PDU Set-Aware Delay-Reliability-Optimal Downlink Scheduling for Extended Reality Applications in 6G

Moyukh Laha*, Gabriel Miro-Muntean*
*School of Electronic Engineering, Dublin City University, Ireland
{moyukh.laha, gabriel.muntean}@dcu.ie

***Abstract*—Extended Reality (XR) is poised to become a flagship application in 6G networks, requiring the joint support of enhanced mobile broadband (eMBB) and ultra-reliable low-latency communications (URLLC)—a service class beyond the capabilities of current 5G systems. To meet these demands, 3GPP Release 18 introduces a Protocol Data Unit (PDU) Set-based quality of service (QoS) framework, necessitating XR scheduling schemes that are both PDU Set-aware and capable of satisfying stringent QoS requirements. This paper presents a downlink scheduling framework tailored for XR services in 6G, incorporating a novel mechanism for tracking PDU Set delay. Employing this framework, we formulate XR scheduling as an optimization problem that enforces PDU Set-specific QoS constraints and solve it optimally to maximize the number of XR users meeting their service requirements. Simulation results demonstrate substantial performance improvements, achieving 20–30% gains over state-of-the-art delay-reliable schedulers and multifold gains over traditional approaches.**

***Index Terms*—Extended reality, PDU-Set, Delay-reliable Scheduling, 6G.**

I. Introduction

Extended Reality (XR)—encompassing Virtual Reality (VR), Augmented Reality (AR), and Mixed Reality (MR)—is set to revolutionize human-human and human-machine interactions. Recognized as a key use case for future 6G systems [1], XR demands an unprecedented combination of communication requirements: high throughput for volumetric data, ultra-low latency for real-time responsiveness, and high reliability to preserve immersive quality. These demands situate XR at the convergence of *enhanced Mobile Broadband* (eMBB) and *Ultra-Reliable Low-Latency Communication* (URLLC) [1], creating significant challenges for next-generation wireless networks.

A key challenge in XR application support lies in its traffic structure: each application-layer frame is split into multiple interdependent *Protocol Data Units* (PDUs), all of which must be delivered within a tight delay bound to enable frame decoding. The loss or delay of even one PDU renders the entire frame unusable, resulting in wasted resources and degraded quality of experience [2]. To address this, 3GPP Release 18 introduced the PDU Set framework [3], [4], which groups all PDUs associated with a single media frame into a PDU Set. This framework extends IP-layer QoS by embedding XR-specific service parameters: the *PDU Set Delay Budget* (PSDB) defines the delivery deadline; the *PDU Set Error Rate* (PSER) constrains permissible losses; the *PDU Set Importance* (PSI) reflects content criticality; and the *PDU Set Integrated Handling Indicator* (PSIHI) signals that incomplete delivery should trigger early discard. Furthermore, 3GPP redefines XR capacity as the number of users that a cell can support [5], moving beyond traditional metrics centered on aggregate throughput. An XR user is deemed satisfied only if at least x% (e.g., 99%) of its PDU sets are received within the PSDB [2], establishing a QoS criterion that jointly enforces delay and reliability—what we refer to as *delay-reliability*. This fundamentally shifts the system objective from maximizing throughput to ensuring frame-level deadline compliance.

To support XR traffic, MAC-layer schedulers must therefore operate at PDU Set granularity and respect strict delay-reliability constraints. However, few scheduling algorithms in the literature satisfy these requirements. Conventional schedulers such as Proportional Fair (PF) treat PDUs independently and are ill-suited for XR traffic. Some XR-aware approaches have been proposed: for instance, [6] models the scheduling problem as a periodic Markov Decision Process but assumes uniform packet sizes and periodic arrivals, which limits its practical use; [7] introduces Frame-Level Integrated Transmission without delay-reliability awareness. Architectural enhancements, such as the application-aware MAC in [8] and AI-assisted provisioning in [9], improve user satisfaction but do not target PDU Set-level XR scheduling. A heuristic scheduler in [2] incorporates PDU Set-awareness but lacks explicit delay tracking and does not solve the problem optimally. The Delay-Reliability Optimal XR (DROX) scheduler [10] introduced a Delay Tracking Queue (DTQ) at the MAC layer to meet delay-reliability constraints. This mechanism was later extended to multisensory scenarios in [11]. However, DROX does not explicitly consider PDU Set awareness or all its associated metrics as defined by 3GPP Release 18.

