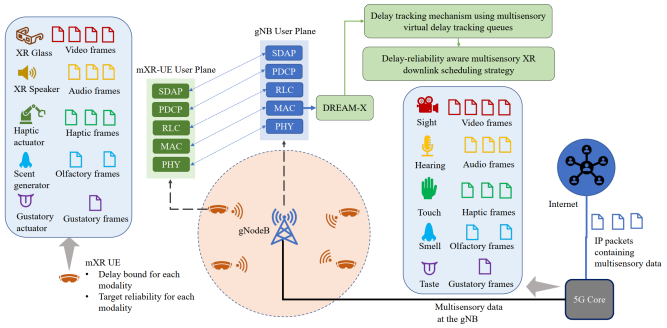


Figure 1. Protocol Stack and System Model with Multi-sensory XR Data.

A. System Model

We adopt the standard 5G protocol stack, as shown in Figure 1. While our proposed scheduling framework addresses the multi-sensory XR requirements expected in 6G and future developments, we incorporate the 5G stack and architecture in the system model because of its clearly defined and standardized protocol layers and network architecture. Multi-sensory Data Units (mDUs), generated at the application layer, traverse through the 5G stack to reach the Medium Access Control Queues (MACQs) at the gNB, with each XR user-modality pair assigned a dedicated queue. These queues buffer mDUs and track their associated Packet Data Convergence Protocol (PDCP) Discard Timers (PDTs), which define their individual delay bounds [31]. DREAM-X operates at the MAC layer and incorporates the delay tracking mechanism and the downlink scheduler. mXR users are assumed to be randomly located within the gNB coverage area. This study focuses solely on the downlink; end-to-end scheduling, including uplink will be considered in future work. We further assume that each XR user runs a single mXR application and is admitted to an exclusive mXR-RAN slice, dedicated to only immersive applications. Mixed-application scenarios are outside the scope of this work and are reserved for future investigation.

B. Multi-modal Delay Tracking Mechanism

We introduce a novel multi-modal delay tracking mechanism that orchestrates heterogeneous delay requirements across five distinct sensory modalities while preserving cross-modal synchronization integrity. The mechanism employs two foundational concepts: *Multi-modal Delay Tracking Queues* (mDTQs) and *Delay Tracking Granularity* (DTG), enabling fine-grained temporal management.

1) Architecture and Design Principles: The mechanism utilizes a multi-modal virtual queue structure, denoted as $\mathbf{Q}[i, m, k]$ that organizes Data Units (DUs) according to user index i, modality type $m \in \{\text{visual, audio, haptic, olfactory, gustatory}\}$, and delay state k. Initially, DUs arrive at modality-specific MAC queues before systematic reorganization into the structured mDTQ framework, as illustrated in Figure 2. This architecture enables independent delay tracking per modality while facilitating coordinated scheduling decisions across multiple sensory streams. We synchronize the DTG with the *Transmission Time Interval* (TTI) duration based on numerology configuration. For instance, under numerology 1 with a

TTI duration of 0.5 ms, the DTG is set to $t_d = 0.5$ ms, thereby ensuring consistent granularity in delay management.

2) Modality-Aware Enqueueing and DU Positioning: Upon DU arrival for user i and modality m, the mechanism determines the appropriate mDTQ position using $k = \min\left(\left\lfloor \frac{D_m}{t_d} \right\rfloor, L_{\max}\right)$. This assignment ensures that a DU placed in position k must be scheduled within k scheduling opportunities to avoid delay violations, with position k directly reflecting the remaining delay budget in TTI units and enabling urgency-based prioritization across modalities. To maintain space efficiency, we limit mDTQ size to $L_{\max}$, effectively over-provisioning modalities with large delay bounds. For instance, olfactory and gustatory streams with 1-second delay bounds would theoretically require 2000 mDTQ positions at numerology 1 (0.5ms TTI), but are instead constrained to $L_{\max}$ positions, forcing earlier scheduling than necessary. This trade-off significantly reduces memory requirements with minimal scheduler overhead, since these modalities generate small, infrequent DUs containing only triggering information rather than substantial content data like video or audio streams.

3) Temporal Evolution and Violation Detection: During each scheduling opportunity, selected DUs are transmitted while unscheduled DUs undergo systematic temporal shifting to reflect their reduced delay budgets. Specifically, unscheduled DUs in position k migrate to position $k-1$ according to the shifting rule: $\mathbf{Q}[i, m, k-1] \leftarrow \mathbf{Q}[i, m, k]$. This process continues iteratively, ensuring continuous reassignment based on updated delay constraints. DUs that reach the position $\mathbf{Q}[i, m, 0]$ no longer have scheduling opportunities available, resulting in delay violations for the specific user-modality combination, as shown in Figure 3. The mechanism maintains violation tracking through the metric $\nu_{i,m} = \frac{V_{i,m}^{\text{total}}}{A_{i,m}^{\text{total}}}$, where $V_{i,m}^{\text{total}}$ and $A_{i,m}^{\text{total}}$ represent cumulative violations and arrivals for user i and modality m, respectively.

4) Cross-Modal Violation Management: Since immersive experience requires synchronized delivery across all modalities, violation in any single modality constitutes a user-level violation. When violations in any modality exceed the target threshold (e.g., 0.1% for 99.9% reliability), the user is deemed unsatisfied. This cross-modal violation management ensures holistic optimization of multi-sensory applications rather than independent modality scheduling.

C. Optimization Problem Formulation

We now describe how employing the multi-modal delay tracking structure, we can formulate the scheduling optimization problem.

1) Objective Function: The optimization objective seeks to minimize the number of users failing to meet delay-reliability requirements across any sensory modality. A user is considered satisfied only when *all* associated modalities meet their respective delay-reliability targets, reflecting the interdependent nature of multi-modal XR experiences:

$$\mathcal{O}: \text{minimize} \sum_{i=1}^{N} \prod_{\sigma \in \{v,a,h,o,g\}} \mathbb{I}\{\gamma_i^\sigma(T) \geq \mu_i^\sigma\} \qquad (1)$$

where $\gamma_i^\sigma(T)$ represents the achieved delay reliability violation for user i and modality $\sigma \in \{v, a, h, o, g\}$ over duration T, μ_i^σ denotes the target violation threshold, and $\mathbb{I}\{\cdot\}$ is the indicator function. The product operator ensures that violation of any single modality's delay-reliability requirement results in user dissatisfaction.

2) Video Modality Constraints: To illustrate the mathematical formulation, we detail the constraints governing the video modality queue dynamics. The video stream is subject to four fundamental constraints that capture DU arrival, service, temporal evolution, and violation detection:

Enqueuing Constraint: New video data units (DUs) enter the delay tracking queue (DTQ) position corresponding to their delay bound: $\mathcal{C}1: Q_i^{K_i^v}(t+1) = a_i^v(t) f_i^v(t), \forall i, t$, where $Q_i^{K_i^v}(t)$ represents the queue content at the maximum delay position $K_i^v = \lfloor D^v / t_d \rfloor$ for user i, $a_i^v(t) \in \{0, 1\}$ indicates video arrival occurrence, and $f_i^v(t)$ denotes the DU size.

Queue Evolution Constraint: DTQ contents undergo temporal progression with service depletion: $\mathcal{C}2: Q_i^{j-1,v}(t+1) = Q_i^{j,v}(t) - \alpha_i(t) x_i^{j,v}(t), \forall i, t, j \in \mathcal{J}_i^v$, where $\mathcal{J}_i^v = \{1, 2, \ldots, K_i^v\}$, $\alpha_i(t)$ represents the achievable data rate per resource block for user i based on channel conditions, and $x_i^{j,v}(t)$ denotes the resource block allocation decision variable.

Violation Measurement Constraint: Delay violations are quantified as the ratio of delay-violated video DUs to total video arrivals: $\mathcal{C}3: \gamma_i^v(T) = \frac{\sum_{t=1}^{T} \mathbb{I}(Q_i^{0,v}(t) > 0)}{\sum_{t=1}^{T} a_i^v(t)}, \forall i$, where $Q_i^{0,v}(t)$ represents delay-bound-violated video DUs.

Service Feasibility Constraint: Amount of data served cannot exceed available queue contents: $\mathcal{C}4: \alpha_i(t) x_i^{j,v}(t) \leq Q_i^{j,v}(t), \forall i, t, j \in \mathcal{J}_i^v$.

3) Multi-modal Extension: The same constraint structure applies to all remaining sensory modalities: audio (a), haptic (h), olfactory (o), and gustatory (g). Each modality maintains separate DTQ positions $K_i^\sigma = \lfloor D^\sigma / t_d \rfloor$ reflecting their distinct delay bounds. We define $\mathcal{J}_i^\sigma = \{1, 2, \ldots, K_i^\sigma\}$ for each modality σ. Therefore the constraint sets considering rests of the modalities become as follows:

4) Complete Constraint Set: Audio Modality Constraints:

$$\begin{aligned}
\mathcal{C}5: &\quad Q_i^{K_i^a}(t+1) = a_i^a(t) f_i^a(t) && \forall i, t \\
\mathcal{C}6: &\quad Q_i^{j-1,a}(t+1) = Q_i^{j,a}(t) - \alpha_i(t) x_i^{j,a}(t) && \forall i, t, j \in \mathcal{J}_i^a \\
\mathcal{C}7: &\quad \gamma_i^a(T) = \frac{\sum_{t=1}^{T} \mathbb{I}(Q_i^{0,a}(t) > 0)}{\sum_{t=1}^{T} a_i^a(t)} && \forall i \\
\mathcal{C}8: &\quad \alpha_i(t) x_i^{j,a}(t) \leq Q_i^{j,a}(t) && \forall i, t, j \in \mathcal{J}_i^a
\end{aligned}$$

Haptic Modality Constraints:

$$\begin{aligned}
\mathcal{C}9: &\quad Q_i^{K_i^h}(t+1) = a_i^h(t) f_i^h(t) && \forall i, t \\
\mathcal{C}10: &\quad Q_i^{j-1,h}(t+1) = Q_i^{j,h}(t) - \alpha_i(t) x_i^{j,h}(t) && \forall i, t, j \in \mathcal{J}_i^h \\
\mathcal{C}11: &\quad \gamma_i^h(T) = \frac{\sum_{t=1}^{T} \mathbb{I}(Q_i^{0,h}(t) > 0)}{\sum_{t=1}^{T} a_i^h(t)} && \forall i \\
\mathcal{C}12: &\quad \alpha_i(t) x_i^{j,h}(t) \leq Q_i^{j,h}(t) && \forall i, t, j \in \mathcal{J}_i^h
\end{aligned}$$

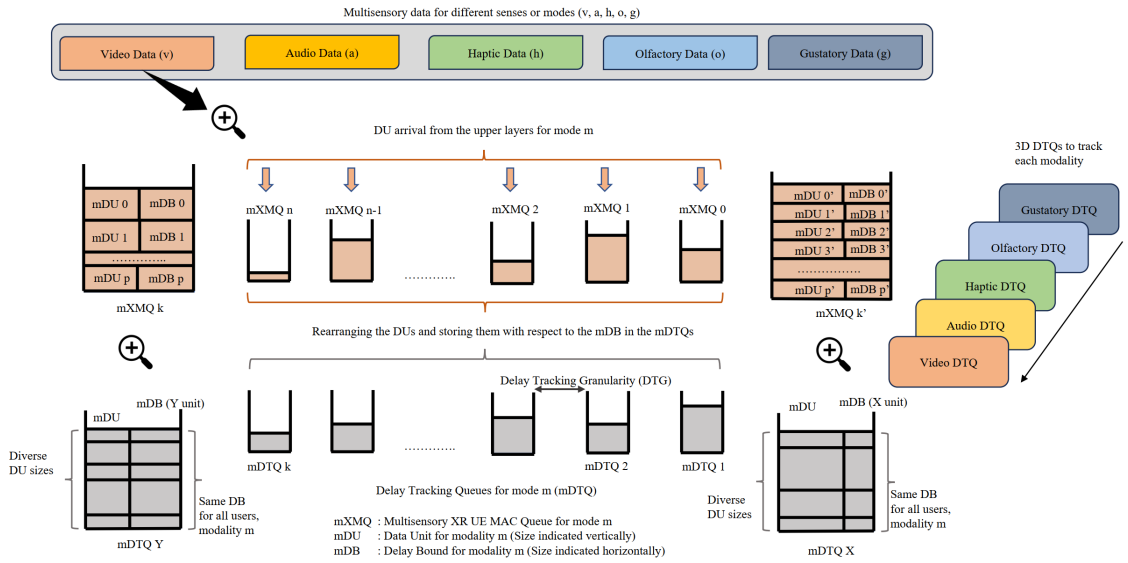


Figure 2. Multi-modal delay tracking mechanism at the gNB MAC using virtual DTQs. mDUs from modality m are enqueued in XR user-specific MAC queues (mXMQs) and mapped to appropriate DTQ positions based on their modal delay bounds (mDBs). Zoomed views highlight DTQ contents for two users and one modality (e.g., video); the same process applies across all modalities.

