Enhancing URLLC Performance in Teleoperation Systems through Dual Prediction and Resource Optimization

Arezoo Ansari

A Thesis

in

The Department

of

Electrical and Computer Engineering

Presented in Partial Fulfillment of the Requirements

For the Degree of

Master of Applied Science (Electrical Engineering) at

Concordia University

Montréal, Québec, Canada

March 2026

© Arezoo Ansari , 2026

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Arezoo Ansari**

Entitled: **Enhancing URLLC Performance in Teleoperation Systems through Dual Prediction and Resource Optimization**

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science (Electrical Engineering)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

_____ Chair
Dr. Wei-ping Zhu

_____ Examiner
Dr. Dongyu Qiu

_____ Supervisor
Dr. Jun Cai

Approved by _____
Wahab (Abdelwahab) Hamou-Lhadj, Chair
Department of Electrical and Computer Engineering

_____ 2026

_____
Amir Asif, Dean
Faculty of Engineering and Computer Science

Abstract

Enhancing URLLC Performance in Teleoperation Systems through Dual Prediction and Resource Optimization

Arezoo Ansari

Ultra-reliable and low-latency communication (URLLC) is a key enabler for mission-critical applications in next-generation wireless networks. However, simultaneously achieving ultra-high reliability and ultra-low latency remains a major challenge. In this work, we propose a novel Dual Prediction Scheme (DPS) for URLLC-based teleoperation systems, where predictive algorithms are deployed at both the transmitter and the receiver to jointly mitigate latency and enhance reliability. In the proposed framework, the receiver reconstructs delayed or lost packets through local prediction, while the transmitter proactively encapsulates multiple predicted future states into a single short packet to safeguard against consecutive losses. To evaluate system reliability and energy efficiency, we formulate a joint optimization problem that minimizes the average transmit power subject to URLLC constraints by jointly optimizing bandwidth allocation and prediction horizons. The resulting problem is non-convex and is efficiently solved via an iterative algorithm. Simulation results verify that the proposed DPS significantly reduces transmit power while satisfying URLLC requirements, demonstrating its strong potential for real-time and energy-efficient wireless teleoperation systems.

Acknowledgments

To my supervisor for his guidance, my colleagues for their camaraderie, my husband and family for their unwavering support, and my newborn son, who gave this journey deeper meaning.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Ultra-reliable and low-latency communication (URLLC) is a key service category in 5G and beyond, targeting mission-critical applications that require extremely small end-to-end (E2E) latency and very high reliability. Prominent examples include cooperative automated driving, industrial automation, smart grids, and particularly Tactile Internet (TI) applications such as remote robotic surgery and immersive teleoperation [3–5]. These applications demand packet delivery within a few milliseconds and error probabilities as low as 10^{-5}–10^{-9}.

TI-based teleoperation forms a closed human–machine control loop [2], in which remote sensory information must reach the operator with minimal delay, and the operator's haptic and control responses must be transmitted back to the teleoperator robot just as quickly. Even small latency violations can degrade stability, reduce transparency, and compromise safety. Although 5G introduces mechanisms that reduce individual delay components, elements such as processing delay, queuing delay under variable traffic, and geographical propagation limits cannot be eliminated in practical systems. These intrinsic constraints create a bottleneck in meeting TI-grade E2E latency and motivate the search for approaches beyond conventional communication optimization.

Prediction has therefore emerged as a key enabler for TI-URLLC. Rather than reducing all delay components physically, prediction compensates for latency by forecasting future commands or states. Predicted samples can mask part of the delay, maintain the continuity of the haptic and control loop, and reduce the impact of packet losses or jitter. Consequently, prediction becomes a central design tool for meeting URLLC requirements in teleoperation. Moreover, prediction can be

most effective when applied jointly at both the transmitter and the receiver, as each side compensates for different types of delay and packet loss.

Building on these insights, this thesis develops a prediction-assisted TI teleoperation framework under URLLC constraints. We consider a multi-user system where each operator–teleoperator pair uses dual-side prediction. Incorporating finite-blocklength coding and queuing-delay principles, we model the experienced delay and reliability when prediction is employed at both ends. We then develop an optimization framework that jointly allocates bandwidth and selects prediction horizons to minimize transmit power while satisfying strict delay and reliability constraints.

1.1 Target Application Scenarios and Practical Motivation for Power Control

The proposed dual-prediction and resource-allocation framework is motivated by TI/URLLC teleoperation deployments in which short packets, delay spikes, and intermittent losses can directly affect control-loop stability and user experience. Representative application scenarios include:

- **Remote robotic surgery (telesurgery):** stringent delay and reliability are required to preserve safety and transparency of the haptic/control loop.
- **Industrial telerobotics in smart factories:** remote manipulation for inspection, assembly, and maintenance, often with multiple simultaneous operator–robot sessions sharing limited wireless resources.
- **Remote operation of heavy machinery (mining, construction, ports):** safety-critical motion control where jitter and packet losses can degrade stability and operator performance.
- **Search-and-rescue teleoperated robots (UGV/UAV):** harsh propagation and congested networks lead to deep fades and bursty losses, making prediction essential to maintain continuity during short outages.
- **Tele-rehabilitation and haptic training:** missing or delayed haptic samples degrade perceived quality, motivating receiver-side reconstruction while controlling resource usage.

Practical motivation for power control. In these scenarios, increasing transmit power is not always feasible due to battery-limited platforms (e.g., mobile robots, UAVs, and wearable haptic devices), thermal constraints, and regulatory/hardware limits. Moreover, in dense deployments and spectrum-reuse settings (e.g., imperfect isolation and reuse across nearby devices/cells), excessive transmit power can increase the interference footprint, which may degrade reliability and experienced delay for other concurrent services. Therefore, minimizing transmit power in this thesis is motivated not only by energy efficiency, but also by practical interference-aware operation that facilitates the coexistence of multiple teleoperation pairs under stringent URLLC constraints.

Motivated by the above application scenarios, the contributions of this thesis form the conceptual and analytical foundation for prediction-aware teleoperation in future 5G/6G networks.

1.2 Motivation and Research Questions

Building on the above discussion, URLLC has been extensively studied using a variety of techniques that target different components of E2E delay. At the physical and MAC layers, 5G introduces mini-slot transmission and flexible numerology to shorten the transmission time interval (TTI) [6]. Finite blocklength coding has been leveraged to characterize and reduce transmission delay while guaranteeing a target block error probability [7–9]. Queuing delay has also been identified as a critical bottleneck, and adaptive blocklength schemes have been proposed to jointly balance transmission and queuing delays [9]. Other solutions, such as relay-assisted transmission, have been explored to further reduce communication delay and enhance reliability [10].

Despite these advances, meeting stringent URLLC requirements in practical networks remains challenging. Certain delay components such as coding delays, processing delays, and backhaul/core-network latency cannot be arbitrarily reduced. Moreover, the finite speed of light fundamentally limits the maximum separation between operator and robot in real-time teleoperation. These physical and architectural constraints make it difficult to achieve TI-grade E2E latency while simultaneously guaranteeing very high reliability.

These limitations have motivated the integration of prediction and edge intelligence into URLLC

systems. Prediction can virtually "compensate" for latency by forecasting future states of the controlled system and thus reducing the experienced delay. For example, [11] proposes a joint prediction–communication framework that optimizes transmitter-side prediction alongside frequency-resource allocation. However, this approach relies solely on transmitter-side prediction and cannot recover from packet losses or deep fades at the receiver. Conversely, receiver-side prediction has been used to synchronize physical devices and digital twins in TI/metaverse scenarios [12], but it does not explicitly enforce URLLC constraints and assumes fixed or known delays, which limits its applicability in dynamic wireless environments.

Prediction can be incorporated at either the transmitter or the receiver side. Transmitter-side prediction is typically more accurate because it relies on real, up-to-date data, and the prediction model can be continuously updated to reflect changes in the operator's behavior or system dynamics. However, if a predicted packet is lost or severely delayed, the receiver has no means to maintain continuity. Receiver-side prediction addresses this limitation by generating predicted samples whenever a packet does not arrive within the required delay threshold, with an adaptively adjusted prediction horizon.

These complementary properties motivate a dual prediction strategy in which both the transmitter and receiver are equipped with synchronized predictors. Such a scheme allows the transmitter to reduce delay and transmission frequency while enabling the receiver to reconstruct missing or delayed packets, thereby improving continuity and robustness.

Dual-side prediction has been explored in wireless sensor networks (WSNs) to reduce data transmission and energy consumption [13, 14], and packetized predictive control (PPC) has been used in networked control systems to improve robustness over unreliable links [15–17]. However, these works (i) do not consider the extreme delay–reliability requirements of URLLC, (ii) do not consider the human–robot interaction in TI teleoperation systems, and (iii) often treat wireless resource consumption in isolation, without capturing multi-user resource coupling and bandwidth allocation. Consequently, there is no unified framework that jointly models dual-side prediction, URLLC-grade reliability, and wireless resource allocation for multi-user TI teleoperation systems.

Furthermore, since multiple teleoperation pairs may coexist, power control is also essential to limit the interference footprint in multi-user scenario settings while meeting stringent URLLC

targets.

In this context, the central challenge of this thesis can be stated as follows:

> *How can we jointly exploit prediction at both the master and slave sides of a TI-enabled teleoperation system, together with wireless resource allocation, to minimize transmit power while satisfying stringent URLLC constraints on experienced delay and reliability for multiple teleoperation pairs?*

To address this challenge and close the identified gaps, this thesis is guided by the following research questions:

- **RQ1: Experienced-delay modelling.** How can we model the E2E experienced delay of a TI-enabled teleoperation system, combining transmission, queuing, and core delays with the delay-mitigation effect of prediction at both the master and slave sides?
- **RQ2: Reliability under finite blocklength.** How can finite blocklength coding be used to jointly optimize transmit power and bandwidth for each teleoperation pair under URLLC reliability constraints?
- **RQ3: Role of dual-side prediction.** How do the transmitter and receiver prediction horizons jointly affect experienced delay, reliability, and robustness to packet losses, and why can dual prediction offer advantages over transmitter-only or receiver-only prediction?
- **RQ4: Power and resource optimization.** Given the above models, how can we formulate and solve an optimization problem that minimizes transmit power subject to (i) experienced-delay constraints, (ii) overall system reliability constraints, and (iii) total bandwidth limitations across multiple teleoperation pairs?

These research questions form the bridge from existing URLLC and prediction-based approaches to the dual prediction scheme and optimization framework developed in this thesis.

1.3 Research Contributions

In summary, the main contributions of this work are:

- We establish a dual prediction scheme in a TI-based teleoperation system to enhance reliability, reduce experienced delay, and achieve URLLC.
- We implement prediction at the receiver side to recover lost or delayed packets, while the transmitter predictor contributes to both reliability and delay improvement, thereby enhancing the receiver predictor's performance.
- We jointly optimize bandwidth allocation and prediction horizons to minimize the average transmit power across all devices.
- We propose an algorithm to find a near-optimal solution to the formulated optimization problem, focusing on power control while satisfying URLLC constraints for all users.
- With extensive simulation, we show that our proposed dual prediction scheme significantly outperforms other existing benchmarks in terms of reliability and delay with limited available bandwidth and power.

1.4 Thesis Outline

This thesis comprises six chapters. Chapter 1 introduces the research problem, motivation, and contributions. Chapter 2 reviews background and related work. Chapter 3 presents the system model and the proposed dual-prediction scheme (DPS) and develops the reliability components with their mathematical formulations. Chapter 4 formulates an optimization problem to minimize average power under ultra-reliable low-latency communication (URLLC) constraints and provides an algorithm in order to solve the optimization problem. Chapter 5 reports comprehensive simulation results. Finally, Chapter 6 concludes the thesis and outlines avenues for future work.

Chapter 2

Background and Literature Review

2.1 URLLC in 5G

Fifth-generation (5G) wireless systems represent a major evolution beyond 4G. Rather than focusing solely on human-centric voice and data services, 5G is designed to support massive connectivity for machines and devices, along with stringent requirements on latency and reliability for mission-critical applications. It can handle very high data rates and connect many different types of devices. Because of this, 5G must meet stricter needs for speed, delay, and reliability. For example, smart meters need very reliable connections but only small amounts of data, while 5G video streaming needs very high data rates but can accept a bit less reliability [1].

To meet the diverse requirements of different 5G use cases and verticals, the International Telecommunication Union (ITU) has defined three main service classes for 5G systems [1]:

- **enhanced Mobile Broadband (eMBB):** Supports high–data-rate applications such as video streaming, web browsing, video conferencing, and virtual/augmented reality.
- **massive Machine-Type Communication (mMTC):** Targets a massive number of Internet of Things (IoT) devices that are typically low-power, sporadically active, and send small data packets.
- **Ultra-Reliable Low-Latency Communication (URLLC):** Enables mission-critical applications that require extremely high reliability and very low delay, such as the tactile internet,

autonomous driving, factory automation, and remote robotic control.

The relationships among the three 5G service classes and their associated applications, as specified in the ITU IMT-2020 framework, are illustrated in Fig. 2.1 . Although there are overlaps among these service classes, URLLC systems are characterized by particularly stringent latency and reliability requirements.

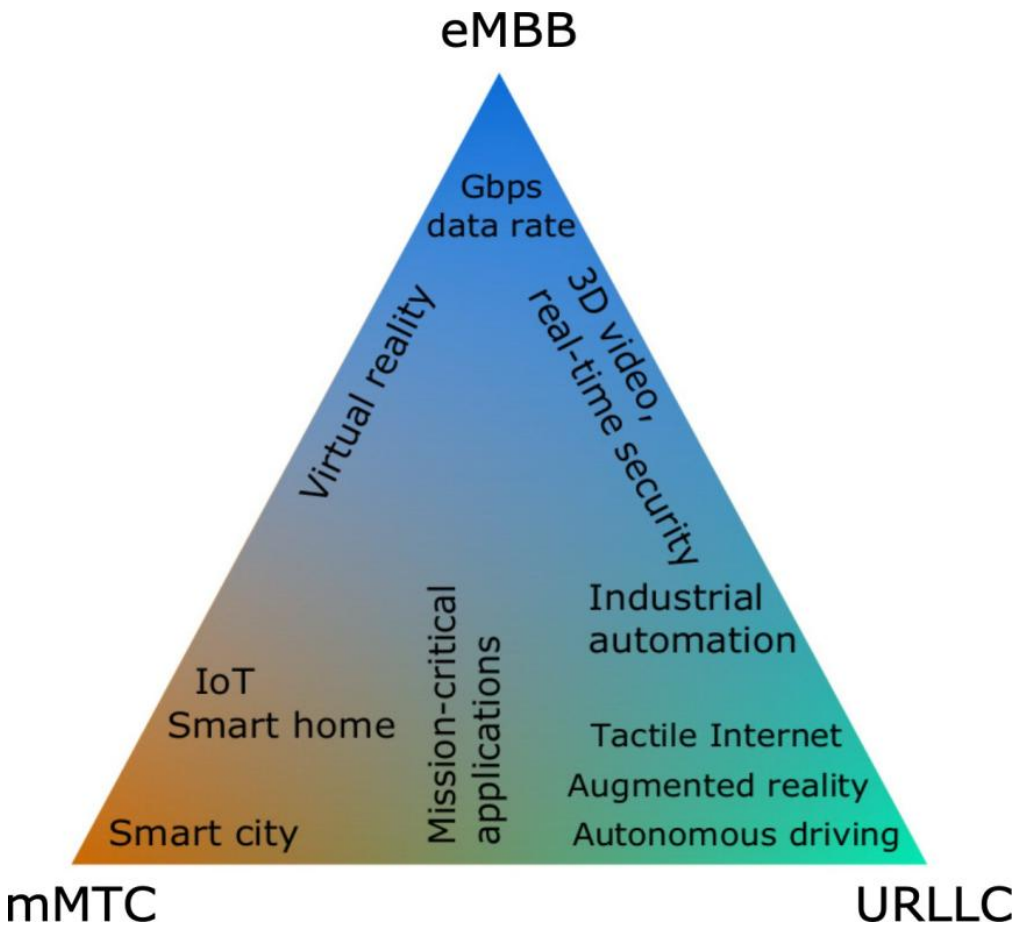

Figure 2.1: ITU IMT2020 use case depicting 3 different service classes for 5G [1].

In the context of URLLC, reliability is defined in [18] as:

> "The percentage value of the amount of sent packets/messages successfully delivered to a given node within the time constraint required by the targeted service, divided by the total number of sent packets/messages."

Similarly, latency is defined in [18] as:

> "The time that it takes to transfer a given piece of information from a source endpoint device to a destination endpoint device, measured at the application service access points, from the moment it is transmitted by the source endpoint device to the moment it is successfully received at the destination endpoint device."

On the standardization side, 3GPP has defined 5G New Radio (NR), where Releases 16–18 detail the URLLC requirements and architectures [19]. Early 5G deployments mainly support eMBB, IoT, mission-critical control, and fixed wireless access, with only limited realization of full mMTC and URLLC capabilities.

Several characteristics make URLLC fundamentally different from traditional system architectures. A key distinction is the packet structure: URLLC typically relies on very short packets to meet ultra-low E2E delay requirements, often on the order of less than 1 ms. At the same time, it must ensure that packets are received correctly with extremely high success probabilities, typically in the range of $(1 - 10^{-5})$ to $(1 - 10^{-9})$. These stringent latency and reliability constraints are among the most challenging aspects of 5G network design. While meeting such requirements at the link layer can be relatively manageable, especially in small-area deployments, achieving them at the network layer over wide-area networks, such as those supporting remote surgery, remains extremely difficult [20, 21].

In wide-area scenarios, the overall latency is composed of multiple components, including uplink and downlink transmission delays, coding and processing delays, queuing delays, and routing delays in the backhaul and core network [22]. Although each individual delay component can often be reduced in isolation, the cumulative effect of all these delays makes URLLC design particularly challenging, as it requires carefully optimizing all contributions to keep the total E2E delay below the required ultra-low latency threshold.

New techniques must be specifically designed for URLLC. In the 3GPP NR standards, latency is treated as the top-priority requirement for URLLC [23]. Reliability, while still essential, is considered secondary because existing tools such as channel coding and space, antenna, and frequency diversity can already provide very high reliability [24]. As mentioned earlier, the most direct way

to reduce latency is to shorten the packet length to just a few bytes (e.g., around 20 bytes or even less) [22]. However, such short packets severely limit the coding gain that can be achieved with traditional error-correcting codes [25]. On the other hand, improving reliability usually requires sending redundant information, either by retransmitting the same packet or by adding extra parity bits for error detection and correction, which increases latency. This trade-off between low latency and high reliability shows that incremental tweaks are not enough; a new design paradigm is needed to fully realize the potential of URLLC.

2.2 Tactile Internet-Based Teleoperation Systems

Tactile Internet (TI) was first introduced by G. Fettweis in 2014 [26]. TI goes beyond the Internet of Things (IoT) and targets latency-stringent applications such as tele-healthcare, smart industry, drone surveillance, and connected cars. While IoT is well suited for machine-to-machine (M2M) or machine-type communication (MTC), it is not designed for human-to-machine (H2M) interaction over high-latency wireless networks. IoT mainly carries audio-visual data, whereas TI is envisioned to convey human skills and actions over the network thanks to its URLLC capabilities. Among the different TI use cases, teleoperation systems are regarded as one of the most prominent, as they directly couple a human operator with a remote physical or virtual environment.

In a TI-enabled teleoperation system, an expert operator interacts with a remote environment through a wireless network and a robotic platform. The system typically adopts a master–slave architecture, where the master domain employs a human–system interface (HSI) to control a robot at the remote site. As illustrated in Fig. 2.2, this TI-enabled teleoperation architecture comprises three main domains: the master domain, the controlled domain, and the network domain.

A. Master Domain

The master domain comprises the human operator and the HSI (master robot), which integrates haptic devices, a video console, and audio equipment to provide multimodal feedback. The HSI converts human actions into control and tactile commands using suitable haptic coding techniques, enabling the operator to touch, feel, and manipulate objects in real or virtual environments while perceiving the remote scene through coordinated visual, auditory, and haptic cues. In multi-user

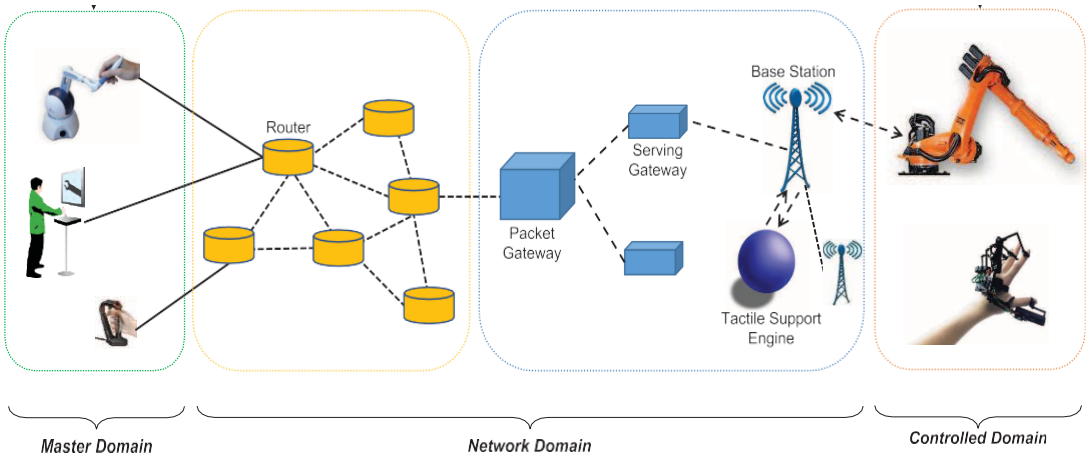

Figure 2.2: Architecture of a TI-enabled teleoperation system [2].

teleoperation scenarios, multiple operators may collaboratively control a single controlled domain. The combination of haptic, auditory, and visual feedback substantially enhances perception, as the human brain naturally fuses multiple sensory modalities [27].