This paper addresses this gap by proposing PDU Set-aware Delay-Reliability Optimal XR Scheduling (PaDXR)—a novel scheduler that fully aligns with the Rel-18 PDU Set framework and exploits 3GPP-defined PSI-based content differentiation to enhance XR capacity. To this end, we propose the *PDU Set Delay Tracking Queue* (PDTQ) mechanism to track precise delay of the PDU sets. Based on this framework, we formulate the XR scheduling problem as a Mixed-Integer Nonlinear Program (MINLP), which we linearize to obtain a Mixed-Integer Linear Program (MILP) solvable using standard solvers.

The main contributions of this paper are as follows:

- We introduce PDTQ, a PDU Set-aware delay tracking mechanism that accurately monitors PDU Set deadlines across different PSI levels.

- We propose PaDXR, a downlink scheduler that leverages PDTQ and formulates a PDU Set-aware optimization problem as a MINLP, which is linearized to a MILP for a tractable, optimal solution.
- Through extensive simulations, we demonstrate that PaDXR significantly outperforms both conventional and state-of-the-art delay-reliability-aware XR schedulers across diverse operating conditions.

The remainder of the paper is structured as follows: Section II details the PDTQ mechanism; Section III presents the PaDXR scheduler and its formulation and solution; Section IV evaluates performance through simulation; and Section V concludes the paper with future research directions.

II. PDU Set Delay Tracking Mechanism

A. System Model and Assumptions

We adopt the standard 5G/6G protocol stack. Data generated by the application layer is buffered in per-user Medium Access Control queues (MACQs) at the gNB, each storing the PDU sets along with their associated parameters, such as the PDU set delay budget (PSDB), PDU set importance (PSI), etc. Our scheme operates at the gNB's MAC layer and focuses solely on downlink scheduling; end-to-end scheduling, including uplink, is left for future work. We consider only video-based XR traffic due to its high bandwidth demands from volumetric content. We assume each XR user runs a single application within a dedicated XR-RAN slice provisioned exclusively for XR services. As per 3GPP Release 18 [12], XR-awareness allows the RAN to manage XR traffic independently of other 5G service classes; mixed-application scenarios are therefore excluded and left for future work. We also assume perfect knowledge of data arrivals and channel conditions for optimal solutions, with practical scenarios involving partial or uncertain information considered in future work.

B. PDTQ Structure and Operation

Building on the delay tracking mechanism in [10], we introduce two key enhancements: (a) a PDU Set-level delay tracking mechanism, and (b) incorporation of PDU Set importance levels defined by the PSI metric [3], [4]. The proposed mechanism consists of *PDU Set Delay Tracking Queues (PDTQs)* and a configurable *Delay Tracking Granularity (DTG)*.

PSI-Awareness: The PDU-Set Importance (PSI), introduced in 3GPP Release 18 for XR-specific enhancements in NR, quantifies the relative importance of PDU sets within a QoS flow. Under congestion, the RAN can prioritize higher-importance PDU sets (e.g., I-frames in video XR) while deferring or dropping lower-importance ones (e.g., B-/P-frames), which impact quality but not decodability. This enables graceful degradation under resource constraints. Unlike uniform reliability enforcement, which allocates resources equally across all content, PSI-aware handling enables strategic allocation—providing strict guarantees to critical *base layer* data and relaxed targets for *enhancement layer* content—thereby maximizing user satisfaction under capacity limits.

Table I
Notation for PaDXR Scheduling

Symbol	Description
N	Number of XR users
T	Scheduling duration (time slots)
$\rho \in \{b, e\}$	PSI level (base, enhancement)
$\gamma_i^{\rho}(T)$	Delay-reliability violation rate for user i, PSI level ρ
μ_i^{ρ}	Target delay-reliability violation threshold, user i, PSI ρ
$Q_i^{j,\rho}(t)$	Queue content at PDTQ-j for user i, PSI level ρ
K_i^{ρ}	Maximum PDTQ index for user i, PSI level ρ
$\mathcal{J}_i^{\rho}$	Set of PDTQ indices $\{1, 2, \dots, K_i^{\rho}\}$
D^{ρ}	PSDB for PSI level ρ
τ	Delay tracking granularity (DTG)
$a_i^{\rho}(t)$	Arrival indicator for user i, PSI ρ at time t
$f_i^{\rho}(t)$	PDU Set size for user i, layer ρ at time t
$\alpha_i(t)$	Achievable data rate per RB for user i at time t
$x_i^{j,\rho}(t)$	RB allocation for PDTQ-j, user i, PSI level ρ
C	Total system capacity (resource blocks)
$\mathbb{I}\{\cdot\}$	Indicator function

