INSTITUTE FOR ADVANCED ECE RESEARCH
RESEARCH ARTICLE

Recent Advances in Next-Generation Wireless Communication Systems
ISSN: 3139-2156 (O), Vol. 1, No. 2, May - Aug 2026 (pp.35-43)

AI-Driven Resource Allocation for Ultra-Reliable Low-Latency Communication in Next-Generation Wireless Systems

Ahmed Al-Zawi

Technological Projects Department, Libyan Center for Remote Sensing and Space Science, Tripoli 21821, Libya

KEYWORDS:

URLLC,
AI-based resource allocation,
latency optimization,
6G networks,
reinforcement learning,
wireless communication

ARTICLE HISTORY:

Submitted : 18.10.2025
Revised : 16.11.2025
Accepted : 08.12.2025

DOI:
https://doi.org/10.17051/RANGWCS/01.02.05

ABSTRACT

Ultra-Reliable Low-Latency Communication (URLLC) has become an important category of services in future 5G and upcoming 6G wireless systems as it supports mission-critical services including autonomous driving, remote surgery, and industrial automation. Nevertheless, stringent latency requirements (sub-millisecond) and ultra-high reliability are a major challenge when using dynamical and heterogeneous networks. Traditional resource allocation algorithms, which are based on either fixed or heuristic based strategies, are not always efficient in changing traffic and channel conditions as it swings swiftly and that is why they achieve suboptimal performance. To overcome this issue, the present paper suggests an AI-based adaptive resource orchestration model, based on machine learning, reinforcement learning, which can be dynamic in optimising network resources online. The proposed strategy is capable of learning continuously based on network states such as channel conditions, traffic demand, and queue dynamics, and making intelligent allocation decisions based on such information in order to minimise latency, having to accept reliability constraints. The framework enables the network to be more responsive to changes in the network operations due to the inclusion of predictive and adaptive, unlike the conventional techniques. Its performance assessment shows that the AI-based scheme proposed results in a considerable decrease of end-to-end latency as compared with traditional allocation methods, especially when the traffic is heavy and dynamic. The findings underscore the ability of smart resource management in fulfilling the URLLC demands. It highlights the possibilities of AI-based optimization as one of the opportunities of the next-generation wireless systems, and opens the path to scalable, efficient, and reliable next-generation communication in 6G networks.

Author's e-mail: Ahm.al.za@yahoo.com

How to cite this article: Al-Zawi A, AI-Driven Resource Allocation for Ultra-Reliable Low-Latency Communication in Next-Generation Wireless Systems, Recent Advances in Next-Generation Wireless Communication Systems , Vol. 1, No. 2, 2026 (pp.35-43).

1. INTRODUCTION

The combination of connected massive number of devices, data intensive applications together with mission-essential services have led to faster developments of wireless communication systems since the fifth-generation (5G) up to sixth-generation (6G) paradigm. One of the service categories previously established in 5G and proposed in 6G, Ultra-Reliable Low-Latency Communication (URLLC) has become an essential enabler of applications of the next generation that demand high Quality of Service (QoS) requirements. URLLC can be used in situations with latency-sensitive and reliability-conscious application like autonomous driving, remote robotic surgery, smart manufacturing, industrial automation as well as in the context of tactile internet services as even the smallest delays during communication or loss of packets can have disastrous implications [2], [9]. In contrast to traditional communication services, URLLC places very strong performance requirements, which usually demand end-to-end latency of less than 1 ms and reliabilities of more than 99.999%. The very nature of stochastic wireless channels, changes in traffic patterns, and scarced radio resources makes it difficult to accomplish these two tasks. In practise, the latency depends on several factors such as transmission delay, queuing delay, processing delay and propagation delay and these factors all depend on the network congestion strategy and the resource

allocation strategies. This consequently makes having an efficient management of resources to be a crucial ingredient to fulfilling the requirements of URLLC [3], [13].

Old resource allocation schemes in wire network are based on either optimization based or heuristics, and these schemes are commonly developed on the basis of either a static or semi-static assumption. Although such approaches can reasonably perform in a steady environment, it does not have the ability to adapt effectively to the quickly changing network conditions. It is in extremely dynamic environments, like in metropolitan areas or industrial Internet of Things, that the static allocation schemes cannot respond to the changes in channel state information and user locomotion or traffic demand, resulting in a higher latency and worse reliability [5], [8]. Also, 6G networks become more heterogeneous, integrating technologies, like edge computing, network slicing, and intelligent reflecting surfaces, which further complicates the resource allocation problem [1]. Artificial Intelligence (AI) in this domain has come in as an innovative technology to empower smarter (intelligent) and adaptive wireless communication systems. The AI-based methods, especially machine learning (ML) and reinforcement learning (RL), provide the opportunity to identify complex network can be modelled, optimal policies can be learned based on information, and real-time decisions do not have to be made using explicit mathematical models. These methods have the capability of dynamically managing the allocation of resources by forecasting the network conditions and optimising performance metrics like latency, reliability and spectral efficiency [11]. The reinforcement learning more specifically can be very effective in issues surrounding resource allocation because the reinforcement learning allows agents to act in the environment and optimise decision-making policies through repeated interactions with the environment based on the observed rewards [4].

