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

Moyukh Laha*, Goutam Das**, Gabriel Miro-Muntean*
*School of Electronic Engineering, Dublin City University, Ireland
**Electronics & Electrical Communication Engineering, Indian Institute of Technology Kharagpur, India
{moyukh.laha, gabriel.muntean}@dcu.ie, gdas@gssst.iitkgp.ac.in

***Abstract*—Extended Reality (XR) stands at the forefront of enabling the Metaverse, promising transformative advancements in human-machine and interpersonal interactions. Achieving seamless XR experiences, however, requires the capabilities of 6G networks, as current 5G solutions fall short of addressing XR's dual demands for enhanced mobile broadband (eMBB) and ultra-reliable low-latency communications (URLLC). Bridging this gap calls for innovative scheduling frameworks tailored to XR's stringent requirements. This paper introduces a delay-reliability-aware optimal downlink scheduling framework for XR services in 6G networks. The proposed approach integrates a novel delay tracking mechanism to optimize the scheduling process, ensuring that a maximum number of XR users meet stringent delay and reliability criteria. Simulation results demonstrate substantial performance gains, with the proposed framework significantly outperforming conventional scheduling techniques, making it a compelling 6G scheduling solution for XR applications.**

***Index Terms*—Extended reality, Metaverse, Delay-reliability aware Scheduling, 6G.**

I. Introduction

Extended Reality (XR), encompassing Virtual Reality (VR), Augmented Reality (AR), and Mixed Reality (MR), represents a transformative technology that merges physical and virtual environments to create immersive digital experiences. XR holds immense promise to revolutionize various industries and serves as a gateway to the Metaverse—a persistent, content-rich, and socially significant digital ecosystem where users can interact, engage, and co-create [1], [2]. This close synergy between XR and the Metaverse has drawn significant interest from leading technology companies, such as Apple, Meta, and Microsoft, as well as from academia due to its market potential and transformative impact.

The adoption of XR, however, hinges on advancements in wireless communication, particularly 6G networks, which are essential for delivering the high-bandwidth, low-latency, and highly reliable connectivity required for immersive experiences [2]–[4]. Although wired solutions exist, their limited mobility and logistical challenges render them impractical for most real-world applications. Consequently, optimizing 6G wireless networks to meet XR's unique requirements is critical for its widespread adoption. While 5G networks address diverse use cases through three core service categories—Enhanced Mobile Broadband (eMBB), Ultra-Reliable Low-Latency Communications (URLLC), and Massive Machine-Type Communications (mMTC)—XR traffic presents a unique challenge. XR applications demand an unprecedented combination of high data rates, ultra-low latency, and stringent reliability, effectively blending eMBB and URLLC requirements. A key metric for XR, termed *delay reliability*, requires 90-99% of XR frames to meet a defined delay threshold [3], [5], [6], necessitating novel network designs and protocols to address this trifecta of demands.

Scheduling plays a pivotal role in meeting XR's stringent delay-reliability requirements. Although 5G scheduling schemes have been widely studied [7], the specific requirements of XR applications remain insufficiently addressed. Existing scheduling mechanisms have been shown to be inadequate for meeting the stringent demands of XR [5], [6]. Many prioritize energy efficiency or rely on impractical assumptions, such as periodic arrivals and uniform packet sizes [8]–[10]. Others address delay or reliability independently but fail to consider them [11], [12] jointly. Machine learning-based and joint scheduling approaches [13], [14] offer adaptability but are not tailored to XR's needs. Consequently, delay-reliability-aware scheduling mechanisms specifically designed for XR are notably absent in the literature.

This paper addresses this gap by introducing a novel Delay Reliability Aware Optimal Scheduling (DROX) framework for XR applications in 6G networks. DROX integrates a novel delay-tracking mechanism with joint delay-reliability optimization, making it uniquely suited for XR. The key contributions are as follows:

- A novel delay-tracking mechanism employing virtual delay-tracking queues at the MAC layer to precisely monitor delay bounds for all users.
- The delay-reliability-aware optimal downlink scheduling problem is formulated as a non-linear integer programming (NLIP) problem and subsequently solved.
- Comprehensive performance evaluations demonstrating the superiority of the proposed framework over popular scheduling schemes.