State-of-the-art haptic devices, such as those from Geomagic and Sensable, are typically linkage-based robotic arms attached to a stylus, capable of tracking position and exerting forces at the tip. For future TI teleoperation systems, further advances are needed, particularly in increasing the degrees of freedom (DoF), improving transparency and stability, and embedding network interfaces for direct or indirect communication with cellular and edge networks.

B. Slave (Controlled) Domain

The controlled domain contains the teleoperator (slave robot), which is driven by command signals from the master domain and interacts with objects in a typically unknown remote environment. It is equipped with a high-definition 3D camera, a high-quality microphone, and tactile/force sensors to capture rich visual, auditory, and haptic information. For example, in telesurgery the surgeon does not physically touch the patient; instead, they control a robotic arm at the patient site through the HSI while monitoring the operation in real time [28]. Through the continuous exchange of command and feedback signals, information and energy flow between the master and controlled domains, forming a global control loop. In TI teleoperation systems, this loop must remain stable and transparent despite time-varying network delays and packet losses, so that the operator experiences

the remote environment as if interacting locally.

C. Network Domain

The network domain provides the medium for bilateral communication between the master and controlled domains, thereby kinesthetically coupling the human operator to the remote environment. Ideally, the operator should feel fully immersed in this remote environment. To enable real-time haptic interaction in teleoperation, TI requires ultra-reliable and ultra-responsive connectivity with very stringent latency and reliability targets.

The underlying 5G-based communication architecture, comprising the Radio Access Network (RAN) and the Core Network (CN), is expected to support these requirements. In TI-enabled teleoperation systems, the key functions of the 5G RAN include: i) efficient support of various radio access technologies (RATs), such as traditional cellular, millimeter-wave, massive MIMO, and full-duplex; ii) tactile quality of experience (QoE) and quality of service (QoS) scheduling and radio resource management for haptic services coexisting with other verticals (e.g., M2M, vehicle-to-vehicle, smart grids); iii) efficient packet delivery through reliable radio protocols and robust physical (PHY) layer design; and iv) effective resolution of air-interface contention via advanced medium access control (MAC) techniques.

The main functionalities of the 5G CN relevant to TI teleoperation are: i) dynamic, application-aware QoS provisioning; ii) support for edge-cloud access and computation offloading to reduce end-to-end latency; and iii) robust security mechanisms to protect sensitive haptic and visual data.

In TI teleoperation systems, both content (e.g., video, audio) and skillset data (e.g., motion trajectories, control policies) are transported over a high-performance 5G core network and next-generation Internet. While advances in hardware, protocols, and architectures are crucial to shrinking end-to-end delay, the ultimate physical limit is imposed by the finite speed of light. To push teleoperation beyond this limit, TI leverages predictive edge artificial intelligence (AI) engines that are cached and executed in real time close to the tactile endpoints. The most critical content to be stored and processed are AI models that predict the haptic/tactile experience, i.e., the future motion on one end and the corresponding force feedback on the other. This prediction capability allows the active and reactive ends of the teleoperation system to be spatially decoupled, since the tactile experience is virtually emulated at either side. As a result, TI-enabled teleoperation can support a

much wider geographic separation between the two ends, going beyond the classical 1 ms-at-speed-of-light limit while maintaining a high quality of haptic interaction.

Taken together, these architectural studies clarify the functional roles of the master, slave, and network domains and outline the enabling 5G mechanisms for TI teleoperation. However, prediction and edge AI are usually discussed as high-level enablers. Existing architectural works do not provide a framework for designing and evaluating prediction for teleoperation systems under strict URLLC constraints. This notion leaves a gap that this thesis aims to address.

2.3 URLLC in Teleoperation Systems

URLLC plays a central role in enabling high-performance teleoperation systems where a human operator interacts with a remote robot through haptic, visual, and control signals. The authors in [29] investigate the performance of URLLC in real robotic teleoperation systems using an experimental prototype. They identify three major teleoperation modes, known as supervisory, unilateral, and bilateral control, each with distinct communication requirements. Bilateral teleoperation, which supports real-time haptic interaction, is the most demanding and requires round-trip E2E guarantees with extremely stringent latency and reliability. In such systems, sensor-to-human links must provide high data rates for visual and haptic feedback, while controller-to-actuator links rely on URLLC to maintain stable motion control. The authors further analyze how communication latency affects control transparency, which is defined as the extent to which the operator perceives the remote environment's actual impedance. Their results show that even small delays break passivity and degrade transparency, and that perfect transparency is unattainable in practical networks due to inevitable latency. Moreover, they demonstrate that small latency fluctuations (jitter) can significantly impact the operator's haptic perception.

The work in [30] extends the teleoperation architecture toward 6G by introducing a multi-connectivity platform that simultaneously leverages several communication interfaces, such as 5G, Wi-Fi 6, and Ethernet. This design enhances reliability through redundancy and link diversity, enabling seamless switching when wireless links experience fading or interference. In addition, the platform integrates edge computing and intelligent control mechanisms to achieve and compensate

for delay variations, resulting in more stable haptic interaction and higher transparency compared to single-link URLLC systems.

The work in [31] examines URLLC requirements from the perspective of control stability in bilateral teleoperation systems. The authors focus on achieving URLLC to maintain stable force and motion interaction between the human operator and the remote robot. In the system architecture, the teleoperation loop relies on a continuous exchange of position, velocity, and force signals between master and slave devices, which makes the communication link a critical component of the control system. The study highlights that unpredictable delay variations in wireless networks can destabilize the control loop, especially when transmitting high-frequency haptic data. To address this, the authors propose a fuzzy sliding mode control (FSMC) strategy that enhances robustness against latency and packet delay variation while ensuring smooth force tracking. They show that their approach reduce tracking error and improve stability under varying wireless delay conditions. The authors show that stable teleoperation over wireless networks requires both stringent URLLC guarantees and delay-aware controller design. Their work emphasizes that communication and control must be co-optimized to meet the requirements of real-time haptic interaction.

The work in [32] investigates how URLLC can be supported in telerobotics through 5G network slicing and Lyapunov-based resource optimization. In their system model, multiple telerobotic operations coexist with bandwidth-hungry eMBB users, sharing the same wireless infrastructure. The authors emphasize that telerobotics requires strict URLLC guarantees for control commands, which is characterized by ultra-low latency, high reliability, and stable timing, while high-rate video feedback is delivered over an eMBB slice. A key insight from the system model is that variations in wireless communication delay directly cause robot tracking drift, since user commands and state feedback may not be synchronized with the robot's control cycle. To address this, the authors formulate a Lyapunov-optimized joint communication–control problem that minimizes tracking error while simultaneously maximizing throughput for eMBB users. Their results show that dynamic 5G slicing can isolate URLLC and eMBB traffic, ensuring that URLLC users consistently meet delay constraints even under varying channel conditions and fluctuating user density. They show that increasing the number of users reduces the data rate per user, but the URLLC slice maintains stable performance by allocating the minimal required radio resources.

The survey in [33] provides a broad taxonomy of URLLC architectures and enabling technologies for 6G-enabled industrial systems, several of which directly relate to teleoperation. The authors highlight that teleoperation and Tactile Internet applications represent some of the most demanding URLLC use cases to maintain stable human–robot interaction. In the architectural overview, 6G URLLC systems are expected to incorporate distributed edge intelligence, multi-connectivity, and joint communication–control loops to support real-time haptic feedback and remote actuation. The paper emphasizes that even sub-millisecond delays or small jitter variations can destabilize teleoperation systems, especially when high-frequency motion or force feedback is involved. This reinforces the need for deterministic communication with bounded delay. The authors further identify several technologies crucial for teleoperation, including multi-layer redundancy, edge computing, predictive control, and AI-driven channel prediction, all of which help mitigate delay and packet loss in dynamic industrial environments. Their taxonomy classifies URLLC service requirements across sensing, control, and actuation loops, showing that teleoperation uniquely spans all three categories and thus demands cross-domain resource coordination. Additionally, this work emphasizes that industrial telerobotics cannot rely solely on higher bandwidth; instead, stability depends on synchronized feedback loops and latency-aware resource management. Overall, this survey reinforces that teleoperation represents one of the most stringent URLLC applications in 6G and requires advanced architectural support, including predictive intelligence, multi-connectivity, and tight control–communication integration.

Overall, these studies establish teleoperation as one of the most demanding URLLC use cases and demonstrate that latency, jitter, and reliability directly affect stability, transparency, and tracking performance. However, a new approach is still needed to cope with latency and its fluctuations so that the teleoperation loop remains stable and transparency is improved. In this thesis, we propose a prediction-based scheme that can virtually reduce the experienced latency to (almost) zero by forecasting future motion and force signals, thereby preserving stability while significantly enhancing transparency.

2.4 Communications in URLLC

A number of methods have been proposed to reduce latency and enhance reliability in communication systems, including information theoretic limits, physical and MAC layer mechanisms, and cross layer design. In particular, [34] laid the foundation for URLLC analysis by introducing finite-blocklength, which characterizes the maximum coding rate for a given blocklength and error probability.

Moreover, ultra-mini slot transmission (UMST) has been proposed as a novel low-latency scheme that is particularly suitable for short-packet transmissions in URLLC scenarios [35]. Unlike traditional slots of 14 OFDM symbols, mini-slots allow flexible scheduling with as few as 1–2 OFDM symbols, enabling much shorter uplink/downlink turnaround times. The work in [36] examines the challenges of short-packet transmission in mini-slot-assisted URLLC, highlighting the tension between channel estimation accuracy, signaling overhead, and latency. In 5G and 6G NR, mini-slots consisting of 2, 4, or 7 OFDM symbols enable rapid scheduling without waiting for slot boundaries. However, coherent detection in such short packets requires pilots for channel estimation, and the pilot patterns defined in 3GPP Release 18 can consume up to 25% of the packet payload in a 2-symbol mini-slot. This pilot overhead increases latency and reduces spectral efficiency. Reducing pilots leads to inaccurate channel estimation, especially under mobility, which severely degrades the reliability of coherent detection. To overcome this limitation, the authors propose integrating differential modulation (DM) with standard coherent detection, which forms an adaptive transmission strategy for mini-slot-based short-packet URLLC. Frequency-domain differential OFDM (FDDi-OFDM) and time-domain differential OFDM (TDDi-OFDM) are presented as pilot-free alternatives that avoid channel estimation overhead and enable low-complexity, non-coherent decoding. Using finite blocklength information-theoretic tools, the authors derive closed-form BLER approximations for both coherent and differential detection and reveal that differential schemes outperform pilot-assisted coherent schemes under high mobility, whereas coherent detection is superior in low-mobility or slow-fading scenarios. This is because DM relies on symbol-to-symbol correlation and is thus robust to rapidly time-varying channels, while coherent detection suffers from outdated pilots. The paper concludes that adaptive switching between these two modes enables near-optimal

BLER–latency performance. This work demonstrates that efficient URLLC communication requires jointly optimizing pilot overhead, detection architecture, and mini-slot structure to maintain reliability and minimize latency in short-packet transmissions.

Complementing the above studies on mini-slot-assisted URLLC, the authors of [37] further investigate frequency-domain differential modulation (FD-DM) as a pilot-free signaling technique for ultra-short packets. In conventional coherent URLLC transmission, pilot symbols are required for channel estimation, but the extremely small size of mini-slots causes pilots to dominate the entire packet overhead. FD-DM bypasses this bottleneck by encoding information in the frequency-domain symbol-to-symbol differences, which enables non-coherent detection without pilots. This design is particularly advantageous in the presence of rapid channel variations, Doppler shifts, or high user mobility, where frequent channel estimation becomes infeasible. The authors show that FD-DM maintains reliability even when the channel coherence time is smaller than a mini-slot, and achieves significantly lower block error rate (BLER) compared to pilot-assisted coherent detection under high mobility or fast fading conditions. Using a finite blocklength analysis, they demonstrate that FD-DM yields a more favorable BLER–latency tradeoff in short-packet transmission, which makes it a promising waveform for 5G/6G mini-slot scheduling. Overall, this work reinforces that effective URLLC communication requires not only flexible mini-slot structures but also waveform and modulation designs that eliminate pilot overhead and remain robust to rapidly time-varying wireless conditions.

At the physical layer, URLLC can be enhanced by employing diversity techniques such as time, frequency, and spatial diversity [38]. In [39], the authors formulated a Lyapunov optimization framework for mmWave-enabled massive MIMO networks to maximize network utility under probabilistic latency and reliability constraints. By leveraging spatial diversity through massive MIMO beamforming, they improve reliability while ensuring guaranteed latency. Release 16 of 3GPP [40] enhances URLLC by introducing redundant transmission, in which user packets are duplicated and delivered simultaneously to the receiver over two disjoint user-plane paths to improve reliability. This kind of diversity is known as k-repetition. The authors of [36] investigate physical–layer and MAC–layer mechanisms for enhancing the reliability and latency of URLLC in industrial control networks. The paper addresses the challenge that conventional OFDM-based

systems are highly vulnerable to interference and jamming, particularly in harsh IoT environments. To overcome these limitations, the authors propose an improved URLLC design that integrates frequency-hopping multi-carrier (FH-MC) signaling at the PHY layer with a mini-slot-based hybrid automatic repeat request (HARQ) re-transmission strategy at the MAC layer. FH-MC enables each subcarrier to hop across frequency slots following a pseudo-random sequence, which allows the transmitter to avoid jammed or poor-quality frequencies. Other techniques, including relay-assisted transmission, have also been proposed to reduce communication delay at the physical layer in URLLC scenarios [10].

The authors of [41] examine URLLC from an interference-management perspective, highlighting that interference is one of the primary obstacles to achieving the stringent reliability targets required in beyond-5G and 6G networks. In some applications such as teleoperation, URLLC traffic often coexists with eMBB and mMTC services. This creates heterogeneous interference patterns that can lead to sudden SNR drops, packet collisions, and unpredictable latency spikes. The paper emphasizes that traditional interference-avoidance methods such as scheduling, fixed power control, or simple frequency reuse are insufficient for UL/DL URLLC transmissions, especially when short packets leave little room for retransmissions or channel estimation overhead. To address this, the authors survey a range of interference-management techniques tailored to URLLC, including robust beamforming, coordinated multi-point transmission, statistical interference prediction, and proactive resource reservation. Their results show that predictive interference control, where future interference levels are estimated using learning-based models, can reduce outage probability and improve the worst-case delay. The authors also discuss how interference management interacts with short-packet communication principles and finite-blocklength constraints. When packets contain only a few channel uses, even small interference bursts can drastically increase the block error rate (BLER). This motivates ultra-fast interference mitigation at both the PHY and MAC layers. For example, the survey highlights that interference-aware mini-slot scheduling and power adaptation can provide substantial reliability gains without increasing latency. However, techniques such as URLLC-specific spatial processing incluing massive MIMO and intelligent reflecting surfaces help maintain consistent SNR margins under dense deployments. Overall, this work reinforces that

achieving URLLC performance requires not only waveform and mini-slot design but also proactive and robust interference-management strategies, particularly in multi-service and high-density environments expected in 6G.

A substantial line of work jointly considers queuing and transmission delays with finite-blocklength constraints [42]. In [9], the authors address the trade-off between transmission and queuing delays through an adaptive blocklength transmission scheme. The authors in [43] investigated queuing-based multichannel scheduling and proposed a Bayesian optimization framework to minimize the queuing delay while meeting stringent reliability requirements. Their results demonstrate that efficient queuing-aware scheduling is essential for URLLC.

For the 5G core network, the conflicting requirements of URLLC are addressed through local hosting of services, enabled by edge computing capabilities [44]. Building on this concept, [45] proposed an edge computing-based architecture that relocates the user plane function to the network edge using control and user plane separation. By bringing computational resources closer to end users and applications, this architecture helps reduce E2E latency. The work [46] studied collaborative task offloading in mobile edge computing. Their approach encourages the participation of mobile users and MEC servers while ensuring computational efficiency. Their proposed optimization problem selects the optimal task executor. In addition, it optimizes communication and computation resource allocation as well as time scheduling under dynamic task arrivals with a maximum delay tolerance. Such advances contribute to reducing delay at the network edge.

In summary, these studies show that URLLC communication can be approached from multiple angles, including finite-blocklength coding, mini-slot and waveform design, diversity and redundancy, interference management, queuing-aware scheduling, and edge computing. Collectively, they demonstrate that transmission and network delays can be pushed toward the millisecond regime while maintaining very low BLER. However, achieving an overall end-to-end delay on the order of 1 ms remains highly challenging, and additional AI-based techniques such as prediction and sending samples in advance are needed. As a result, there is limited understanding of how to jointly exploit predictive mechanisms in teleoperation and URLLC communication to both satisfy stability and transparency requirements and reduce URLLC resource consumption, which motivates the prediction-centric URLLC design developed in this thesis.

2.5 Communication Resource Allocation and Consumption in URLLC

Since the advent of 5G, it has become clear that guaranteeing ultra-high reliability and ultra-low latency typically entails significant consumption of communication resources. Time, frequency, and spatial redundancy (e.g., short transmission time intervals, frequency diversity, multi-connectivity, and massive MIMO) can improve reliability and reduce delay, but inevitably increase bandwidth usage and transmit power. Given that wireless spectrum and energy are limited, URLLC system design must rely on efficient resource allocation and scheduling strategies that jointly balance stringent delay–reliability requirements with resource efficiency. In this context, power control is a primary lever for energy efficiency, because in the finite-blocklength regime the required transmit power is tightly coupled with bandwidth, latency budget, and target error probability. As a result, power minimization must co-design these variables rather than tune them in isolation. This motivates a rich body of work on energy-efficient bandwidth, power, and transmission-time optimization under URLLC constraints.

In general, resource allocation refers to assigning bandwidth, power, and frequency channels to users or devices so as to improve network capacity, reduce packet loss, and lower energy consumption under QoS constraints. For instance, the scheme proposed by [47] enhances a V2X framework by exploiting resource awareness at the terminals: user equipments monitor specific channels within a resource pool and, after an initial selection, re-examine the remaining subchannels to avoid collisions, thereby improving reliability and reducing power consumption. Similarly, [48] proposes an energy-efficient resource allocation and power-control algorithm for IoT systems that first prioritizes subchannels with high channel gain and then allocates transmit power across these subchannels to minimize uplink energy consumption while satisfying QoS requirements. Simulation results show that, as the number of IoT terminals and average task load increase, such joint subchannel–power optimization can substantially reduce the average system energy consumption. In real-time cyber-physical systems, packetized predictive control (PPC) has been identified as an effective way to co-design control and communication over unreliable wireless links [17]. By transmitting multiple future control commands in a single packet, PPC reduces retransmissions and protocol overhead, and the predictive horizon is tuned to minimize wireless resource consumption while maintaining

control performance.

In parallel, resource scheduling mechanisms have been developed to dynamically allocate communication resources over time in response to changing traffic and network conditions. Effective scheduling is essential in URLLC to prevent congestion, limit queue build-up, and maintain stable latency under heterogeneous or bursty traffic. For example, in the IoT context, a network resource scheduling and mapping mechanism for multi-task scenarios is proposed in [49], which improves resource utilization and load balancing while reducing network energy consumption and task completion time. A dynamic resource allocation scheme for clustered machine-to-machine (M2M) communication with URLLC guarantees is presented in [50], achieving higher throughput, improved resource efficiency, and lower access delay for mixed delay-sensitive and delay-tolerant services.

The work in [51] addresses the challenge of supporting heterogeneous vehicular applications under URLLC constraints by proposing an intelligent resource-allocation framework tailored to vehicular edge computing (VEC). In fact, vehicular networks generate diverse tasks ranging from safety-critical sensing and cooperative perception to delay-tolerant infotainment each with different latency and reliability requirements. The authors highlight three main obstacles: high mobility, rapidly fluctuating V2X channel quality, and limited onboard computing resources, all of which can lead to excessive queuing delays, task failures, or violations of strict URLLC deadlines. To mitigate these issues, the paper proposes a joint communication–computation optimization framework that dynamically selects task-offloading destinations (local vehicle, roadside MEC, or remote cloud), allocates wireless bandwidth, and adjusts computational resources based on the task's reliability requirement and current network conditions. The authors use probabilistic latency and reliability models including queueing delay bounds and wireless outage probabilities to guide URLLC-aware offloading decisions. Their priority-driven scheduling mechanism ensures that safety-critical tasks with strict URLLC requirements are preferentially allocated compute and communication resources.