In more recent studies, it has been shown that the AI-centric frameworks may be used to boost the URLLC behaviour in multi-access edge computing (MEC) and heterogeneous network settings. As an example, deep reinforcement learning methods have been used to schedule and achieve optimal resource scheduling and resource allocation in multi-edge networks to perform better under changing traffic conditions [4]. Anthropologically, more advanced AI models, such as diffusion-based and hybrid deep learning models, have been suggested to deal with the issues of scalability and adaptation of 6G resource management [7], [10]. The state of AI implementation into URLLC systems is still an open field of research, even though a number of challenges are open and involve the real-time implementation, model complexity, and reliability guarantees.

Key Contributions

To mitigate these issues, the proposed paper proposes a new AI-powered resource allocation framework of URLLC to next-generation wireless systems. The most important works of this work are as follows:

- Creation of an AI-based adaptive resource allocation system, which uses reinforcement to make real-time and dynamic decisions in wireless networks.
- The latency minimization formulation as the main optimization goal, which guarantees the high quality URLLC requirements.
- Consideration of various network individuals, such as the channel state information, traffic need, and the queue dynamics, in the decision making process to ensure enhanced flexibility.
- End to end performance measurement of a significant improvement in latency, relative to traditional allocation of resources in dynamic networks conditions.
- Offering of insights into how AI might be applied in the future in 6G systems, and how it might be used to introduce scalability, efficiency, and intelligence to real-life implementations.

2. Background And Related Work

The decentral issue of the efficient allocation of resources in the wireless communication has been focused on since the times of Ultra- Reliable Low-Latency Communication (URLLC) where both strict latency and reliability demands are to be fulfilled at the same time. Conventional methods of resources allocation are more or less grounded on heuristic scheduling and optimization based methods. Priority-based scheduling, round-robin allocation, and proportional fairness are heuristic techniques that are popular because of their ease and minimal computational cost. Nevertheless, these schemes are based on fixed guideline and do not have the cleverness to accommodate the dynamical network condition which is not suitable in URLLC where hard latency is a mandatory requirement in a fluctuating traffic blocking [5]. Simultaneously, methods grounded on optimization have been widely researched to enhance the efficiency of resource allocation. These approaches present the allocation problem as a mathematical optimization problem, the goal of which is to maximise the throughput, reduce delay, or minimise energy usage. Convex optimization, linear programming and stochastic optimization techniques may be used to obtain close-optimal solutions when the system models are defined. However, they are not applicable in real-life URLLC setting because of their high level of computational invalidity and the need to model the system correctly. These approaches cannot support real-time decision-making as network size and heterogeneity heighten in 6G networks, which

deteriorates the performance of latency-sensitive applications [8], [9].

The drawbacks of the traditional solutions are what have triggered the move towards Artificial Intelligence (AI) in the resources scheduling of the next-generation wireless network. The AI-based approaches have used data-driven learning to simulate complex and dynamic network dynamics in order to promote adaptive and real-time decision-making. Prediction of network states, traffic patterns and channel conditions has been done through supervised learning techniques hence aiding in making more informed allocation decisions. Nevertheless, these methods need huge labelled datasets and are not that resilient when subjected to unknown or high dynamics [11]. Reinforcement learning (RL) has become one of the most promising paradigms to be applied in the area of the wireless resource allocation because it enables learning the optimal policies by obtaining feedback and updating them continuously through the interaction with the environment. An agent in RL-based models sees the current state of the network and transparently assigns resources to achieve a cumulative reward, which it can be used to minimise such performance measures as latency or reliability of the system. The latest research has established that deep reinforcement learning can be used to optimise the resource allocation in URLLC and multi-access edge computing systems with better scalability and performance than more classic methods [4], [7]. Moreover, to overcome the scalability and complexity challenges of a 6G network, new AI models, such as hybrid deep learning and diffusion-based models are suggested and expand the opportunities of AI-based resource management possibilities [10].

Nevertheless, even with these developments a number of serious issues are yet to be dealt with. Much of the research literature to date is on optimal throughput or spectral optimization in which the very high latency requests that come with URLLC applications are frequently neglected. Additionally, despite the flexibility of reinforcement learning, most of the current models lack the ability to explicitly plan with references to latency-sensitive goals, which constrains their applicability in applications with delay sensitivity. Also, the vast majority of AI-based solutions are tested in constrained simulation settings and fail to consider the real-time implementation issues of computational demands and decision times.

Research Gap:

The resource allocation techniques, both conventional and AI-based currently, do not provide a unified framework, which ensures real-time flexibility, guarantees the strict minimization of latency, and reliability simultaneously in highly dynamic and heterogeneous 6G scenarios. Specifically, the latency-aware, reinforcement learning-based models of resource allocation are urgently needed that would be able to deliver effectiveness in practical deployment conditions as well as respond to the high demands of URLLC applications.