The remainder of the paper is organized as follows: Section II reviews related work. Section III presents the **D**elay **R**eliability Aware **O**ptimal Scheduling framework for **X**R (DROX), detailing its components, including the delay-tracking mechanism and scheduling optimization. Section IV evaluates the proposed scheme's performance compared to existing methods. Finally, Section V concludes the paper and discusses future research directions.

II. Related Work

Research on scheduling specifically designed for XR applications remains limited. In [8], the authors propose an energy-efficient strategy utilizing Discontinuous Reception (DRX), while [9] employs playout buffers and priority-based classification to minimize power consumption. However, these methods rely on unrealistic traffic assumptions, such as a Poisson arrival process, which fails to capture XR's dynamic traffic patterns. Theoretical advancements include [10], which models multi-user XR scheduling in 6G as a periodic Markov Decision Process (MDP) and approximates it as a nonlinear Knapsack Problem. Although this approach provides valuable theoretical insights, it assumes periodic packet arrivals and uniform packet sizes, limiting its practical applicability. A subsequent extension in [15] incorporates heterogeneous arrivals but retains periodicity assumptions and does not address joint delay-reliability metrics, which are critical for XR traffic. Other works focus on heuristic solutions. For instance, [11] introduces a heuristic bin-covering scheduler that prioritizes Protocol Data Unit (PDU) sets based on delay budgets and data requirements, enhancing XR capacity but lacking real-time delay-tracking mechanisms. Similarly, [12] mitigates inter-cell interference from eMBB users through weighted round-robin (WRR) scheduling, and [16] proposes a Frame-Level Integrated Transmission scheme for XR. Architectural adaptations also contribute to the field. Application-aware MAC enhancements in [17] and AI-assisted service provisioning in [6] improve user satisfaction but do not incorporate delay-reliability-aware scheduling. [18] introduces an XR loopback system utilizing edge servers to dynamically optimize traffic parameters via Adjust Packet Size (APS) and Adjust Frames Per Second (AFPS). While this approach increases adaptability, it lacks robust mechanisms for delay tracking and reliability assessment. In contrast to these prior efforts, our study presents a comprehensive solution tailored specifically for XR scheduling in 6G networks. By integrating a novel delay-tracking mechanism and considering the critical delay-reliability metric, our approach fills a significant gap in the literature on XR scheduling in 6G.

III. DROX: Delay Reliability Aware Optimal Scheduling for XR

We propose a novel downlink scheduling scheme designed to maximize the number of XR users meeting their stringent delay-reliability requirements in 6G networks. XR applications impose unique Quality-of-Service (QoS) constraints, characterized by a predefined delay bound within which their Data Units (DUs)—a generalized term for data packets, frames, or Service Data Units (SDUs)—must be successfully transmitted. Failure to deliver DUs within this delay bound results in delay violations, degrading the immersive user experience. Additionally, each XR user has a reliability threshold, representing the percentage of DUs that must be delivered within the delay bound to ensure an acceptable Quality of Experience (QoE) [3], [5], [6]. Meeting these dual constraints—delay and reliability—for a diverse user base while optimizing network resource utilization presents a significant challenge. The objective of the proposed scheme is to maximize the number of XR users who meet their specific delay-reliability targets.

To achieve this, the scheduling scheme comprises two key components: *Delay Tracking Mechanism*: This mechanism monitors the delay status of each DU in real time, precisely tracking the remaining time to meet the delay bound. This fine-grained delay information is essential for informed and efficient scheduling decisions. *Optimized Downlink Scheduling Strategy*: Using the delay tracking mechanism, this component dynamically prioritizes DUs based on their remaining delay budgets and user-specific reliability thresholds. By integrating these components, the scheme adapts dynamically to real-time network conditions, ensuring optimal resource allocation to satisfy the stringent delay-reliability demands.