The survey in [52] provides a comprehensive overview of resource allocation mechanisms that support the coexistence of enhanced Mobile Broadband (eMBB) and URLLC services in beyond-5G and 6G networks. In teleoperation, URLLC and eMBB users must share the same time–frequency resources, leading to strong cross-service interference and unpredictable latency if not properly

coordinated. The paper highlights that URLLC traffic requires deterministic low latency and ultra-high reliability, while eMBB traffic demands high throughput. This creates a fundamental tradeoff in multi-service vehicular scenarios where vehicles simultaneously run safety-critical perception tasks (URLLC) and bandwidth-intensive services (eMBB). To address this challenge, the survey discusses several techniques including puncturing, superposition coding, NOMA, dynamic mini-slot scheduling, and grant-free access that allocate resources adaptively based on URLLC arrival patterns, packet deadlines, and traffic load. This work emphasizes the importance of URLLC-aware resource slicing, where network resources are partitioned dynamically so that URLLC services receive strict guarantees without starving eMBB traffic. Resource allocation strategies must be latency-aware, reliability-aware, and interference-aware to prevent URLLC failures during high vehicular mobility or sudden traffic surges. The survey also highlights that probabilistic delay bounds, which are derived from finite blocklength analysis and queueing theory, are essential tools for characterizing deadline violation probabilities in highly dynamic vehicular environments. This work reinforces that heterogeneous VEC systems must employ URLLC-prioritized scheduling, dynamic network slicing, and cross-layer resource optimization to ensure stable service delivery for both teleoperation applications and high-throughput infotainment services.

The authors in [53] study a UAV-assisted system that jointly optimizes transmit power, bandwidth allocation, and 3-D UAV deployment under URLLC constraints. In this work, the reliability constraint is equivalently reformulated as a transmit-power threshold to enable minimizing the average transmit power subject to latency and bandwidth limits. In [54], the authors introduced an energy-efficient packet delivery mechanism and jointly optimized bandwidth allocation and power control for uplink and downlink transmissions. This mechanism proactively drops certain packets during deep fading conditions to reduce resource consumption. Their approach minimizes the average total power while satisfying the QoS requirements of URLLC, thereby addressing the challenge of maintaining target reliability under severe channel fading.

In summary, existing works on URLLC-oriented resource allocation and scheduling show that careful joint optimization of bandwidth and power can substantially reduce energy consumption. Nevertheless, most of these studies investigate resource allocation and energy usage under fixed latency and reliability constraints, without exploiting prediction mechanisms to further relax the

communication burden. In particular, they rarely consider how prediction in teleoperation could reshape resource-allocation decisions and lower the required transmit power. In contrast, this thesis demonstrates that a dual-prediction scheme can significantly reduce transmit power while improving the efficiency of URLLC resource allocation.

2.6 Predictions in URLLC

Meeting the stringent requirements of URLLC, especially under the unpredictable and time-varying conditions of wireless environments, calls for accurate and real-time prediction mechanisms. Consequently, in recent years, prediction has emerged as a promising approach to meet URLLC requirements. In such systems, prediction plays three important roles: (i) proactively mitigating the impact of channel fading, network congestion, and mobility; (ii) enabling anticipatory resource allocation and control strategies that prevent latency violations and packet losses before they occur [55]; (iii) predicting the information in advance so it can be reconstructed if the transmitted data is lost or delayed [11, 12, 56].

The work in [57] provides a comprehensive discussion of how prediction and learning mechanisms can support URLLC in future 6G networks. The authors show that conventional model-based prediction techniques such as time-series models, Markov methods, and Kalman filters are limited in URLLC settings because they rely on simplified assumptions and cannot exploit long-term temporal dependencies. The paper highlights the role of deep learning, particularly recurrent neural networks (RNNs) and long short-term memory (LSTM) models, in predicting traffic loads, mobility patterns, and channel dynamics with higher accuracy and robustness. This work demonstrates that deep learning can achieve extremely low prediction error probabilities, reaching the URLLC-level reliability of 10^{-5} even for prediction horizons up to 10–20 ms. These findings confirm that accurate prediction, whether of mobility, traffic state, or channel variation, can be directly exploited to mask communication delays and reduce user-experienced latency. The authors also emphasize that prediction alone is not sufficient in non-stationary wireless environments. Since deep models trained offline may fail when network conditions change, the paper proposes a multi-level 6G architecture in which device-level prediction is enhanced through edge and cloud intelligence. Techniques such

as deep transfer learning and federated learning allow prediction models to be updated efficiently with limited samples. This maintains high reliability despite dynamic traffic and channel conditions. Overall, the paper shows that prediction is a central component of URLLC, enabling delay compensation, improving scheduling and resource allocation, and providing the foundation for intelligent, latency-aware communication-control co-design in 6G systems.

Traditional model-driven optimization techniques often fall short of URLLC requirements because they rely on idealized assumptions and are difficult to apply in real time. Key delay components, such as queuing, access, and processing delays are stochastic and highly sensitive to network load and topology [58]. Moreover, conventional methods struggle with scalability and adaptation in non-stationary environments, which is critical in high-mobility scenarios like vehicular networks or telesurgical systems. In contrast, data-driven approaches, particularly deep learning, can approximate complex control policies and predict future system states from historical and contextual information [59].

However, generic deep learning models typically require large training datasets and may generalize poorly when deployed in environments whose statistics differ from those seen during training. To address this limitation, [60] advocates integrating domain knowledge, such as information-theoretic bounds, queuing models, and cross-layer dependencies, into the learning process. This hybrid model and data-driven paradigm improves learning efficiency, accelerates convergence, and enhances interpretability, enabling URLLC systems to deliver accurate predictions while satisfying strict QoS constraints.

In [11], the authors introduced a joint prediction and communication framework to mitigate experienced delay by optimizing frequency resources and prediction horizons. Although the approach effectively balances the trade-off between delay and reliability, it relies solely on transmitter-side prediction, which limits the system's ability to recover packet loss. The authors in [61] proposed a temporal adaptive algorithm that combines different prediction techniques, with different capabilities and complexities, based on the channel condition of the communication system. Moreover, the authors introduce multi-service edge-intelligence that integrates wireless access, multi-access edge computing (MEC), and machine learning (ML). Prediction at the receiver has shown potential in enhancing synchronization between human and remote devices in TI. In [12], prediction is employed

to synchronize a physical device (transmitter) with its digital twin (receiver) in the metaverse. However, this work does not guarantee URLLC requirements and assumes fixed or known delay, which limits its applicability in dynamic wireless environments. In the context of telesurgery, [62] proposed a machine learning framework to compensate for delayed or lost packets by predicting haptic feedback at the patient side. This approach demonstrates the potential of intelligent techniques for enhancing user experience, but it operates primarily at the application layer and does not investigate communication resources or prediction horizons. Another study [63] focused on predicting the head and body motions of a human (users) in advance to render virtual reality (VR) videos. However, this work neither accounts for prediction errors nor explicitly addresses URLLC reliability constraints. The idea of model-mediated tele-operation approach was mentioned in [2]. By predicting the movement or the force feedback, the user experienced delay can be reduced.

The authors in [64] employed a prediction method for three-dimensional position and force data based on an advanced first-order autoregressive (AR) model. After an initialization and training phase, the adaptive coefficients of the model are computed to generate the predicted values. The algorithm then decides whether to update the training data using the predicted samples or the current real measurements. A related line of work proposes applying prediction at the receiver side to mitigate latency in haptic communication systems. The authors of [65] develop an LSTM-based prediction model that estimates future haptic samples and forwards them in advance. This reduces the experienced delay by a prediction window. Their system predicts multi-dimensional haptic trajectories and transmits the predicted samples ahead of time, which allows the receiver to experience a reduced effective delay. To improve reliability, they combine this predictive mechanism with a k-repetition short-packet transmission scheme, which lowers the transmission error probability without increasing the packet duration. They also incorporate queuing delay through an effective bandwidth model and formulate an optimization problem that jointly adjusts the base-station power and prediction power to balance latency, reliability, and resource consumption. While this work provides an initial attempt at receiver-side prediction under URLLC constraints, it remains limited in several aspects. First, the prediction model is treated as a black box, and the paper does not analyze how prediction accuracy affects end-to-end reliability, even though the authors acknowledge that

prediction errors dominate the total error probability. Second, the optimization relies on approximate expressions for transmission errors and does not incorporate the dynamics of non-stationary haptic signals or variable network conditions. Finally, the framework considers only receiver-side prediction and does not exploit the complementary strengths of transmitter-side models, which typically achieve higher accuracy by using real, up-to-date data. Despite these limitations, the work reinforces the potential of prediction to reduce latency in haptic communications, while highlighting the need for deeper modeling of prediction errors and more principled communication–prediction co-design.

Another relevant work examines the use of prediction to improve information freshness in cognitive IoT networks. The authors of [66] propose a prediction-assisted framework in which the receiver predicts future status updates when a packet is not delivered on time, thereby reducing the Age of Information (AoI). In the system model, the device generates real updates while the base station uses a prediction mechanism to infer missing samples based on historical observations. The authors formulate an optimization problem that adapts the transmission policy under spectrum-sharing constraints, and they demonstrate that prediction can significantly reduce both instantaneous and long-term AoI compared to conventional non-predictive strategies. Their results also show that the benefits of prediction increase when the primary user's activity limits channel availability, since prediction compensates for longer intervals without successful transmissions. Although this work highlights the potential of prediction to improve timeliness under unreliable or intermittent channels, it remains limited in several ways. First, the prediction model itself is not explicitly designed or evaluated; prediction accuracy is assumed to be known rather than derived from a concrete machine-learning model. As a result, the framework does not quantify how prediction errors propagate into the AoI performance. Second, the approach focuses solely on information freshness and does not consider stringent URLLC metrics such as reliability, haptic stability, or delay decomposition. Finally, the method employs receiver-side prediction exclusively and does not explore transmitter-side prediction or joint prediction strategies, which are essential for reducing experienced delay in real-time teleoperation systems. Despite these limitations, the paper demonstrates that prediction can effectively enhance timeliness in dynamic wireless environments and motivates the need for more rigorous prediction modeling within URLLC applications.

The authors of [67] propose a receiver-side prediction framework that reconstructs missing or delayed haptic feedback during needle insertion. Their system operates entirely at the surgeon console. It predicts the next force and torque sample whenever the corresponding packet fails to arrive within the 1 ms deadline. The prediction model is trained offline using expert demonstrations and encoded through a Hidden Markov Model (HMM), which captures temporal dependencies in the force trajectories. A Gaussian Mixture Regression (GMR) layer then generates the predicted haptic profile online. They show that HMM/GMR achieves lower error and faster prediction time compared to GMM/GMR, and consequently meeting the 1 ms latency target for up to four hidden states.

Different from command or mobility predictions in control systems, predicting some other features of traffic or performance of communications is also helpful. In [68], based on the predicted traffic state, a bandwidth reservation scheme was proposed to improve the spectral efficiency of URLLC. By exploiting the correlation among different nodes, the behavior of different users can be predicted [69]. Then, by reserving resources according to the predicted behavior, the access delay can be reduced. A fast HARQ protocol was proposed in [70], prediction is used to omit some HARQ feedback signals and successive message decodings, so that the expected delay can be improved by 27 percent to 60 compared with standard HARQ. In [71], the outcome of the decoding was predicted before the end of the transmission. With the predicted result, there is no need to wait for the acknowledgment feedback, and thus the E2E delay can be reduced. the work in [72] introduces a data-driven, context-aware approach for traffic modeling and a traffic predictor to support resource reservation in the tactile internet for bursty traffic. The traffic state is defined as the number of arriving packets in the next time window, which is modeled and predicted based on historical context information, which refers to user motion commands or haptic feedback. The proposed method provides a general framework for multi-state applications. In [73], the authors provide a detailed discussion of next-generation technologies and network intelligence in 5G NR to support URLLC requirements. They highlight federated reinforcement learning (FRL) as a promising machine learning framework for 5G NR URLLC and present it as a potential solution for meeting stringent reliability and latency targets. The paper also offers an in-depth analysis of MAC-layer channel access mechanisms that enable URLLC in 5G NR for Tactile Internet (TI) applications.

However, the focus is primarily on FRL as a candidate technique for addressing 5G NR URLLC requirements.

The work in [74] demonstrates how traffic prediction can be leveraged to improve URLLC performance in mixed-service environments. The authors study the coexistence of eMBB and randomly arriving URLLC traffic under a hybrid transmission-time-interval (TTI) structure, where eMBB operates over long TTIs and URLLC is scheduled at mini-slot granularity. In this paper, the system employs two schedulers, an eMBB scheduler (slot-level) and a URLLC scheduler (mini-slot-level), whose interaction is highly sensitive to unpredictable URLLC arrivals. To address the uncertainty inherent in random URLLC traffic, the paper proposes a prediction-based framework in which the scheduler estimates the statistical distribution of future URLLC arrivals using queueing theory and Poisson-based modeling. These predictions are then used to configure the coding redundancy of eMBB code blocks before each slot, ensuring that eMBB transmissions remain reliable even under intensive mini-slot preemption. A key insight of the paper is that predicting the expected preemption pattern enables accurate BLER prediction for eMBB, allowing the scheduler to choose robustness levels that satisfy reliability constraints. This is achieved through an estimated preemption distribution and a threshold-based BLER model. This paper shows that incorporating traffic prediction allows the system to achieve high throughput while maintaining low BLER, which outperforms static or non-predictive baselines. While the prediction mechanism is statistical rather than learning-based, the work clearly demonstrates that anticipating future traffic rather than reacting to it can substantially improve reliability and resource efficiency under URLLC constraints. More broadly, this reinforces the importance of predictive mechanisms in URLLC systems, both for reducing uncertainty in resource allocation and for enabling proactive, rather than reactive, latency control.

In networked control systems under unreliable wireless links, prediction is used as packetized predictive control (PPC) where multiple control commands are sent in a single transmission to enhance their reliability and robustness [15–17]. The authors in [16] and [17] utilized this method to minimize resource consumption in a real-time cyber-physical system. The model treats wireless resource consumption in isolation, without considering multi-user scenarios with interference. The authors minimize wireless resource consumption, but the delay, which is critical for URLLC, is not

explicitly optimized or constrained.

In summary, prediction has been applied to URLLC and related systems at multiple levels, including channel and traffic prediction, HARQ and access-control optimization, mobility and command forecasting, and haptic-signal reconstruction. These works demonstrate that anticipating future states can mask communication delay, improve information freshness, and reduce wireless resource consumption. However, many of them either ignore strict URLLC guarantees, treat prediction models as black boxes without explicitly linking prediction errors to end-to-end reliability, or consider prediction and communication design in isolation. Existing frameworks typically adopt either transmitter-side or receiver-side prediction alone and do not fully exploit joint, dual-sided prediction to cope with packet loss, and variable delay. As a result, there is still no framework that jointly designs dual prediction and URLLC communication while explicitly modeling prediction errors and optimizing horizons for haptic signals, which motivates the dual-prediction, URLLC-aware teleoperation scheme developed in this thesis.

2.7 Analytical Foundations for Predictive URLLC

The analytical foundations of predictive URLLC aim to characterize the performance limits and behavior of communication systems under stringent latency and reliability constraints. They form the backbone of prediction mechanisms by providing tractable models that anticipate performance metrics and guide learning algorithms in low-latency environments. At the core of this analytical toolkit are short-blocklength information theory, queuing theory, and stochastic geometry, each capturing essential aspects of URLLC operation.

Short Blocklength Information Theory: Short-blocklength information theory refines classical Shannon capacity, which assumes infinite blocklength and vanishing error probability, to better match URLLC scenarios where packets are short and must be delivered within a finite delay. In [34], an approximation is derived for the maximum coding rate over AWGN channels given a fixed blocklength and target error probability, revealing a fundamental tradeoff among rate, reliability, and latency. This finite-blocklength regime is crucial for predicting achievable throughput and required bandwidth in URLLC applications, especially under rapidly varying channel conditions.

It also enables proactive resource allocation based on predicted reliability for different coding and modulation schemes.

Queuing Theory: Queuing theory supports predictive URLLC through models that characterize queuing delay and buffer occupancy. While average-delay results such as Little's Law provide basic insight, URLLC requires statistical delay guarantees, in particular the violation probability of a given latency threshold. Tools such as effective bandwidth and effective capacity map arrival and service processes to exponential tail bounds on delay distributions, offering predictive guidance for resource provisioning and traffic shaping [75]. In addition, the Age of Information (AoI) metric captures the freshness of received data and is critical in predictive control systems, where stale information can severely degrade real-time decision making [76].

Stochastic Geometry: Stochastic geometry extends predictive analysis to large-scale networks with randomly distributed users and base stations. By modeling nodes as spatial point processes, it enables statistical predictions of link availability, interference, and access delay under varying user densities. Although early work mainly focused on average performance, recent studies have incorporated delay distributions and coverage probabilities tailored to URLLC [77]. For example, characterizing the probability of delay outage under different network topologies supports predictive scheduling and handover strategies that avoid latency violations.

Finally, cross-layer optimization acts as a unifying framework that links these theories across the physical, link, and network layers. For prediction, cross-layer models estimate end-to-end performance based on contextual parameters such as channel state, queue status, and mobility patterns. However, such models are often analytically intractable due to nonconvexity and high dimensionality. Consequently, recent work uses these analytical tools to shape and initialize learning-based predictors, restrict them to feasible policy spaces, and reduce training time and error rates [78].

2.8 Deep Learning for Prediction in URLLC

Deep learning has become a powerful tool for predictive modeling in URLLC systems. It provides flexible, data-driven methods to learn complex relationships between high-dimensional network states and control decisions. Unlike traditional optimization algorithms, which require accurate analytical models and are often too slow for real-time use, trained deep neural networks can produce near-optimal decisions with very small online computation time, making them suitable for the sub-millisecond latency required in URLLC [79]. In particular, deep-learning-based time-series prediction of channel conditions, traffic load, and user mobility enables proactive resource allocation, early detection of latency violations, and anticipatory handover. This improves both reliability and responsiveness in highly dynamic wireless environments. For such predictive tasks in URLLC, three main classes of deep learning techniques are commonly used: supervised learning, which is the main approach for time-series forecasting, unsupervised learning for representation learning and anomaly detection, and deep reinforcement learning (DRL) for sequential decision-making under uncertainty. In teleoperation, trajectory prediction is essential for mitigating delay and achieving zero experienced delay at both the operator and teleoperator sides. In the literature, three neural-network-based models are commonly used for trajectory prediction [56]: recurrent neural networks (RNNs), long short-term memory (LSTM) networks, and convolutional neural networks (CNNs). In all cases, the input is a sequence of past observations, and the output is a sequence of predicted samples over a future horizon [56].

RNN for Prediction

At time slot t, the input to an RNN cell consists of the current feature vector k_t and the hidden state from the previous time slot, h_{t-1}. The cell then updates its output and hidden state according to

$$o_t = \sigma\big(W_o[h_{t-1}, k_t] + b_o\big), \tag{1}$$

$$h_t = \sigma\big(W_h o_t + b_h\big), \tag{2}$$

where o_t is the cell output, h_t is the new hidden state, W_o and W_h are weight matrices, b_o and b_h are bias vectors, and $\sigma(\cdot)$ denotes an activation function.

LSTM for Prediction

Each LSTM cell takes three inputs at time slot t: the current feature vector k_t, the previous cell state L_{t-1} (long-term memory), and the previous hidden state h_{t-1} (short-term memory). The cell then updates its gates, cell state, and hidden state as

$$f_t = \sigma\big(W_f[h_{t-1}, k_t] + b_f\big), \tag{3}$$

$$i_t = \sigma\big(W_i[h_{t-1}, k_t] + b_i\big), \tag{4}$$

$$\tilde{L}_t = \tanh\big(W_k[h_{t-1}, k_t] + b_k\big), \tag{5}$$

$$L_t = f_t \odot L_{t-1} + i_t \odot \tilde{L}_t, \tag{6}$$

$$o_t = \sigma\big(W_o[h_{t-1}, k_t] + b_o\big), \tag{7}$$

$$h_t = o_t \odot \tanh(L_t), \tag{8}$$

where f_t is the forget gate controlling how much of L_{t-1} is retained, i_t is the input gate controlling how much new information is added, $\tilde{L}_t$ is the candidate cell state, o_t is the output gate, L_t is the updated cell state, h_t is the updated hidden state, W_f, W_i, W_k, W_o are weight matrices, b_f, b_i, b_k, b_o are bias vectors, $\sigma(\cdot)$ is an activation function, $\tanh(\cdot)$ is the hyperbolic tangent, and $\odot$ denotes element-wise multiplication.

CNN for Prediction

The CNN-based predictor consists of several convolutional layers, optional pooling layers, and a final fully connected layer. The convolutional layers extract local temporal features from the input sequence, the pooling layers reduce the feature dimension, and the fully connected layer maps the extracted features to the predicted trajectory.

In the convolutional layer, the feature map is obtained by convolving the input with a kernel and then applying a non-linear activation. For input patch $k^t_{i,j}$ and kernel ψ_t, the intermediate and

output features at location (i, j) are given by

$$Z_{i,j}^{t} = W_{\psi_t} * k_{i,j}^{t} + b_{\psi_t}, \tag{9}$$

$$Y_{i,j}^{t} = \Phi\big(Z_{i,j}^{t}\big), \tag{10}$$

where W_{ψ_t} and b_{ψ_t} denote the kernel weights and bias, respectively, $*$ represents the convolution operator, and $\Phi(\cdot)$ is a non-linear activation function. A pooling layer (e.g., max-pooling) is then applied to reduce the spatial/temporal resolution of $Y_{i,j}^{t}$. Finally, the resulting feature maps are flattened and passed through a fully connected layer that outputs the future trajectory samples.