3. System Model

In this section, the system architecture and assumptions of the model on how the available resources can be allocated by AI will be described in a next-generation wireless network with URLLC capabilities. The targeted system reflects the main features of dynamic and multi-user environments that have severe requirements towards latency and reliability. We take a multi-user wireless communication system, which is comprised of a centralised base station (BS) attached to various user devices and edge computing nodes. The base station will coordinate resource allocation and edge nodes will offer low-latency processing and computing services to delay sensitive applications. The system has a heterogeneous network environment where different users create various traffic patterns and have different channel conditions. The general structure of the discussed system model is shown in Fig. 1 that represents the communication between the base station, edge nodes and user devices.

In this network, different radio resources are share ably operated to maximise the network performance. The available bandwidth, the transmission power and the time slot allocation form the major resource parameters. Widely spread bandwidth environmental variable dictate the amount of data that is transmitted by a user and power distribution defines the signal quality and reliability. Transmission of users and access to shared communication resources is scheduled using time slots. The allocation of these resources efficiently is very vital to reduce latency levels and guarantee credible communication in dynamic network settings. Strict Quality of Service (QoS) limitations are placed on the system in order to comply with URLLC requirements. The main performance measure that is taken into account in this work is the end-to-end latency, which has to be minimal to meet the need to achieve sub-milliseconds of a delay. Transmission delay, queuing delay, and processing delay affect the latency in the system and they are all dependent on how resources allocation strategies perform. Besides the latency, reliability is a vital constraint that can guarantee the possibility of effective data transmission to ultra-high reliability standards (usually over 99.999%). These QoS requirements require smart and responsive resource management systems that can work real time.

The traffic model, which is used in this work, is representative of the nature of the URLLC applications which is mainly dynamic and bursty. The URLLC

traffic, in contrast to regular traffic patterns, is sporadic in the appearance of packets, very small packets, and has strict deadlines in terms of delay. Such boisterous behaviour poses a big problem in terms of resource allocation, when traffic demand is suddenly surveyed it may result to congestion and a rise in queuing time. As such, the system needs to be constantly responsive to changes in both load and network conditions to ensure the system can sustain a constant performance. All in all, the introduced system model reflects the key peculiarities of the next-generation wireless networks, such as multi-user interaction, heterogeneous infrastructure, and strict demands of the QoS. The given model underlies the creation of the AI-powered resource allocation model that is discussed in the following section.

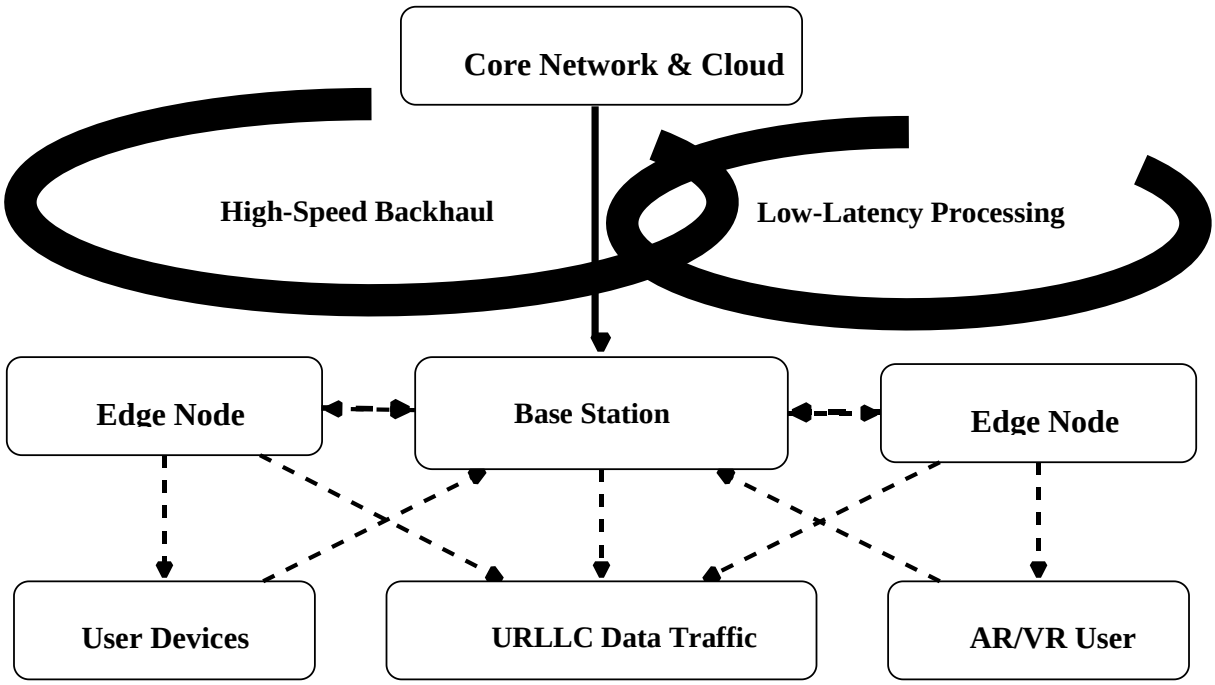

Fig. 1. AI-Driven URLLC System Model for Next-Generation Wireless Networks.