PDTQ Architecture: PSI-tagged PDU sets are buffered in per-user MAC queues and mapped to virtual PDTQs based on their remaining PSDB (Fig. 1). Each PDTQ index ϕ hosts two sub-queues: one for the base layer (critical PDU sets, high PSI) and one for the enhancement layer (non-critical, low PSI). Delay Tracking Granularity (DTG) τ, aligned with the TTI (e.g., 0.5 ms for numerology 1), defines the resolution. For a PDU set with remaining PSDB Δ_i and PSI level $\rho \in$ base : b, enhancement : e, the PDTQ index is $\phi = \lfloor(\Delta_i - \tau)/\tau\rfloor$. Based on PSI, each PDU set is placed into the corresponding sub-PDTQ at index ϕ (Fig. 1). Scheduling is performed at the PDU Set level within each sub-PDTQ. This dual-queue structure exposes both temporal urgency (via PDTQ index) and content importance (via sub-queue), enabling informed scheduling. While the current design supports two PSI levels, it can be extended with additional sub-queues.

PDTQ Dynamics: At each scheduling interval, PDU sets are selected from both sub-PDTQs (Section III). Unscheduled PDUs shift to the next lower-indexed sub-PDTQ within the same priority class (PDTQ-j $\rightarrow$ PDTQ-$(j-1)$). PDU sets remaining in PDTQ-1 after the scheduling are dropped due to delay violation. As shown in Fig. 2, a violation in either sub-PDTQ leads to a user-level violation, as both base and enhancement PDU sets must meet their delay-reliability targets. The shifting mechanism ensures accurate tracking of remaining delay budgets.

III. PDU-Set Aware Delay Reliable Optimal Scheduling

We now formulate the scheduling problem employing the PDTQ framework. Table I summarizes the key symbols.

1) Objective Function: The optimization minimizes XR users failing to meet delay-reliability requirements. A user is deemed satisfied only when both PSI level PDU sets meet their respective delay-reliability targets:

$$\mathcal{O} : \text{minimize} \sum_{i=1}^{N} \mathbb{I}\left\{ \bigvee_{\rho \in \{b,e\}} \gamma_i^{\rho}(T) \geq \mu_i^{\rho} \right\} \tag{1}$$

where the logical OR ($\bigvee$) indicates that violation of either PSI level results in user dissatisfaction.

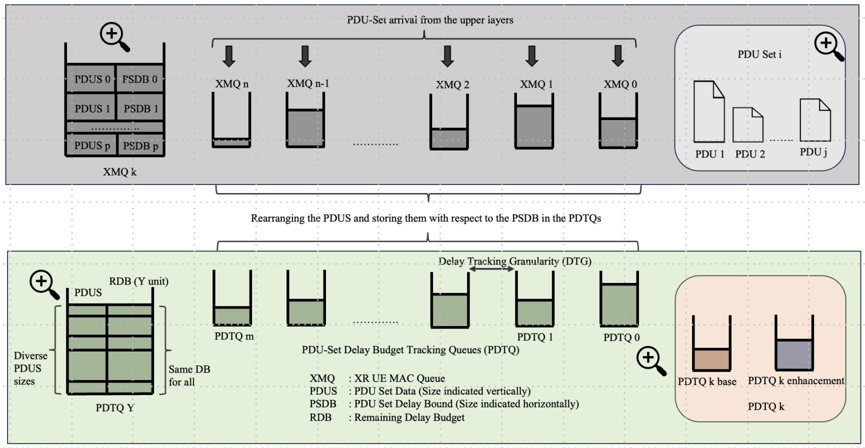

Figure 1. PDTQ mechanism: PDU sets from the upper layers come to the individual XR user MAC queue, which are then placed in the appropriate sub-PDTQs based on their PSDB. We zoom in on a random XR MAC queue and one PDTQ with its sub-queues to further illustrate the concept.

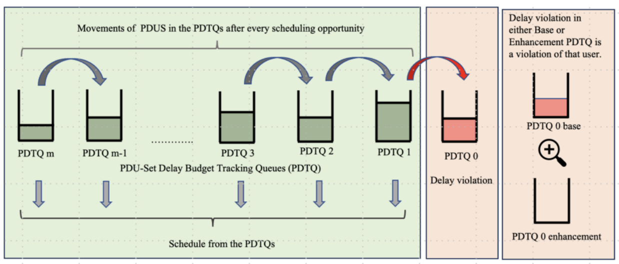

Figure 2. PDU Set movements through the PDTQs: precise delay tracking.