A. System Model

Our system model adopts the protocol stack architecture of the 5G protocol stack, as shown in Fig. 1. In this setup, Data Units (DUs) generated by the upper layers are buffered in Medium Access Control (MAC) queues (MACQs) at the gNB, with a separate queue allocated for each XR User Equipment (UE). These queues store the DUs along with their associated delay bounds. Each DU is tagged with a unique UE ID, identifiable through the corresponding MAC address. The proposed DROX scheme operates at the MAC layer of the gNB, as depicted in Fig. 1. We assume XR users are randomly distributed within the coverage area of the gNB. The analysis targets video-based XR traffic, as it constitutes the most bandwidth-intensive and challenging aspect of XR communications. Fully immersive XR systems that integrate multisensory data, while important, fall outside the current scope and will be addressed in future research. To simplify the problem, we assume that each XR user runs a single XR application exclusively within the XR-RAN slice. Therefore, scenarios involving mixed applications, such as eMBB, URLLC, or mMTC traffic, are excluded from this study and left for future exploration. Additionally, the model assumes that future DU arrivals and channel conditions are known to the scheduler, enabling optimal problem-solving. More practical solutions with relaxed assumptions will be explored in subsequent work.

B. Delay Tracking Mechanism

The delay tracking mechanism, a key component of the proposed scheduling scheme, relies on two foundational concepts: *Delay Tracking Queues (DTQs)* and *Delay Tracking Granularity (DTG)*. Upon receipt from the upper layers, Data Units (DUs) are initially stored in user-specific MAC queues (XMQs) and then reorganized into DTQs according to their delay bounds, as shown in Fig. 2. The DTG plays a crucial role in this process, dictating the granularity of delay tracking. It is synchronized with the Transmission Time Interval (TTI) duration, which varies depending on the numerology in use—for example, 1 ms for numerology 0. This synchronization ensures precise and consistent delay tracking across diverse numerology settings.

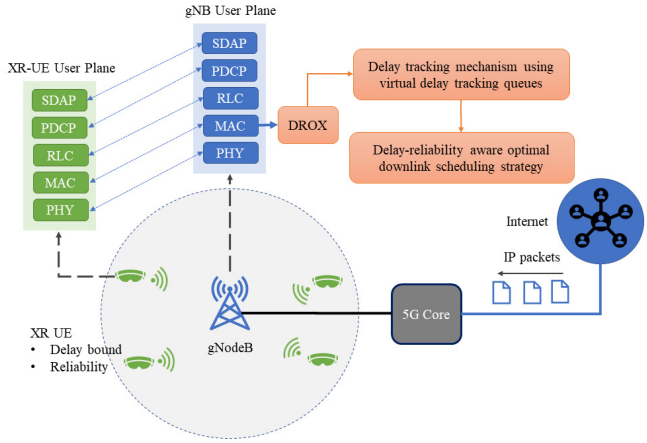

Figure 1. 5G Protocol Stack and System Model

To allocate a DU with a delay bound D_i for user i to the appropriate DTQ, the DTG value (t_d) is utilized. The DTQ index x is computed as:

$$x = \left\lfloor \frac{D_i - t_d}{t_d} \right\rfloor \tag{1}$$

DUs in DTQ x must be scheduled within x scheduling opportunities to avoid exceeding their delay bounds, resulting in a delay violation. Fig. 2 illustrates this mechanism, focusing on the flow from individual MAC queues (e.g., XMQ_k and $XMQ_{k'}$) to DTQs. While DUs for a single user often share similar delay bounds (DBs), the model accommodates diverse DBs and varying DU sizes across users. DUs with identical DBs but differing DU sizes can be grouped into the same DTQ, ensuring flexible and efficient handling.