2.9 Dual Prediction Scheme

The Dual Prediction Scheme (DPS) is used in Wireless Sensor Networks (WSNs) to reduce data transmission. This novel method conserves energy and extends the network's lifespan [13, 14]. In [13], both the transmitter (e.g., sensor nodes in a WSN) and the receiver (such as the cloud) operate with an identical prediction model. Sensor nodes use sampled data to evaluate the predicted values and check the error margin. If the prediction meets accuracy requirements, data transmission is skipped, and thus, significantly lowering system energy consumption. However, if the error exceeds a predefined threshold, the actual data is sent to the gateway and prompts the cloud to use the real value and update its prediction model accordingly [14]. The authors in [80] investigate the impact of using DPS for reducing the number of sensor nodes in a WSN. They characterize the theoretical gains of processing data in sensors and conditioning its transmission to the predictions' accuracy. These work rely on identical predictors primarily for energy savings and do not recognize dual prediction's potential for meeting URLLC constraints. In contrast, our proposed DPS is explicitly designed for URLLC services with stringent end-to-end latency and reliability constraints. Rather than exploiting prediction only as an energy-saving mechanism, we reinterpret dual prediction as an additional degree of freedom in the URLLC design space.

2.10 Conclusion

This chapter has reviewed the main lines of research related to URLLC communications, dual prediction schemes, and power control. Existing URLLC studies have established finite-blocklength information-theoretic limits, latency–reliability trade-offs, and diversity techniques, and have incorporated queuing and transmission delays as well as edge computing to reduce E2E latency. These advances significantly improve URLLC performance, but achieving an overall delay on the order of 1 ms remains highly challenging. In practical networks, certain delay components, such as coding delay, processing delays in computing systems, and transmission latency in the backhaul and core networks, are intrinsic and difficult to further reduce.

Prediction has emerged as a powerful tool to enhance wireless communication performance, especially in teleoperation systems. However, most existing URLLC-related prediction schemes either do not explicitly consider predicting and sending commands or samples in advance, or rely solely on transmitter-side prediction. In contrast, we propose a dual-prediction scheme that leverages advanced algorithms to anticipate the transmitter state in advance. Similar concepts have been explored in WSNs, where identical predictors at both transmitter and receiver are primarily used to reduce data transmissions and conserve energy. However, these works do not explicitly address URLLC requirements or power–latency–reliability trade-offs, nor do they optimize or verify the prediction horizons.

Taken together, the existing literature reveals a clear gap: there is no framework that jointly designs dual prediction at both transmitter and receiver with URLLC-aware power control under finite-blocklength and delay constraints, particularly for teleoperation scenarios in the tactile internet. This gap motivates the next chapter, where we introduce a dual prediction-based URLLC communication framework and formulate a power-minimization problem that explicitly couples prediction horizons, communication resources, and reliability–latency requirements.

Chapter 3

System Model and Problem Formulation

As illustrated in Fig. 3.1, we consider a communication network consisting of multiple adjacent base stations (BSs) interconnected via fiber links. Within this system, there are M independent teleoperation systems, each consisting of a user (human operator/master side) remotely controlling and manipulating a dedicated robotic arm (teleoperator/slave side). In this system model, the dual prediction scheme (DPS) is adopted, where both the master and slave sides of each teleoperation system are equipped with prediction methods to predict future states. In the network, each master that intends to communicate with its corresponding slave first samples the control command and performs prediction up to the desired horizon. All predicted samples, along with the current sample, are then encapsulated into a single packet and transmitted to its serving BS, which forwards it to the slave's access point (AP). At the AP, packets are buffered and delivered to the receiver on a first-come, first-served (FCFS) basis. Upon reception, the corresponding slave decodes the control command and, based on both the newly received data and previously stored data, performs prediction up to a horizon that satisfies the URLLC requirement. The details of the packet structure and the communication procedure for a typical teleoperation system are explained in subsections A and B, respectively. Moreover, the main notation used in this thesis is summarized in Table 3.1.

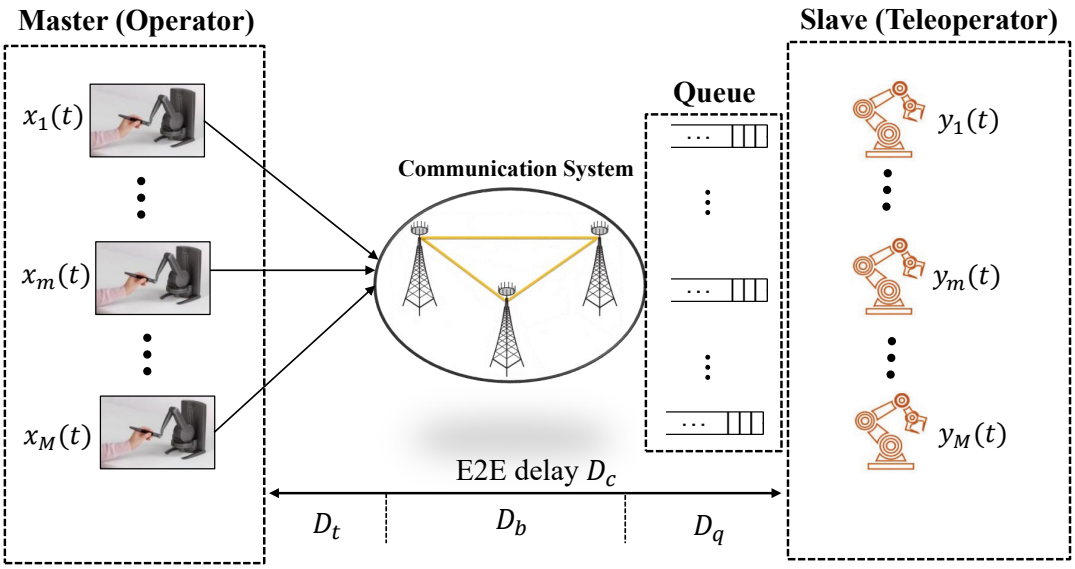

Figure 3.1: System model of the proposed URLLC-enabled teleoperation framework.

Table 3.1: Main notation used in the system model and analysis

Symbol	Description
Indices and basic quantities	
m	Index of teleoperation pair (master–slave), $m = 1, \ldots, M$
M	Number of teleoperation pairs
t	Discrete time-slot (sampling) index
T_s	Time-slot duration
Delays and URLLC constraints	
D_t	Transmission delay
D_q	Queuing delay at the BS / edge server
D_b	Backhaul / core-network delay
D_c	Communication delay, $D_c = D_t + D_q + D_b$
D_e^m	Experienced (round-trip) delay of pair m
$D_{\max}$	Maximum allowable delay (URLLC deadline)
ϵ_t^m	Transmission error probability of user m
ϵ_q^m	Queuing-delay violation probability of user m

Continued on next page

Table 3.1 (continued)

Symbol	Description
$\epsilon_{\mathrm{Rx}}^{m}$	Prediction error probability at receiver m
ϵ_{o}^{m}	Overall error probability of pair m
$\epsilon_{\max}$	Target overall error (URLLC reliability requirement)
Dual prediction parameters	
H_{Tx}	Transmitter prediction horizon
H_{Rx}	Receiver prediction horizon
H_{Tx}^{m}	Transmitter horizon for user m
H_{Rx}^{m}	Receiver horizon for user m
$H_{\mathrm{Tx,max}}$	Maximum transmitter horizon (packet length constraint)
δ_j	Prediction-error threshold for feature j
Physical-layer and traffic parameters	
P_m	Transmit power of user m
P_{ave}	Average transmit power over users
B_m	Bandwidth allocated to user m
$B_{\max}$	Total available system bandwidth
l	Length of one sample
L	Packet length, $L = l\,(H_{\mathrm{Tx}} + 1)$
λ	Mean packet arrival rate (packets per slot/frame)

3.1 Packet Transmission Scheme

In each time slot T_s, the current sample at time slot t and the predicted samples up to horizon H_{Tx}, i.e., $\bar{x}(t+1), \ldots, \bar{x}(t+H_{Tx})$, are encoded into a single packet and transmitted. This approach has two advantages: First, it reduces the prediction horizon at the receiver, thereby lowering the prediction error probability. Second, it provides a safeguard against consecutive packet losses. Fig. 3.2 illustrates the packet structure. If the sample length at the transmitter side is l, then the bit length of one packet is $L = l(H_{Tx} + 1)$. Note that encoding multiple control commands in a single packet is reasonable, as in teleoperation systems, the control command sent from the master side

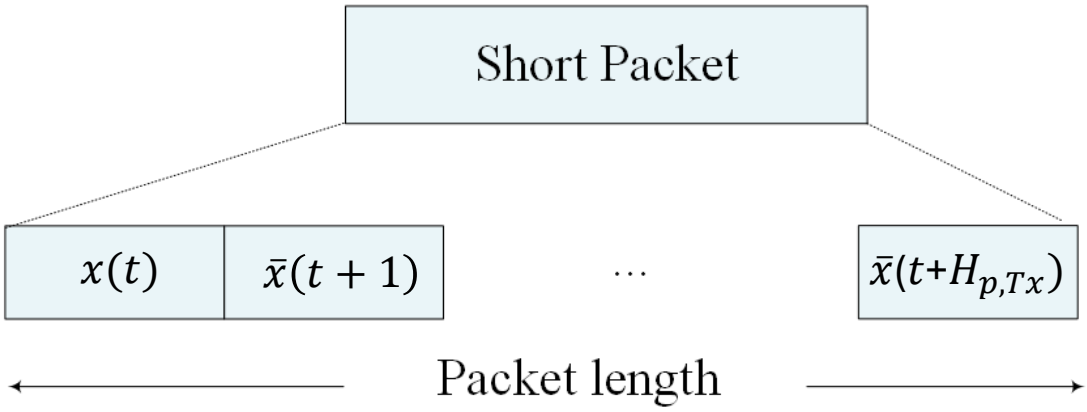

Figure 3.2: Packet structure.

typically consists of around 20 bits each, whereas the information bits in short packet transmissions are on the order of 200 bits [81]. Therefore, it is possible to encapsulate and encode multiple commands from the master side within each packet transmission. In addition, the authors in [56] highlighted that packet length in short packet transmissions cannot be excessively small.

3.2 Dual Prediction Scheme

Fig. 3.3 illustrates the proposed DPS. In the first step, the haptic interface samples the control command, $x(t)$. Next, the predictor generates future samples up to the prediction horizon H_{Tx}. The short-packet transmitter then encapsulates the sequence $\{x(t), \bar{x}(t+1), \ldots, \bar{x}(t+H_{Tx})\}$ into a packet for transmission. The transmission introduces a delay D_c due to communication latency. Upon receiving the predicted states, the slave side performs its own prediction. It adjusts for the communication delay D_c, and estimates the future state $\tilde{x}(t - D_c + H_{Tx} + H_{Rx})$ based on the received information and its previous observations, where H_{Rx} denotes the reception horizon.

In the proposed DPS, the performance and function of the transmitter (Tx) predictor are different from that of the receiver (Rx). Specifically, the Rx predictor is used to meet the required experienced delay, while the Tx predictor can help with both delay and reliability. Since the prediction horizon at the master is short (typically 3 to 5 time slots), the transmitter's prediction error probability is neglected, while the receiver's prediction has an error probability denoted as ϵ_{Rx}.

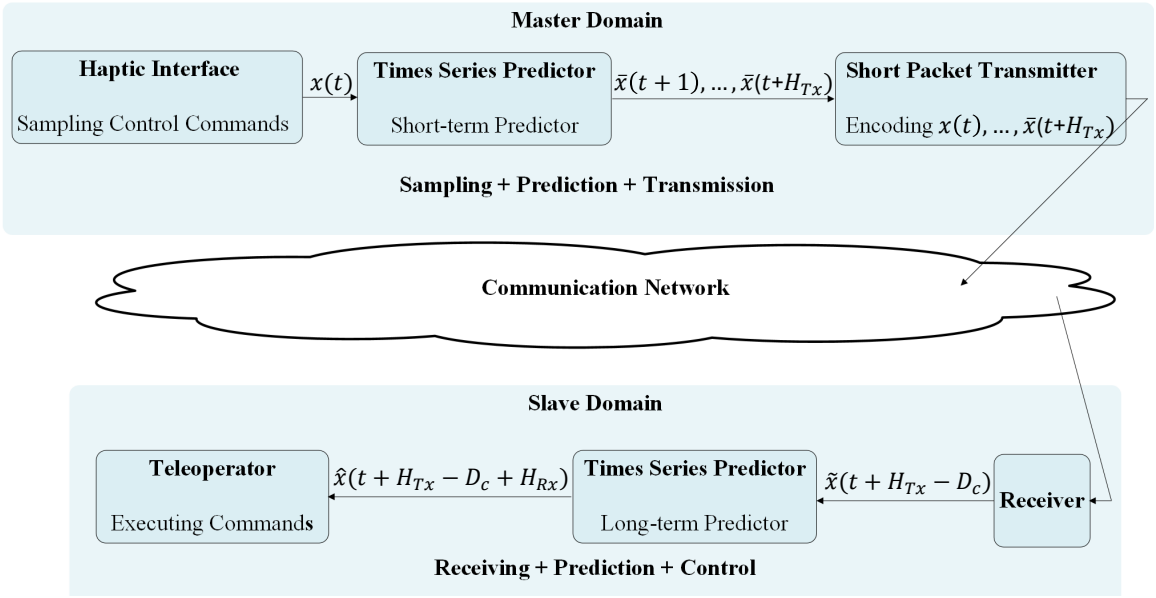

Figure 3.3: Illustration of the proposed Dual Prediction Scheme.

3.3 Overall Delay and Experienced Delay

In the proposed communication system, the overall communication delay D_c comprises multiple components. First, the master side transmits a packet to its serving BS via wireless communication. Next, it is forwarded via fiber links to the BS of the target slave, where it enters the queue before being transmitted to the slave device. Thus, as shown in Fig. 3.1, the total communication delay D_c consists of the transmission delay D_t, queuing delay D_q, and backhaul/core network delay D_b. Based on the proposed DPS, the experienced delay D_e is defined as follows:

$$D_e^m = D_t^m + D_q^m + D_b - T_s(H_{Rx}^m + H_{Tx}^m) \tag{11}$$

where T_s is the time slot duration, and the superscript m denotes the mth master device.

3.4 Overall Reliability

In this system, the communication reliability is determined by the transmission error probability ϵ_t^m and the queuing delay bound violation probability ϵ_q^m. Thus, the communication error

probability can be expressed as:

$$\epsilon_c^m = 1 - \big(1 - \epsilon_t^m\big)\big(1 - \epsilon_q^m\big). \tag{12}$$

According to URLLC requirements, ϵ_t^m and ϵ_q^m must be kept below 10^{-5}. Thus, (12) can be approximated as $\epsilon_c^m \approx \epsilon_t^m + \epsilon_q^m$. Since the transmitter encapsulates predicted future samples into a single packet to safeguard against consecutive packet losses, ϵ_c^m can be rewritten as

$$\epsilon_c^m = \big(\epsilon_t^m + \epsilon_q^m\big)^{H_{Tx}^m}. \tag{13}$$

In (13), packet transmissions are assumed independent. Given that $\epsilon_t^m + \epsilon_q^m < 1$, increasing the transmitter prediction horizon H_{Tx}^m enhances system reliability. Additionally, a larger H_{Tx}^m reduces the required prediction horizon at the receiver side, thereby decreasing the receiver's prediction error probability, ϵ_{Rx}^m. However, a large H_{Tx}^m can adversely impact the transmission error probability, as discussed in the next sections. Taking into account the error probability of the receiver-side predictor ϵ_{Rx}^m, the overall error probability ϵ_o^m is

$$\epsilon_o^m = 1 - \big(1 - \epsilon_c^m\big)\big(1 - \epsilon_{Rx}^m\big). \tag{14}$$

As before, given that ϵ_{Rx}^m and ϵ_c^m are each $< 10^{-5}$, ϵ_o^m can be approximated as

$$\epsilon_o^m \approx \big(\epsilon_t^m + \epsilon_q^m\big)^{H_{Tx}^m} + \epsilon_{Rx}^m. \tag{15}$$

The following sections describe each component of the reliability in detail.

3.5 Transmission Error Probability

For short packet transmission, the achievable rate can be approximated as [34]

$$R(P, B, D_t, \epsilon_t) = \log_2(1 + \frac{gP}{N_0 B}) - \sqrt{\frac{V}{D_t B}} Q^{-1}(\epsilon_t), \tag{16}$$

where P denotes the transmission power, g is the channel gain, N_0 represents the noise power spectral density, B is the bandwidth, and D_t is the transmission delay. $\gamma = \frac{gP}{N_0 B}$ represents the signal-to-noise ratio (SNR). V is the channel dispersion and is given by

$$V = (\log_2 e)^2 \left(1 - \frac{1}{(1+\gamma)^2}\right),$$

which measures the variability of channel capacity with respect to block length. In URLLC scenarios, where SNR typically exceeds 5 dB, V can be approximated as $(\log_2 e)^2$ [25]. Consequently, the transmission error probability is expressed as

$$\epsilon_t = Q\left(\frac{D_t B \log_2(1+\gamma) - l(H_{Tx}+1) + \log_2(D_t B)/2}{\log_2 e \sqrt{D_t B}}\right), \tag{17}$$

where $l(H_{Tx}+1) = D_t B R$ denotes the total number of information bits contained in each packet, as determined by the packet structure under the dual prediction scheme. For fixed B and P, the transmission error probability ϵ_t increases as the prediction horizon H_{Tx} grows. This highlights the importance of carefully selecting the prediction horizon to optimize system reliability.

3.6 Queuing Delay Violation Probability

In this thesis, the packet arrival process is assumed to follow a Poisson distribution with an average arrival rate of λ packets per frame [68]. To analyze queuing delay, effective bandwidth is used, which characterizes the minimum constant service rate needed to support a random arrival process while satisfying a queuing delay bound D_q and a violation probability. According to [22], the effective bandwidth can be expressed as

$$E_B = \frac{\ln(1/\epsilon_q)}{D_q \ln\left(\frac{\ln(1/\epsilon_q)}{\lambda D_q} + 1\right)} \text{ (packets/slot)}. \tag{18}$$

To satisfy the queuing delay bound D_q and the violation probability ϵ_q, the constant packet

service rate must be equal to the effective bandwidth, which yields the following constraint:

$$\frac{1}{D_t} = E_B. \tag{19}$$

Accordingly, the queuing delay violation probability can be expressed as

$$\epsilon_q = \exp\left(D_q\left[\frac{1}{D_t}W_{-1}\left(-\lambda D_t e^{-\lambda D_t}\right) + \lambda\right]\right), \tag{20}$$

where $W_{-1}(\cdot)$ refers to the "-1" branch of the Lambert W-function, which is the inverse of $f(x) = xe^x$. Equation (20) shows that under fixed λ and E_B, ϵ_q strictly decreases as the queuing delay D_q increases [11].

3.7 Prediction Error Probability

In the proposed DPS system model, the receiver-side predictor is prone to errors, with the error probability being a function of the prediction horizon. Let $X(t) = [x_1(t), x_2(t), \ldots, x_F(t)]^T$ denote the state of a device at the t-th time slot, where F is the number of features. Consider that the device's state follows Kalman filter with the following state transition function:

$$X(t+1) = \Phi X(t) + W(t), \tag{21}$$

where $\Phi = [\phi_{i,j}]_{F\times F}, i, j = 1, 2, \cdots, F$, is the constant state transition matrix and $W(t) = [w_i(t)]_{F\times 1}, i = 1, 2, \cdots, F$, is the transition noise. In this work, $W(t)$ are independent random variables following Gaussian distributions with zero mean and variances $\sigma_1^2, \sigma_2^2, \cdots, \sigma_F^2$. Then, using the methods in [11], the prediction error probability can be calculated as follows:

$$\begin{aligned}\epsilon_{Rx}(H_{Rx}) = 1 - \prod_{j=1}^{F}\Bigg[1- \\ \psi\left(\frac{-\delta_j}{\sqrt{\sigma_j^2 + \sum_{i=1}^{H_{Rx}-1}\sum_{k=1}^{F}(\phi_{j,k,H_{Rx}-i})^2\sigma_k^2}}\right)\Bigg],\end{aligned} \tag{22}$$

where H_{Rx} denotes the prediction horizon, δ_j is the permissible error threshold between the actual and predicted value of feature j, $\psi(\cdot)$ is the cumulative distribution function (CDF) of a standard Gaussian distribution with zero mean and unit variance, and $\phi_{j,k,H_{Rx}-i}$ represents the element of $(\Phi)^{H_{Rx}-i}$ at the j-th row and k-th column. It can be shown that ϵ_{Rx} increases with H_{Rx}.

Chapter 4

URLLC with DPS: A Prediction and Communication Co-Design

4.1 Problem Formulation

In wireless communication systems, the primary resources, power, bandwidth, and time, must be efficiently managed to meet performance and operational requirements. In this work, we propose an optimization framework that jointly allocates power and bandwidth with the objective of minimizing the system's average transmit power:

$$P_1 : \min_{P^m, B^m, D_q^m, H_{Rx}^m, H_{Tx}^m} \; P_{\text{ave}} = \frac{1}{M} \sum_{m=1}^{M} P^m \tag{23}$$

$$\text{s.t. } (\epsilon_t^m + \epsilon_q^m)^{H_{Tx}^m} + \epsilon_{Rx}^m \leq \epsilon_{\max}, \tag{23a}$$

$$D_t^m + D_q^m + D_b - T_s(H_{Rx}^m + H_{Tx}^m) \leq D_{\max}, \tag{23b}$$

$$\sum_{m=1}^{M} B^m \leq B_{\max}, \tag{23c}$$

$$H_{Tx}^m \leq H_{Tx,\max}, \tag{23d}$$

where constraints (23a) and (23b) represent the maximum allowable reliability and delay, respectively, which are URLLC requirements, and $B_{\max}$ denotes the maximum available bandwidth.