4. Proposed AI-Driven Resource Allocation Framework

This part provides the suggested AI-based resource allocation model, which is going to fulfil the high-latency criteria of Ultra-Reliable Low-Latency Communication (URLLC) in next-generation wireless networks. The framework builds on the reinforcement learning (RL) which facilitates adaptability and real-time decision making when conditions are dynamic within the network. The general layout of the suggested framework comprises three key elements, namely sensing layer, AI decision engine, and resource allocation controller. As seen in Fig. 2, sensing layer takes the duty of gathering real-time network data such as channel conditions, traffic demands and queue status of user devices and edge nodes. This information keeps on being updated to capture the dynamism of the network environment. The data obtained will then be inputted into the AIs decision engine, which will process the input and come up with the best resource allocation plans. Lastly, the allocation controller implements the decisions by assigning resources to users within the network including resources like bandwidth, transmission power and time slots.

The reinforcement learning model is the central element of the suggested framework, which allows managing the resources intelligently and in an adaptive manner. The resource allocation problem is structured as an allowable Markov Decision Process (MDP) where the system state is the present network state comprising of channel state information, traffic load and queue dynamics. This state is noticed by the RL agent and an action based on a particular resource allocation strategy is chosen. Depending on the result of the action the agent gets a reward which signals the network performance especially with regard to the reduction of latency. With time, the agent will optimise on a good policy that reduces the latency but still keeps the constraints of reliability. The AI model receives the input of several network parameters that represent the real-time position of the system. These are channel state information which is a measure of the quality of wireless connexions; traffic load which is the strength of data transmission requests; and queue status which is the number of packets awaiting transmission. With these added inputs, the model is

capable of responding to changes in the network conditions and making judgement of allocation.

The deliverable of the framework in question is the optimal resource allocation policy which will dynamically assign network resources to the users. With the use of this policy, the distribution of bandwidth, power, and time slots among users is achieved to reach the desired performance goals. The proposed approach, in contrast to the traditional approaches that rely on a static approach, is constantly ensuring that policies are updated according to the observed network conditions so as to enable efficient use of resources and better responsiveness to traffic changes. The main goal of the suggested framework is to shorten the end-to-end latency that is the key performance indicator of the URLLC systems. The rewarding or punishment feature by the reinforcement learning model is meant to punish high latency and motivate efficient use of resources. The framework guarantees achievement of high reliability on delivery of delay-sensitive applications with high rate requirements by maximising this goal. On the whole, the suggested AI-enhanced resources allocation framework is a scaleable and adjustable resource management system to address the resources in the next-generation wireless networks. The framework helps allow the URLLC systems to be well behaved in very dynamic and heterogeneous environments by incorporating real time sensing, intelligent decisions and dynamical control.

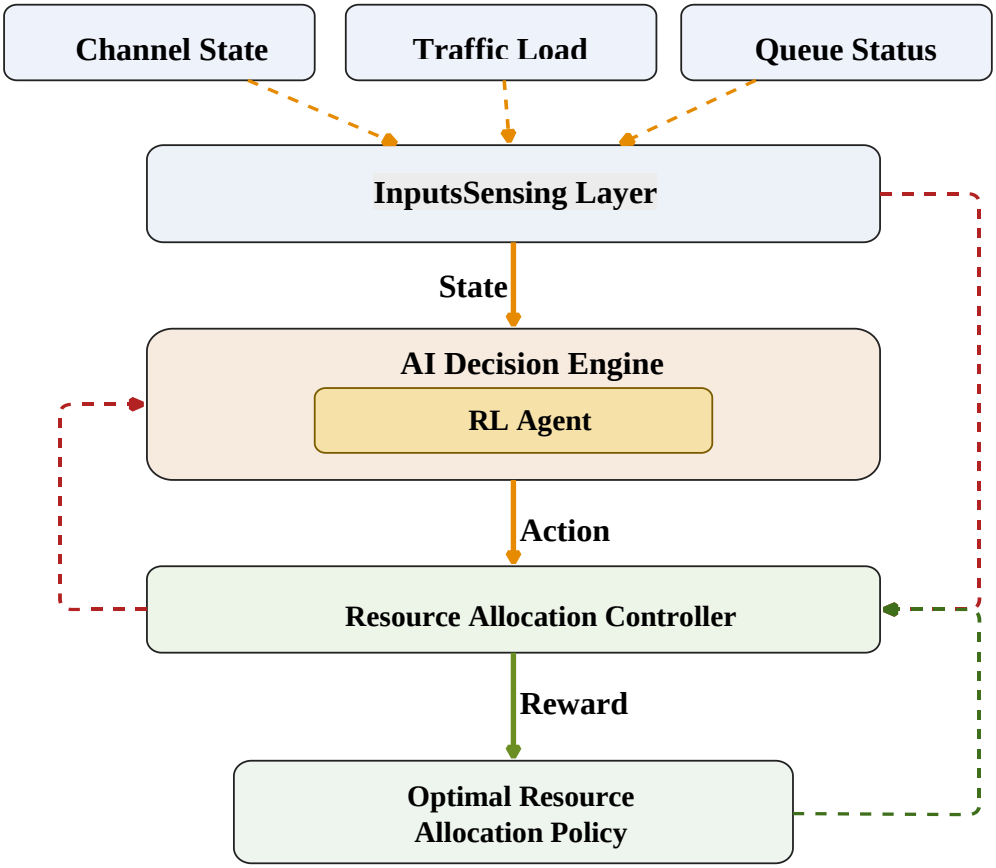

Fig. 2. AI-Based Resource Allocation Controller.