2) Base Layer Constraints: The base layer is subject to constraints governing base PSI level PDU set arrival, service, temporal evolution, and violation detection and measurements as follows:

Base Layer Enqueuing: New base layer PDU sets enter the corresponding base sub-PDTQ:

$$C1 : Q_i^{K_i^b, b}(t+1) = a_i^b(t) f_i^b(t), \quad \forall i, t \tag{2}$$

where $K_i^b = \lfloor D^b / \tau \rfloor$ represents the maximum PDTQ index for base layer PDU sets (i.e., PSI level: *base*: high importance).

Base Layer Queue Evolution: Base sub-PDTQ contents undergo temporal progression with service as follows:

$$C2 : Q_i^{j-1,b}(t+1) = Q_i^{j,b}(t) - \alpha_i(t) x_i^{j,b}(t), \quad \forall i, t, j \in \mathcal{J}_i^b \tag{3}$$

Base Layer Delay-reliability Violation Measurement: It is quantified as ratio of delay-reliability violated to total arrivals:

$$C3 : \gamma_i^b(T) = \frac{\sum_{t=1}^{T} \mathbb{I}(Q_i^{0,b}(t) > 0)}{\sum_{t=1}^{T} a_i^b(t)}, \quad \forall i \tag{4}$$

Base Layer Service Feasibility: Scheduled data cannot exceed corresponding queue size:

$$C4 : \alpha_i(t) x_i^{j,b}(t) \le Q_i^{j,b}(t), \quad \forall i, t, j \in \mathcal{J}_i^b \tag{5}$$

3) Enhancement Layer Constraints: The enhancement layer follows identical constraint structure with distinct parameters:

Enhancement Layer Enqueuing:

$$C5 : Q_i^{K_i^e, e}(t+1) = a_i^e(t) f_i^e(t), \quad \forall i, t \tag{6}$$

where $K_i^e = \lfloor D^e / \tau \rfloor$ represents the maximum PDTQ index for enhancement layer PDU sets.

Enhancement Layer Queue Evolution:

$$C6 : Q_i^{j-1,e}(t+1) = Q_i^{j,e}(t) - \alpha_i(t) x_i^{j,e}(t), \quad \forall i, t, j \in \mathcal{J}_i^e \tag{7}$$

Enhancement Layer Delay-reliability Violation Measurement:

$$C7 : \gamma_i^e(T) = \frac{\sum_{t=1}^{T} \mathbb{I}(Q_i^{0,e}(t) > 0)}{\sum_{t=1}^{T} a_i^e(t)}, \quad \forall i \tag{8}$$

Enhancement Layer Service Feasibility:

$$C8 : \alpha_i(t) x_i^{j,e}(t) \le Q_i^{j,e}(t), \quad \forall i, t, j \in \mathcal{J}_i^e \tag{9}$$

4) System-Wide Constraints: Capacity Constraint: Total allocated resource blocks across all users and layers must not exceed system capacity:

$$C9 : \sum_{i=1}^{N} \sum_{\rho \in \{b,e\}} \sum_{j=1}^{K_i^\rho} x_i^{j,\rho}(t) \le C, \quad \forall t \tag{10}$$

Integer Decision Variable Constraints: Allocated number of resource blocks must be integers:

$$C10 : \quad x_i^{j,b}(t) \in \{0, 1, \dots, C\}, \quad \forall i, t, j \in \mathcal{J}_i^b \tag{11}$$

$$C11 : \quad x_i^{j,e}(t) \in \{0, 1, \dots, C\}, \quad \forall i, t, j \in \mathcal{J}_i^e \tag{12}$$

Channel Parameter Constraint: Size of the data served per RB must be non-negative:

$$C12 : \alpha_i(t) \ge 0, \quad \forall i, t \tag{13}$$

5) Complete Formulation: Combining all these, the complete optimization problem is thus formulated as follows:

$$\boxed{\begin{aligned} \mathcal{O} : \quad & \text{minimize} \sum_{i=1}^{N} \mathbb{I}\left\{ \bigvee_{\rho \in \{b,e\}} \gamma_i^\rho(T) \ge \mu_i^\rho \right\} \\ & \text{subject to:} \quad C1 - C12 \end{aligned}} \tag{14}$$

The integer decision variables, non-linear indicator functions, and their relationships (both layers must be delay-reliability satisfied simultaneously) make this problem a Mixed Integer Non-linear Program (MINLP), which is extremely difficult to solve. We therefore linearize it as follows.