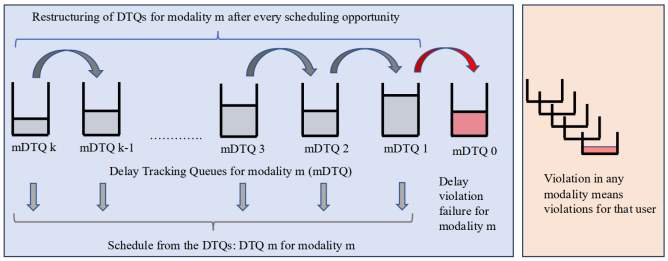


Figure 3. Multi-modal delay tracking via DTQ shifting. After each scheduling opportunity, DUs are shifted; those reaching the final queue incur delay violations. Any modality-level violation constitutes a user-level violation.

Olfactory Modality Constraints:

$$\mathcal{C}13: \quad Q_i^{K_i^o}(t+1) = a_i^o(t) f_i^o(t) \qquad \forall i,t$$

$$\mathcal{C}14: \quad Q_i^{j-1,o}(t+1) = Q_i^{j,o}(t) - \alpha_i(t) x_i^{j,o}(t) \qquad \forall i,t,j \in \mathcal{J}_i^o$$

$$\mathcal{C}15: \quad \gamma_i^o(T) = \frac{\sum_{t=1}^{T} \mathbb{I}(Q_i^{0,o}(t) > 0)}{\sum_{t=1}^{T} a_i^o(t)} \qquad \forall i$$

$$\mathcal{C}16: \quad \alpha_i(t) x_i^{j,o}(t) \le Q_i^{j,o}(t) \qquad \forall i,t,j \in \mathcal{J}_i^o$$

Gustatory Modality Constraints:

$$\mathcal{C}17: \quad Q_i^{K_i^g}(t+1) = a_i^g(t) f_i^g(t) \qquad \forall i,t$$

$$\mathcal{C}18: \quad Q_i^{j-1,g}(t+1) = Q_i^{j,g}(t) - \alpha_i(t) x_i^{j,g}(t) \qquad \forall i,t,j \in \mathcal{J}_i^g$$

$$\mathcal{C}19: \quad \gamma_i^g(T) = \frac{\sum_{t=1}^{T} \mathbb{I}(Q_i^{0,g}(t) > 0)}{\sum_{t=1}^{T} a_i^g(t)} \qquad \forall i$$

$$\mathcal{C}20: \quad \alpha_i(t) x_i^{j,g}(t) \le Q_i^{j,g}(t) \qquad \forall i,t,j \in \mathcal{J}_i^g$$

5) System-Wide Resource Constraints: Capacity Constraint:

$$\mathcal{C}21: \quad \sum_{i=1}^{N} \sum_{\sigma \in \{v,a,h,o,g\}} \sum_{j=1}^{K_i^\sigma} x_i^{j,\sigma}(t) \le C, \quad \forall t$$

Integer Decision Variable Constraints:

$$\mathcal{C}22: \quad x_i^{j,v}(t) \in \{0,1,\ldots,C\} \qquad \forall i,t,j \in \mathcal{J}_i^v$$

$$\mathcal{C}23: \quad x_i^{j,a}(t) \in \{0,1,\ldots,C\} \qquad \forall i,t,j \in \mathcal{J}_i^a$$

$$\mathcal{C}24: \quad x_i^{j,h}(t) \in \{0,1,\ldots,C\} \qquad \forall i,t,j \in \mathcal{J}_i^h$$

$$\mathcal{C}25: \quad x_i^{j,o}(t) \in \{0,1,\ldots,C\} \qquad \forall i,t,j \in \mathcal{J}_i^o$$

$$\mathcal{C}26: \quad x_i^{j,g}(t) \in \{0,1,\ldots,C\} \qquad \forall i,t,j \in \mathcal{J}_i^g$$

Channel Parameter Constraint:

$$\mathcal{C}27: \quad \alpha_i(t) \ge 0, \quad \forall i,t$$

6) Complete Problem Formulation: The complete multi-modal XR scheduling optimization problem is formulated as:

$$\mathcal{O}: \quad \text{minimize} \sum_{i=1}^{N} \prod_{\sigma \in \{v,a,h,o,g\}} \mathbb{I}\{\gamma_i^\sigma(T) \ge \mu_i^\sigma\}$$

$$\text{subject to:} \quad \mathcal{C}1 - \mathcal{C}27$$

This formulation maximizes the number of fully satisfied users, in line with the 3GPP specifications and the stringent demands of immersive XR, where delay-reliability targets are minimum thresholds for usability. When resources are limited, the scheduler prioritizes fully serving a subset of users (achieving ≥99% delay-reliability across all modalities) rather than partially serving all at degraded quality, which would make the XR experience unacceptable for everyone.

D. A Discussion on Inter-modal Synchronization

Inter-modal synchronization is critical for multi-sensory XR, as temporal misalignment across modalities significantly degrades QoE. From a scheduling perspective, this can be addressed in two main ways: (a) the *anchor-based approach*, where one modality (e.g., video) serves as the anchor and others (e.g., audio, haptics) are scheduled relative to it using explicit inter-modal delay constraints (e.g., audio packets must arrive within Δ_{AV} of the corresponding video frames); and

(b) the *joint scheduling approach*, as adopted in DREAM-X, where all modalities are jointly scheduled within their respective delay budgets defined in 3GPP TS 22.261 [6]. Our objective function models an AND relationship across modalities, such that a delay violation in any modality results in user-level failure. This stricter formulation creates tight coupling via joint optimization. A user is considered successfully served only when all modalities meet their individual delay constraints. Since the per-modal budgets encapsulate end-to-end requirements, jointly satisfying them inherently preserves inter-modal alignment, allowing the receiver's upper layers to synchronize modalities without QoE degradation. Further discussion is provided in Section IV.

E. Problem Complexity and Practical Challenges

The optimization problem presents intrinsic complexities that preclude standard solution methods, requiring tailored approaches for practical implementation. The problem constitutes a Nonlinear Integer Program driven by products of indicator functions $\prod_\sigma \mathbb{I}\{\gamma_i^\sigma(T) \geq \mu_i^\sigma\}$ in the objective and violation measurement constraints $\gamma_i^\sigma(T) = \frac{\sum_{t=1}^{T} \mathbb{I}(Q_i^{0,\sigma}(t)>0)}{\sum_{t=1}^{T} a_i^\sigma(t)}$, introducing nonlinearity and NP-hardness with exponential solution spaces that render exact solutions infeasible for realistic user populations. Additionally, optimal allocation requires complete foresight of future traffic arrivals and channel states in dynamic wireless settings with inherently stochastic conditions. Finally, practical solutions demand sub-millisecond scheduling, impossible to achieve for the global optimal, necessitating efficient strategies that balance performance with runtime feasibility.

F. Model Predictive Control Framework and Problem Reformulation

To address these fundamental challenges, we employ Model Predictive Control (MPC) framework that transforms the intractable infinite-horizon optimization into a sequence of manageable *finite-horizon subproblems* solved in *rolling horizon* manner. At each time slot t, the system formulates and solves a finite-horizon optimization over prediction horizon H, incorporating current DTQ state $\mathbf{Q}(t)$ and predicted future conditions $\{\hat{a}_i^\sigma(t+\tau), \hat{\alpha}_i(t+\tau)\}_{\tau=1}^{H}$. The finite-horizon optimization yields resource allocation decisions $\mathbf{X}_{t:t+H}$ for the entire horizon, but the system implements only the immediate slot allocation $\mathbf{X}_t$. Following implementation, the system updates all state variables—DTQ contents $\mathbf{Q}(t+1)$, violation statistics, and user satisfaction indicators—based on latest system dynamics. This closed-loop feedback mechanism ensures the MPC framework adapts to real network conditions enabling robust performance under uncertainty.

The critical innovation enabling tractable finite-horizon solutions lies in reformulating the complex nonlinear subproblem by exploiting the structure of the DTQ and transforming nonlinear delay constraints into computationally manageable linear inequalities. This reformulation consists of three key innovations as follows:

1) DTQ Structure and Demand Preprocessing: The multi-modal DTQ structure provides the foundation for problem reformulation through its inherent encoding of temporal scheduling flexibility. Each DTQ position j represents the number of remaining scheduling opportunities before deadline expiration: position $j = 0$ contains deadline-expired data (violations), while position $j \geq 1$ contains data units with exactly j remaining time slots for successful transmission. This structure enables direct transformation from queue state to resource demands. For each DTQ position $j \geq 1$ containing data bits $Q_i^{j,\sigma}(t)$, we compute the required resource blocks as:

$$d_i^{j,\sigma}(t) = \begin{cases} \left\lceil \frac{Q_i^{j,\sigma}(t)}{\alpha_i(t)} \right\rceil & \text{if } Q_i^{j,\sigma}(t) > 0 \\ 0 & \text{otherwise} \end{cases}$$

where $\alpha_i(t)$ represents the achievable data rate per resource block (RB) for user i at time t, and the ceiling operation $\lceil \cdot \rceil$ ensures integer resource block allocation. This preprocessing transforms the bit-level queue state into discrete resource block requirements while preserving all temporal deadline information.

2) Cumulative Deadline Constraint Innovation: The fundamental insight enabling linear reformulation lies in recognizing that DTQ position j encodes cumulative scheduling flexibility: data at position j can be served across any combination of the next j time slots before deadline expiration. This flexibility translates into cumulative deadline constraints that replace the complex indicator functions of the original problem with linear inequalities. For demand $d_i^{j,\sigma}$ representing data at DTQ position j, the constraint becomes:

$$\sum_{\tau=0}^{j-1} x_i^{\tau,\sigma} \geq d_i^{j,\sigma} \cdot y_i \quad \forall i, \sigma, j : d_i^{j,\sigma} > 0$$

This formulation captures the essential insight: if user i is to be satisfied ($y_i = 1$), then the cumulative resource allocation over time slots 0 through $j-1$ must meet or exceed the demand at position j. The constraint automatically handles the temporal flexibility while maintaining deadline guarantees.

3) Historical Reliability Integration: To ensure continuity in user satisfaction and maximize the overall number of satisfied mXR users, we introduce historical reliability tracking through indicator $z_i(t)$ that maintains awareness of past performance without requiring infinite memory:

$$z_i(t) = \begin{cases} 1 & \text{if } \max_\sigma \nu_i^\sigma(t) \leq \mu_i^\sigma \\ 0 & \text{otherwise} \end{cases}$$

where $\nu_i^\sigma(t) = \frac{V_i^\sigma(t)}{A_i^\sigma(t)}$ represents the achieved violation rate for user i and modality σ, computed over a sliding window of length W. This mechanism addresses a critical challenge in MPC-based scheduling: if each finite-horizon subproblem is solved independently without historical context, the scheduler may satisfy different sets of users across time slots, potentially reducing the overall number of users who maintain satisfaction throughout the evaluation period. The historical reliability indicator $z_i(t)$ provides decision continuity by prioritizing users who are already meeting their delay-reliability targets in case

of resource contention, ensuring that previously satisfied users continue to receive adequate service. This approach maximizes the cumulative number of satisfied mXR users by maintaining consistency rather than repeatedly switching between different user sets. The mechanism provides: (a) decision continuity ensuring consistent user satisfaction across MPC iterations, (b) bounded memory requirements preventing infinite state accumulation, and (c) adaptive performance tracking responding to changing conditions.