Moreover, the limited packet length imposes an upper bound on the transmitter prediction horizon, as expressed in (23d).

Note that in problem P_1, while P^m explicitly appears only in the transmission error probability ϵ_t^m, it influences all other factors indirectly through the URLLC constraints. Reducing P^m impacts ϵ_t^m, and thus to satisfy the overall error probability requirement, ϵ_q^m must be decreased. However, decreasing ϵ_q^m necessitates an increase in D_q^m, which may violate the overall delay constraints. Furthermore, an increase in ϵ_t^m requires a reduction in ϵ_{Rx}^m, which in turn demands a reduction in H_{Rx}^m. This adjustment could affect the maximum allowable delay $D_{\max}$. Additionally, reducing P^m might necessitate a higher bandwidth allocation to maintain the same level of ϵ_t^m. This impacts the total bandwidth constraint and potentially introduces resource allocation challenges within the URLLC framework.

The formulated optimization problem is nonconvex because the URLLC constraints are nonlinear. Specifically, the transmission error probability ϵ_t is defined via the Gaussian Q-function in (17) and the queuing violation probability ϵ_q involves the Lambert $W^{-1}(\cdot)$ in (20). These depend nonlinearly on the decision variables of $P^m, B^m, D_t^m, H_{\mathrm{Tx}}^m$ and D_q^m. In addition, decision variables H_{Tx}^m and H_{Rx}^m are discrete integers, which further increase the problem's complexity. Moreover, constraint (23c) couples all users, which adds another layer of difficulty to the optimization. In the following section, we outline our approach to addressing this problem.

4.2 Algorithm to Solve Problem P_1

To address the problem, we decompose it into two subproblems. The first subproblem is solved for a given B^m and P_{ave} is minimized as a function of B^m. By doing so, the optimal horizons are obtained. In the second subproblem, given the horizons, the optimal bandwidth that minimizes the transmit power is found. The two subproblems are solved iteratively until convergence is reached. The main steps for addressing the optimization problem are illustrated in the block diagram of Fig. 4.1.

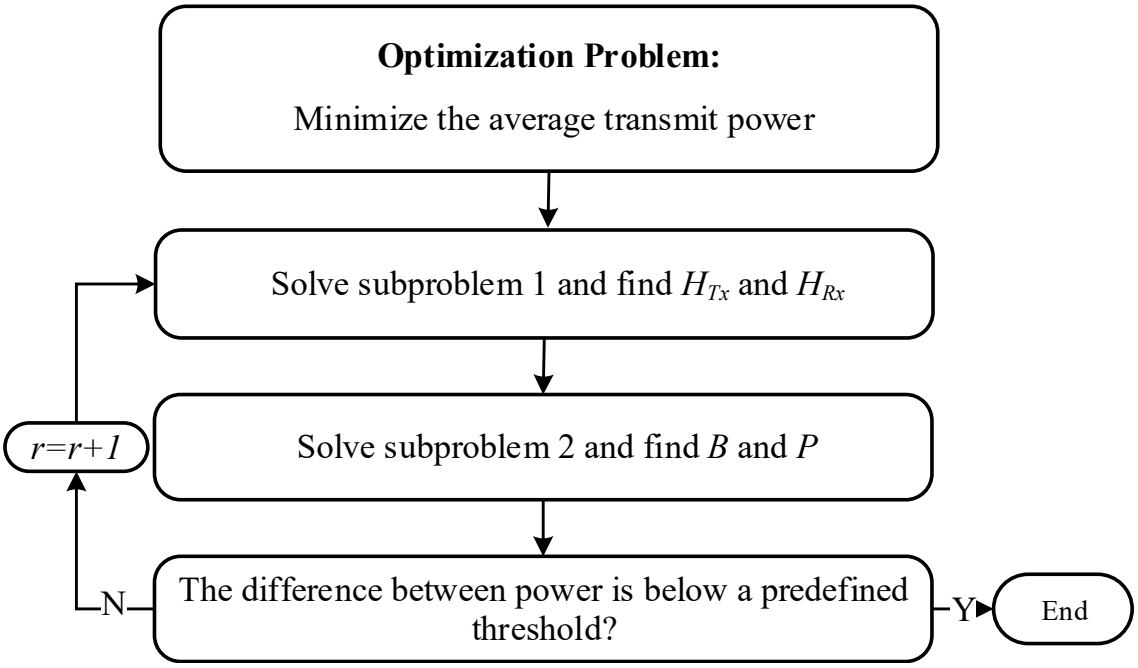

Figure 4.1: Block diagram for the proposed solution.

4.2.1 Sub-problem 1

We first attempt to decouple the users in order to formulate the first subproblem. The only coupling arises from constraint (23c). To address this, we initially assign tentative bandwidth values to each master such that constraint (23c) is satisfied. Thus, the problem can be reformulated as M single-user problems without the constraint on total bandwidth, which is written in problem (24).

$$\min_{P^m, H_{Rx}^m, H_{Tx}^m, D_q^m} P^m \tag{24}$$

$$\text{s.t. (23a), (23b), (23d).}$$

Problem (24) minimizes the transmit power of each individual user that can satisfy a certain overall reliability, $P_{\min}^m(\epsilon_o^m)$. However, determining this minimum power remains highly challenging, as the optimization problem is inherently non-convex and involves integer-valued decision variables. Motivated by [11], we first minimize ϵ_o^m for a given P^m, and then we will use binary search to

determine the minimum power needed to satisfy $\epsilon_o^m \leq \epsilon_{\max}$. Since URLLC constraints are strictly decreasing in transmit power, the optimal solution for minimizing the overall error probability can be achieved when the equality in constraint (23b) is satisfied, that is:

$$D_q^m = D_{\max} + T_s(H_{Rx}^m + H_{Tx}^m) - D_t^m - D_b. \tag{25}$$

By leveraging this equality, D_q^m will be a function of the prediction horizons and consequently, there will be fewer independent optimization variables. Accordingly, the minimum overall error probability can be achieved by optimizing the prediction horizons H_{Tx}^m and H_{Rx}^m in the simplified problem that is formulated as (26).

$$\min_{H_{Rx}^m, H_{Tx}^m} \epsilon_o^m = (\epsilon_q^m + \epsilon_t^m)^{H_{Tx}^m} + \epsilon_{Rx}^m \tag{26}$$

$$\text{s.t. } D_t^m + D_q^m + D_b - T_s(H_{Rx}^m + H_{Tx}^m) = D_{\max}, \tag{26a}$$

$$H_{Tx}^m \leq H_{Tx,\max}. \tag{26b}$$

To solve this problem, following the approaches in [11] and [22], a near-optimum solution can be achieved by imposing the following constraints. According to [11], this simplification leads to negligible performance loss.

$$(\epsilon_t^m + \epsilon_q^m)^{H_{Tx}^m} = \epsilon_{Rx}^m \tag{27}$$

$$\epsilon_t^m = \epsilon_q^m \tag{28}$$

These assumptions reduce the complexity of the optimization problem by balancing the error probabilities between transmission, queuing, and reception prediction. Thus, the optimal values of H_{Tx}^m and H_{Rx}^m can be determined using numerical approach. We denote these optimal values as H_{Tx}^{m*} and H_{Rx}^{m*}, respectively. The corresponding minimum overall error probability is denoted by $\epsilon_{o,\min}^{m*}$. Thus, the following optimization problem can be used to determine the minimal power needed to

ensure overall reliability:

$$\min_{P^m} \quad P^m \tag{29}$$
$$\text{s.t.} \quad \epsilon^{m*}_{o,\min} \leq \epsilon_{\max}.$$

In the optimization problem, ϵ^m_{o} is a decreasing function of the transmit power P^m, as illustrated in our simulation results. Consequently, reducing P^m leads to an increase in ϵ^m_{o}. The minimum power is therefore achieved when $\epsilon^m_{\text{o,min}} = \epsilon_{\max}$. Now, after obtaining the optimal transmit power P^{m*}, we determine the corresponding horizons, i.e., $H^{m*}_{Tx} = H_{Tx}(P^{m*})$ and $H^{m*}_{Rx} = H_{Rx}(P^{m*})$. By substituting these optimal horizons into the queuing delay expression for D_q in (25), we obtain all decision variables for the first subproblem, given the initially selected bandwidth values.

4.2.2 Sub-problem 2

We now proceed to the second subproblem, where, based on the decision variables obtained from the first subproblem, the optimal bandwidth allocation for each master is determined. Our objective is now to minimize the total average power consumption across all users, subject to the total available bandwidth constraint. Therefore, the second subproblem can be reformulated as follows:

$$P_2 : \min_{P^m,\, B^m} \; P_{\text{ave}} = \frac{1}{M} \sum_{m=1}^{M} P^m \tag{30}$$

$$\text{s.t. } (\epsilon^m_t + \epsilon^m_q)^{H^m_{Tx}} + \epsilon^m_{Rx} \leq \epsilon_{\max}, \tag{30a}$$
$$\sum_{m=1}^{M} B^m \leq B_{\max}. \tag{30b}$$

Since ϵ^m_o is a decreasing function of P^m, the optimal solution of P_2 is attained when the equality in (30a) holds. Furthermore, the unknown parameters in this subproblem are only power and bandwidth. Thus, in constraint (30a), all terms remain constant except for ϵ^m_t. By performing the

necessary mathematical manipulations, shifting the constant values to the other side, and representing them as $\epsilon^m_{t,\max}$, we obtain:

$$\min_{P^m,\, B^m} \frac{1}{M} \sum_{m=1}^{M} P^m \tag{31}$$

$$\text{s.t. } \epsilon_t^m = \epsilon^m_{t,\max}, \tag{31a}$$

$$\sum_{m=1}^{M} B^m \leq B_{\max}, \tag{31b}$$

where $\epsilon^m_{t,max}$ is a constant value obtained as follows:

$$\epsilon^m_{t,max} = (\epsilon_{\max} - \epsilon^m_{Rx})^{(1/H^m_{Tx})} - \epsilon^m_q \tag{32}$$

By applying the transmission error probability formula and performing the necessary mathematical derivations, the transmit power can be expressed as a function of bandwidth as follows:

$$P^m = \frac{N_0 B^m}{g^m} \left\{ \exp\left[\frac{l(H^m_{Tx}+1)\ln 2}{B^m D^m_t} + Q^{-1}(\epsilon^m_{t,\max}) \sqrt{\frac{1}{B^m D^m_t}} \right] - 1 \right\} \triangleq P^m_{\text{th}}. \tag{33}$$

Based on (33), P^m is nonconvex with respect to B^m.

Lemma 1: P^m_{th} initially decreases and then increases with respect to B^m, and there exists a unique solution B^m_{th} that minimizes P^m_{th}. Moreover, P^m_{th} is strictly convex with respect to B^m for $0 \leq B^m \leq B^m_{\text{th}}$.

The proof of this lemma is provided in [53]. Thus, problem P_2 can be simplified as follows:

$$P_3 : \min_{B^m} \ \frac{1}{M} \sum_{m=1}^{M} P^m_{\text{th}} \tag{34}$$

$$\text{s.t.} \sum_{m=1}^{M} B^m \leq B_{\max}, \tag{34a}$$

$$0 \leq B^m \leq B_{\text{th}}^m. \tag{34b}$$

The objective function of P_3 is convex because, according to Lemma 1, P_{th}^m is strictly convex with respect to B^m for $0 \leq B^m \leq B_{\text{th}}^m$. Since the sum of convex functions remains convex, the overall objective function is convex. Moreover, the total bandwidth constraint, $\sum_{m=1}^{M} B^m \leq B_{\max}$, is linear and thus convex. The box constraints, $0 \leq B^m \leq B_{\text{th}}^m$, also define a convex set. Since both the objective function and the constraints are convex, the problem is convex.

Slater's constraint qualification (SCQ) states that if a convex optimization problem has a strictly feasible point in the interior of its feasible region, then strong duality holds. Under this condition, problem P_3 can be solved using the Lagrange dual method. To confirm Slater's condition, we check for the existence of a strictly feasible point. The feasible region includes all B^m values where $0 < B^m < B_{\text{th}}^m$ and $\sum_{m=1}^{M} B^m < B_{\max}$. Since B_{th}^m is finite, there is always a strictly viable solution within this range. Thus, the optimization problem meets Slater's constraint qualification. As a result, strong duality applies, meaning the problem can be effectively solved using the Lagrange dual method.

Let λ be the nonnegative Lagrange multiplier associated with constraints (34a). The Lagrange function of problem (34) is

$$\mathcal{L}(B^m, \lambda) = \frac{1}{M} \sum_{m=1}^{M} P_{\text{th}}^m + \lambda \left(\sum_{m=1}^{M} B^m - B_{\max} \right) \tag{35}$$

The dual problem is obtained by maximizing the Lagrangian with respect to the Lagrange multiplier λ while minimizing with respect to the primal variables B^m. Thus, the Lagrange dual function can be written as

$$g(\lambda) = \min_{B} \mathcal{L}(B, \lambda) \tag{36}$$

$$\text{s.t. } 0 \leq B^m \leq B_{\text{th}}^m. \tag{36a}$$

Therefore, the dual problem can be expressed as

$$\max\ g(\lambda) \tag{37}$$

$$\text{s.t. } \lambda \geq 0. \tag{37a}$$

Next, using the dual variable (λ), we obtain $\min_{B^m} \mathcal{L}(B^m, \lambda)$. The dual problem is then solved to obtain the optimal λ. These two steps are explained below.

Step 1: Solution of the Lagrange Dual Function

First, with the given dual variables, problem (35) is decomposed into M subproblems as follows:

$$\min_{B^m} \frac{1}{M} P_{\text{th}}^m + \lambda B^m, \tag{38}$$

which has been shown that the above problem is convex with respect to B^m within the feasible region. Hence, the optimal solution is obtained by applying the Karush–Kuhn–Tucker (KKT) conditions. To this end, the derivative of (38) with regard to B^m is obtained as

$$\begin{aligned}\frac{\partial L}{\partial B^m} = \lambda + \frac{1}{M}\Bigg[&- \frac{N_0}{g^m} + \frac{N_0}{g^m}\Bigg(1 - \frac{l(H_{Tx}^m + 1)\ln 2}{B^m D_t^m} \\ &- \frac{Q^{-1}(\epsilon_{t,\max}^m)}{2\sqrt{B^m D_t^m}}\Bigg) \exp\left(\frac{l(H_{Tx}^m + 1)\ln 2}{B^m D_t^m}\right. \\ &\left. + \frac{Q^{-1}(\epsilon_{t,\max}^m)}{\sqrt{B^m D_t^m}}\right)\Bigg].\end{aligned} \tag{39}$$

The optimum bandwidth, which is denoted by B^{m*}, is obtained by solving the following equation.

$$\left(\frac{\partial L}{\partial B^m}\right)\Big|_{B^m = B^{m*}} = 0 \tag{40}$$

Using Lemma 1, there exists a unique solution B^{m*} that satisfies $\frac{\partial L}{\partial B^m} = 0$. Due to the complexity of equation (40), the derivative $\frac{\partial L}{\partial B^m} = 0$ may not have a closed-form solution. Instead, we can use numerical methods to solve for B^{m*}. To address this, a binary search algorithm is introduced based

Algorithm 1 Algorithm for Solving the Optimization Problem

Input: Number of users M, total available bandwidth $B_{\max}$, maximum allowable error $\epsilon_{\max}$, maximum delay $D_{\max}$, transmission delay D_t^m, backhaul delay D_b, max horizon $H_{Tx,\max}$, arrival rate λ^m, sample length l, channel gain g^m, noise power N_0, noise σ_j, threshold δ_j.
Output: $\{P^{m*}, B^{m*}, H_{Tx}^{m*}, H_{Rx}^{m*}\}$: Optimal power, bandwidth, and horizons.
Initialize B_0^m, P_0^m, and set $r = 0$.
repeat
  Solve subproblem 1 with $\{P_r^m, B_r^m\}$ to get H_{Tx}^{m*} and H_{Rx}^{m*}
  Set $H_{Tx,r+1}^m = H_{Tx}^{m*}$ and $H_{Rx,r+1}^m = H_{Rx}^{m*}$
  Solve problem P2 to get P^* and B^*; set $P_{r+1} = P^*$, $B_{r+1} = B^*$
  Update $r = r + 1$
until Fractional decrease in the objective is below threshold
return $\{P_r^m, B_r^m, H_{Tx,r}^m, H_{Rx,r}^m\}$

on Lemma 1 to determine B^{m*}.

Step 2: Solution of the Dual Problem

Once the optimal bandwidth is obtained, the dual problem in (37) is solved to find the optimal dual variable that maximizes $g(\lambda)$. To this end, we adopt a subgradient-based method [53] to solve the dual problem and update the Lagrangian multiplier λ as mentioned below:

$$\lambda_{t+1} = \lambda_t + \theta_t \left(\sum_{m=1}^{M} B^m - B_{\max} \right), \tag{41}$$

where t is the iterative index and θ_t^m is the update step size.

After updating the Lagrangian multipliers, the original problem is re-solved. Steps 1 and 2 are then repeated until Subproblem 2 converges. The overall solution procedure is summarized in Algorithm 1 and illustrated in Fig. 4.1.

4.2.3 Proof for the Convergence of Algorithm 1

In this section, we provide the proof of convergence for Algorithm 1.

Let $\phi\big(P^m, B^m, D_q^m, H_{\mathrm{Rx}}^m, H_{\mathrm{Tx}}^m\big)$ denote the objective of problem (23). First, we set the bandwidth so that constraint (23c), which couples the teleoperation pairs, holds. The problem then reduces to optimization (24), which we treat as *Subproblem 1*.

Subproblem 1. In (24), we fix power P^m (the objective of (24)) and focus on the constraints so that, for each fixed P^m, the best $D_q^m, H_{\mathrm{Rx}}^m, H_{\mathrm{Tx}}^m$ minimizing the left-hand side of constraint (23a) are obtained. Because power is a decreasing function of delay, the best P^m is achieved when the total delay equals $D_{\max}$ (i.e., equality holds in (23b)). Hence, D_q^m becomes a function of H_{Rx}^m and H_{Tx}^m. For each fixed P^m, we choose H_{Rx}^m and H_{Tx}^m to minimize the left-hand side of (23a) by solving (26). Following [56] and [22], (26) is solved via the two equations (27) and (28), which yields H_{Rx}^m and H_{Tx}^m as functions of P^m. With all variables expressed in terms of P^m, we obtain the optimal power by solving (29). Therefore, after Subproblem 1,

$$\begin{aligned}&\phi\big(p^m(r+1), B^m(r), D_q^m(r+1), H_{\mathrm{Rx}}^m(r+1), H_{\mathrm{Tx}}^m(r+1)\big)\\ &\leq\ \phi\big(p^m(r), B^m(r), D_q^m(r), H_{\mathrm{Rx}}^m(r), H_{\mathrm{Tx}}^m(r)\big).\end{aligned} \tag{42}$$

Subproblem 2. In (30), we substitute the optimal values of the decision variables from Subproblem 1 and optimize the bandwidth B^m to further decrease the objective. After Subproblem 2, we have

$$\begin{aligned}&\phi\big(p^m(r+1), B^m(r+1), D_q^m(r+1), H_{\mathrm{Rx}}^m(r+1), H_{\mathrm{Tx}}^m(r+1)\big)\\ &\leq\ \phi\big(p^m(r+1), B^m(r), D_q^m(r+1), H_{\mathrm{Rx}}^m(r+1), H_{\mathrm{Tx}}^m(r+1)\big).\end{aligned} \tag{43}$$

From 42 and 43 we have:

$$\begin{aligned}&\phi\big(p^m(r+1), B^m(r+1), D_q^m(r+1), H_{\mathrm{Rx}}^m(r+1), H_{\mathrm{Tx}}^m(r+1)\big)\\ &\leq\ \phi\big(p^m(r), B^m(r), D_q^m(r), H_{\mathrm{Rx}}^m(r), H_{\mathrm{Tx}}^m(r)\big).\end{aligned} \tag{44}$$

Thus, in each iteration the objective value is nonincreasing. Since $\phi(\cdot) \geq 0$, the sequence $\{\phi_r\}_{r\geq 0}$ is bounded below and monotonically nonincreasing; hence it converges. This establishes the convergence of Algorithm 1.

Chapter 5

Simulation Results

This chapter presents simulation studies to evaluate the performance of the proposed dual prediction scheme. A one-dimensional mobility model is adopted to represent the movement of the master side, and the prediction model in (22) is employed to generate prediction samples and compute the corresponding prediction error probability. For this one-dimensional movement, the state transition function in (21) is given by [82]

$$\begin{bmatrix} r(k+1) \\ v(k+1) \\ a(k+1) \end{bmatrix} = \begin{bmatrix} 1 & \Delta t & \frac{1}{2}\Delta t^2 \\ 0 & 1 & \Delta t \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r(k) \\ v(k) \\ a(k) \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ w(k) \end{bmatrix}, \tag{45}$$

where $r(k)$, $v(k)$, and $a(k)$ denote the position, velocity, and acceleration in the kth time slot, respectively. Δt is the slot duration, and $w(k)$ is zero-mean Gaussian noise added to the acceleration. Using this model, we can implement the proposed prediction scheme and analyze how DPS improves the trade-off between URLLC and resource utilization. For comparison, the performance of two baseline schemes is also presented: one without prediction and another with prediction applied only at the receiver side. The simulation parameters used throughout are summarized in Table 5.1, unless noted otherwise. In all simulations, the channel gain is modeled as $10\log_{10}(g^m) = 35.3 + 37.6\log_{10}(d^m) + S^m$, where d^m denotes the distance between user m and the access point (AP), and $S^m \sim \mathcal{N}(0, 8^2)$ (dB) represents log-normal shadowing.