A. Linearization

We now transform the MINLP into a Mixed-Integer Linear Programming (MILP) problem solvable by standard optimization solvers.

1) Linearizing the Objective Function: We introduce binary decision variables to replace non-linear elements:

User Satisfaction Variable: Define $y_i \in \{0,1\}$ where:

$$y_i = \begin{cases} 1, & \text{if user } i \text{ violates any layer} \\ 0, & \text{otherwise} \end{cases} \tag{15}$$

Layer Violation Variables: For each layer $\rho \in \{b, e\}$, define $z_i^\rho \in \{0,1\}$ where:

$$z_i^\rho = \begin{cases} 1, & \text{if } \gamma_i^\rho(T) \geq \mu_i^\rho \\ 0, & \text{otherwise} \end{cases} \tag{16}$$

The objective function thus becomes:

$$O_L : \text{minimize} \sum_{i=1}^{N} y_i \tag{17}$$

To enforce the logical OR relationship (user dissatisfied if either layer fails), we introduce:

$$C13: \quad y_i \geq z_i^b, \quad \forall i \tag{18}$$

$$C14: \quad y_i \geq z_i^e, \quad \forall i \tag{19}$$

$$C15: \quad y_i \leq z_i^b + z_i^e, \quad \forall i \tag{20}$$

2) Linearizing Delay Reliability Constraints: The delay reliability calculation $\gamma_i^\rho(T)$ contains indicator functions and fractions that must be linearized.

Delay-reliability Violation Indicators: We introduce binary variables $p_i^\rho(t) \in \{0,1\}$ to replace $\mathbb{I}(Q_i^{0,\rho}(t) > 0)$:

$$p_i^\rho(t) = \begin{cases} 1, & \text{if } Q_i^{0,\rho}(t) > 0 \\ 0, & \text{otherwise} \end{cases} \tag{21}$$

Using the big-M method with sufficiently large constant M_1:

$$C16: \quad p_i^b(t) \geq \frac{Q_i^{0,b}(t)}{M_1}, \quad \forall i, t \tag{22}$$

$$C17: \quad p_i^b(t) \leq Q_i^{0,b}(t), \quad \forall i, t \tag{23}$$

$$C18: \quad p_i^e(t) \geq \frac{Q_i^{0,e}(t)}{M_1}, \quad \forall i, t \tag{24}$$

$$C19: \quad p_i^e(t) \leq Q_i^{0,e}(t), \quad \forall i, t \tag{25}$$

Fraction Linearization: Introduce auxiliary variables β_i^ρ:

$$C20: \quad \beta_i^b = \frac{1}{\sum_{t=1}^{T} a_i^b(t)}, \quad \forall i \tag{26}$$

$$C21: \quad \beta_i^e = \frac{1}{\sum_{t=1}^{T} a_i^e(t)}, \quad \forall i \tag{27}$$

The delay reliability thus becomes:

$$C22: \quad \gamma_i^b(T) = \beta_i^b \sum_{t=1}^{T} p_i^b(t), \quad \forall i \tag{28}$$

$$C23: \quad \gamma_i^e(T) = \beta_i^e \sum_{t=1}^{T} p_i^e(t), \quad \forall i \tag{29}$$

3) Linking Layer Violations to Binary Variables: To connect $\gamma_i^\rho(T) \geq \mu_i^\rho$ with binary variables z_i^ρ, we use the big-M method with constant M_2:

For base layer:

$$C24: \quad \beta_i^b \sum_{t=1}^{T} p_i^b(t) - \mu_i^b \geq -M_2(1 - z_i^b), \quad \forall i \tag{30}$$

$$C25: \quad \beta_i^b \sum_{t=1}^{T} p_i^b(t) - \mu_i^b \leq M_2 z_i^b, \quad \forall i \tag{31}$$

For enhancement layer:

$$C26: \quad \beta_i^e \sum_{t=1}^{T} p_i^e(t) - \mu_i^e \geq -M_2(1 - z_i^e), \quad \forall i \tag{32}$$

$$C27: \quad \beta_i^e \sum_{t=1}^{T} p_i^e(t) - \mu_i^e \leq M_2 z_i^e, \quad \forall i \tag{33}$$