At each scheduling opportunity, DUs are selected from DTQs for transmission (selection criteria are discussed in the next subsection). If a DTQ is only partially scheduled, the unscheduled DUs have their delay bounds reduced by one scheduling interval, effectively moving them to the next DTQ. For instance, unscheduled DUs in DTQ m are shifted to DTQ $m-1$, with $m-1$ scheduling opportunities remaining to meet their delay requirements. This iterative process ensures that all DUs are reassigned to appropriate DTQs based on their updated delay bounds. DUs that remain in DTQ 1 after a scheduling attempt are deemed to have a delay bound of zero, indicating a delay violation. These DUs are discarded, as they have failed to meet the target delay requirements. This systematic approach ensures that only DUs within their allowable delay bounds are retained for transmission, preserving the integrity of the delay tracking process. The entire mechanism, including the iterative shifting of DUs across DTQs, is depicted in Fig. 3.

C. Optimization Problem Formulation

While the delay tracking mechanism is essential, it must be integrated into the scheduling process to achieve delay-reliability-aware scheduling. We now formulate the optimization problem using this delay-tracking framework.

Objective: The goal is to maximize the number of XR users meeting their delay-reliability targets, equivalent to minimizing users violating these targets. Mathematically, this is expressed as:

$$\text{minimize} \sum_{i=1}^{N} I\{\gamma_i(T) \geq \mu_i\} \tag{2}$$

where $\gamma_i(T)$ is the achieved delay reliability violation for user i over time T, and μ_i is the allowed delay reliability violation threshold.

Constraints: The problem is subject to the following constraints: *DU Enqueuing:* DUs are enqueued into the last DTQ of each user:

$$Q_i^{K_i}(t+1) = a_i(t) f_i(t), \forall i, t \tag{3}$$

where $a_i(t)$ is the DU arrival indicator and $f_i(t)$ is the DU size.

DTQ Updates: Unscheduled DUs shifts among DTQs based on remaining delay bounds:

$$Q_i^{j-1}(t+1) = Q_i^j(t) - \alpha_i(t) x_i^j(t), \forall i, j, t \tag{4}$$

where $\alpha_i(t)$ is the data served per resource block (RB) (channel parameter), and $x_i^j(t)$ is the RB allocation, which is also the decision variable.

Delay Reliability: Delay violations are tracked as:

$$\gamma_i(T) = \frac{\sum_{t=1}^{T} I(Q_i^0(t) > 0)}{\sum_{t=1}^{T} a_i(t)}, \forall i \tag{5}$$

where $Q_i^0(t) > 0$ indicates a delay violation.

System Capacity: The total allocated resources must not exceed system capacity:

$$\sum_{i=1}^{N} \sum_{j=1}^{K_i} x_i^j(t) \leq C, \forall t \tag{6}$$

Queue Size: Scheduled data cannot exceed the queue size:

$$\alpha_i(t) x_i^j(t) \leq Q_i^j(t), \forall i, j, t \tag{7}$$

Combining these, the optimization problem is formalized as:

$$O: \quad \text{minimize} \sum_{i=1}^{N} I\{\gamma_i(T) \geq \mu_i\} \tag{8}$$

Subject to:

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

$$C2: Q_i^{j-1}(t+1) = Q_i^j(t) - \alpha_i(t) x_i^j(t), \forall i, j, t$$

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

$$C4: \sum_{i=1}^{N} \sum_{j=1}^{K_i} x_i^j(t) \leq C, \forall t$$

$$C5: \alpha_i(t) x_i^j(t) \leq Q_i^j(t), \forall i, j, t$$

$$C6: x_i^j(t) \in \{0, 1, 2, ..., C\}, \forall i, j, t$$

The integer nature of the decision variables ($x_i^j(t)$) and the non-linear indicator functions make this a Non-Linear Integer Programming (NLIP) problem, which is challenging to solve.