Table 5.1: Simulation Parameters

Parameter	Value
Slot duration T_s	1 ms
Transmission duration D_t	1 ms
Backhaul delay D_b	12 ms
Average packet arrival rate λ	100 packets/s
Noise power spectral density N_0	-174 dBm/Hz
Standard deviation of acceleration noise σ_w	$0.01\ \mathrm{m/s^2}$
Threshold δ	0.1 m
Sample length l	16

Fig. 5.1 shows prediction error probability, which is calculated based on the previously described state transition and prediction model. As seen in the simulation result, the error is zero up to a horizon of 6 ms, which validates the assumption that the transmitter prediction error is zero. Therefore, a value of 6 is used as $H_{Tx,max}$ in the simulation for the multi-user scenario. After this value, the error remains negligible up to a prediction horizon of 18. Beyond this point, it increases sharply and follows an exponential growth trend.

Fig. 5.2 illustrates Lemma 1. The required transmit power P_{th} is plotted versus the bandwidth for a fixed distance between the teleoperation master side and the access point, under different sample lengths. The results demonstrate that P_{th} initially decreases and then increases with respect to the bandwidth. This behavior indicates the existence of a unique optimal bandwidth B_{th} that minimizes P_{th}. Moreover, P_{th} is strictly convex in B over the interval $0 \leq B \leq B_{th}$.

Fig. 5.3 illustrates the convergence behavior of the proposed algorithm in a scenario with 50 teleoperation systems. It is observed that the minimum transmit power required for URLLC scenarios drops rapidly across iterations, and the algorithm converges within 7 iterations.

In the remainder of the simulation section, we evaluate the performance of the proposed DPS scheme under two different scenarios: a single-user scenario and a multi-user scenario.

5.1 Simulation in Single-User Scenario

In the single-user scenario, the distance between the user and the access point (AP) is fixed. First, we fix the receiver's prediction horizon. Under the delay constraint $D_{\max} = 0$ ms and the reliability constraint $\epsilon_{\max} = 10^{-5}$, the transmitter's prediction horizon H_{Tx}^{m} is optimized to minimize

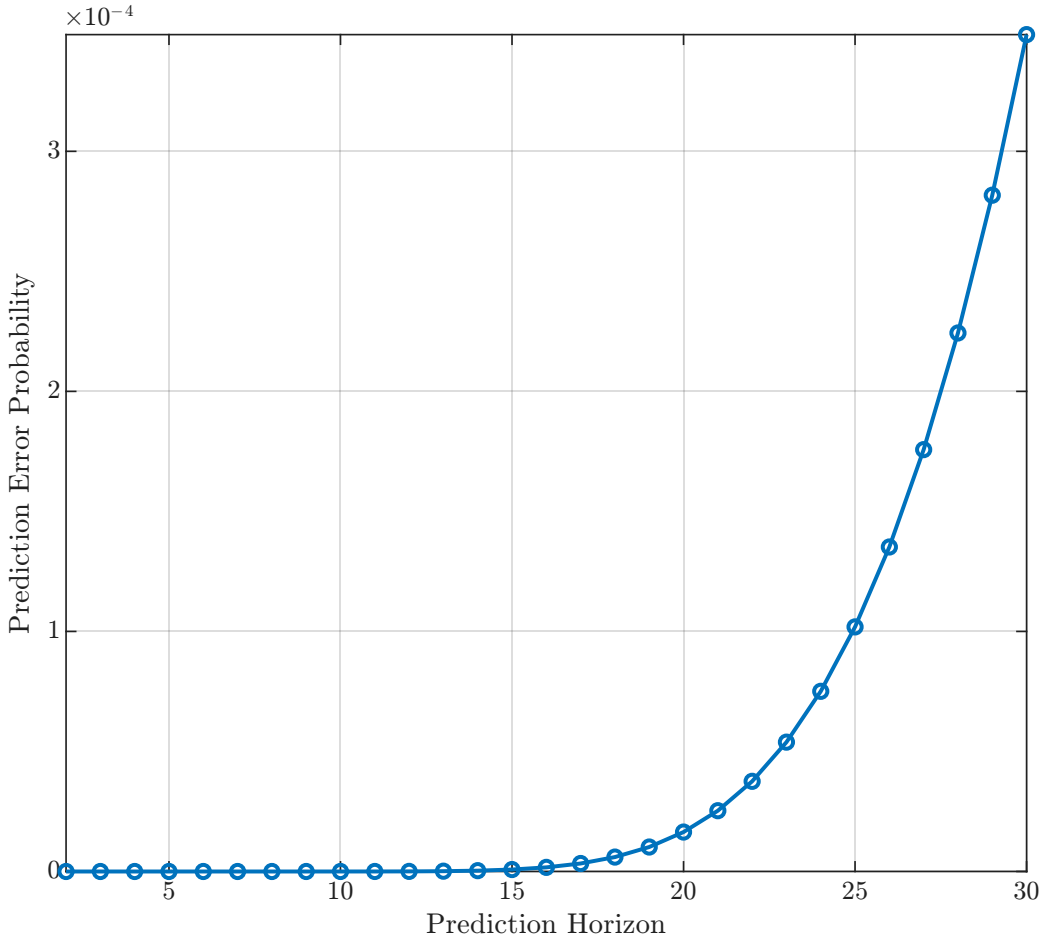

Figure 5.1: Prediction error probability as a function of prediction horizon.

the required transmit power P^m. Here, $D_{\max} = 0$ denotes zero experienced delay due to prediction, even though raw communication components sum to a positive delay.

Fig. 5.4 illustrates the relationship between transmit power and prediction horizon for two different sample sizes: $l = 16$ and $l = 32$ bits. Fig. 5.5 presents similar results under varying bandwidth conditions while fixing the sample length at $l = 32$ bits. In all four plots, the power exhibits a non-monotonic behavior: it initially decreases and then increases as the prediction horizon grows. In fact, initially, due to the exponential term in the reliability expression, increasing H_{Tx} leads to a reduction in the required transmit power. However, as H_{Tx} continues to grow, the corresponding increase in packet length necessitates more power to maintain the target reliability. This behavior can also be verified mathematically by the fomula of P_{th} in (33). This optimal value of H_{Tx} depends on system parameters including bandwidth B_m, delay D_t, and sample length l.

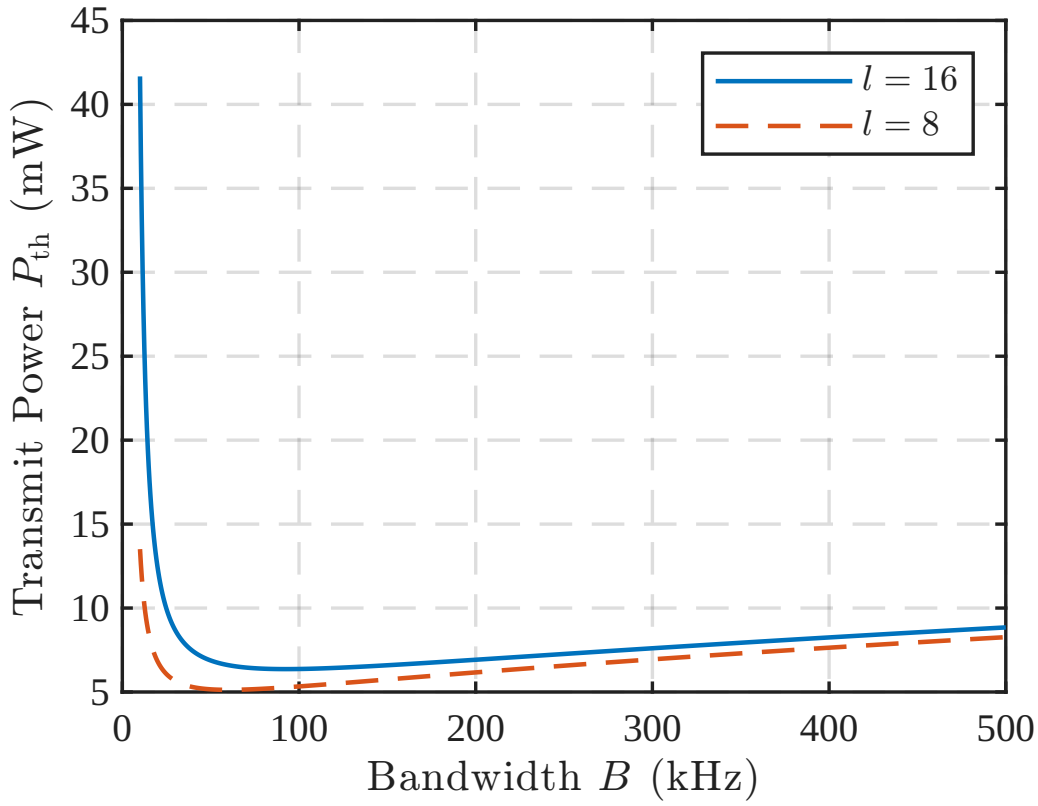

Figure 5.2: Transmit power threshold P_{th} versus the bandwidth.

From Fig. 5.4 and Fig. 5.5, it is observed that when the sample length is small or the available bandwidth is large, multiple samples (e.g., three) can be efficiently encapsulated into a single packet without violating reliability or delay constraints. However, as the sample length increases or the bandwidth decreases, including additional samples in the same packet necessitates higher transmit power to satisfy the overall reliability and latency requirements. These results highlight that both the available bandwidth and the sample length significantly impact the optimal prediction horizon and the minimum required transmit power.

In addition, Fig. 5.5 shows a crossover between the two power curves for different bandwidth values. This occurs due to the opposing effects of bandwidth on power consumption. At small values of H_{Tx}, the linear scaling factor $\frac{N_0 B}{g_m}$ dominates, which results in higher transmit power for larger bandwidth. However, as H_{Tx} increases, the packet size grows, and the rate requirement term $\frac{l(H_{Tx}+1)}{BD_t}$ begins to dominate. This term, which is located in the exponent, causes the transmit power to rise more rapidly for smaller bandwidth values. Consequently, although the lower bandwidth

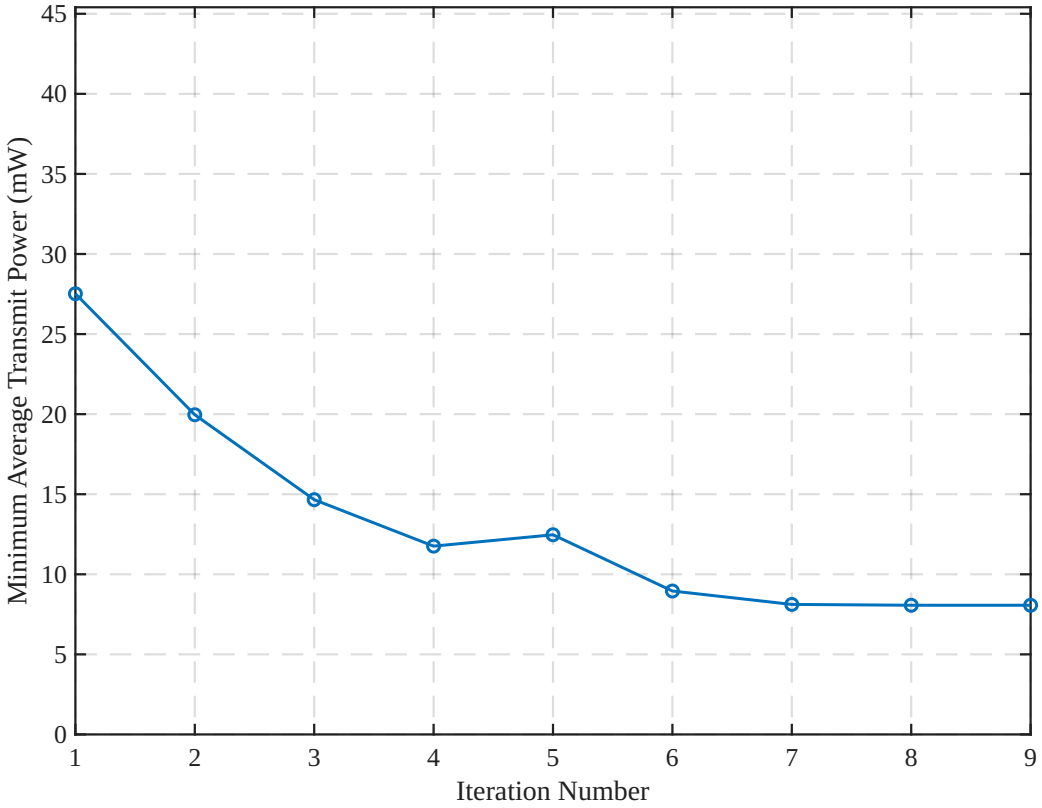

Figure 5.3: Convergence behavior of the proposed algorithm.

yields lower power at small horizons, at higher horizons, due to the limited packet length, much more power is required. This behavior results in a crossover point between the two curves.

Fig. 5.6 shows the minimum required transmit power as a function of the receiver's prediction horizon H_{Rx} for different values of the error threshold δ. The algorithm selects the smallest transmit power that satisfies the overall reliability constraint $\epsilon_{\max} = 10^{-5}$. For example, when $\delta = 0.1$, the maximum allowable H_{Rx} is 18, beyond which the reliability constraint is violated. Additionally, values of H_{Rx} below 12 do not meet the delay constraint and are therefore excluded by the algorithm. The optimal value of H_{Rx} in this case is 13, after which a slight increase in power is observed due to the need to compensate for the growing total prediction error. In fact, increasing the receiver's prediction horizon H_{Rx} provides more flexibility in aligning the communication resources, which allows the system to reduce the required transmit power. However, after a certain point, due to the increasing prediction error, additional transmit power is required to offset the increase in overall error probability and ensure the target reliability is maintained.

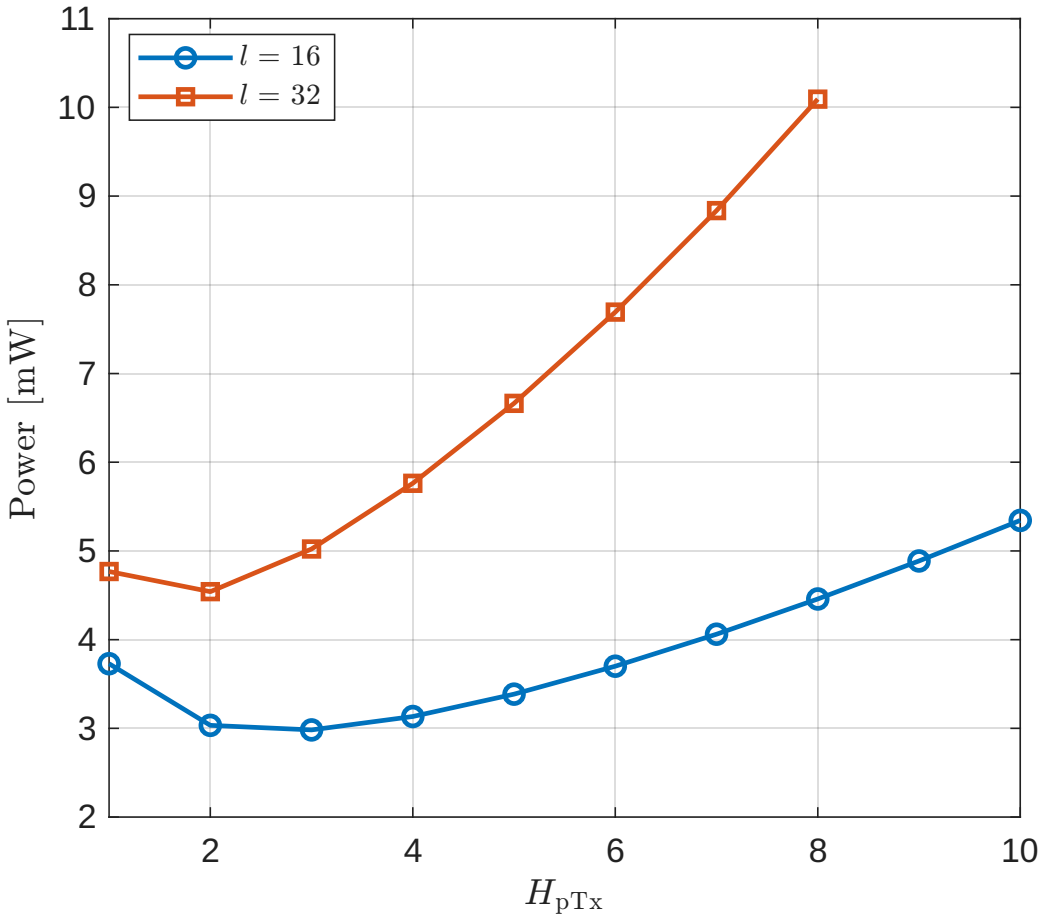

Figure 5.4: Transmit power versus transmitter prediction horizon at bandwidth $B = 100$ kHz.

Fig. 5.7 illustrates the optimal values of the transmitter and receiver prediction horizons as a function of transmit power. When the available power is low, the transmitter's prediction horizon H_{Tx} remains small, as it cannot accommodate a large number of samples per block while still satisfying the reliability constraint. In this regime, the receiver's predictor compensates for delay management. As the transmit power increases, H_{Tx} also increases, enabling more samples to be included in each transmission. However, beyond a certain power level, the horizons saturate and no longer increase. This plateau is attributed to the limited bandwidth, which becomes the dominant factor influencing the optimal horizon distribution.

The relationship between overall error probability and transmit power, based on the optimal values of the prediction horizons in Fig. 5.7, is illustrated in Fig. 5.8. The results show that increasing the transmit power leads to a rapid reduction in the overall error probability. In particular, higher

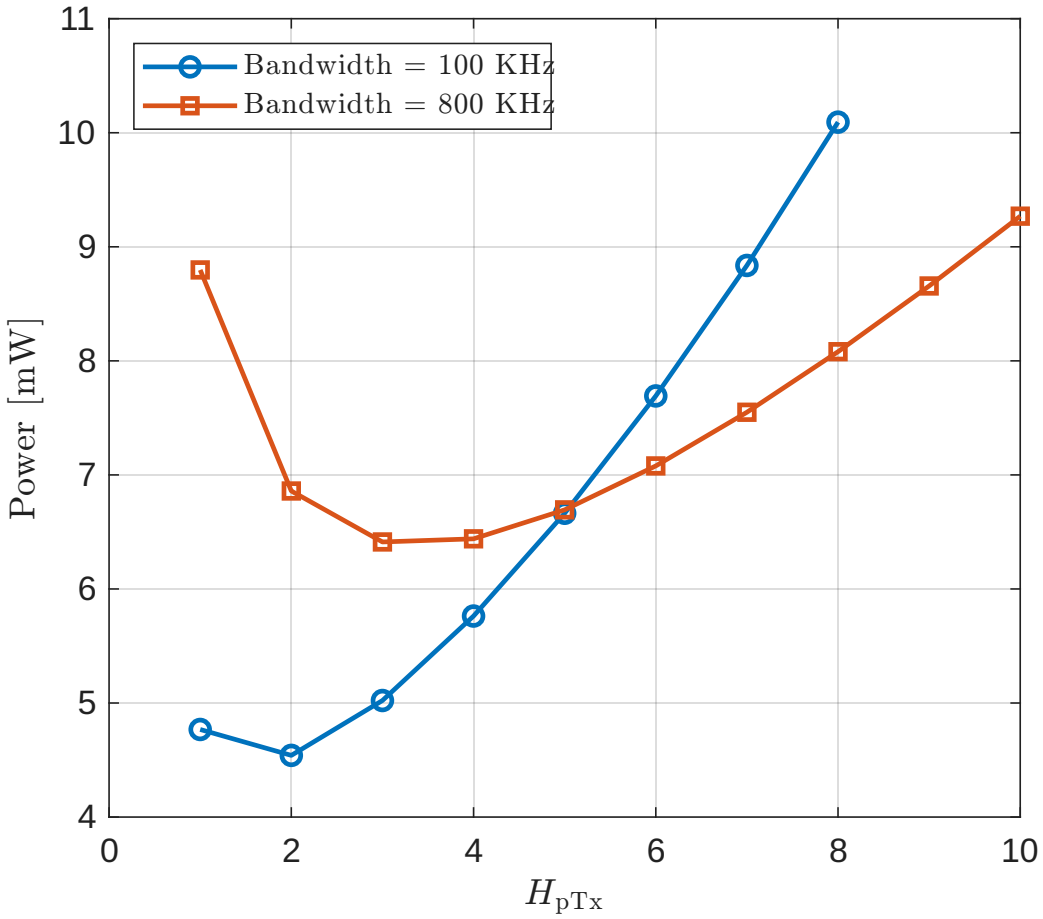

Figure 5.5: Transmit power versus transmitter prediction horizon for sample length $l = 32$ bits.

power allows for a larger transmitter prediction horizon H_{Tx}, which further accelerates the decline in ϵ_o due to the exponential form $(\epsilon_t + \epsilon_q)^{H_{\mathrm{Tx}}}$. Since both ϵ_t and ϵ_q are typically small, their compounded effect decreases sharply with increasing H_{Tx}. As a result, the system achieves significantly improved reliability with relatively modest increases in prediction horizon, until ϵ_o approaches the receiver-side prediction error ϵ_{Rx}, beyond which further improvements become negligible.