4) Complete Linearized Formulation: In summary, the complete MILP formulation is as follows:

$$\boxed{\begin{aligned} O_L: \quad & \text{minimize} \sum_{i=1}^{N} y_i \\ \text{subject to:} \quad & C1 - C12 \text{ (Original constraints)} \\ & C13 - C27 \text{ (Linearization constraints)} \\ & y_i, z_i^b, z_i^e, p_i^b(t), p_i^e(t) \in \{0,1\}, \quad \forall i, t \end{aligned}} \tag{34}$$

B. Optimal Solution

The linearized PDU set-aware delay-reliability optimal scheduling formulation is presented in Eqn. 34. This MILP can be solved using standard optimization solvers like CPLEX or Gurobi. However, due to the recursive structure of the equations, they cannot be directly fed into standard solvers. We therefore unroll the recursion over time before inputting it to the Gurobi solver to obtain optimal solutions.

IV. Performance Evaluation

A. Simulation Setup

We implement a 3GPP-compliant system-level custom simulation framework in Python to evaluate the performance of PaDXR. The simulation environment comprises a single gNB serving multiple XR users, following the system parameters specified in Table II. Users are uniformly distributed within the cell coverage area, and each 1-second simulation run is averaged over 25 independent trials with different random seeds. For benchmarking, we compare PaDXR against standard schedulers—*Proportional Fair* (PF), *Round Robin* (RR), and

Table II
Simulation Parameters

Parameter	Value	Parameter	Value
Carrier freq.	3.5 GHz	Channel model	3GPP 38.901 [14]
BS tx power	43 dBm	UE tx power	23 dBm
Data rate (d_r)	30 Mbps	Frame rate (f_r)	60 fps
Numerology (μ)	1	Max bandwidth	100 MHz
PSDB	2.5–10 ms	Target delay-reliability	b: 99%; e: 90%
Mobility model	Stationary users	Mean PDU Set size ($\bar{S}$)	$\frac{d_r \times 10^6}{f_r \times 8}$ [Bytes]
Std. dev. PDU Set size (σ_S)	$0.105 \times \bar{S}$	PDU Set truncation range	$[0.5 \times \bar{S}, 1.5 \times \bar{S}]$
Mean inter-arrival time ($\bar{T}$)	$\frac{1}{f_r}$ [s]	Jitter mean ($\bar{J}$)	0 ms
Jitter std. dev. (σ_J)	2 ms	Cell range (R)	500m

Maximum Rate (MR)—as well as the delay-reliability-aware scheduler DROX [10].

Traffic modeling adopts the 3GPP XR framework [13], where video traffic comprises pseudo-periodic frames. PDU set sizes and arrival jitter are drawn from truncated Gaussian distributions, consistent with 3GPP guidelines. PSI levels (base and enhancement) are generated via a two-state Markov chain, whose transition matrix reflects realistic correlation patterns: base layers switch to enhancement layers with 80% probability, emulating typical codec behavior where base frames serve as keyframes; enhancement layers exhibit moderate persistence (60%) or revert to base layers (40%). Notably, the proposed solution remains agnostic to the Markov model and is compatible with alternative traffic models.

Performance is measured using the metric *Satisfied XR User Count*, defined as the number of users receiving at least 99% of base layer PDU sets and 90% of enhancement layer PDU sets within their respective PSDB. This aligns with the 3GPP definition of XR capacity, which quantifies it as the number of XR users supportable per cell [5].

B. Results and Discussions

We now examine the impact of various parameters on the performance of the proposed PaDXR scheduling scheme in comparison to existing approaches:

1) Effect of Input XR Users: Fig. 3 compares scheduler performance under varying user loads in terms of the number of satisfied XR users. Traditional schedulers (PF, RR, MR) perform poorly across all loads, satisfying at most one user. They lack PDU Set awareness and, more critically, a precise PDTQ mechanism. As a result, significant data arrives after PSDB expiration or is only partially delivered within the PSDB, leading to poor performance. This mismatch underscores their unsuitability for XR traffic and highlights the need for PDU Set-aware scheduling. DROX performs well due to its delay tracking mechanism, confirming that accurate delay tracking is vital for XR scheduling. However, PaDXR consistently outperforms DROX at all loads. The performance gap is small at lower loads (5–10 users) but becomes substantial at higher loads

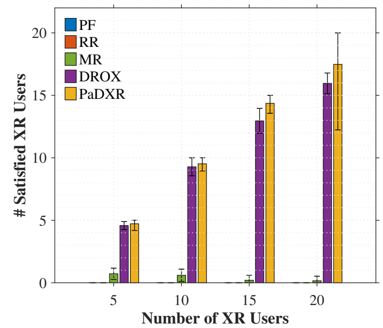

Figure 3. Effect of the user count: d_r = 30 Mbps; f_r = 60 fps; PSDB = 5 ms; target delay-reliability: base PDU sets = 99%, enhancement PDU sets = 90%; f_c = 3.5 GHz; BW = 100 MHz; μ = 1; R = 500 m; mobility: stationary.