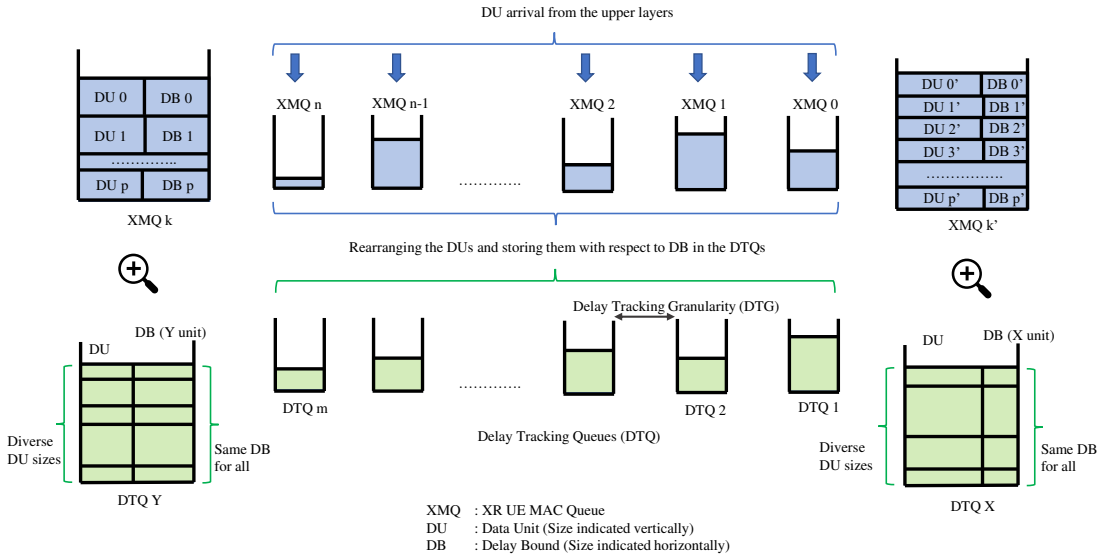

Figure 2. The delay tracking mechanism at the MAC of the gNB. The data units (DU) from the upper layers come to the individual XR user MAC queue, which is then placed to the virtual delay tracking queues (DTQ) based on their Delay bounds. Two random XR MAC queues and DTQs are zoomed in to show their inside contents.

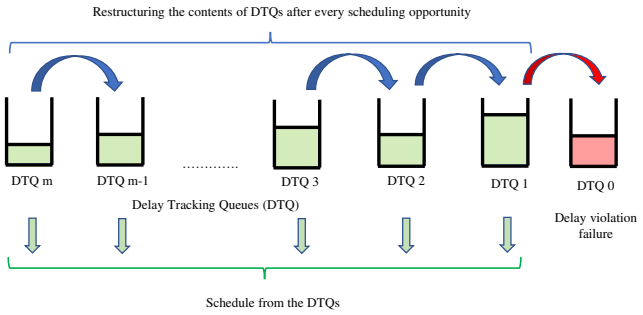

Figure 3. The shifting operations in the DTQs to track the delay bound of the DUs accurately.

D. Linearization for Optimal Solutions

The non-linear nature of the original problem makes it extremely challenging to solve. To address this, we linearize it by introducing binary variables and auxiliary transformations, converting the NLIP into a Mixed-Integer Linear Program (MILP) that is solvable using standard optimization solvers. Key transformations include: The objective $\sum_{i=1}^{N} I\{\gamma_i(T) \geq \mu_i\}$ is replaced by $\sum_{i=1}^{N} y_i$, where y_i indicates whether user i violates its delay-reliability target, using the big-M method. The indicator function $I(Q_i^0(t) > 0)$ is replaced with a binary variable $p_i(t)$, controlled by constraints C5-C6. The fraction in $\gamma_i(T)$ is linearized using an auxiliary variable β_i, with constraints C6-C8 ensuring consistency.