Complete Mathematical Reformulation: Integrating the DTQ-based demand computation, cumulative deadline constraints, and historical reliability tracking, each reformulated finite-horizon subproblem at time t becomes:

$$\mathcal{O}': \quad \text{maximize} \sum_{i=1}^{N} y_i \cdot (N + z_i) \tag{2}$$

$$\text{subject to:} \quad \sum_{\tau=0}^{j-1} x_i^{\tau,\sigma} \geq d_i^{j,\sigma} \cdot y_i \quad \forall i, \sigma, j : d_i^{j,\sigma} > 0 \tag{3}$$

$$\sum_{i=1}^{N} \sum_{\sigma} x_i^{\tau,\sigma} \leq C \quad \forall \tau \in \{0, 1, \ldots, H-1\} \tag{4}$$

$$y_i \leq z_i \quad \forall i \tag{5}$$

$$x_i^{\tau,\sigma} \in \mathbb{Z}_+, \quad y_i \in \{0, 1\} \quad \forall i, \tau, \sigma \tag{6}$$

The objective (2) achieves dual goals: primarily maximizing satisfied users (coefficient N dominates), and secondarily providing continuity preference for users already meeting their reliability targets (coefficient z_i) to maintain overall system performance. Constraint (3) implements the cumulative deadline flexibility, (4) enforces system capacity limits, (5) ensures only historically reliable users can be targeted for satisfaction in the current subproblem, and (6) ensures proper variable domains.

This transformation makes the subproblems Mixed Integer Linear Program (MILP) that can be solved with standard MILP solvers with convergence guarantees.

G. Finite-Horizon Subproblem Solution Methods

The reformulated finite-horizon optimization, while structurally linear, remains computationally challenging due to integer constraints and potentially large solution spaces. DREAM-X addresses this through a comprehensive solution framework providing both optimal and heuristic approaches for solving each finite-horizon subproblem.

1) Optimal MILP Solution for Finite-Horizon Subproblems: For scenarios requiring theoretical optimality, we employ Mixed Integer Linear Programming using commercial solvers to solve each finite-horizon subproblem optimally. We configure the solver with appropriate time limits and optimality gaps for real-time operation, providing global optimality guarantees for each finite-horizon decision within specified bounds.

2) Delay-Reliability-Aware Earliest Deadline First (DRA-EDF) Heuristic: For practical deployment requiring computational efficiency, we propose DRA-EDF, a greedy heuristic that captures the essential optimization principles of each finite-horizon subproblem while maintaining polynomial complexity. The algorithm operates through a three-phase strategy as detailed in Algorithm 1.

Algorithm 1 Delay-Reliability-Aware Earliest Deadline First Heuristic

function DRA-EDF($\mathbf{D}, C, H, \mathbf{z}$, $\mathbf{D}$: RB Demand matrix from prepossessing)
  Phase 1: Delay-Reliability-Based User Partitioning
  $\mathcal{U}_H \leftarrow \{i : z_i = 1\}, \mathcal{U}_L \leftarrow \{i : z_i = 0\}$ ▷ Partition by reliability
  Phase 2: Temporal Priority Ordering
  for $\mathcal{U} \in \{\mathcal{U}_H, \mathcal{U}_L\}$ **do** ▷ Sort within each tier
    Sort users in $\mathcal{U}$ by earliest deadline: $\min_{\sigma,j:d_i^{j,\sigma}>0} j$
  end for
  $\mathcal{U}_{\text{sorted}} \leftarrow \mathcal{U}_H \cup \mathcal{U}_L$ ▷ Combine with priority order
  Phase 3: All-or-Nothing Resource Allocation
  $\mathbf{X} \leftarrow \mathbf{0}^{N \times |\mathcal{M}| \times H}$, $\mathbf{C}_{\text{rem}} \leftarrow C \cdot \mathbf{1}^H$ ▷ Initialize
  for $i \in \mathcal{U}_{\text{sorted}}$ **do** ▷ Process users in priority order
    total_demand $\leftarrow \sum_{\sigma,j} d_i^{j,\sigma}$ ▷ Total user demand
    if total_demand $\leq \sum_{\tau=0}^{H-1} \mathbf{C}_{\text{rem}}[\tau]$ **then** ▷ Admission control
      AllocateUserDemands($i, \mathbf{D}, \mathbf{X}, \mathbf{C}_{\text{rem}}$) ▷ Algorithm 2
    end if
  end for
  return $\mathbf{X}$
end function

The all-or-nothing allocation strategy ensures complete user satisfaction rather than partial service, which is critical for multi-modal XR applications where incomplete modality delivery severely degrades user experience. The detailed demand allocation process is described in Algorithm 2.

Algorithm 2 RBs Allocation with Deadline Awareness

function AllocateUserDemands($i, \mathbf{D}, \mathbf{X}, \mathbf{C}_{\text{rem}}$)
  Collect all demands: $\mathcal{D}_i \leftarrow \{(m, j, d_i^{j,m}) : d_i^{j,m} > 0\}$
  Sort $\mathcal{D}_i$ by deadline: $(m, j, d) \prec (m', j', d')$ if $j < j'$ ▷ Earliest deadline first
  for $(m, j, d) \in \mathcal{D}_i$ **do** ▷ Process demands by deadline
    remaining_demand $\leftarrow d$
    for $\tau = 0, \ldots, j-1$ **do** ▷ Allocate within deadline window
      if remaining_demand > 0 **and** $\mathbf{C}_{\text{rem}}[\tau] > 0$ **then**
        allocation $\leftarrow \min(\text{remaining\_demand}, \mathbf{C}_{\text{rem}}[\tau])$
        $\mathbf{X}[i, m, \tau] \leftarrow \mathbf{X}[i, m, \tau] + \text{allocation}$
        $\mathbf{C}_{\text{rem}}[\tau] \leftarrow \mathbf{C}_{\text{rem}}[\tau] - \text{allocation}$
        remaining_demand $\leftarrow$ remaining_demand $-$ allocation
      end if
    end for
    if remaining_demand > 0 **then** ▷ Cannot satisfy this user
      Rollback all allocations for user i and **return**
    end if
  end for
end function

H. Integrated DREAM-X Framework

The complete DREAM-X solution integrates the following modules:

1) DTQ Management and State Evolution: The DTQ management system handles mDU arrivals, temporal evolution, and violation detection through integrated algorithms. Algo-

rithm 3 details the core DTQ operations including enqueueing new arrivals and managing temporal evolution.

Algorithm 3 DTQ Enqueuing and Temporal Management

function ENQUEUEARRIVALS($\mathbf{Q}, \mathbf{A}_t, t$)
  for $i \in \mathcal{N}, \sigma \in \{v, a, h, o, g\}$ **do**
    if $A_t[i, \sigma] > 0$ **then** ▷ New arrival for user i, modality σ
      delay_bound $\leftarrow D^\sigma$ ▷ Modality-specific delay bound
      enqueue_pos $\leftarrow \min(\lfloor \text{delay\_bound}/t_d \rfloor, L_{\max} - 1)$ ▷ Calculate position
      $\mathbf{Q}[i, \sigma, \text{enqueue\_pos}] \leftarrow \mathbf{Q}[i, \sigma, \text{enqueue\_pos}] + A_t[i, \sigma]$
    end if
  end for
end function
function TEMPORALEVOLUTION($\mathbf{Q}, \mathbf{V}$)
  for $i \in \mathcal{N}, \sigma \in \{v, a, h, o, g\}$ **do**
    if $\mathbf{Q}[i, \sigma, 1] > 0$ **then** ▷ Data moving to violation bin
      $\mathbf{V}[i, \sigma] \leftarrow \mathbf{V}[i, \sigma] + 1$ ▷ Record violation
    end if
    for $j = 0, \ldots, L_{\max} - 2$ **do** ▷ Shift DTQ positions
      $\mathbf{Q}[i, \sigma, j] \leftarrow \mathbf{Q}[i, \sigma, j + 1]$ ▷ Temporal advancement
    end for
    $\mathbf{Q}[i, \sigma, L_{\max} - 1] \leftarrow 0$ ▷ Clear arrival position
  end for
  return $\mathbf{Q}, \mathbf{V}$
end function

2) Demand Computation Preprocessing with Predictions: The demand computation integrates current DTQ state with future arrivals and channels predictions till horizon for enhanced optimization as shown in Algorithm 4.

Algorithm 4 Integrated Demand Computation with Prediction

function COMPUTEDEMANDS($\mathbf{Q}, t, H$, use_prediction)
  if $H > 0$ **and** use_prediction **then** ▷ Predictive mode
    $\hat{\mathbf{A}}_{t:t+H} \leftarrow \text{PredictArrivals}(t, H)$ ▷ Generate predictions
    $\mathbf{Q}_{\text{ext}} \leftarrow \text{ExtendDTQ}(\mathbf{Q}, \hat{\mathbf{A}}_{t:t+H}, H)$ ▷ Extend DTQ
  else
    $\mathbf{Q}_{\text{ext}} \leftarrow \mathbf{Q}$ ▷ Use current DTQ only
  end if
  $\mathbf{D} \leftarrow \mathbf{0}^{N \times |\mathcal{M}| \times \text{size}(\mathbf{Q}_{\text{ext}})}$ ▷ Initialize demand matrix
  for $i \in \mathcal{N}, \sigma \in \{v, a, h, o, g\}, j = 1, \ldots, \text{size}(\mathbf{Q}_{\text{ext}})$ **do**
    if $\mathbf{Q}_{\text{ext}}[i, \sigma, j] > 0$ **then** ▷ Non-empty position
      rate $\leftarrow \hat{\alpha}_i(t + \max(0, j - L_{\max}))$ ▷ Predicted channel rate
      $\mathbf{D}[i, \sigma, j] \leftarrow \lceil \mathbf{Q}_{\text{ext}}[i, \sigma, j]/\text{rate} \rceil$ ▷ RB demand
    end if
  end for
  return $\mathbf{D}$
end function
function EXTENDDTQ($\mathbf{Q}, \hat{\mathbf{A}}_{t:t+H}, H$)
  $\mathbf{Q}_{\text{ext}} \leftarrow \mathbf{0}^{N \times |\mathcal{M}| \times (L_{\max} + H)}$ ▷ Extended DTQ
  $\mathbf{Q}_{\text{ext}}[:, :, 0 : L_{\max}] \leftarrow \mathbf{Q}$ ▷ Copy current state
  for $\tau = 1, \ldots, H$ **do** ▷ Add predicted arrivals
    for $i \in \mathcal{N}, \sigma \in \{v, a, h, o, g\}$ **do**
      if $\hat{\mathbf{A}}[i, \sigma, t + \tau] > 0$ **then**
        pos $\leftarrow \min(L_{\max} + \tau - 1, L_{\max} + H - 1)$ ▷ Future position
        $\mathbf{Q}_{\text{ext}}[i, \sigma, \text{pos}] \leftarrow \mathbf{Q}_{\text{ext}}[i, \sigma, \text{pos}] + \hat{\mathbf{A}}[i, \sigma, t + \tau]$
      end if
    end for
  end for
  return $\mathbf{Q}_{\text{ext}}$
end function

3) Solution Application and System Update: The solution application converts optimization results into system actions and updates state variables as described in Algorithm 5.

Algorithm 5 Solution Application and State Update

function APPLYALLOCATION($\mathbf{Q}, \mathbf{X}_t, \boldsymbol{\alpha}_t$)
  for $i \in \mathcal{N}, \sigma \in \{v, a, h, o, g\}$ **do**
    allocated_rbs $\leftarrow \mathbf{X}_t[i, \sigma]$ ▷ Current slot allocation
    served_bits $\leftarrow$ allocated_rbs $\cdot\, \alpha_i(t)$ ▷ Convert to bits
    remaining_service $\leftarrow$ served_bits
    for $j = 1, \ldots, L_{\max}$ **do** ▷ Serve from most urgent positions
      if $\mathbf{Q}[i, \sigma, j] > 0$ **and** remaining_service > 0 **then**
        service $\leftarrow \min(\mathbf{Q}[i, \sigma, j], \text{remaining\_service})$
        $\mathbf{Q}[i, \sigma, j] \leftarrow \mathbf{Q}[i, \sigma, j] - \text{service}$
        remaining_service $\leftarrow$ remaining_service $-$ service
      end if
    end for
  end for
  return $\mathbf{Q}$
end function
function UPDATEDELAYRELIABILITYINDICATORS($\mathbf{V}, \mathbf{A}, \boldsymbol{\mu}$)
  for $i \in \mathcal{N}$ **do**
    max_violation_rate $\leftarrow 0$
    for $\sigma \in \{v, a, h, o, g\}$ **do**
      if $\mathbf{A}[i, \sigma] > 0$ **then** ▷ User has traffic for this modality
        violation_rate $\leftarrow \mathbf{V}[i, \sigma]/\mathbf{A}[i, \sigma]$
        max_violation_rate $\leftarrow \max(\text{max\_violation\_rate}, \text{violation\_rate})$
      end if
    end for
    $z_i \leftarrow \mathbb{I}\{\text{max\_violation\_rate} \leq \mu_i\}$ ▷ Update delay-reliability status
  end for
  return $\mathbf{z}$
end function

4) Main DREAM-X Control Loop: The complete DREAM-X algorithm integrates all components into a unified MPC framework as outlined in Algorithm 6.