5. Performance Metric

The main performance parameter that has been taken into account in the present work is the latency because applications of Ultra-Reliable Low-Latency Communication (URLLC) require very high and reliable data transmission rates. In the next generation wireless systems, the latency is the total amount of time taken by a data packet to pass through the network to its destination and be effectively processed successfully. Because the key aim of the proposed resource allocation framework is to minimise the delay of communication under the conditions of the dynamic network, latency is chosen as the most important evaluation measure. The total end to end latency can be formulated as follows.

$$Latency = Transmission\ Delay + Queueing\ Delay + Processing\ Delay \quad (1)$$

In which the different parts are used as the various phases of the processing of packets in the network. Transmission delay is associated with the time it takes

to transmit data across the wireless channel and it greatly depends upon the bandwidth that it allocates to a user, quality of channel and transmission rate. In cases where allocation of bandwidth is inadequate or the bandwidth conditions are not good, the transmission time is more, leading to increased overall latency. Thus, effective bandwidth allocation is an important factor in reducing this element. Queueing delay is defined as the delay that the packets have to wait before being scheduled to be transmitted over the network. This delay is known to be critical in URLLC case since the traffic tends to be bursty and dynamic where the arrival of packets can suddenly increase, leading to overflow in buffers and scheduling queues. In turn, effective scheduling mechanism is also tightly connected to queuing delay. The strategy with a resource allocation that quickly responds to changes in traffic might decrease the number of queues and increase the timeliness of the packet delivery.

Processing delay is the amount of time needed to process data, compute and make a control decision in the base station or edge node. This latency in the proposed model is based on the efficiency of the calculation procedure and the receptiveness of the AI decision engine. Because the next-generation URLLC systems progressively depend on edge-assisted intelligent control, minimal processing overhead is needed in ensuring a high degree of latency requirements. The speed of resource allocation can be boosted greatly using a lightweight and adaptive decision model which will consequently minimise this part of total delay. Regarding a system perspective, the given AI-based framework is expected to collaboratively minimise the three latency components by making real-time decisions regarding the resource allocation towards channels, traffic load, and the queue state. The controller based on the reinforcement learning will streamline transmission efficiency, discourage the worst scenario of the queuing, and enhance the speed of decision making by dynamically allocating bandwidth, power, and time slots based on the prevailing network factors. It is especially true in the URLLC context where delay constraints are significantly higher compared to traditional wireless services. The aim of the suggested model is as follows: to realise sub-milliseconds latency, which is one of the primary requirements of a URLLC-based application in the area of industrial automation, remote surgery, autonomous transportation, and immersive AR/VR. The fact that the framework met this latency threshold, which is held by next-generation services being delay-sensitive, indicates that the framework can accommodate delay-sensitive services and at the same time be responsive to network dynamics. Latency reduction, therefore, is the main measure of evaluation as well as the key parameter of the feasibility of the suggested AI-based resource allocation policy.

6. Results And Performance Evaluation

The latency is a central measure of assessment of the performance of the proposed AI-driven resource allocation framework in varying network conditions. The findings are contrasted with conservative methods of scheduling and optimization based allocation to prove the aptitude of the proposed method in satisfying the Ultra-Reliable Low-Latency Communication (URLLC) needs. The test is conducted with the objective of reducing latencies, system scalability with the increase in the number of users, and stability under dynamic problem conditions.

6.1 Latency Comparison

The proposed AI-based framework latency performance is compared to the traditional methods, and the results are shown in Figure 3. The heuristic scheduling method has an average latency of 1.25 ms and optimization based method latency is about 0.95 ms. Granted, the suggested resource allocation framework based on AI has much lower latency of 0.48 ms, which is significantly lower than the requirement of sub-milliseconds delay in URLLC. This is a 61.6 percent decrease in latency over heuristic scheduling, and 49.5 percent decrease over optimization-based allocation. The enhancement is mainly explained by the adaptive learning feature of the reinforcement learning model which responded dynamically to the real time network conditions by dynamically changing the resources allocation. The AI-based framework is able to constantly optimise bandwidth, power and scheduling decisions unlike a traditional approach which was based on fixed or predetermined policies which lead to more transmission and queuing delays. As it can be seen in Figure 3, the proposed method performs much better in terms of latency and it would be more appropriate in the case of delay-sensitive URLLC applications.