The MILP formulation is as follows:

$$O: \quad \text{minimize} \sum_{i=1}^{N} y_i \tag{9}$$

Subject to:

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

$$C2: \quad Q_i^{j-1}(t+1) = Q_i^j(t) - \alpha_i(t) x_i^j(t), \forall i, j, t$$

$$C3: \quad \sum_{i=1}^{N} \sum_{j=1}^{K_i} x_i^j(t) \leq C, \forall t$$

$$C4: \quad \alpha_i(t) x_i^j(t) \leq Q_i^j(t), \forall i, j, t$$

$$C5: \quad p_i(t) \geq \frac{Q_i^0(t)}{M_1}, \; p_i(t) \leq Q_i^0(t), \forall i, t$$

$$C6: \quad \beta_i = \frac{1}{\sum_{t=1}^{T} a_i(t)}, \; \gamma_i(T) = \beta_i \sum_{t=1}^{T} p_i(t), \forall i$$

$$C7: \quad \beta_i \sum_{t=1}^{T} p_i(t) - \mu_i \geq -M_2(1 - y_i), \forall i$$

$$C8: \quad \beta_i \sum_{t=1}^{T} p_i(t) - \mu_i \leq M_2 y_i, \forall i$$

$$C9: \quad y_i \in \{0, 1\}, \; p_i(t) \in \{0, 1\}, \forall i, t$$

$$C10: \quad x_i^j(t) \in \{0, 1, 2, ..., C\}, \forall i, j, t$$

Here, y_i and $p_i(t)$ represent binary decision variables. This linearization transforms the original NLIP into a mixed-integer linear program (MILP), making it solvable using standard optimization solvers.

E. Optimal Solution

The delay-reliability-aware optimal scheduling problem after linearization is described in Eqn. 9. This is a MILP that can be solved using solvers such as CPLEX or Gurobi. However, these equations are recursive in nature; therefore, we have unwrapped these equations over time and then fed them to the Gurobi solver for optimal solutions.

IV. Performance Evaluation

A. Simulation Environment

We developed a custom 5G NR simulation environment in Python to evaluate the proposed scheme under realistic conditions. The scenario involves multiple XR users served by a gNB, with the scheme operating at the MAC layer (Fig. 1). Simulation parameters adhere to 3GPP specifications, summarized in Table I. XR users are randomly distributed within the coverage area, and each simulation runs for 1 second, with results averaged over 20 iterations using different seeds. The proposed scheme is compared against well-known schedulers: *Proportional Fair (PF)*, *Round Robin (RR)*, and *Maximum Rate (MR)*. Optimal solutions are found using Pyomo framework with the Gurobi solver.

Table I
Simulation and XR Traffic Parameters

Parameter	**Value**	**Parameter**	**Value**
Carrier freq.	3.5 GHz	Channel model	3GPP 38.901 [19]
BS max power	44 dBm	UE tx power	23 dBm
Data rate (d_r)	30 Mbps	Frame rate (f_r)	60 fps
Numerology (μ)	1	Max bandwidth	10 MHz
Delay bound	1–7.5 ms	Delay violations	1–10%
Mobility model	Stationary users	Mean DU size ($\bar{S}$)	$\frac{d_r \times 10^6}{f_r \times 8}$ [Bytes]
Std. dev. DU size (σ_S)	$0.105 \times \bar{S}$	Packet 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

The XR traffic model, based on 3GPP specifications [20], focuses on video traffic represented as pseudo-periodic frames. Frame sizes and jitter are drawn from truncated Gaussian distributions, ensuring realistic variability. Two metrics assess the performance: *Delay-Reliability Satisfied Users:* The number of users meeting their delay-reliability targets. *Delay-Reliability Throughput:* The percentage of data delivered within the specified delay bound, emphasizing usability in XR scenarios. This setup provides a realistic evaluation framework for the proposed scheme.