Algorithm 6 DREAM-X Main Control Loop

Require: $N, |\mathcal{M}|, L_{\max}, H, C, T, \boldsymbol{\mu}$ ▷ System parameters
Ensure: Resource allocation schedule $\mathbf{X} \in \mathbb{Z}^{N \times |\mathcal{M}| \times T}$
Initialize $\mathbf{Q} \leftarrow \mathbf{0}^{N \times |\mathcal{M}| \times L_{\max}}$, $\mathbf{V} \leftarrow \mathbf{0}^{N \times |\mathcal{M}|}$, $\mathbf{z} \leftarrow \mathbf{1}^N$
for $t = 0, \ldots, T - 1$ **do** ▷ Main scheduling loop
  $\mathbf{A}_t \leftarrow \text{Arrivals}(t)$ ▷ Traffic arrivals
  ENQUEUEARRIVALS($\mathbf{Q}, \mathbf{A}_t, t$) ▷ Process arrivals (Algorithm 3)
  $\mathbf{D} \leftarrow \text{ComputeDemands}(\mathbf{Q}, t, H, \text{prediction\_enabled})$ ▷ Algorithm 4
  if optimal_solver_enabled **then** ▷ Solution method selection
    $\mathbf{X}_{t:t+H} \leftarrow \text{SolveOptimalMILP}(\mathbf{D}, C, H, \mathbf{z})$
  else
    $\mathbf{X}_{t:t+H} \leftarrow \text{DRA-EDF}(\mathbf{D}, C, H, \mathbf{z})$ ▷ Algorithm 1
  end if
  $\mathbf{X}_t \leftarrow \mathbf{X}_{t:t+H}[:, :, 0]$ ▷ Extract current slot solution
  $\mathbf{Q} \leftarrow \text{ApplyAllocation}(\mathbf{Q}, \mathbf{X}_t, \boldsymbol{\alpha}_t)$ ▷ Algorithm 5
  $\mathbf{Q}, \mathbf{V} \leftarrow \text{TemporalEvolution}(\mathbf{Q}, \mathbf{V})$ ▷ Algorithm 3
  $\mathbf{z} \leftarrow \text{UpdateDelayReliabilityIndicators}(\mathbf{V}, \mathbf{A}, \boldsymbol{\mu})$ ▷ Algorithm 5
end for
return $\mathbf{X}$ ▷ Complete allocation schedule

The integrated framework shows how DREAM-X transforms the intractable multi-modal XR scheduling problem into a practical real-time implementable solution.

I. Prediction Framework

DREAM-X employs a predictive framework that forecasts future multi-modal DU arrivals and channel states, denoted

by $\hat{\mathbf{A}}_{t:t+H}$ and $\hat{\boldsymbol{\alpha}}_{t:t+H}$, over a prediction horizon H (Algorithm 7). The design prioritizes computational efficiency over accuracy due to the stringent constraints of real-time scheduling. While advanced machine learning models could offer higher accuracy, their computational cost may exceed TTI deadlines. Accordingly, lightweight statistical models are adopted to balance efficiency with acceptable prediction quality.

Each sensory modality exhibits distinct traffic characteristics, enabling modality-specific statistical prediction: video follows a truncated Gaussian distribution for DU sizes and inter-arrival jitter; audio follows a CBR pattern with minimal variation; haptic traffic is periodic, with uniform size variation and controlled jitter; and olfactory/gustatory traffic follows sporadic Poisson processes with exponential inter-arrival times [5], [6], [32]. The predictor uses this distributional knowledge, and the distribution parameters are adapted via exponential moving averages over sliding windows of length W. For channel prediction, a persistence model is used: $\hat{\alpha}_i(t+\tau) = \alpha_i(t)$ for $\tau \in [1, H]$. The MPC framework provides inherent robustness to prediction errors. At each time slot, the scheduler solves the optimization problem based on the current states, then recalibrates using updated DTQ states and refreshed predictions. This closed-loop feedback prevents error accumulation despite imperfect forecasts. The design yields two key advantages: first, the lightweight models introduce negligible delay, ensuring responsiveness within TTI deadlines and enabling practical real-time deployment; second, integration with rolling horizon optimization creates a predictor-scheduler co-design—where re-optimization at each slot mitigates prediction errors and their accumulation, and modestly accurate predictions enhance the MPC optimizer's decision quality.

Algorithm 7 Future Arrivals and Channels Prediction

function PREDICTTRAFFICARRIVALS($t, H, \mathbf{A}_{1:t}$)
  for $i \in \mathcal{N}, \sigma \in \{v, a, h, o, g\}$ **do**
    Extract recent observations: $\mathbf{A}_{i,\sigma}^{\text{recent}} \leftarrow \mathbf{A}_{i,\sigma}[t-W:t]$
    if $\sigma =$ video **then**
      DU sizes: $\hat{f}_{i,\sigma} \sim \mathcal{TN}(\hat{\mu}_{\text{size}}, \hat{\sigma}^2_{\text{size}}, [f_{\min}, f_{\max}])$
      Inter-arrivals: $\hat{\tau}_{i,\sigma} \sim \mathcal{TN}(\hat{\mu}_{\text{time}}, \hat{\sigma}^2_{\text{time}}, [\tau_{\min}, \tau_{\max}])$
    else if $\sigma =$ audio **then**
      DU sizes: $\hat{f}_{i,\sigma} = \mu_{\text{CBR}} + \mathcal{N}(0, \sigma^2_{\text{small}})$
      Inter-arrivals: $\hat{\tau}_{i,\sigma} = \tau_{\text{fixed}} + \mathcal{N}(0, \sigma^2_{\text{jitter}})$
    else if $\sigma =$ haptic **then**
      DU sizes: $\hat{f}_{i,\sigma} = \mu_{\text{haptic}} + \mathcal{U}(-\delta, \delta)$
      Inter-arrivals: $\hat{\tau}_{i,\sigma} = \tau_{\text{periodic}} + \mathcal{N}(0, \sigma^2_{\text{micro}})$
    else if $\sigma \in \{$olfactory, gustatory$\}$ **then**
      Inter-arrivals: $\hat{\tau}_{i,\sigma} \sim \text{Exp}(\hat{\lambda})$, DU sizes: $\hat{f}_{i,\sigma} \sim \text{Exp}(\hat{\mu})$
    end if
    Adapt all distribution parameters using sliding window W
    Generate predictions: $\hat{\mathbf{A}}_{i,\sigma}[t+1:t+H]$
  end for
  Channel Prediction: $\hat{\alpha}_i(\tau) \leftarrow \alpha_i(t), \forall \tau \in [t+1, t+H]$
  return $\hat{\mathbf{A}}_{t:t+H}, \hat{\boldsymbol{\alpha}}_{t:t+H}$
end function

J. Complexity Analysis

DREAM-X exhibits distinct complexity characteristics that vary across its algorithmic components. The DTQ management operations: enqueuing requires $O(N \cdot |\mathcal{M}|)$, predictive extension demands $O(H \cdot N \cdot |\mathcal{M}|)$, while demand computation grows as $O(N \cdot |\mathcal{M}| \cdot (L+H))$, where N represents users, $|\mathcal{M}|$ denotes modalities, H indicates prediction horizon, and L captures maximum delay tracking depth. The optimization phase, however, presents different complexity profiles depending on our chosen solution method. The optimal MILP approach faces exponential complexity of $O(2^N \cdot C^{N \cdot |\mathcal{M}| \cdot H} \cdot \text{poly}(N \cdot |\mathcal{M}| \cdot L \cdot H))$ due to N binary satisfaction variables and $N \cdot |\mathcal{M}| \cdot H$ integer resource allocation variables. Whereas, our DRA-EDF heuristic achieves polynomial complexity of $O(N \log N + N \cdot |\mathcal{M}| \cdot L \cdot H)$. Both solution application and temporal evolution operations require $O(N \cdot |\mathcal{M}| \cdot L)$ computations for DTQ updates and violation tracking. When we consider the overall DREAM-X complexity across T scheduling intervals, the choice of optimization approach becomes crucial. Using optimal MILP results in $T \cdot O(2^N \cdot C^{N \cdot |\mathcal{M}| \cdot H})$ complexity, which limits practical deployment to small scenarios. Our DRA-EDF heuristic, however, achieves $T \cdot O(N \log N + N \cdot |\mathcal{M}| \cdot L \cdot H)$ complexity, making real-time operation feasible.

It is important to note that DREAM-X operates within the existing 3GPP-standard architecture and requires no hardware modifications. The DTQ framework is fully implementable through software-based queue management, reorganizing DUs according to their remaining delay budgets using basic array operations and arithmetic. Since downlink scheduling is performed at the gNB, which possesses sufficient computational and memory resources, both the polynomial time complexity of the DRA-EDF heuristic and the space complexity of $O(N \cdot |\mathcal{M}| \cdot (L+H))$ are operationally feasible. Moreover, the prediction framework utilizes lightweight statistical models suitable for real-time execution. This design ensures minimal computational overhead, enabling the scalable deployment of DREAM-X without hardware bottlenecks.

IV. PERFORMANCE EVALUATION

A. Simulation Framework and Baseline Algorithms

We developed a custom system-level simulation framework in Python. The framework adheres to the standard 5G protocol stack. Multi-modal XR traffic is generated based on 3GPP-compliant models (Table I); abstracted PDCP and RLC layers handle packet processing, where PDCP discard timers define per-modality delay bounds. The MAC layer implements the complete DREAM-X architecture, including DTQ mechanisms. The physical layer adopts the 3GPP TR 38.901 channel model and supports modulation schemes from QPSK to 256-QAM based on MCS levels. Each resource block's capacity is determined by the number of subcarriers, OFDM symbols, modulation order, and code rate corresponding to the selected MCS index. The simulation runs 20 independent Monte Carlo iterations with different seeds for statistical significance.

DREAM-X operates in four variants: *DREAM-X with perfect knowledge of future arrivals and channels using the optimal solver (DK-MIP)*, *DREAM-X with predictions of arrivals and channels using the optimal solver (DP-MIP)*, *DREAM-X with perfect knowledge using heuristic (DK-HEU)*, and *DREAM-X with predictions using heuristic (DP-HEU)*. We

employed the Pyomo framework with the Gurobi solver for optimal solutions.

To establish performance baselines, we extend classical scheduling algorithms to accommodate multi-modal XR traffic while preserving their fundamental principles. We develop a unified parameterized framework through a generalized performance metric for user i and modality m: $\Phi_{i,m}(\alpha,\beta) = w_m \cdot \frac{(R_i^{\text{instant}})^\alpha}{(\bar{R}_{i,m})^\beta}$, where w_m represents modality-specific QoS weights (set to 1 for equal importance across all modalities), R_i^{instant} denotes instantaneous achievable data rate, $\bar{R}_{i,m}$ captures time-averaged rates per user-modality combination, and (α,β) control algorithmic behavior. The scheduling decision selects $(\text{user}^*, \text{modality}^*) = \arg\max_{i\in\mathcal{N},m\in\mathcal{M}, Q_{i,m}>0} \Phi_{i,m}(\alpha,\beta)$. This framework instantiates three baseline schedulers: *Multi-modal Proportional Fair (mPF)* with $(\alpha=1,\beta=1)$ balancing throughput and fairness, *Multi-modal Maximum Rate (mMR)* with $(\alpha=1,\beta=0)$ for pure throughput maximization, and *Multi-modal Round Robin (mRR)* with $(\alpha=0,\beta=1)$, ensuring systematic traversal regardless of channel conditions. Additionally, we implement the XR-aware PDU-Set Heuristic (PSH) [14], extending it to a multi-modal version (mPSH) by applying its original priority computation across all sensory modalities, thereby providing a more competitive, XR-specific baseline for a fairer comparison.