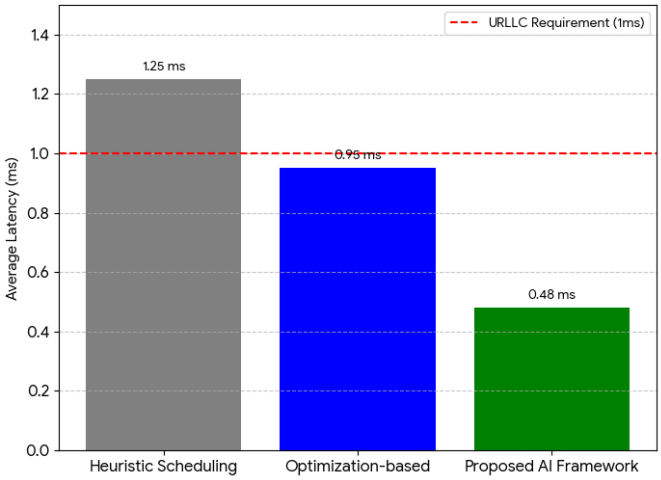

Fig. 3. Latency Comparison (Proposed vs Existing Methods)

6.2 System Behavior under Load

The latency performance is measured against the number of network users to determine scalability as seen in Figure 4. The latencies of the heuristic, optimization-based, and AI-based approaches to the problem are 0.60 ms, 0.52 ms, and 0.35 ms, respectively, when the number of users is rather small (10 users). When the users make a point of 30 users, the latency of heuristic scheduling, optimization based allocation and the proposed AI model are 1.10 ms, 0.88 ms and 0.48 ms respectively. The difference in performance is further evident at the increased network loads (50 users). The heuristic approach attains a latency of 1.75 ms which is unacceptable by the URLLC standards whilst the optimization-based one attains a latency of 1.30 ms. On the contrary, the AI-based framework keeps the latency around 0.63 ms, meaning that it does not go out of the sub-millisecond target. This means that the proposed model is scalable and can support the growth of the traffic without major performance losses. The findings in Figure 4 reveal that in traditional approaches there is a steep rise in latency with the number of users because of poor allocation of resources and longer queuing time. Nevertheless, the AI-based strategy shows slower rise, which makes it smarter in terms of resource allocation dynamically and efficient network operations also in the conditions of heavy loading.

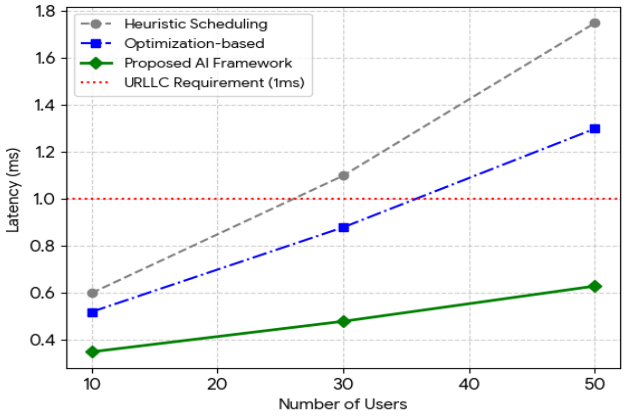

Fig. 4. Latency vs Number of Users

6.3 Stability Analysis

The proposed framework is tested in dynamic conditions of the network where a load of traffic and channel quality change with time. Figure 5 has shown the variation in latency across several time instances. The fluctuation of the heuristic approach is high with the latency ranging between 0.70 ms and 1.65 ms, meaning that the approach is not very stable when the conditions are dynamic. Likewise, the optimization-based approach is also moderate in variations, and the latency lies within standard deviations of 0.60 ms and 1.25 ms. conversely, the suggested AI-based system ensures that the latency is kept within a limited scope of 0.40 ms to 0.52 ms in all time samples. This shows a far better stabilised and steady performance of this kind when compared to the traditional methods. This decreased variability can be attributed to the nature of the repeat adaptive control to the varied needs of the network and proactive resource allocation choice that the reinforcement learning agent makes. The obtained results presented in Figure 5 confirm that the proposed framework guarantees the reduction of the average latency, as well as good performance in the time, which is imperative to the URLLC applications where the variability of delay could affect the reliability of the system considerably. Its enhanced stability renders the AI-based solution very appropriate to practise in the real world, in the dynamic wireless setting.

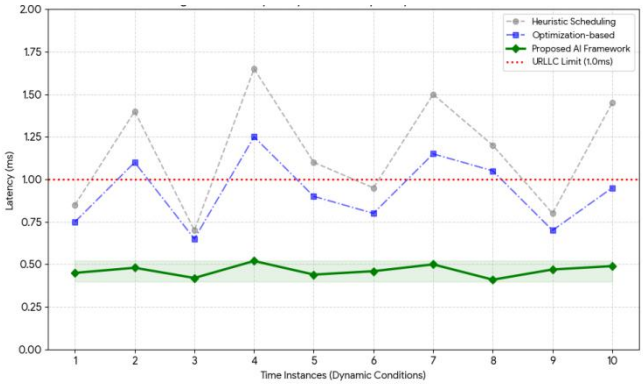

Fig. 5. Latency vs Traffic Load.