B. Results and Discussions

We now examine the influence of various parameters on the proposed DROX scheduling scheme compared to existing approaches.

1) Effect of Input XR Users: The impact of the number of input XR users on scheduling performance is depicted in Fig. 4. With parameters set to a data rate (d_r) of 30 Mbps, a frame rate (f_r) of 60 fps, a delay bound of 5 ms, and a target reliability of 95%, the findings reveal distinct trends. For a low number of XR users, existing schemes such as PF, RR, and MR demonstrate reasonable performance. However, as the user count rises, their effectiveness diminishes sharply. For instance, when there are 10 users, PF and RR fail to support any user, while MR supports just one user on average. In contrast, DROX accommodates all input users up to a threshold of 23 users. Beyond this limit, its performance stabilizes or marginally declines. In the reference scenario with a bandwidth of 10 MHz, DROX sustains up to 23 XR users effectively.

A similar pattern emerges for delay-reliable throughput. The average throughput remains near the target delay reliability of 95% for up to 23 users but drops beyond this threshold due to network congestion. These observations highlight DROX's superiority, with the number of supported users exceeding existing schemes by more than 20 times on average and further improvements achievable through bandwidth allocations.

Admission Control: The analysis also points toward implications for admission control mechanisms. Given the system's ability to reliably support up to 23 XR users under the specified parameters, admission control strategies could limit network access to this threshold to maintain performance consistency. Beyond this limit, user experience deteriorates, impacting all connected users. Additional bandwidth would be necessary to accommodate more users.

2) Effect of Delay Bound: The influence of the delay bound (DB) on performance is shown in Fig. 5. Results indicate that increasing the DB enhances DROX's performance. A larger DB allows more flexibility in scheduling data units (DUs) without breaching delay constraints, thereby improving overall efficiency. Despite the resource constraints, DROX supports a significant number of users, even with a stringent delay bound of 1 ms, which can be increased even further by leveraging techniques such as increased bandwidth or higher numerology.

3) Effect of Violation Percentage: The effect of the target violation percentage is presented in Fig. 6. A higher violation percentage corresponds to a less stringent delay reliability target. In such cases, DROX's performance improves as relaxed reliability requirements allow for less strict scheduling decisions, simplifying the problem and enhancing target achievement.

The results affirm the robustness and superiority of the proposed DROX scheme. By outperforming existing methods across various parameters, DROX offers a promising solution for next-generation XR applications.

V. Conclusions

This paper presented a delay-reliability-aware optimal downlink scheduling scheme designed specifically for XR applications in 6G networks. To meet the stringent requirements of XR traffic, we proposed a precise delay tracking mechanism and formulated the scheduling task as an optimization problem, which was subsequently linearized and solved. Through

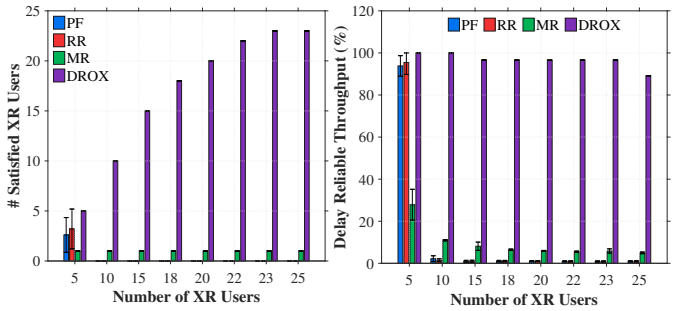

Figure 4. Effect of the number of input XR users on different schemes: d_r = 30Mbps, f_r = 60fps, DB = 5ms, Target Reliability = 95%, f_c = 3.5 GHz, FR1, BW = 10MHz, μ = 1, R = 500m, Mobility: Stationary.