B. Traffic Modeling and System Configuration

Each sensory modality exhibits distinct traffic models as detailed in Table I. Video-XR traffic follows truncated Gaussian distributions, audio maintains near-constant bit rates (CBR) with minimal variation, haptic data features periodic patterns with uniform size variations, while olfactory and gustatory modalities follow Poisson processes with exponential distributions reflecting their sporadic nature [5], [6], [32]. Although olfactory and gustatory streams typically require 100 ms-1 s delay bounds, they are constrained to 10 ms through the $L_{\max}$ over-provisioning for space efficiency, as described earlier.

C. Performance Metrics

We employ six key performance metrics for comprehensive evaluation: *Satisfied Users* counts users meeting target delay-reliability across all subscribed modalities, representing the primary optimization objective. *Raw Delay Reliability* measures DU delivery success rates across all users, providing baseline performance assessment. *Effective Delay Reliability* computes delay-reliability exclusively for satisfied users, revealing quality experienced by successful XR sessions. *Experience Continuity* quantifies the percentage of simulation time users maintain uninterrupted satisfaction across all modalities, capturing temporal stability crucial for immersive experiences. *Jain's Fairness Index* evaluates fairness in delay reliability, mean delay, and jitter among satisfied users, with values approaching unity indicating superior fairness. *Delay Statistics* provide comprehensive DU delay analysis including mean, variance, and percentile distributions for satisfied users, enabling detailed QoS characterization. Note that all scheduling schemes listed in Table I are evaluated across all setups; however, for clarity, only those schemes that support a non-zero number of users are shown in the figures.

Table I
MULTI-SENSORY XR TRAFFIC MODELS AND SYSTEM CONFIGURATION

Video Modality	
Data Rate, Frame Rate, Delay Bound	30-60 Mbps, 60-120 fps, 5-10 ms
DU Size Dist.	Truncated Gaussian: $\mathcal{N}_T(\mu,(0.105\mu)^2,[0.5\mu,1.5\mu])$
Inter-Arrival Time Dist.	$\frac{1000}{\text{Frame Rate}} + \mathcal{N}_T(0,4,[-4,4])$ ms
Audio Modality	
Data Rate, Frame Rate, Delay Bound	0.256 Mbps, 50 fps, 5-10 ms
DU Size Dist.	CBR with minimal variation: $\mu + \mathcal{N}(0,(0.02\mu)^2)$
Inter-Arrival Time	Periodic
Haptic Modality	
Data Rate, Frame Rate, Delay Bound	1 Mbps, 1000 fps, 2.5-5 ms
DU Size Dist.	Uniform variation: $\mu + \mathcal{U}(-0.05\mu, 0.05\mu)$
Inter-Arrival Time Dist.	$1 + \mathcal{N}(0, 0.04)$ ms
Olfactory/Gustatory Modalities	
Data Rate, Frame Rate, Delay Bound	0.1 Mbps, 2 fps (average), 100 ms
DU Size Dist.	Exponential: Exp(μ) where μ = mean DU size
Inter-Arrival Time Dist.	Poisson process with $\lambda = 2$ Hz
5G NR System Configuration	
Carrier Frequency	3.5 GHz (FR1)
Transmission Power	43 dBm (UMa), 24 dBm (InH)
Bandwidth	50-100 MHz
Numerology	0, 1, 2 (15, 30, 60 kHz SCS)
Channel Model	3GPP TR 38.901 UMa and InH
Deployment Area	500m radius (UMa), 50m radius (InH)
Antenna Heights	h_{BS} = 25m (UMa), 3m (InH); h_{UE} = 1.5m
UE Mobility	Stationary
UE Distribution	Uniform random within cell coverage
Modulation Schemes	QPSK - 256QAM
Channel Estimation	Perfect CSI at current slot
DREAM-X Scheduler Parameters	
DTQ Depth ($L_{\max}$)	10 ms (truncated), 100 ms (full)
Delay-Reliability (ρ)	99-99.9%
Scheduling Schemes	mPF, mMR, mRR, mPSH [14], DREAM-X variants

D. Results and Discussions

DREAM-X's performance is evaluated with various system parameters against existing baselines, as follows.

1) Impact of Input mXR Users: Figure 4 shows how various scheduling schemes handle increasing user loads. We observe that even with just 5 users, traditional schemes struggle significantly—mMR satisfies only 1 user by greedily allocating all resources to the user with the best channel condition, while mPF and mRR fail entirely to support even a single user. Even mPSH [14], designed for video-XR applications, does not support a single mXR user (see the next subsection for a detailed discussion). DREAM-X variants intelligently distribute RBs based on actual requirements, satisfying nearly all users with minimal differences between optimal (MIP) and heuristic solvers. At 10 users, the performance gap widens: traditional schemes maintain their poor performance, while DREAM-X variants support approx 9.8-8.5 users. Here, subtle differences emerge between DREAM-X variants: DK-

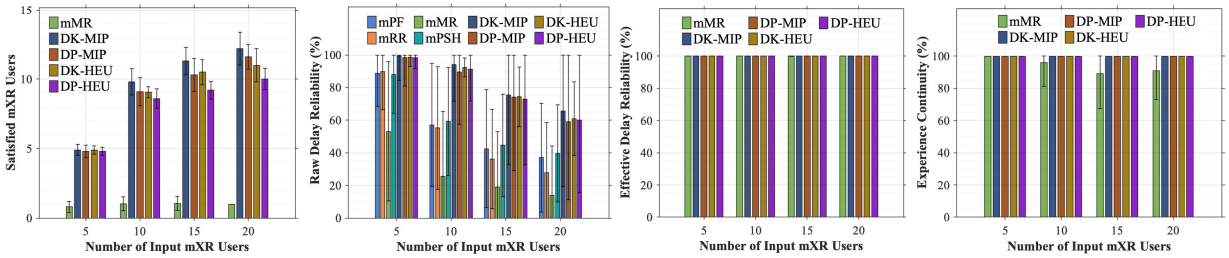


Figure 4. Effect of input mXR users: video data rate v_{d_r} = 30Mbps; video frame rate v_{f_r} = 60fps; V/A/O/G Delay Bound (DB) = 10ms; Haptic DB = 5ms; Target Delay-Reliability (ρ) = 99%; f_c = 3.5 GHz; BW = 100MHz; Numerology μ = 1; Environment = UMa.

MIP achieves the highest performance supporting almost 10 users, whereas DK-HEU supports slightly lesser, proving our heuristic's effectiveness. Prediction-based variants (DP-MIP, DP-HEU) perform slightly lower due to prediction errors, but the gap remains small because we re-optimize at each time slot with updated information following the MPC based rolling-horizon framework, preventing error accumulation. The same performance patterns persist at 15-20 users, but satisfaction plateaus around 10-12 users—revealing the network's fundamental capacity limit.

Raw delay-reliability metrics, averaging across all users (satisfied and unsatisfied), decrease with load and show mPF/mRR outperforming mMR since they serve multiple user-modality pairs rather than focusing on good channel conditions. mPSH performs slightly better than traditional schemes due to its specifically designed heuristic for XR applications. However, since mXR requires $> 99\%$ delay-reliability across all modalities simultaneously, mMR's focused approach actually yields better user satisfaction than other scheme's scattered efforts. All DREAM-X variants achieve $\sim 99.9\%$ *effective delay-reliability* and *experience continuity* for satisfied users, demonstrating scheme effectiveness.

2) Discussion on mPSH Scheduling: The previous subsection showed that mPSH, despite being state of the art, fails to support even a single user under standard conditions, underperforming even mMR. Here, we examine this behavior under varying configurations. mPSH computes priority as $m_{k,i} = e^{\alpha_{k,i}}/\beta_{k,i}$, where $\alpha_{k,i}$ denotes transmission progress and $\beta_{k,i}$ reflects normalized urgency. While temporally aware, this approach contrasts with our mDTQ-based explicit delay tracking, which proves critical under tight timing constraints.

Figure 5 systematically evaluates mPSH under four configurations. Under favorable conditions (5 users, 100 MHz bandwidth, and relaxed delay bounds of 20 ms for V/A/O/G and 10 ms for H, which is double the standard), mPSH supports approximately 4 users, compared to 5 supported by DREAM-X variants (subplot a). However, performance degrades sharply under resource constraints. At 50 MHz bandwidth (subplot b), mPSH supports only one user, while DREAM-X maintains support for approximately 4.5. Increasing the network load to 10 users at 100 MHz (subplot c) causes complete failure of mPSH, whereas DREAM-X continues to support nearly all users. Finally, under standard tight delay bounds (10/5 ms in accordance with 3GPP specifications, subplot d), mPSH fails entirely, while DREAM-X supports between 9 and 10 users. These outcomes demonstrate the critical importance of explicit

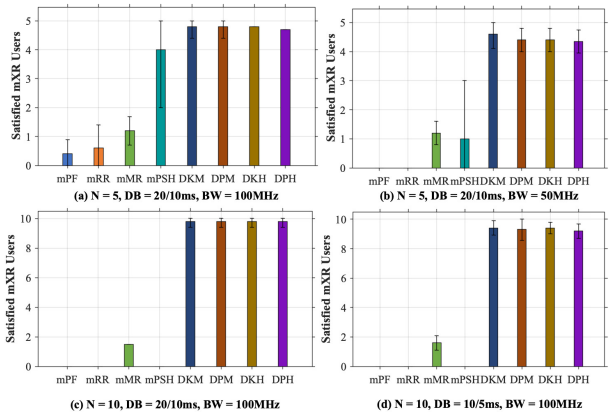


Figure 5. Comparative evaluation of mPSH with other schemes under varying network conditions. (DKM : DK-MIP, DPM : DP-MIP, DKH : DK-HEU, DPH : DP-HEU).

delay tracking for multi-modal XR scheduling.

This performance gap highlights two key limitations. First, lacking real-time visibility into delay budgets, mPSH cannot distinguish delay-critical data from data with temporal slack, instead relying solely on completion ratios. Moreover, it does not incorporate urgency into an optimization framework, unlike DREAM-X. Second, under constrained resources (such as limited system bandwidth), mPSH continues to allocate resources across all users, resulting in widespread partial delivery and zero satisfaction. In contrast, DTQ enables explicit delay tracking, which DREAM-X integrates into an optimized solution. Its design supports strategic abandonment, focusing resources on feasible users while sacrificing others, thereby implementing implicit admission control that transforms uncontrolled degradation into managed capacity with guaranteed QoS. These results underscore the need for explicit delay tracking and targeted allocation for delay-reliable scheduling.

3) Delay Analysis: Figure 6 reveals the delay characteristics under increasing network congestion (considering only satisfied users). Under light loads (5 users), mMR achieves low delays through opportunistic scheduling, but satisfies very few users. Our DREAM-X variants maintain consistent median delays around 5.5ms—higher than mMR but within target requirements. As load increases to 10 users, DREAM-X demonstrates remarkable stability with median delays between 5.5-6.5ms and 95th percentile delays below 7.5ms. Under heavy loads, DREAM-X continues to deliver consistent performance with median delays stabilizing around 6-7ms, while 95th percentile delays never exceed 9ms—ensuring delay-

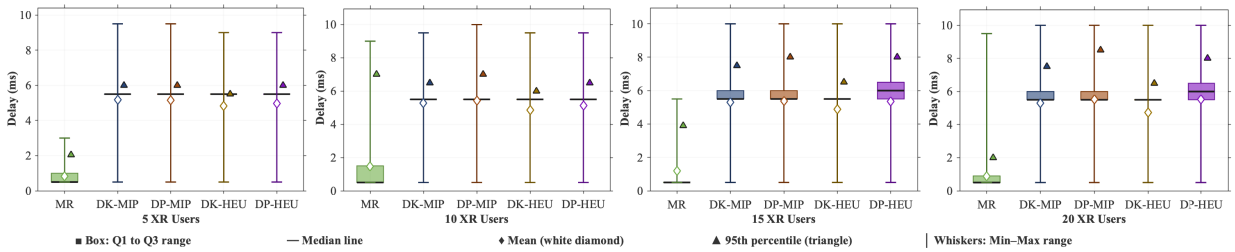


Figure 6. Delay statistics: $N = 20$; v_{d_r} = 30Mbps; v_{f_r} = 60fps; DB = 10ms(V/A/O/G), 5ms(H); ρ = 99%; FR1; BW = 100MHz; UMa.