7. Discussion

As has been evidenced in the above section, the usefulness of the proposed AI-based resource allocation framework was quite evident, and reduced latency and enhanced system performance in URLLC-enabled wireless networks. The major factor that has provoked this enhancement is the capacity of the AI model, specifically, the reinforcement learning (RL) agent to monitor network conditions all the time, as well as modify its decisions dynamically. As compared to the traditional approaches used to allocate resources based on fixed criteria or pre-determined optimization criteria, the given approach tries to dynamically regulate the bandwidth, the scheduling and the transmission parameters on the ground of the current system condition. This flexibility enables the framework to be more responsive to variation in traffic load and channel conditions and therefore queue build up and the overall latency is minimised. The main strengths of the proposed framework are that it is much faster in terms of end-to-end latency that is crucial to URLLC applications. Considering such intelligent resource allocation, the system guarantees packets a small time of waiting to be sent and seen, even in the case of high traffic in the network. Besides latency reduction, the framework had excellent dynamic adaptable behaviour as well, which ensures that it can preserve performance in an environment

that is very dynamic. Such flexibility directly leads to the enhanced Quality of Service (QoS), since the system will be able to fulfil the harsh delay-specifications and ensure the consistency of the communication, at the same time. Our findings of consistent latency during dynamic conditions also indicate the strength of the proposed strategy as compared to the traditional strategies.

Regardless of these benefits, the suggested AI-based framework poses some trade-offs, which should be taken into consideration. Reinforcement learning and real time decision-making mechanism does augment computational complexity, specifically in the locations of the edge nodes or base station where the AI model is installed. This extra processing cost can also impose extra processing capacity and energy cost which might be a factor in a resource-limited setting. Such a trade-off is however justified by the high level of performance in regards to latency as well as flexibility of the system. In addition, newer developments in the field of edge computing and hardware acceleration will help reduce such challenges, and AI-based solutions will become more adequate to use in the field.

The extent to which the suggested framework can be applicable is to an extensive list of next-generation wireless applications where ultra-low latency and high reliability are needed. In the autonomous vehicle systems, such as in the case of real-time decision-making and collision-avoidance, fast and dependable communication is definitely needed. The suggested model is capable of facilitating these needs by ensuring that the network has minimum communication lag and uniform performance under different conditions of traffic. Likewise, in the field of remote robotic surgery, where even the milliseconds can influence the accuracy and safety, the capability to achieve a sub-milliseconds of the latency is critical. The framework also has a big promise in the field of industrial automation and intelligent manufacturing environments, where machine-to-machine communication requires reliability and low delay in order to operate effectively. On the whole, the discussion shows that the proposed AI-based resource allocation system results in a balanced solution that will effectively resolve the problems of latency of URLLC systems and still stay adaptable and reliable. Despite the existing trade-offs between computational overhead and the advantages in both performance and scalability, the approach remains a good contender in future 6G wireless communication systems.

8. Limitations

Although the suggested AI-controlled resource allocation model showed positive results in the area of reducing latency and enhancing flexibility, some shortcomings have to be admitted. To begin with, a model derived through the reinforcement learning technique is largely reliant on the quality and the varieties of training data. Unless the training setting captures the real world through the dynamics of the network, the learnt policy might not generalise well to unexplored conditions hence the policy can achieve high performances in real world applications. The other important restrictive measure is the calculation burden of AI-based decision-making. The reinforcement learning is also associated with extra processing overhead when it comes to situations that may need real-time inference and continuous learning. Even though edge computing will be able to alleviate this problem, resource-constrained devices might not experience straightforward implementation of the complex models under a very strict latency limit.

Moreover, the assessment of the offered framework is mainly performed based on simulation settings. Simulations can be considered useful in giving insight into the performance of systems but simplified assumptions are usually based and may not necessarily reflect the uncertainties of the performance of the real world wireless networks. The use of hardware constraints, variation in interference, and specific constraints of deployment are aspects that are hard to model and can lead to performance divergence in the simulation and actual implementation. Lastly, the practical implementation of AI-motivated resource allocation models is associated with a number of practical challenges, such as the integration of systems, scalability, and reliability. Stability of performance in a wide range of network environments, model stability and security and privacy issues are still not resolved. These issues imply that more research and development is required to fill the gap between theory and practise of URLLC implementations.

9. Future Work

Based on the suggested framework, a number of research directions are available to explore to make AI-guided allocation of resources more adequate to performance and functionality of next-generation wireless systems. A potential avenue is edge AI integration, in which the learning and decision-making are decentralised into edge nodes instead of centralised. The strategy may greatly minimise processing delay, enhance responsiveness, and hence it is more favourable in applications of URLLC that require latency. The second avenue worth adopting is federated learning to manage the distributed resources. Throughout this paradigm several network parties jointly train a common model without sharing raw data and thus maintain privacy and minimise communication costs. Decentralised and scalable control of large-scale 6G networks can be achieved using federated learning, but it is also flexible enough to adapt networks to the local contexts.