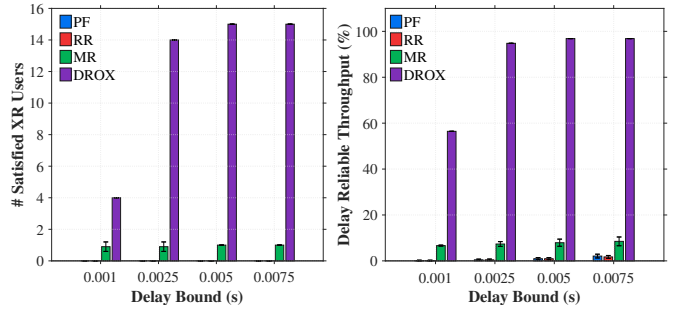

Figure 5. Effect of delay bound on different schemes: N = 15, d_r = 30Mbps, f_r = 60fps, DB = 5ms, Target Reliability = 95%, f_c = 3.5 GHz, FR1, BW = 10MHz, μ = 1, R = 500m, Mobility: Stationary.

extensive simulations, the proposed scheme demonstrated significant superiority over existing approaches, achieving multi-fold improvements in delay reliability. Nonetheless, the inherent scalability limitations of the ILP-based solution pose a challenge. Future work will address this by developing scalable, heuristic-based designs and integrating predictive frameworks for traffic arrivals and channel conditions to enhance practicality and efficiency.

Acknowledgements

This publication has emanated from research jointly funded by Taighde Éireann – Research Ireland under Grant number 13/RC/2094_2, and co-funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

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. [Online]. Available: https://www.frontiersin.org/journals/computer-science/articles/10.3389/fcomp.2022.1068478

[3] "3gpp technical report tr 26.928, "extended reality (xr) in 5g."

[4] T. Taleb, A. Boudi, L. Rosa, L. Cordeiro, T. Theodoropoulos, K. Tserpes, P. Dazzi, A. I. Protopsaltis, and R. Li, "Toward supporting xr services: Architecture and enablers," *IEEE Internet of Things Journal*, vol. 10, no. 4, pp. 3567–3586, 2023.

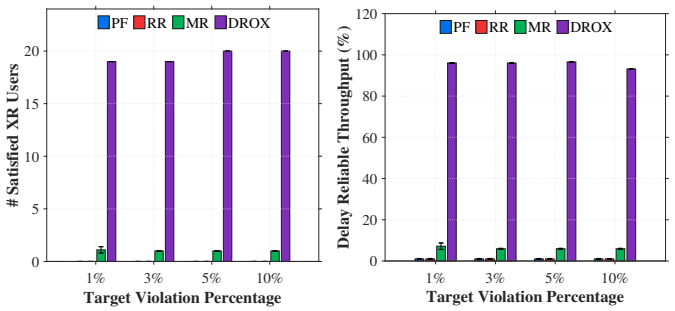

Figure 6. Effect of target violation percentage on different schemes: N = 20, d_r = 30Mbps, f_r = 60fps, DB = 5ms, Target Reliability = 95%, f_c = 3.5 GHz, FR1, BW = 10MHz, μ = 1, R = 500m, Mobility: Stationary.

[5] F. Alriksson, D. H. Kang, C. Phillips, J. L. Pradas, and A. Zaidi, "Xr and 5g: Extended reality at scale with time-critical communication," *Ericsson Technology Review*, vol. 2021, no. 8, pp. 2–13, 2021.

[6] 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.

[7] 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. 51 643–51 661, 2022.

[8] 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)*, 2024, pp. 546–549.

[9] J. Xin, S. Xu, H. Xu, and H. Zhang, "Joint resource allocation and scheduling optimization for xr traffic," in *2023 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, 2023, pp. 1–6.

[10] 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*, 2023, pp. 4074–4079.

[11] 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.

[12] 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. 113 324–113 338, 2023.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] S. Dutta, D. Roy, and G. Das, "Modified split-rendering architecture to enable ai-assisted application-aware mac for xr slice," *IEEE Networking Letters*, pp. 1–1, 2023.

[18] 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.

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

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