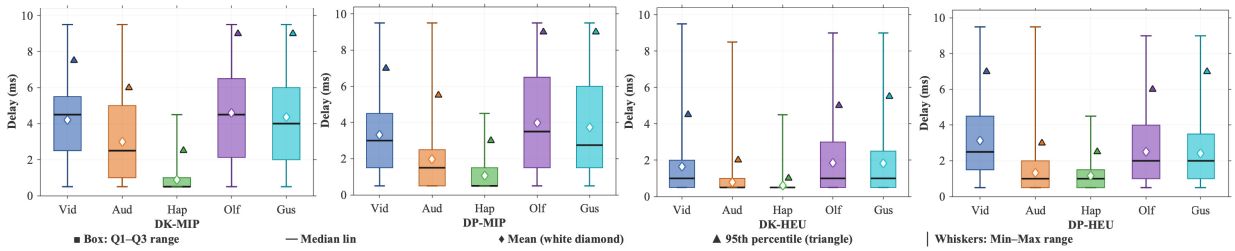


Figure 7. Multimodal Delay Statistics: $N = 20$; v_{d_r} = 30Mbps; v_{f_r} = 60fps; DB = 10ms(V/A/O/G), 5ms(H); ρ = 99%; FR1; BW = 100MHz; UMa.

reliability essential for immersive experiences.

4) Inter-modal Synchronization: Beyond overall delay, multisensory XR demands inter-modal synchronization across all modalities. Figure 7 presents a per-modality delay analysis for satisfied users, illustrating how DREAM-X accommodates diverse temporal requirements and synchronization across five sensory streams. Delivery of video data, the most bandwidth-intensive modality, records median delays of 3–5 ms and 95th percentiles of 7–9 ms, well within the 10 ms bound; these results reflect the success of scheduling large video-DUs over multiple slots. Audio data exhibits tighter control, with medians of 2–3 ms and 95th percentiles below 6 ms, aided by small packet sizes that allow flexible scheduling. Haptic traffic, constrained by the strictest bound of 5 ms due to tactile feedback demands, achieves the lowest delays, with medians of 0.5–1 ms and 95th percentiles below 3 ms, as DREAM-X's DTQ accurately prioritizes these packets. Transmission of olfactory and gustatory modalities, governed by relaxed 10 ms limits, exhibits medians of 2–6 ms and 95th percentiles near 9 ms, showing effective exploitation of their temporal slack to prioritize the more constrained stream types.

The consistency across DREAM-X variants underscores the framework's robustness. These results offer key insights into inter-modal alignment. Since our formulation enforces joint scheduling with an AND relationship, requiring all modalities to meet their delay bounds for a user to be satisfied, tight delay control naturally ensures synchronization. For satisfied users, all modal data arrives within narrow temporal windows (0.5–6 ms medians), bounding inter-modal skew by differences in delay distributions. For example, the worst-case audio-video skew is approximately 3 ms (5 ms video median minus 2 ms audio median), well within the 30 ms psychophysical threshold for AV sync. This demonstrates that our implicit synchronization, achieved through joint optimization under strict per-modality constraints, maintains the inter-modal alignment essential for high-fidelity multisensory XR.

5) Fairness Analysis: Figure 8 shows *Jain's Fairness Index* (JFI), computed over satisfied users only. We consider only the satisfied users to correctly reflect the binary nature of XR service, where users either meet the 99% delay-reliability target or do not. Maximizing satisfied users under resource constraints requires fully serving a feasible subset while excluding others, effectively implementing implicit admission control. Including all users in the JFI calculation would misrepresent the optimization objective. DREAM-X variants exhibit high fairness across delay reliability, mean delay, and jitter under all loads. In contrast, mMR shows high fairness for reliability but poor fairness for delay and jitter due to opportunistic scheduling and the absence of DTQ mechanisms. It transmits under favorable channel conditions, resulting in inconsistent delays even among the few satisfied users.

6) Impact of System Bandwidth: Figure 9 analyzes the impact of available spectrum on performance. At 50 MHz, schemes such as mPF, mRR, and mPSH fail completely, while mMR supports only one user. DREAM-X variants remain resilient, supporting about 5 users through efficient resource optimization. Doubling the bandwidth to 100 MHz yields minimal gains for existing approaches, but DREAM-X variants fully exploit the additional spectrum. DK-MIP scales up to 12 users, while others reach 10 to 11. This highlights a key insight: additional bandwidth alone offers limited benefit without intelligent scheduling.

7) Impact of Data Rate: Figure 10 examines the impact of video data rate. At 30 Mbps, DREAM-X variants support 10 to 12 users. Increasing the rate to 60 Mbps reveals a clear capacity trade-off, with performance tapering to 7 to 8 users. Higher rates consume more resource blocks per user, limiting the number of concurrent sessions even for advanced schedulers. While video is the most bandwidth-intensive modality, similar trends are observed across all modalities.

8) Impact of Frame Rate: Figure 11 shows the impact of video frame rate variations. At 60 fps, DREAM-X variants face challenging conditions with DK-MIP supporting 6 users

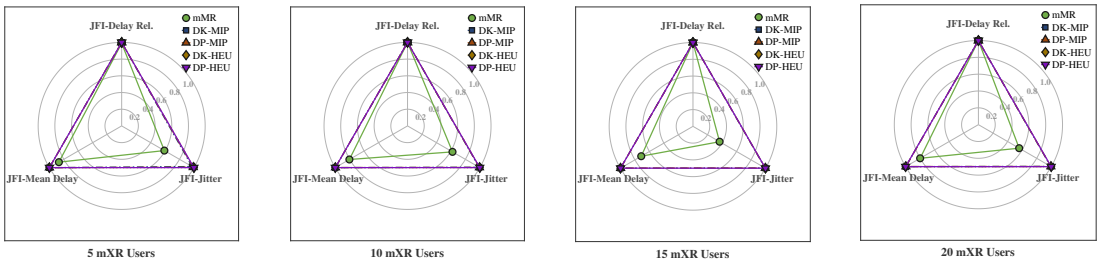


Figure 8. Jain's Fairness Index (JFI) of Delay reliability, Mean Delay and Jitter with respect to input mXR users: v_{d_r} = 30Mbps; v_{f_r} = 60fps; DB = 10ms(V/A/O/G), 5ms(H); ρ = 99%; FR1; BW = 100MHz; UMa.

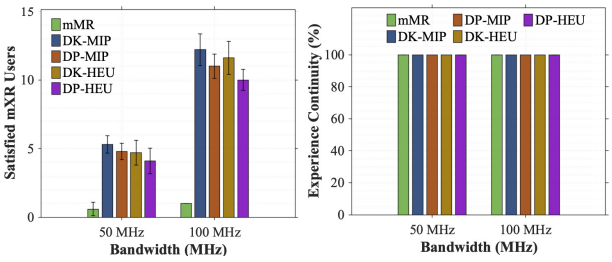


Figure 9. Effect of Bandwidth: $N = 20$; v_{d_r} = 30Mbps; v_{f_r} = 60fps; DB = 10ms(V/A/O/G), 5ms(H); ρ = 99%; FR1; UMa.

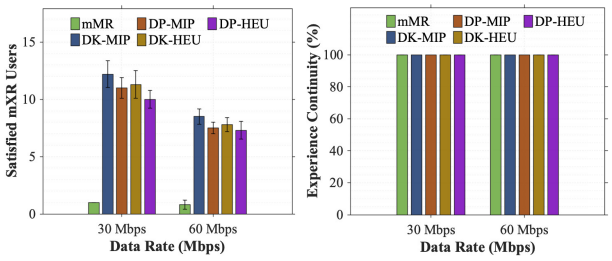


Figure 10. Effect of Data Rate: $N = 20$; v_{d_r} = 30, 60 Mbps; v_{f_r} = 60, 120fps; DB = 10ms/5ms; ρ = 99%; FR1; BW = 100MHz; UMa.

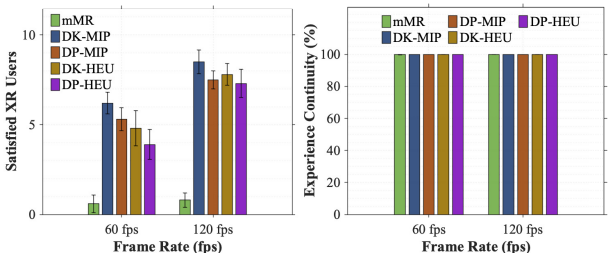


Figure 11. Effect of Frame Rate: $N = 20$; v_{d_r} = 60Mbps; v_{f_r} = 60, 120fps; DB = 10ms(V/A/O/G), 5ms(H); ρ = 99%; FR1; BW = 100MHz; UMa.

while other variants accommodate around 5-4 users. This is because at 60 fps, video DUs arrive in larger, more bursty manner that create scheduling bottlenecks. When multiple users experience simultaneous bursts, the scheduler struggles to serve large DUs within strict delay bounds. Conversely, 120 fps distributes the same total data rate across smaller, more frequent DUs, creating a relative uniform traffic flows that are easier to schedule. Our results highlight that traffic uniformity, not just total load, critically determines performance.

9) Impact of Delay Bound Requirements: Figure 12 examines how delay bounds affect performance. When we impose stricter delay bounds (5/2.5 ms for video/haptic),

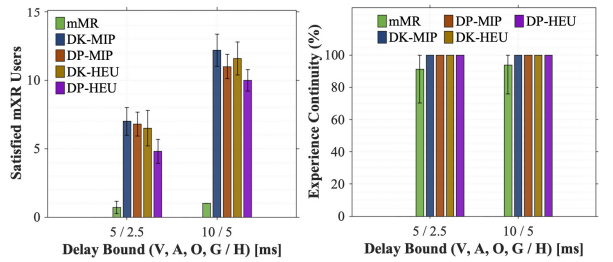


Figure 12. Effect of Delay Bound: $N = 20$, v_{d_r} = 30Mbps, v_{f_r} = 60fps, Delay-Reliability = 99%, f_c = 3.5 GHz, FR1, BW = 100MHz, μ = 1, UMa.

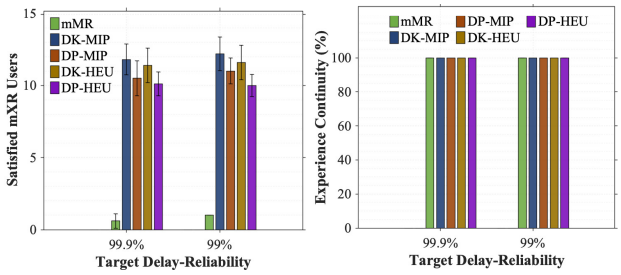


Figure 13. Effect of Delay-Reliability: $N = 20$, v_{d_r} = 30Mbps, v_{f_r} = 60fps, V/A/O/G DB = 10ms, H DB = 5ms, FR1, BW = 100MHz, μ = 1, UMa.

even DREAM-X faces significant challenges, exposing the fundamental difficulty of meeting sub-5ms latencies for complex multi-sensory traffic. With relaxed delay bounds (10/5 ms), DREAM-X variants unlock substantial improvements. Traditional schemes show minimal gains, with mMR still capped at one user. This asymmetric response reveals that DREAM-X can effectively exploit relaxed timing constraints, while existing schemes remain limited.

10) Impact of Target Delay-Reliability: Figure 13 demonstrates system performance under varying delay-reliability targets. The results reveal minimal performance differences between these stringent requirements—DREAM-X variants support 11-12 users under 99% targets with slight degradation to 10-11 users at 99.9%, while traditional schemes remain ineffective regardless of the reliability threshold. All DREAM-X variants maintain near-perfect continuity ($\sim$100%) under both delay-reliability targets, confirming that satisfied users receive consistently excellent service.