The history of evolving to 6G networks presents another opportunity to use advanced communication technology like Terahertz (THz) communication. The bandwidth in THz frequencies is extremely high that it can be used to transmit data at very high speeds, whilst this is very convenient to URLLC with regards to the latency needs. The potential area of future research is the exploration of how AI-based resource allocation schemes can be modified to utilise the peculiarities of THz communication channels. Moreover, real-time hardware-based AI solutions implementation is also an essential step in real deployment. In the real world, rigorous latency requirements will need the creation of highly efficient and runtime lightweight models that could be run on special hardware accelerators, either on a GPU or edge AI chip. The validity of the proposed approach can be further validated by means of experimental implementations with the use of testbeds and prototype systems. Altogether, further developments in the future ought to be aimed at closing the gap between the virtual and real worlds using scaling and optimization of computational heavyloads and utilising new technologies. Such developments will be important to achieving the potential of AI-based resource allocation of URLLC in the next generation 6G wireless communication systems.

CONCLUSION

The resource allocation framework proposed is a significant enhancement of the performance of ultra-reliable low-latency communication (URLLC) by smartly allocating bandwidth, scheduling, and processing as well as optimising them. The system is able to reduce the overall latency of the system significantly due to effective minimization of transmission, queuing and processing delays through adaptive learning mechanisms. The near sub-millisecond latency obtained points to the ability of the achieved framework to address the requirements of URLLC. This development is especially relevant to next-generation wireless applications, where the real-time responsiveness, reliability, and scalability are paramount and include autonomous systems, healthcare of mission-critical importance, and industrial automation setting.

REFERENCES

1. Alhashimi, H. F., Hindia, M. N., Dimyati, K., Hanafi, E. B., Alden, F. Z., Qamar, F., & Nguyen, Q. N. (2025). Survey on AI-enabled resource management for 6G heterogeneous networks: recent research, challenges, and future trends. *Computers, Materials, & Continua*, *83*(3), 3585.
2. Bennis, M., Debbah, M., & Poor, H. V. (2018). Ultrareliable and low-latency wireless communication: Tail, risk, and scale. *Proceedings of the IEEE*, *106*(10), 1834-1853.
3. Chen, S., Qin, Z., Hu, S., Shi, Y., & Li, T. (2020). A vision of 6G wireless communications: ultra-reliable low-latency and AI-driven networking. *IEEE Commun. Mag*, *58*(3), 36-42.
4. Dawoud, H., Ansari, S., Mohamed, A., Imran, M., & Popoola, O. (2024, December). DURLLCON: Deep Reinforcement Learning for URLLC Optimization in Multi-edge Networks. In *International Conference on Web Information Systems Engineering* (pp. 159-174). Singapore: Springer Nature Singapore.
5. Filali, A., Mlika, Z., Cherkaoui, S., & Kobbane, A. (2022). Dynamic SDN-based radio access network slicing with deep reinforcement learning for URLLC and eMBB services. *IEEE Transactions on Network Science and Engineering*, *9*(4), 2174-2187.
6. Gomathi, T., Panigrahi, S., Rajendran, N. K., Kapoor, P., Goyal, S., & Jadhav, Y. (2025). Edge AI for ultra-reliable low latency communication (URLLC) in 6G networks. *National Journal of Antennas and Propagation*, *7*(1), 54-61.
7. Nouri, S., Karbalaeimotaleb, M., Shah-Mansouri, V., & Taleb, T. (2025). Generative Resource Allocation for 6G O-RAN with Diffusion Policies. *arXiv preprint arXiv:2506.07880*.
8. Pase, F., Giordani, M., Cuozzo, G., Cavallero, S., Eichinger, J., Verdone, R., & Zorzi, M. (2022, December). Distributed resource allocation for URLLC in IIoT scenarios: A multi-armed bandit approach. In *2022 IEEE Globecom Workshops (GC Wkshps)* (pp. 383-388). IEEE.
9. Popovski, P., Trillingsgaard, K. F., Simeone, O., & Durisi, G. (2018). 5G wireless network slicing for eMBB, URLLC, and mMTC: A communication-theoretic view. *Ieee Access*, *6*, 55765-55779.
10. Radha, B., & Sarojini, R. (2025). Advanced AI techniques for optimizing resource allocation in 6G networks: A hybrid diffusion-enhanced meta-deep-reinforcement-learning framework. *AIP Adv.*, *15*, 115306.
11. Sun, Y., Peng, M., Zhou, Y., Huang, Y., & Mao, S. (2019). Application of machine learning in wireless networks: Key techniques and open issues. *IEEE Communications Surveys & Tutorials*, *21*(4), 3072-3108.
12. Yang, P., Xiao, Y., Xiao, M., & Li, S. (2019). 6G wireless communications: Vision and potential techniques. *IEEE network*, *33*(4), 70-75.
13. Zhang, Z., Xiao, Y., Ma, Z., Xiao, M., Ding, Z., Lei, X., & Fan, P. (2019). 6G wireless networks: Vision, requirements, architecture, and key technologies. *IEEE vehicular technology magazine*, *14*(3), 28-41.