(15–20 users), where PaDXR supports 15–20% more users. This gain stems from PaDXR's PSI-aware differentiation. While DROX applies uniform 99% reliability to all PDU sets, PaDXR leverages the layered structure of video streams by applying 99% reliability to base layers and only 90% to enhancement layers. This relaxation mitigates resource over-provisioning, enabling PaDXR to support a larger number of users under the same capacity constraints. At higher load levels, where resources become scarce, the content-aware allocation becomes critical, allowing PaDXR to accommodate users that DROX is unable to serve.

2) Effect of System Bandwidth: Fig. 4 evaluates scheduler performance under varying bandwidth constraints. Traditional schedulers remain ineffective across all bandwidth levels, satisfying fewer than one user even at 100 MHz, due to previously discussed limitations. Both DROX and PaDXR scale their performance with available bandwidth. Under severe constraints (25 MHz), DROX satisfies approximately 3.5 users, while PaDXR reaches 4.7, representing a 34% improvement. This result demonstrates that, as resource scarcity increases, the importance of PSI-aware differentiation becomes more pronounced. At moderate bandwidth (50 MHz), DROX supports 8.7 users, compared to 9.4 for PaDXR, yielding an 8% improvement. At 100 MHz, both converge, satisfying nearly all 10 users, as the abundance of resources eliminates contention. This bandwidth-dependent performance gap highlights a key insight: PSI-aware scheduling provides the greatest benefit under constrained conditions. These findings carry practical implications. In bandwidth-limited environments, such as dense urban deployments or periods of peak usage, PaDXR's PSI-awareness offers significant capacity gains. Conversely, when bandwidth is sufficient, both schedulers achieve comparable performance.

3) Effect of PDU Set Delay Budget: Fig. 5 analyzes PSDB sensitivity with 10 XR users, illustrating how delay bound constraints influence performance. Traditional schedulers remain ineffective across all PSDB values, satisfying fewer than one user even under relaxed 10 ms bounds. In contrast, both DROX and PaDXR scale their performance with PSDB constraints. Under stringent PSDB conditions (2.5 ms), DROX satisfies approximately one user, whereas PaDXR supports two, representing the largest relative gain. At moderate PSDB (5

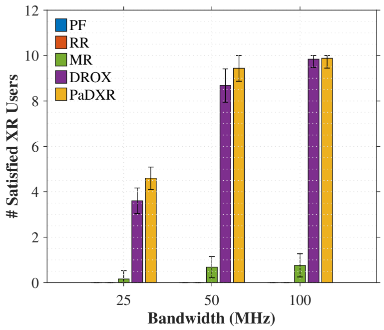

Figure 4. Effect of system bandwidth: N = 10; d_r = 30Mbps; f_r = 60fps; PSDB = 10ms; target delay-reliability: base PDU sets = 99%, enhancement PDU sets = 90%; f_c = 3.5 GHz; μ = 1; R = 500m; Mobility: Stationary.

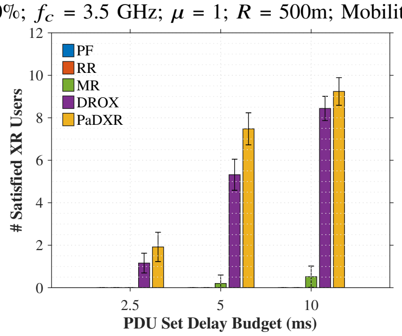

Figure 5. Effect of PSDB: d_r = 30Mbps; f_r = 60fps; target delay-reliability: base PDU sets = 99%, enhancement PDU sets = 90%; f_c = 3.5 GHz; BW = 100MHz; μ = 1; R = 500m; Mobility: Stationary.

ms), DROX serves 5.3 users while PaDXR reaches 7.5, yielding a 42% improvement. At relaxed PSDB (10 ms), DROX satisfies 8.4 users compared to 9.2 for PaDXR, reducing the gap to 10%. These results indicate that PaDXR offers the most significant benefit under tight deadline constraints. At 2.5 ms, the strict timing requirements require aggressive resource allocation. PaDXR's selective reliability, which maintains 99% for base layers and relaxes enhancement layers to 90%, facilitates strategic resource redistribution and enables support for an additional user that DROX cannot accommodate. Overall, PaDXR delivers substantial improvements ranging from 40% to 100% for latency-sensitive XR applications with tight PSDBs (2.5 to 5 ms), such as interactive VR or remote surgery. For delay-tolerant applications (10 ms or higher), both schedulers perform similarly, although PaDXR exhibits a slight advantage.