11) Impact of Prediction Horizon: Figure 14 reveals the relationship between prediction depth and scheduling performance. At zero horizon, DREAM-X operates purely on current DTQ knowledge, supporting 9-10 users. While effective, this myopic approach limits the scheduler's ability to anticipate

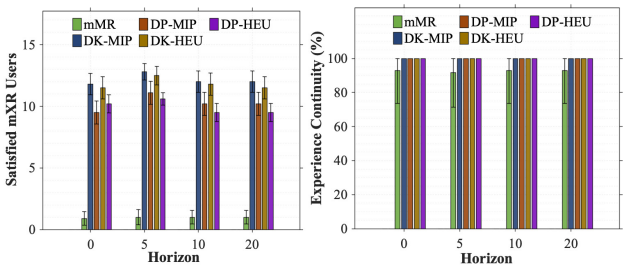


Figure 14. Effect of Prediction Horizon: $N = 20$; v_{d_r} = 30Mbps; v_{f_r} = 60fps; DB = 10ms(V/A/O/G), 5ms(H); ρ = 99%; FR1; BW = 100MHz; UMa.

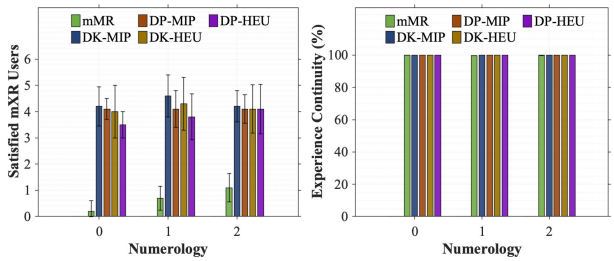


Figure 15. Effect of Numerology: $N = 20$; v_{d_r} = 30Mbps; v_{f_r} = 60fps; DB = 10ms(V/A/O/G), 5ms(H); ρ = 99%; FR1; BW = 100MHz; UMa.

future conditions. When we extend the horizon to 5 time slots, all DREAM-X variants reach peak capacity of 12-10 users, unlocking significant performance gains. This sweet spot demonstrates how modest future visibility enables superior planning. However, further horizon extension to 10 and 20 slots yields diminishing returns, with performance plateauing or slightly degrading due to accumulated prediction errors.

12) Impact of Numerology Configuration: Figure 15 analyzes the impact of numerology configuration on scheduling performance. At numerology 0, DREAM-X variants support approximately 4 users, whereas mPF, mRR, and mPSH fail to support any users, and mMR accommodates only 0.3 users. At numerology 1, DREAM-X achieves its peak capacity of 4.5–4.6 users, while mMR improves to 0.7 users. However, at numerology 2, a critical divergence appears: mMR continues its monotonic improvement, whereas DREAM-X exhibits a slight capacity reduction to approximately 4–4.1 users.

This reduction arises from fundamental differences in scheduling behavior. Traditional mMR consistently benefits from increased numerology due to its opportunistic approach; higher scheduling frequency, offering four times more opportunities at numerology 2 compared to numerology 0, enhances its ability to exploit favorable channel conditions for a single, greedily selected user. In contrast, DREAM-X's advanced multi-user, multi-modality optimization reaches its optimum at numerology 1, where temporal and spectral granularity are best balanced. The performance decline at numerology 2 results from inherent trade-offs. Doubling the subcarrier spacing from 30 kHz to 60 kHz halves the number of available RBs while doubling the scheduling frequency. For DREAM-X, which allocates RBs across multiple user-modality pairs, this reduction imposes substantial constraints. With fewer RBs, the scheduler has reduced flexibility to assign RBs to user-modality pairs, particularly given the heterogeneous channel conditions and stringent delay requirements, along with the inability to allocate fractional RBs. This limits the efficiency of multi-user RB allocation, even under perfect delay tracking. Although the increased scheduling frequency provides finer temporal resolution, it does not compensate for the reduced RB availability. Numerology 1 thus emerges as the optimal point, providing adequate temporal responsiveness for sub-10 ms delay bounds while preserving sufficient RB resources for effective allocation across user-modality pairs. Experience continuity remains nearly perfect ($\sim$100%) across all numerologies for DREAM-X, confirming consistent service quality for satisfied users.

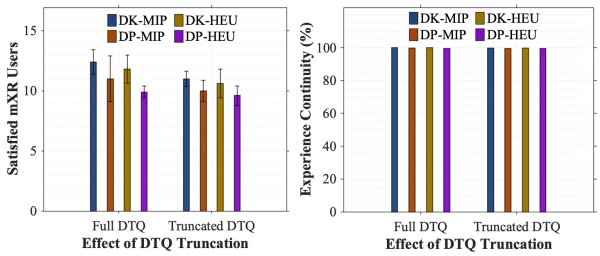


Figure 16. Effect of DTQ Truncation: $N = 20$; v_{d_r} = 30Mbps; v_{f_r} = 60fps; DB = 10ms(V/A/O/G), 5ms(H); ρ = 99%; FR1; BW = 100MHz; UMa.

These insights hold important implications for 6G standardization. While 5G determines numerology primarily based on deployment scenarios and frequency range, future 6G systems targeting immersive XR may benefit from incorporating scheduler complexity into numerology selection. In the system configuration studied here, numerology 1 yields the best performance by balancing temporal granularity with RB availability. However, additional experimentation and field validation across varied operational settings are required to refine these conclusions. Addressing this systematically remains a key direction for future work.

13) Impact of DTQ Truncation: Figure 16 quantifies the performance impact of DTQ truncation, addressing concerns about potential scheduling bias when olfactory and gustatory modalities, which have longer delay bounds (e.g., 100 ms), are artificially constrained to a truncated queue length ($L_{\max}$ = 10 ms). Results reveal modest but measurable degradation: full DTQ representation enables DK-MIP to support approximately 12.5 users, compared to 11 users with truncation, representing a reduction of 1 to 1.5 users across all DREAM-X variants. This degradation arises from the truncation's inability to accurately capture the full delay distributions of long-bound modalities, resulting in suboptimal decisions based on compressed delay information. Importantly, experience continuity remains nearly perfect ($\sim$100%) under both configurations, confirming consistently high-quality service for satisfied users. These findings demonstrate that while full DTQ representation offers better performance, it comes at the cost of greater complexity, whereas the truncated implementation provides a practical, memory-efficient alternative with acceptable trade-offs for resource-constrained deployments.

14) Performance in InH Setup: Figure 17 examines scheduling performance across deployment environments, comparing Urban Macrocell (UMa) and Indoor Hotspot (InH) scenarios. The InH scenario reveals significant differences

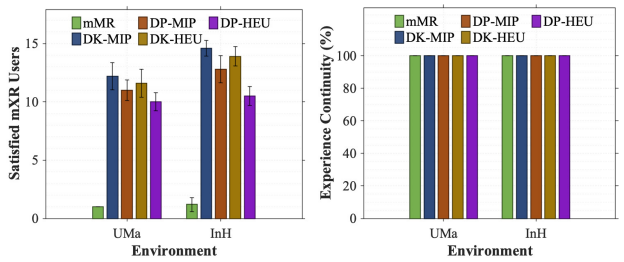


Figure 17. Effect of Environment: $N = 20$; v_{d_r} = 30Mbps; v_{f_r} = 60fps; DB = 10ms(V/A/O/G), 5ms(H); ρ = 99%; FR1; BW = 100MHz.

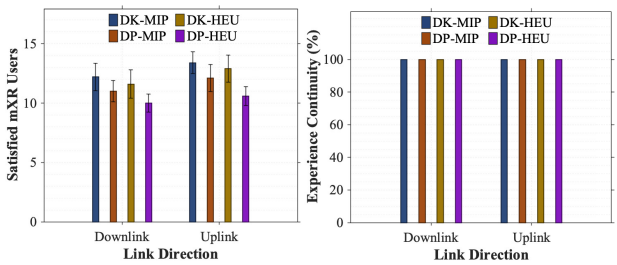


Figure 18. Uplink-Downlink Comparison: $N = 20$, V/A/O/G DB = 10ms, H DB = 5ms, ρ = 99%, f_c = 3.5 GHz, FR1, BW = 100MHz.

due to favorable indoor propagation characteristics. Traditional mMR shows minimal improvement, supporting approximately one user in UMa and barely 1.5 users in InH, failing to leverage the advantages of indoor channel conditions. mPF, mRR, and mPSH do not support even a single mXR user in either scenario. In stark contrast, DREAM-X variants demonstrate substantial capacity gains in the InH environment, supporting 12–10 users in UMa versus 14–11 users in InH, reflecting a 15–20% improvement. This enhancement is attributed to reduced path loss and higher line-of-sight (LOS) probability in indoor deployments, which enable more reliable high-rate transmissions under identical bandwidth constraints. All DREAM-X variants maintain near-perfect experience continuity across both deployment scenarios, confirming consistently high service quality for satisfied users regardless of the environment. These results demonstrate DREAM-X's capability to intelligently exploit scenario-specific channel characteristics while preserving robust delay-reliability guarantees.

15) Guide to Uplink mXR Scheduling: This paper explicitly targets downlink scheduling; however, uplink and end-to-end scheduling are equally critical. We now discuss the implications for uplink scheduling and how the proposed approach may extend to end-to-end scheduling, encompassing both uplink and downlink. Figure 18 provides preliminary insights through a simplified evaluation. We partition the end-to-end air-interface delay budgets as follows: for audio and video (25 ms total), 10 ms for downlink, 10 ms for uplink, and 5 ms for Scheduling Request (SR) plus Buffer Status Report (BSR)/Delay Status Report (DSR) signaling overhead; for haptic traffic (15 ms total), 5 ms each for downlink and uplink, and 5 ms for SR/BSR/DSR signaling.

Additionally, to reflect practical usage, we model asymmetric traffic patterns. In the downlink, traffic is video-heavy (30 Mbps video, 1 Mbps haptic), whereas in the uplink, there is more haptic data (25 Mbps video, 5 Mbps haptic). Results reveal a notable capacity asymmetry. Under the original downlink-focused evaluation, DREAM-X variants support 10 to 12 users. In the uplink scenario with more haptic traffic, DREAM-X supports 13 to 14 users. Despite similar aggregate data rates in both directions (approximately 30 to 31 Mbps), the uplink achieves higher capacity due to reduced video traffic, which mitigates the bursty patterns that typically challenge scheduling under tight delay constraints. When multiple users experience simultaneous video bursts, the scheduler struggles to deliver large DUs within strict bounds. In contrast, haptic traffic exhibits a more uniform temporal profile, which facilitates more efficient scheduling. We emphasize that these results provide directional insights but are not a substitute for comprehensive uplink scheduling analysis. Rigorous evaluation incorporating full BSR/DSR modeling, SR-induced stochastic delays, and coordinated uplink-downlink scheduling mechanisms remains critical future work.

E. Standardization Implications

This section discusses standardization requirements for guaranteeing multi-sensory XR KPIs in NextG networks, building upon 3GPP Release 18's PDU Set awareness framework [33]. 3GPP introduced PDU Sets to address video-XR requirements: application-layer frames are split into inter-dependent PDUs, all of which must be delivered within tight delay bounds to enable frame decoding. The standard defines the *PDU Set Delay Budget* (PSDB) for delivery deadlines, *PDU Set Error Rate* (PSER) for loss constraints, *PDU Set Importance* (PSI) for content criticality, and *PDU Set Integrated Handling Indicator* (PSIHI) for early discard signaling. In our framework, a Data Unit (DU) is analogous to PDU Sets; the delay bounds of the DUs correspond to PSDB, the target delay-reliability maps correspond to PSER, and modality-specific criticality relates to PSI.

Multi-Modal PDU-Set Extensions: The current standard primarily targets video-XR with single-modality optimization. Extending to multi-sensory XR requires additional parameters, such as *inter-modal synchronization parameters*, to specify temporal alignment requirements across modalities. For example, *inter-modal skew* (Δ_{m_i,m_j}): the maximum tolerable temporal misalignment between corresponding PDU sets of modalities m_i and m_j. Additionally, these parameters should support *dynamic cross-modal adjustment*: for example, when one modality violates its delay constraints, synchronized data from other modalities become perceptually useless, necessitating dynamic PSI reduction to avoid wasting resources. Furthermore, SDAP [30] must be extended with enhanced indicators that encode per-DU deadline information for each modality. Current MAC logical channel prioritization cannot differentiate packet criticality within QoS flows; new MAC Control Elements (CE) [29] are needed for real-time, per-modality delay budget signaling at the scheduler level.