Moreover, Fig. 5.8 illustrates the overall error probability ϵ_o across various transmit power levels for two different bandwidth settings, comparing three scenarios: (i) without the prediction mechanism, (ii) with single-side (receiver-only) prediction, and (iii) the proposed DPS method. The results clearly demonstrate the effectiveness of the proposed DPS approach in improving overall reliability across a wide range of power levels. Compared to the baseline without prediction, the DPS scheme consistently achieves lower error probabilities, particularly in the moderate to high power

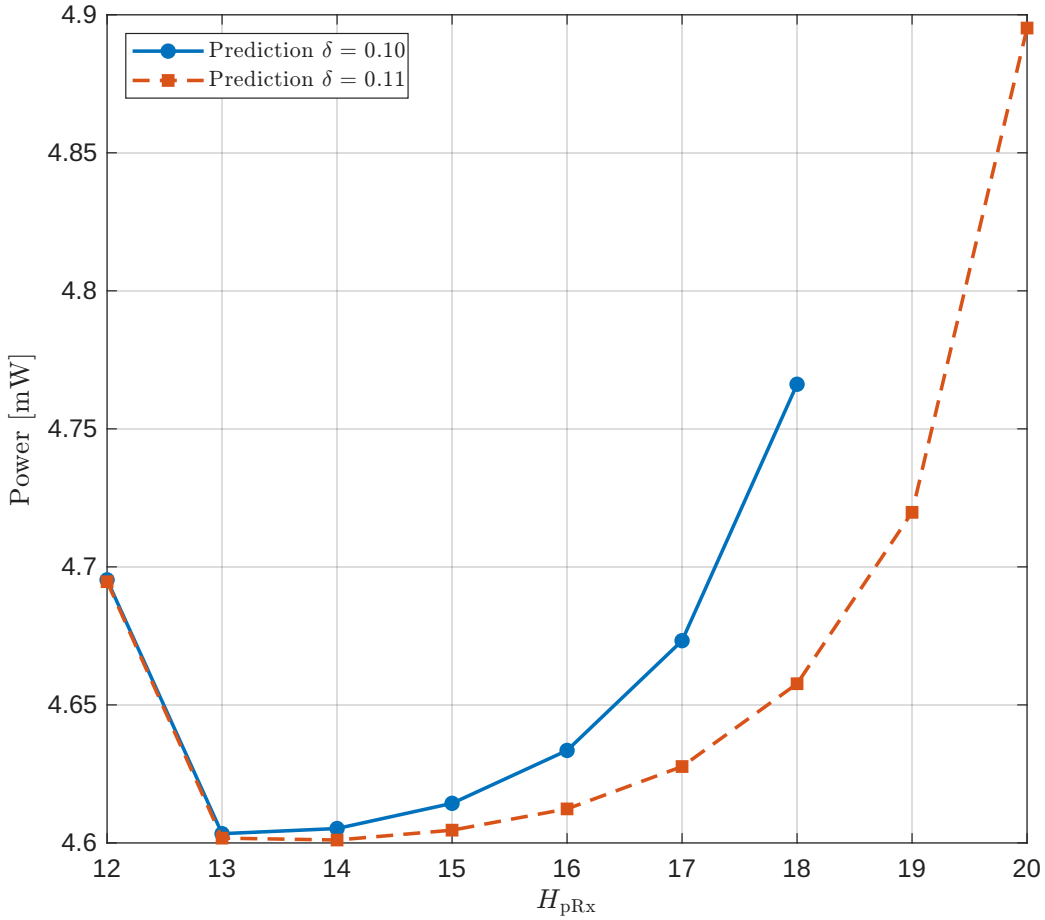

Figure 5.6: Transmit power versus receiver's prediction horizon at bandwidth $B = 100$ kHz.

regimes. The improvement becomes more pronounced at higher bandwidth levels, where the DPS method attains ultra-reliable performance. The gain over single-side prediction is also noticeable, especially when the available bandwidth is large. However, under constrained bandwidth conditions, the advantage of DPS becomes limited, as the short packet length limits the ability to include additional predicted samples. In fact, when both power and bandwidth are limited, the DPS scheme exhibits similar performance to single-side prediction, as it becomes infeasible to encapsulate more samples in a single packet. As the bandwidth or power increases, the difference between DPS and single-side prediction becomes substantial. These findings highlight that the proposed DPS scheme not only satisfies stringent reliability requirements for short-packet transmissions but also enables significant power savings by maintaining ultra-low error probabilities without requiring excessive transmit power.

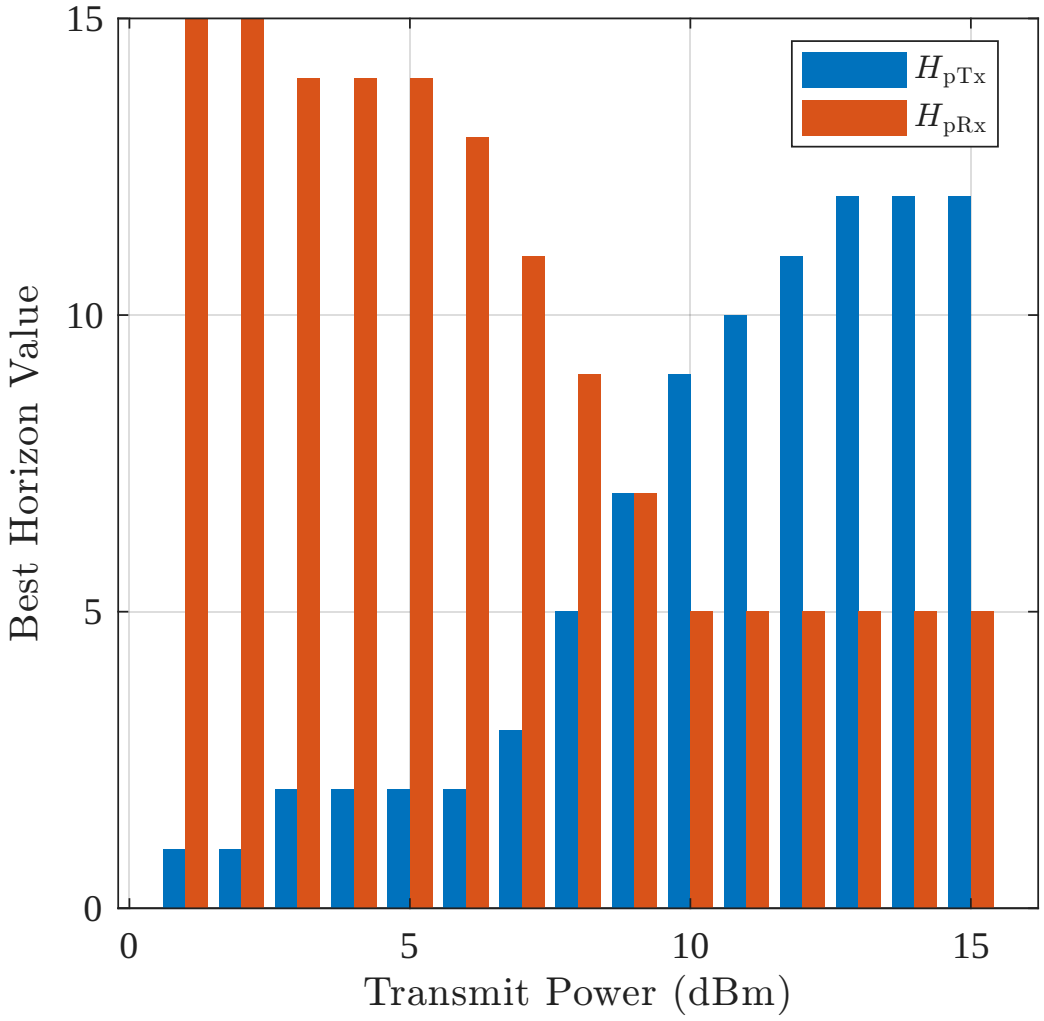

Figure 5.7: Optimal values of the transmitter and receiver prediction horizons vs. available transmit power at bandwidth $B = 100$ kHz and $l = 16$.

5.2 Simulation in Multiple-User Scenarios

In this subsection, we evaluate joint power control and resource allocation for a multi-user URLLC scenario involving 50 teleoperation systems, which are randomly distributed within a region ranging from 100 to 2000 meters.

Fig. 5.9 illustrates the minimal average transmit power versus the maximum available bandwidth for two different reliability requirements. As observed, the average transmit power decreases and eventually converges as the total available bandwidth increases. When bandwidth is limited (e.g., $B_{\max} = 2$ MHz), the minimal average transmit power rises significantly as the overall reliability

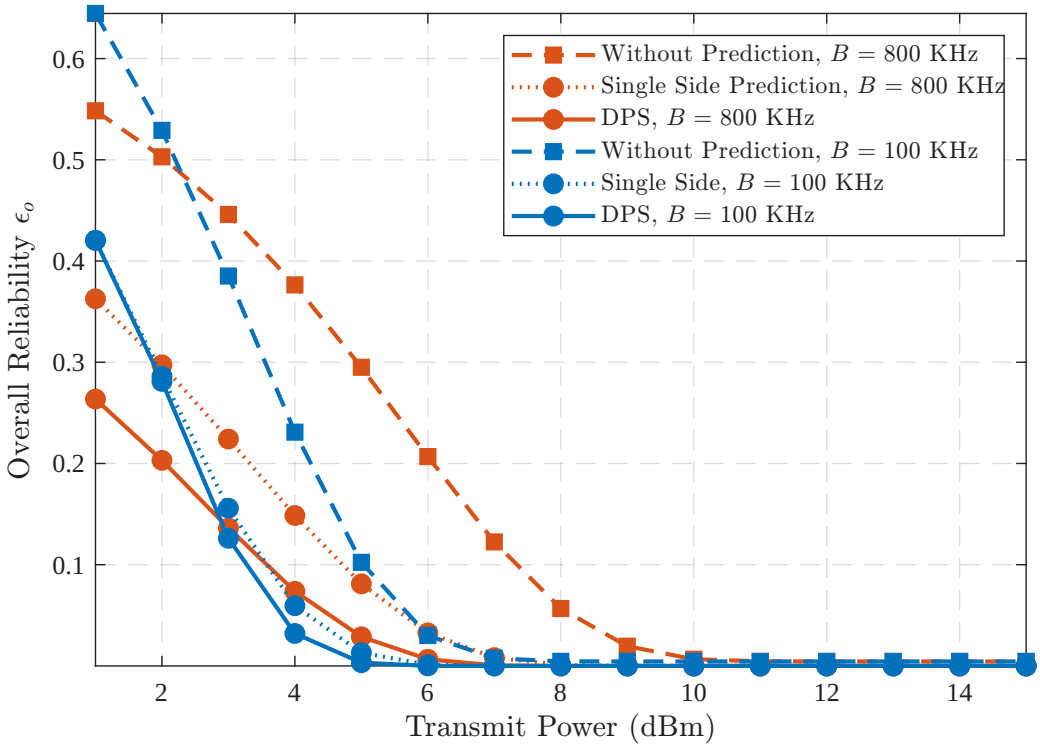

Figure 5.8: Comparison of reliability–power curves between the DPS-based method and the baseline without prediction and with single-side prediction.

requirement becomes more stringent (from 10^{-5} to 10^{-6}). However, when sufficient bandwidth is available ($B_{\max} \geq 12$ MHz), the gap in minimal average transmit power between different reliability thresholds becomes negligible. Moreover, a lower target reliability threshold consistently necessitates higher transmit power, emphasizing the trade-off between reliability and energy efficiency in URLLC systems.

In Fig. 5.10, the minimum average transmit power required for URLLC is compared between two scenarios: one employing the proposed DPS, and the other using prediction only at the receiver side. The results demonstrate that the proposed DPS significantly reduces the transmit power. Furthermore, as the available bandwidth increases beyond 4 MHz, the gap between the two scenarios becomes more pronounced. This is because, under high bandwidth conditions, DPS benefits from the ability to encapsulate more predicted samples in a single packet, thereby enhancing system efficiency. Eventually, the average transmit power converges to approximately 8 mW with DPS, compared to around 26 mW when prediction is performed only at the receiver.

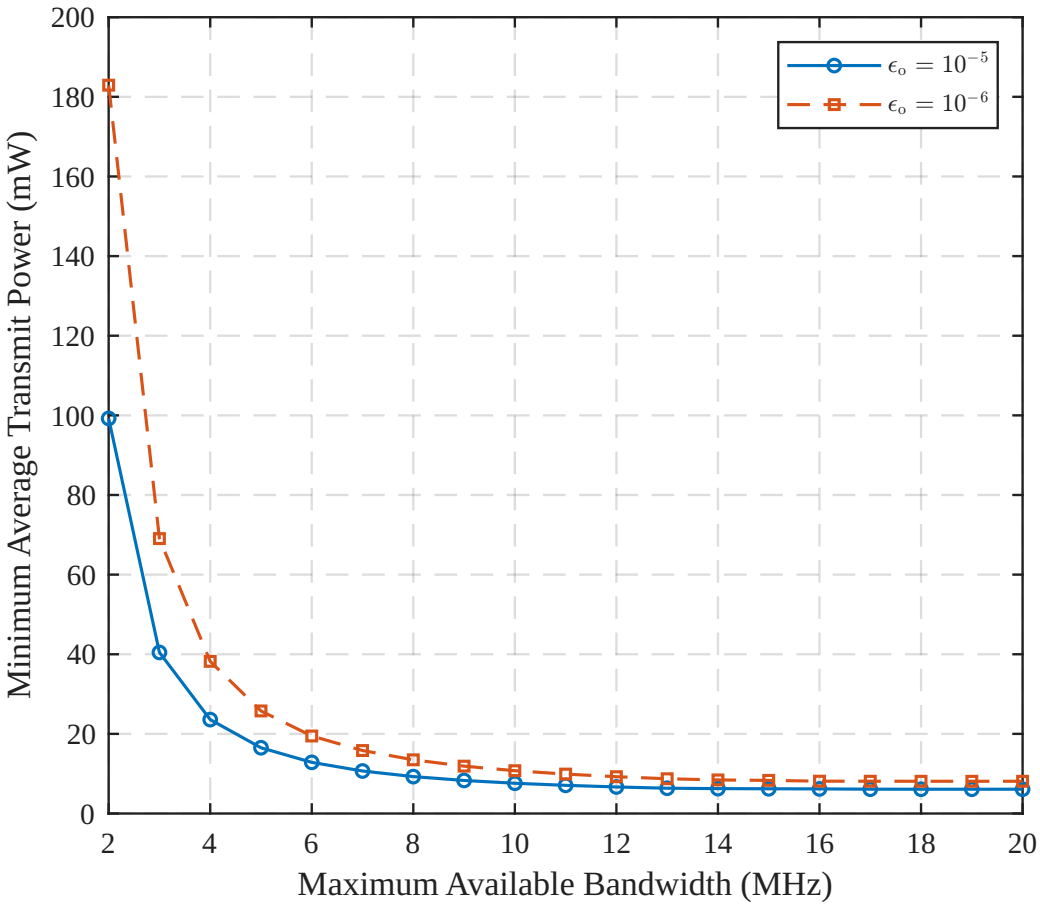

Figure 5.9: Minimum average transmit power versus maximum available bandwidth.

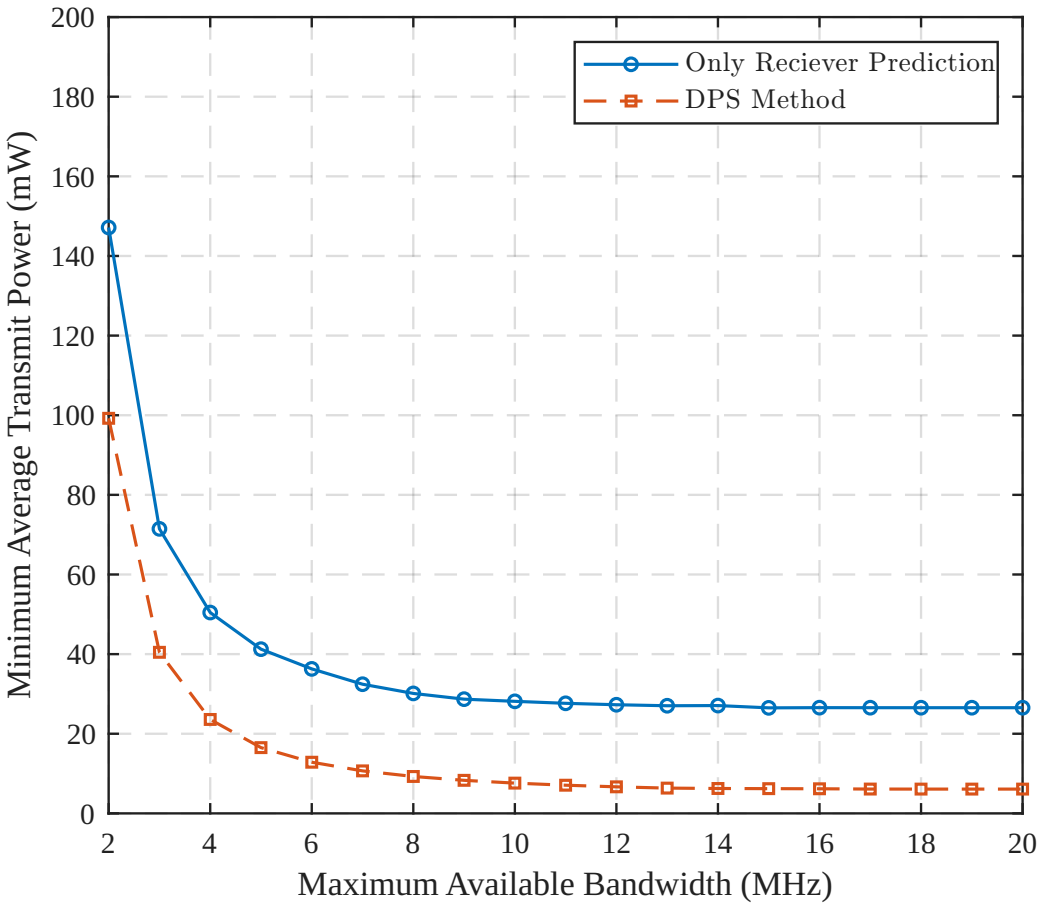

Figure 5.10: Comparison of the minimum average transmit power versus maximum available bandwidth between the proposed DPS-based method and the baseline scheme with prediction applied at only one side.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

This thesis investigated how prediction can be integrated with wireless resource allocation to support URLLC in TI teleoperation systems. We considered a multi-user scenario in which several operator–teleoperator pairs share a wireless URLLC link and must meet stringent constraints on experienced delay and reliability. Building on finite-blocklength information theory and queueing-delay analysis, we developed a model that captures both the total error probability and the E2E latency experienced in the teleoperation system.

Within this framework, we proposed a Dual Prediction Scheme (DPS) in which both the transmitter (master side) and receiver (slave side) run complementary prediction algorithms. Transmitter-side prediction generates future control commands and encapsulates them into a single packet. Receiver-side prediction, in turn, reconstructs missing or delayed samples, which preserves continuity of the haptic/control signals in case packets are lost. The combined effect is to satisfy URLLC constraint , while optimizing transmit power and bandwidth allocation.

We formulated an optimization problem that minimizes the average transmit power subject to constraints on experienced delay, finite-blocklength reliability, and total bandwidth. The decision variables include the prediction horizons at both ends and the bandwidth allocated to each teleoperation pair, so that prediction is explicitly co-designed with wireless resource allocation. Our

results show that relying on single-side prediction (either transmitter-only or receiver-only) is fundamentally suboptimal. With transmitter-only prediction, the receiver cannot reconstruct the state when packets are lost or excessively delayed, and the prediction horizon at the transmitter cannot be increased arbitrarily without violating reliability constraints in the finite blocklength regime. In contrast, the proposed DPS achieves the same delay–reliability targets with significantly lower average transmit power.

Simulation results confirmed these insights. Compared to baseline schemes without prediction or with transmitter-only prediction, the DPS-based design significantly reduces the required transmit power for a given delay–reliability target. The results also revealed clear trade-offs between prediction horizons, bandwidth allocation, and reliability. We showed that a carefully chosen pair of horizons, jointly optimized with bandwidth, achieves substantial power savings without violating URLLC constraints. These findings provide concrete design guidelines for prediction-aware URLLC teleoperation, showing that prediction is most effective when co-designed with resource allocation rather than treated as separate systems.

6.2 Future Work

This thesis establishes a first optimization framework for dual-side prediction in TI teleoperation under URLLC constraints. However, several extensions remain open and merit further investigation:

- **Realistic bilateral teleoperation communication.** The current analysis focuses on a simplified traffic model with a one-way teleoperation loop. In real remote teleoperation, however, communication is inherently bilateral: the operator sends control commands while simultaneously receiving high-volume feedback streams (video, audio, and haptic data). This bidirectional coupling makes the delay constraints significantly more stringent. For maximum transparency, the round-trip delay between an operator's action and the corresponding feedback should ideally be imperceptible; any additional latency or jitter degrades transparency and can induce cybersickness. In practice, this delay cannot be reduced to zero, so prediction mechanisms become essential for anticipating future states and partially masking network latency. A natural next step is therefore to incorporate detailed models of haptic, audio, and

video traffic into the bilateral teleoperation framework. This would enable a tighter co-design of communication, control, and prediction to enhance transparency, stability margins, and human perception thresholds.