V. Conclusion and Future Work

This paper addressed the problem of PDU set-aware delay-reliability optimal scheduling for XR traffic in 6G networks. Conventional schedulers are fundamentally unsuited for XR traffic, and existing delay-reliability-aware approaches impose uniform reliability requirements, failing to leverage the PSI parameter standardized in 3GPP Release 18. We proposed PaDXR, introducing the PDTQ mechanism, which maintains separate sub-queues for base and enhancement PDU sets with differentiated reliability targets. The scheduling problem was formulated as a MINLP and subsequently linearized to a MILP for an optimal solution. Extensive simulations show that PaDXR achieves a 15–20% capacity improvement over state-of-the-art *delay-reliable* schedulers under normal conditions and more than 30–40% gains in resource-constrained scenarios. To obtain the optimal solution, we assume perfect knowledge of future arrivals and channel conditions, which is impractical. Additionally, the problem exhibits exponential complexity, limiting scalability. Future work will address these challenges by incorporating prediction mechanisms for arrivals and channel states and by developing scalable heuristic algorithms for practical deployment.

Acknowledgments

This work was funded by Taighde Éireann-Research Ireland via grants 13/RC/2094_2 (Lero) and 21/FFP-P/10244 (FRADIS) and co-funded by the European Union (SyMeCo, Grant 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.

References

[1] C.-X. Wang, X. You, X. Gao, X. Zhu, Z. Li, C. Zhang, H. Wang, Y. Huang, Y. Chen, H. Haas, J. S. Thompson, E. G. Larsson, M. D. Renzo, W. Tong, P. Zhu, X. Shen, H. V. Poor, and L. Hanzo, "On the road to 6g: Visions, requirements, key technologies, and testbeds," *IEEE Communications Surveys Tutorials*, vol. 25, no. 2, pp. 905–974, 2023.

[2] 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*, 2024, pp. 758–763.

[3] 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.

[4] M. Gapeyenko, S. Paris, M. Isomaki, B. Yanakiev, A. Amiri, B. Sébire, J. Kaikkonen, C. Wu, and K. I. Pedersen, "Overview of nr enhancements for extended reality (xr) in 3gpp 5g-advanced," *IEEE Access*, vol. 12, pp. 186 941–186 956, 2024.

[5] 3rd Generation Partnership Project (3GPP), "Study on XR (Extended Reality) Evaluations for NR (Release 17)," 3GPP, 3GPP Technical Report TR 38.838, Dec. 2021, available at: https://www.3gpp.org/ftp/Specs/archive/38_series/38.838/.

[6] 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.

[7] 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)*, 2021, pp. 1–6.

[8] S. Dutta, D. Roy, M. Laha, and G. Das, "Application-aware mac scheduling for xr over epon-based 6g-backhaul," *Optical Fiber Technology*, vol. 98, p. 104545, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1068520025004201

[9] 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.

[10] M. Laha, G. Das, and G.-M. Muntean, "Delay-reliability aware optimal downlink scheduling for extended reality applications in 6g," in *IEEE INFOCOM 2025 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2025, pp. 1–6.

[11] M. Laha, G. Das, and G. Muntean, "Delay-reliability aware downlink scheduling for multi-sensory immersive extended reality in nextg," *IEEE Journal on Selected Areas in Communications*, vol. 44, pp. 3427–3443, 2026.

[12] A. Cedergren, F. de Laval, S. Falahati, L. Falk, D. H. Kang, R. S. Karlsson, Y. Lyazidi, A. Mirzaei, J. Olsson, J. L. Pradas, P. Schliwa-Bertling, N. Shi, B. Singh, R. Tano, and A. M. Voicu, "5g networks evolution for extended reality," *IEEE Communications Standards Magazine*, vol. 8, no. 3, pp. 54–59, 2024.

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

[14] "Study on channel model for frequencies from 0.5 to 100 ghz," in *(3GPP), TR 38.901, V16.1.0*, Jan.2020.