Cross-Layer Control Signaling: Protocol layers must exchange mXR-specific requirements and performance feedback. RRC requires new signaling to carry the mXR configuration from the core network to the gNB, including per-modality delay bounds and delay-reliability thresholds [29], [31]. Standardized feedback mechanisms are essential, whereby the

MAC layer scheduler reports QoS violations (such as delay bound violations) to upper layers, enabling dynamic service adaptation. Current standards provide limited support for such fine-grained, multi-modal QoS signaling. 6G needs to standardize these capabilities to enable the cross-layer information exchange necessary for transforming multi-sensory XR from a conceptual vision into a deployable reality.

V. Conclusions and Future Work

This work addressed a fundamental challenge in realizing the IMT-2030 vision for immersive communication: enabling seamless multi-sensory XR experiences in NextG networks. DREAM-X, a comprehensive delay-reliability-aware downlink scheduling framework that transforms the intractable challenge of joint optimization across five human sensory modalities into a practical real-time solution, was introduced. Our key innovation, the Multi-modal Delay Tracking Queue architecture combined with Model Predictive Control, enables linear reformulation of complex nonlinear constraints, while balancing performance with computational tractability essential for sub-millisecond scheduling decisions. Extensive evaluation demonstrated a multi-fold performance improvement in terms of the number of satisfied mXR users with more than 99% delay-reliability, directly aligning with 6G KPIs. Additionally, DREAM-X provides critical insights for 3GPP standardization—from packet-level QoS differentiation to cross-layer signaling mechanisms.

Future work will explore how this framework can be extended to uplink or joint uplink-downlink scheduling.

References

[1] O. Hashash, C. Chaccour, W. Saad, T. Yu, K. Sakaguchi, and M. Debbah, "The seven worlds and experiences of the wireless metaverse: Challenges and opportunities," *IEEE Communications Magazine*, pp. 1–8, 2024.

[2] X. S. Shen, J. Gao, M. Li, C. Zhou, S. Hu, M. He, and W. Zhuang, "Toward immersive communications in 6g," *Frontiers in Computer Science*, vol. 4, 2023.

[3] ITU-R WP5, "IMT-2030: Framework and objectives for mobile telecommunications beyond 2030." https://www.itu.int/en/itu-r/study-groups/rsg5/rwp5d/imt-2030/pages/default.aspx, 2023. Accessed: 2025-06-09.

[4] X. Lin, "A tale of two mobile generations: 5g-advanced and 6g in 3gpp release 20," 2025.

[5] I. F. Akyildiz, H. Guo, R. Dai, and W. Gerstacker, "Mulsemedia communication research challenges for metaverse in 6g wireless systems," *ITU Journal on Future and Evolving Technologies*, vol. 4, p. 562–579, Nov. 2023.

[6] 3GPP, "Service requirements for the 5G system; Stage 1," Technical Specification TS-22.261, 3GPP, Dec. 2024. Release-20.

[7] C.-X. Wang *et al.*, "On the road to 6g: Visions, requirements, key technologies, and testbeds," *IEEE Communications Surveys Tutorials*, vol. 25, no. 2, pp. 905–974, 2023.

[8] A. Mamane, M. Fattah, M. E. Ghazi, M. E. Bekkali, Y. Balboul, and S. Mazer, "Scheduling algorithms for 5g networks and beyond: Classification and survey," *IEEE Access*, vol. 10, pp. 51643–51661, 2022.

[9] M. E. Haque, F. Tariq, M. R. A. Khandaker, K.-K. Wong, and Y. Zhang, "A survey of scheduling in 5g urllc and outlook for emerging 6g systems," *IEEE Access*, vol. 11, pp. 34372–34396, 2023.

[10] X. Wang, H. Yao, T. Mai, S. Guo, and Y. Liu, "Reinforcement learning-based particle swarm optimization for end-to-end traffic scheduling in tsn-5g networks," *IEEE/ACM Transactions on Networking*, vol. 31, no. 6, pp. 3254–3268, 2023.

[11] K. Min, Y. Kim, and H.-S. Lee, "Meta-scheduling framework with cooperative learning toward beyond 5g," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 6, pp. 1810–1824, 2023.

[12] W. Zhang, M. Derakhshani, G. Zheng, and S. Lambotharan, "Constrained risk-sensitive deep reinforcement learning for embb-urllc joint scheduling," *IEEE Transactions on Wireless Communications*, vol. 23, no. 9, pp. 10608–10624, 2024.

[13] V. K. Shrivastava, S. Rajendran, A. K. Abraham, and R. Rajadurai, "Enhanced scheduling strategy and energy efficiency for extended reality in 5g advanced," in *2024 IEEE 21st Consumer Communications Networking Conference (CCNC)*, pp. 546–549, 2024.

[14] P. Paymard, S. Paris, A. Amiri, T. E. Kolding, F. S. Moya, and K. I. Pedersen, "Pdu-set scheduling algorithm for xr traffic in multi-service 5g-advanced networks," in *ICC 2024 - IEEE International Conference on Communications*, pp. 758–763, 2024.

[15] E. Chen, S. Dou, S. Wang, Y. Cao, and S. Liao, "Frame-level integrated transmission for extended reality over 5g and beyond," in *2021 IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6, 2021.

[16] Y. Huang, S. Li, Y. T. Hou, and W. Lou, "Gpf+: A novel ultrafast gpu-based proportional fair scheduler for 5g nr," *IEEE/ACM Transactions on Networking*, vol. 30, no. 2, pp. 601–615, 2022.

[17] Y. Huang, S. Li, C. Li, Y. T. Hou, and W. Lou, "A deep-reinforcement-learning-based approach to dynamic embb/urllc multiplexing in 5g nr," *IEEE Internet of Things Journal*, vol. 7, no. 7, pp. 6439–6456, 2020.

[18] R. Dong, C. She, W. Hardjawana, Y. Li, and B. Vucetic, "Deep learning for radio resource allocation with diverse quality-of-service requirements in 5g," *IEEE Transactions on Wireless Communications*, vol. 20, no. 4, pp. 2309–2324, 2021.

[19] X. Zhao, Y.-J. A. Zhang, M. Wang, X. Chen, and Y. Li, "Online multi-user scheduling for extended reality transmissions with hard-latency constraint," in *GLOBECOM 2023 - 2023 IEEE Global Communications Conference*, pp. 4074–4079, 2023.

[20] X. Zhao, Y.-J. A. Zhang, M. Wang, X. Chen, and Y. Li, "Online multi-user scheduling for xr transmissions with hard-latency constraint: Performance analysis and practical design," *IEEE Transactions on Communications*, vol. 72, no. 7, pp. 4055–4071, 2024.

[21] P. Paymard, A. Amiri, T. E. Kolding, and K. I. Pedersen, "Optimizing mixed capacity of extended reality and mobile broadband services in 5g-advanced networks," *IEEE Access*, vol. 11, pp. 113324–113338, 2023.

[22] M. Laha, D. Roy, S. Dutta, and G. Das, "Ai-assisted improved service provisioning for low-latency xr over 5g nr," *IEEE Networking Letters*, vol. 6, no. 1, pp. 31–35, 2024.

[23] D. G. Morín, D. Medda, A. Iossifides, P. Chatzimisios, A. G. Armada, A. Villegas, and P. Peréz, "An extended reality offloading ip traffic dataset and models," *IEEE Transactions on Mobile Computing*, vol. 23, no. 6, pp. 6820–6834, 2024.

[24] F. Chiariotti, M. Drago, P. Testolina, M. Lecci, A. Zanella, and M. Zorzi, "Temporal characterization and prediction of vr traffic: A network slicing use case," *IEEE Transactions on Mobile Computing*, vol. 23, no. 5, pp. 3890–3908, 2024.

[25] N. H. Chu, D. T. Hoang, D. N. Nguyen, K. T. Phan, E. Dutkiewicz, D. Niyato, and T. Shu, "Metaslicing: A novel resource allocation framework for metaverse," *IEEE Transactions on Mobile Computing*, vol. 23, no. 5, pp. 4145–4162, 2024.

[26] B. Bojović, S. Lagén, K. Koutlia, X. Zhang, P. Wang, and L. Yu, "Enhancing 5g qos management for xr traffic through xr loopback mechanism," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 6, pp. 1772–1786, 2023.

[27] X. Zhang, Q. Zhu, and H. V. Poor, "Neyman-pearson criterion driven nfv-sdn architectures and optimal resource-allocations for statistical-qos based murllc over next- generation metaverse mobile networks using fbc," *IEEE Journal on Selected Areas in Communications*, vol. 42, no. 3, pp. 570–587, 2024.

[28] D. Liu, C. Huang, L. Xue, W. Zhuang, X. Shen, and B. Ying, "Collaborative and verifiable vnf management for metaverse with efficient modular designs," *IEEE Journal on Selected Areas in Communications*, vol. 42, no. 3, pp. 616–628, 2024.

[29] "Nr; medium access control (mac) protocol specification," in *3GPP TS 38.321 V18.5.0*, Mar. 2025.

[30] "E-utra and nr; service data adaptation protocol (sdap) specification," in *3GPP TS 37.324 V18.0.0*, Mar. 2024.

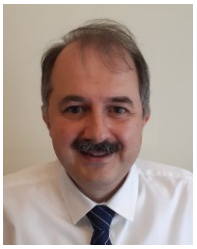
[31] "Nr; packet data convergence protocol (pdcp) specification," in *3GPP TS 38.323 V18.5.0*, Mar. 2025.

[32] "Traffic models and quality evaluation methods for media and xr services in 5g systems," in *3GPP TR 26.926 V0.1.0*, Apr. 2021.

[33] S. Gül, S. Ahsan, S. Paris, and I. D. D. Curcio, "Pdu set-based qos handling in 3gpp: Release 18 overview and future directions," *IEEE Communications Standards Magazine*, vol. 9, no. 3, pp. 104–110, 2025.

Moyukh Laha (Member, IEEE) is a SyMeCo Marie Curie Postdoctoral Fellow at the School of Electronic Engineering, Dublin City University, Ireland. He obtained his Ph.D. from the Department of Electronics and Electrical Communication Engineering at the Indian Institute of Technology (IIT) Kharagpur in 2022. After completing his PhD, he was employed by the 5G Standardization team at IIT Kharagpur, where he contributed to 3GPP and other standardization organizations before joining DCU. His research interests encompass 5G and 6G communication networks, XR/metaverse integration within 5G/6G, and vehicular networks.

Goutam Das is a Professor in the Department of Electronics and Electrical Communication Engineering at the Indian Institute of Technology Kharagpur, India. He obtained the M.Tech. degree from the Indian Institute of Technology Kharagpur, India, in 2001, and the Ph.D. degree from the University of Melbourne, Australia, in 2008. He served as a Senior Research Engineer at the General Electric Research and Development Center for 3.5 years, from 2001 to 2004. He served as a Postdoctoral Fellow at Ghent University, Belgium, from 2009 to 2011. His research focus includes optical and wireless networking.



Gabriel-Miro Muntean (Fellow, IEEE) received his Ph.D. degree from Dublin City University (DCU) Ireland in 2004 for research on quality-oriented adaptive multimedia transmissions. He is currently a Professor with the School of Electronic Engineering and the Co-Director of the Performance Engineering Laboratory at DCU. He has authored over 500 papers in prestigious international journals and conferences, written four books and 30 book chapters, and edited six additional volumes. He coordinated the EU Horizon 2020 project NEWTON, directed the DCU team in the EU Horizon Europe projects TRACTION and HEAT and led other significant Irish research projects, such as eStream and FRADIS. His research interests include quality, performance, and energy efficiency associated with multimedia and multisensory media delivery, technology-enhanced learning, and various data transfers across heterogeneous networks. He was chairperson, reviewer, and technical committee member of other top international conferences, journals, and funding organizations. Professor Muntean serves as an Associate Editor for the IEEE TRANSACTIONS ON BROADCASTING and IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING and he is the Multimedia Communications Area Editor of IEEE COMMUNICATIONS SURVEYS AND TUTORIALS.