- **Using prediction to enable joint URLLC–eMBB support in bilateral teleoperation.** In realistic bilateral teleoperation, both URLLC and eMBB services are required: URLLC for time-critical control and haptic feedback, and eMBB for high-rate video and audio streams. A natural extension of this work is to exploit prediction in managing heterogeneous traffic where URLLC and eMBB flows coexist and share the same resources. This would enable the study of joint resource allocation and prediction mechanisms that strictly satisfy URLLC constraints while still providing sufficient throughput for eMBB traffic.

- **Two-way prediction and adaptive resource allocation.** As future work, the proposed framework can be extended to a two-way prediction scheme in which the predictors also forecast network conditions and dynamically adapt resource allocation. The associated decision-making can be realized by training a Deep Reinforcement Learning (DRL) agent that leverages expert knowledge together with real-time side information, such as channel state information (CSI), to jointly optimize prediction thresholds and resource usage under URLLC constraints.

- **Online learning of prediction models and horizons.** In this work, the prediction models are assumed to be known and optimized offline. In practice, model mismatch and time-varying dynamics can reduce prediction accuracy. Future work could investigate online learning schemes that adapt the prediction models based on observed tracking error, delay, and channel conditions. This points toward reinforcement-learning or adaptive control approaches that jointly tune prediction and resource allocation in real time.

- **Robustness to uncertainty and model errors.** The current optimization assumes that channel statistics are accurately known. In reality, channel conditions, traffic patterns, and even human operator behavior can be highly uncertain. A research direction is to develop robust formulations that explicitly account for uncertainty in these models, ensuring that URLLC

constraints are satisfied with high probability even under mismatched prediction or unexpected traffic bursts.

Addressing these directions will deepen our understanding of prediction-aware URLLC design and broaden the applicability of dual prediction schemes beyond teleoperation to other time-sensitive 6G services, such as cooperative robotics, extended reality, and cyber-physical production systems.

Bibliography

[1] M. E. Haque, F. Tariq, M. R. Khandaker, M. S. Hossain, M. A. Imran, and K.-K. Wong, "A comprehensive survey of 5g urllc and challenges in the 6g era," *arXiv preprint arXiv:2508.20205*, 2025.

[2] M. Simsek, A. Aijaz, M. Dohler, J. Sachs, and G. Fettweis, "5g-enabled tactile internet," *IEEE Journal on selected areas in communications*, vol. 34, no. 3, pp. 460–473, 2016.

[3] N. A. Mohammedali, T. Kanakis, and M. O. Agyeman, "Survey on the impact of ai, robotics and 6g networks on the remote surgery," in *2024 International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*. IEEE, 2024, pp. 1–6.

[4] B. S. Chaudhari, "Enabling tactile internet via 6g: Application characteristics, requirements, and design considerations," *Future Internet*, vol. 17, no. 3, p. 122, 2025.

[5] S. Bera, H. Das, S. Nayak, and R. Patgiri, "Future tactile internet: issues, challenges and applications," in *2021 6th International Conference on Signal Processing, Computing and Control (ISPCC)*. IEEE, 2021, pp. 625–630.

[6] W. Stallings, *5G Wireless: A Comprehensive Introduction*. Pearson, 2021.

[7] W. Huang, S. Xiao, L. Wu, C. Kai, S. He, and C. Li, "Achievable rate region for urllc interference channel with finite blocklength transmission," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 7, pp. 8857–8868, 2023.

[8] M. Shirvanimoghaddam, M. S. Mohammadi, R. Abbas, A. Minja, C. Yue, B. Matuz, G. Han, Z. Lin, W. Liu, Y. Li *et al.*, "Short block-length codes for ultra-reliable low latency communications," *IEEE Communications Magazine*, vol. 57, no. 2, pp. 130–137, 2018.

[9] W. Cheng, Y. Xiao, S. Zhang, and J. Wang, "Adaptive finite blocklength for ultra-low latency in wireless communications," *IEEE Transactions on Wireless Communications*, vol. 21, no. 6, pp. 4450–4463, 2021.

[10] M. Akbari and F. Ashtiani, "Distribution of the peak age of information in a random access relay network," *IEEE Transactions on Vehicular Technology*, vol. 73, no. 5, pp. 7262–7275, 2023.

[11] Z. Hou, C. She, Y. Li, L. Zhuo, and B. Vucetic, "Prediction and communication co-design for ultra-reliable and low-latency communications," *IEEE Transactions on Wireless Communications*, vol. 19, no. 2, pp. 1196–1209, 2019.

[12] Z. Meng, C. She, G. Zhao, and D. De Martini, "Sampling, communication, and prediction co-design for synchronizing the real-world device and digital model in metaverse," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 1, pp. 288–300, 2022.

[13] F. Wu, Y. Chen, X. Chen, W. Fan, and Y. Liu, "An adaptive dual prediction scheme based on edge intelligence," *IEEE Internet of Things Journal*, vol. 7, no. 10, pp. 9481–9493, 2020.

[14] T. Shu, J. Chen, V. K. Bhargava, and C. W. de Silva, "An energy-efficient dual prediction scheme using lms filter and lstm in wireless sensor networks for environment monitoring," *IEEE Internet of Things Journal*, vol. 6, no. 4, pp. 6736–6747, 2019.

[15] R. Kouki, A. Boe, T. Vantroys, and F. Bouani, "Autonomous internet of things predictive control application based on wireless networked multi-agent topology and embedded operating system," *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering*, vol. 234, no. 5, pp. 577–595, 2020.

[16] Z. Pang, C. Bai, G. Liu, Q. Han, and X. Zhang, "A novel networked predictive control method for systems with random communication constraints," *Journal of Systems Science and Complexity*, vol. 34, no. 4, pp. 1364–1378, 2021.

[17] X. Tong, G. Zhao, M. A. Imran, Z. Pang, and Z. Chen, "Minimizing wireless resource consumption for packetized predictive control in real-time cyber physical systems," in *2018 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, 2018, pp. 1–6.

[18] N. Alliance, "Verticals urllc use cases and requirements," *NGMN Alliance*, 2019.

[19] S. Husain, A. Kunz *et al.*, "3gpp 5g core network: An overview and future directions." *Journal of Information & Communication Convergence Engineering*, vol. 20, no. 1, 2022.

[20] J. M. C. Brito, "Trends in wireless communications towards 5g networks—the influence of e-health and iot applications," in *2016 International Multidisciplinary Conference on Computer and Energy Science (SpliTech)*. IEEE, 2016, pp. 1–7.

[21] Y. Ding, S. Wang, R. Lan, W. Lin, X. Liu, and W. He, "Telerobotic surgery: a comprehensive two-decade evolution and the integration of emerging technologies," *International Journal of Surgery*, pp. 1081–1097.

[22] C. She, C. Yang, and T. Q. Quek, "Cross-layer optimization for ultra-reliable and low-latency radio access networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 1, pp. 127–141, 2017.

[23] H. Ji, S. Park, J. Yeo, Y. Kim, J. Lee, and B. Shim, "Ultra-reliable and low-latency communications in 5g downlink: Physical layer aspects," *IEEE Wireless Communications*, vol. 25, no. 3, pp. 124–130, 2018.

[24] N. A. Johansson, Y.-P. E. Wang, E. Eriksson, and M. Hessler, "Radio access for ultra-reliable and low-latency 5g communications," in *2015 IEEE International Conference on Communication Workshop (ICCW)*. IEEE, 2015, pp. 1184–1189.

[25] G. Durisi, T. Koch, and P. Popovski, "Toward massive, ultrareliable, and low-latency wireless communication with short packets," *Proceedings of the IEEE*, vol. 104, no. 9, pp. 1711–1726, 2016.

[26] I. Budhiraja, S. Tyagi, S. Tanwar, N. Kumar, and J. J. Rodrigues, "Tactile internet for smart communities in 5g: An insight for noma-based solutions," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 5, pp. 3104–3112, 2019.

[27] M. O. Ernst and M. S. Banks, "Humans integrate visual and haptic information in a statistically optimal fashion," *Nature*, vol. 415, no. 6870, pp. 429–433, 2002.

[28] R. Gupta, S. Tanwar, S. Tyagi, and N. Kumar, "Tactile-internet-based telesurgery system for healthcare 4.0: An architecture, research challenges, and future directions," *IEEE network*, vol. 33, no. 6, pp. 22–29, 2019.

[29] C. Xu, H. Yu, P. Zeng, and Y. Li, "Towards critical industrial wireless control: Prototype implementation and experimental evaluation on urllc," *IEEE Communications Magazine*, vol. 61, no. 9, pp. 193–199, 2023.

[30] E. Mittag, R. Klose, J. Herbst, M. Rüb, J. Petershans, C. König, M. Hollick, and H. D. Schotten, "Introducing a 6g-enabled multi-connectivity robotic teleoperation platform," in *2025 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, 2025, pp. 1851–1857.

[31] S. Hassouna, J. Kaur, B. Kizilkaya, J. u. R. Kazim, S. Ansari, A. A. Kherani, B. Lall, Q. H. Abbasi, and M. Imran, "Development of open radio access networks (o-ran) for real-time robotic teleoperation," *Communications Engineering*, vol. 4, no. 1, p. 176, 2025.

[32] N. Golmohammadi, M. M. Rayguru, and S. Baidya, "Lyapunov-optimized 5g-sliced communications for telerobotic applications," in *IEEE INFOCOM 2024-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE, 2024, pp. 1–6.

[33] A. M. Ibrahim, R. Nordin, Y. S. Khamayseh, A. Amphawan, and M. B. Jasser, "Urllc for 6g enabled industry 5.0: A taxonomy of architectures, cross layer techniques, and time critical applications," *arXiv preprint arXiv:2510.08080*, 2025.

[34] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2307–2359, 2010.

[35] Y. Yang, C. Chen, and P. Zhu, "Enhanced codebook of sparse vector coding based on mean-variance trade-off model for urllc," *IEEE Communications Letters*, 2025.

[36] C. Zheng, F.-C. Zheng, J. Luo, P. Zhu, X. You, and D. Feng, "Mini-slot-assisted short packet urllc: Differential or coherent detection?" *arXiv preprint arXiv:2408.14089*, 2024.

[37] C. Zheng, F.-C. Zheng, and J. Luo, "Frequency domain differential modulation for mini-slot-assisted short packet urllc," in *2025 IEEE 101st Vehicular Technology Conference (VTC2025-Spring)*. IEEE, 2025, pp. 1–5.

[38] P. Popovski, Č. Stefanović, J. J. Nielsen, E. De Carvalho, M. Angjelichinoski, K. F. Trillingsgaard, and A.-S. Bana, "Wireless access in ultra-reliable low-latency communication (urllc)," *IEEE Transactions on Communications*, vol. 67, no. 8, pp. 5783–5801, 2019.

[39] T. K. Vu, C.-F. Liu, M. Bennis, M. Debbah, M. Latva-Aho, and C. S. Hong, "Ultra-reliable and low latency communication in mmwave-enabled massive mimo networks," *IEEE Communications Letters*, vol. 21, no. 9, pp. 2041–2044, 2017.

[40] 3GPP, "Study on New Radio (NR) access technology (Release 16)," 3rd Generation Partnership Project (3GPP), TR 38.912, Jul. 2020, available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3107.

[41] M. U. A. Siddiqui, H. Abumarshoud, L. Bariah, S. Muhaidat, M. A. Imran, and L. Mohjazi, "Urllc in beyond 5g and 6g networks: An interference management perspective," *IEEe Access*, vol. 11, pp. 54 639–54 663, 2023.

[42] M. E. Haque, F. Tariq, M. R. 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. 34 372–34 396, 2023.

[43] W. Zhang, M. Derakhshani, G. Zheng, C. S. Chen, and S. Lambotharan, "Bayesian optimization of queuing-based multichannel urllc scheduling," *IEEE Transactions on Wireless Communications*, vol. 22, no. 3, pp. 1763–1778, 2022.

[44] 3GPP, "NR; Overall description; Stage-2 (Release 15)," 3rd Generation Partnership Project (3GPP), TS 38.300, Jun. 2018, available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3196.

[45] N. Makondo, H. I. Kobo, T. E. Mathonsi, D. Du Plessis, T. M. Makhosa, and L. Mamushiane, "An efficient architecture for latency optimisation in 5g using edge computing for urllc use cases," in *2024 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)*. IEEE, 2024, pp. 1–7.

[46] G. Li and J. Cai, "An online incentive mechanism for collaborative task offloading in mobile edge computing," *IEEE Transactions on Wireless Communications*, vol. 19, no. 1, pp. 624–636, 2019.

[47] T. Kim, G. Noh, J. Kim, H. Chung, and I. Kim, "Enhanced resource allocation method for 5g v2x communications," in *2021 International Conference on Information and Communication Technology Convergence (ICTC)*. IEEE, 2021, pp. 621–623.

[48] A. Pradhan, S. Das, and M. J. Piran, "Blocklength optimization and power allocation for energy-efficient and secure urllc in industrial iot," *IEEE Internet of Things Journal*, vol. 11, no. 6, pp. 9420–9431, 2023.

[49] J. Xu, K. Li, Y. Chen, and J. Huang, "Optimal task scheduling and resource allocation for self-powered sensors in internet of things: An energy efficient approach," *IEEE Transactions on Network and Service Management*, vol. 21, no. 4, pp. 4410–4420, 2024.

[50] S. B. Prathiba, K. Raja, R. Saiabirami, and G. Kannan, "An energy-aware tailored resource management for cellular-based zero-touch deterministic industrial m2m networks," *IEEE Access*, vol. 12, pp. 33 613–33 627, 2024.

[51] Q. Wu, W. Wang, P. Fan, Q. Fan, J. Wang, and K. B. Letaief, "Urllc-awared resource allocation for heterogeneous vehicular edge computing," *IEEE Transactions on Vehicular Technology*, vol. 73, no. 8, pp. 11 789–11 805, 2024.

[52] B. Adhikari, M. Jaseemuddin, and A. Anpalagan, "Resource allocation for co-existence of embb and urllc services in 6g wireless networks: A survey," *IEEE Access*, vol. 12, pp. 552–581, 2023.

[53] K. Chen, Y. Wang, J. Zhao, X. Wang, and Z. Fei, "Urllc-oriented joint power control and resource allocation in uav-assisted networks," *IEEE Internet of Things Journal*, vol. 8, no. 12, pp. 10 103–10 116, 2021.

[54] K. Li, P. Zhu, Y. Wang, F.-C. Zheng, and X. You, "Joint uplink and downlink resource allocation toward energy-efficient transmission for urllc," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 7, pp. 2176–2192, 2023.

[55] M. Bennis, M. Debbah, and H. V. Poor, "Ultrareliable and low-latency wireless communication: Tail, risk, and scale," *Proceedings of the IEEE*, vol. 106, no. 10, pp. 1834–1853, 2018.

[56] B. Kizilkaya, C. She, G. Zhao, and M. A. Imran, "Task-oriented prediction and communication co-design for haptic communications," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 7, pp. 8987–9001, 2023.

[57] C. She, R. Dong, Z. Gu, Z. Hou, Y. Li, W. Hardjawana, C. Yang, L. Song, and B. Vucetic, "Deep learning for ultra-reliable and low-latency communications in 6g networks," *IEEE network*, vol. 34, no. 5, pp. 219–225, 2020.

[58] Z. Liu, X. Chen, H. Wu, Z. Wang, X. Chen, D. Niyato, and K. Huang, "Integrated sensing and edge ai: Realizing intelligent perception in 6g," *IEEE Communications Surveys & Tutorials*, 2025.

[59] J. Zhao, C. Liu, J. Liao, and D. Wang, "Deep learning in wireless communications for physical layer," *Physical Communication*, vol. 67, p. 102503, 2024.

[60] C. She, C. Sun, Z. Gu, Y. Li, C. Yang, H. V. Poor, and B. Vucetic, "A tutorial on ultrareliable and low-latency communications in 6g: Integrating domain knowledge into deep learning," *Proceedings of the IEEE*, vol. 109, no. 3, pp. 204–246, 2021.

[61] A. Aijaz, N. Jiang, and A. Khan, "Toward multi-service edge-intelligence paradigm: Temporal-adaptive prediction for time-critical control over wireless," *IEEE Internet of Things Magazine*, vol. 6, no. 1, pp. 96–101, 2023.

[62] F. Boabang, A. Ebrahimzadeh, R. H. Glitho, H. Elbiaze, M. Maier, and F. Belqasmi, "A machine learning framework for handling delayed/lost packets in tactile internet remote robotic surgery," *IEEE Transactions on Network and Service Management*, vol. 18, no. 4, pp. 4829–4845, 2021.

[63] X. Hou and S. Dey, "Motion prediction and pre-rendering at the edge to enable ultra-low latency mobile 6dof experiences," *IEEE Open Journal of the Communications Society*, vol. 1, pp. 1674–1690, 2020.

[64] N. Sakr, N. D. Georganas, J. Zhao, and X. Shen, "Motion and force prediction in haptic media," in *2007 IEEE International Conference on Multimedia and Expo*. IEEE, 2007, pp. 2242–2245.

[65] S. Wang, M. Wang, Y. Wu, and G. Fang, "Realize ultra-reliability and low-latency in haptic communication through prediction," in *2021 13th International Conference on Wireless Communications and Signal Processing (WCSP)*. IEEE, 2021, pp. 1–5.

[66] B. Yu, Y. Cai, Y. Zou, B. Li, and Y. Chen, "Can we improve the information freshness with prediction for cognitive iot?" *IEEE Internet of Things Journal*, vol. 9, no. 18, pp. 17 577–17 591, 2022.

[67] F. Boabang, R. Glitho, H. Elbiaze, F. Belqami, and O. Alfandi, "A framework for predicting haptic feedback in needle insertion in 5g remote robotic surgery," in *2020 IEEE 17th Annual Consumer Communications & Networking Conference (CCNC)*. IEEE, 2020, pp. 1–6.

[68] Z. Hou, C. She, Y. Li, T. Q. Quek, and B. Vucetic, "Burstiness-aware bandwidth reservation for ultra-reliable and low-latency communications in tactile internet," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 11, pp. 2401–2410, 2018.

[69] M. Li, X. Guan, C. Hua, C. Chen, and L. Lyu, "Predictive pre-allocation for low-latency uplink access in industrial wireless networks," in *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*. IEEE, 2018, pp. 306–314.

[70] B. Makki, T. Svensson, G. Caire, and M. Zorzi, "Fast harq over finite blocklength codes: A technique for low-latency reliable communication," *IEEE Transactions on Wireless Communications*, vol. 18, no. 1, pp. 194–209, 2018.

[71] N. Strodthoff, B. Göktepe, T. Schierl, C. Hellge, and W. Samek, "Enhanced machine learning techniques for early harq feedback prediction in 5g," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 11, pp. 2573–2587, 2019.

[72] Y. Yeh, S. N. Özsert, D. Prattichizzo, V. S. Varma, and S. E. Elayoubi, "Data-driven context-aware traffic prediction and modeling for tactile internet," 2025.

[73] R. Ali, Y. B. Zikria, A. K. Bashir, S. Garg, and H. S. Kim, "Urllc for 5g and beyond: Requirements, enabling incumbent technologies and network intelligence," *IEEE Access*, vol. 9, pp. 67 064–67 095, 2021.

[74] Y. Jiang, X. Zhang, X. Zhong, and S. Zhou, "A dynamic resource scheduling algorithm based on traffic prediction for coexistence of embb and random arrival urllc," in *2024 IEEE 24th International Conference on Communication Technology (ICCT)*. IEEE, 2024, pp. 253–258.

[75] H. Liu, G. Li, X. Li, Y. Liu, G. Huang, and Z. Ding, "Effective capacity analysis of star-ris-assisted noma networks," *IEEE Wireless Communications Letters*, vol. 11, no. 9, pp. 1930–1934, 2022.

[76] Q. Xiong, X. Zhu, Y. Jiang, J. Cao, X. Xiong, and H. Wang, "Status prediction and data aggregation for aoi-oriented short-packet transmission in industrial iot," *IEEE Transactions on Communications*, vol. 71, no. 1, pp. 611–625, 2022.

[77] G. Sun, F. Baccelli, K. Feng, L. U. Garcia, and S. Paris, "A stochastic geometry framework for performance analysis of ris-assisted ofdm cellular networks," *IEEE Transactions on Wireless Communications*, 2025.

[78] A. A. Shamsabadi, A. Yadav, Y. Gadallah, and H. Yanikomeroglu, "Exploring the 6g potentials: Immersive, hyperreliable, and low-latency communication," *IEEE Vehicular Technology Magazine*, 2025.

[79] M. Ghous, T. L. Nguyen, G. Kaddoum *et al.*, "Deep transfer learning-based performance prediction of urllc in independent and not necessarily identically distributed interference networks," *IEEE Access*, vol. 12, pp. 99 071–99 093, 2024.

[80] G. M. Dias, B. Bellalta, and S. Oechsner, "The impact of dual prediction schemes on the reduction of the number of transmissions in sensor networks," *Computer Communications*, vol. 112, pp. 58–72, 2017.

[81] Z. Hou, C. She, Y. Li, D. Niyato, M. Dohler, and B. Vucetic, "Intelligent communications for tactile internet in 6g: Requirements, technologies, and challenges," *IEEE Communications Magazine*, vol. 59, no. 12, pp. 82–88, 2021.

[82] S. M. Kay, *Fundamentals of statistical signal processing: estimation theory*. Prentice-Hall, Inc., 1993.