# CYGNUS: Minimalist 5G Core for URLL Applications

Muhammad Shayan Nazeer
University of Massachusetts Amherst
Amherst, MA, USA
manzeer@umass.edu

Byung Woong (Alvin) Ko
University of Massachusetts Amherst
Amherst, MA, USA
byungwoongko@umass.edu

Jangrae (William) Jo
University of Massachusetts Amherst
Amherst, MA, USA
jangraejo@umass.edu

## ABSTRACT

Inspired by software-defined networks (SDNs), 3GPP, the standard body for 5G, proposed 5G Control and User Plane Separation (CUPS) for architectural enhancements. This separation aimed for faster, seamless, and improved performance compared to previous generations. CUPS allowed network operators to deploy user plane near the edge to enhance the overall experience. A key aspect of this new architecture was to ensure that the control logic operated independently of user plane procedures. However, it has been observed that certain control procedures, such as charging functions and reflective QoS, involve collaboration between edge and core modules. The impact of this interaction is more observable when it comes to ultra-low latency and high throughput applications like online gaming and MR/AR devices. It has also been observed that the 5G procedures like RAU (Registration Area Update) or paging are unnecessary for such applications. These procedures add control signaling overhead and make 5G core energy inefficient. In this paper, we argue that the current 5G core is inefficient for low latency and high throughput applications. Given the restricted mobility inherent to such applications we suggest a radical redesign by eliminating mobility management and associated procedures while pushing control plane procedures like the charging function to the user plane on the network edge. We propose *CYGNUS*, a lightweight, energy-efficient controller designed to operate in a truly SDN-like fashion within the 5G core. *CYGNUS* only controls policy management and device authentication, minimizing the control and user plane interaction. Once a data session is established, *CYGNUS* allows uninterrupted session operation until a policy change is required.

## 1 INTRODUCTION

The landscape of cellular networks has evolved beyond just mobile phones. We are witnessing a surge in diverse device types connected to the 5G network. Projections suggest that by 2030, over 30 billion IoT devices will be online. Which shows the growing trend of connectivity. Figure 3 illustrates various devices expected to utilize the 5G network. Each device has its unique network requirements. This diversity implies the necessity for specialized versions of 5G to cater to specific device categories. Recognizing this need, the 3GPP, a standard body for 5G, introduced NB-IoT [2], a tailored version of 5G designed explicitly for IoT devices.

In this paper, we will focus on another class of devices connected to the 5G network: URLL devices [6]. URLL devices, characterized by their demand for low latency and high throughput, encompass a range of applications including augmented reality/mixed reality (AR/MR) devices, online gaming, remote robotics, and more. This paper addresses the unique requirements and challenges associated with URLL devices in the 5G ecosystem.

3GPP outlined the specifications of the 5G network, drawing inspiration from software-defined networks (SDNs) [3]. The functionality of the 5G network was categorized into different network functions, some control-based and others related to the data plane. 3GPP defined 5G core into a control plane (with control-based network functions) and data plane. They defined 5G core abstractly into a centralized controller and a forwarding plane as defined in SDNs. However, our research revealed that the implementation did not fully embrace the principles of SDNs, as the control and data planes are still interconnected and it has a notable impact on high-performance devices, particularly those requiring ultra-reliable low-latency (URLL) capabilities.

In this project, we have demonstrated that the current 5G control plane is inefficient for ultra-reliable low-latency (URLL) applications. Through our analysis, we have identified inefficiencies and redundant procedures within the control plane architecture and illustrated the impact on high-performance applications. These redundancies not only increase latency but also compromise the overall performance and responsiveness of the network.

Addressing the shortcomings of the current 5G control plane, we introduce CYGNUS, a novel decentralized [8] and minimalist control plane tailored specifically for ultra-reliable low-latency (URLL) applications. For this control plane we assumed that URLL applications are stationary (don't require handover) and they do not require periodic update when in idle mode. Keeping these assumptions in the view we

optimized the control plane and pushed some control functionality to the edge. In the CYGNUS control architecture, a minimalist control plane is deployed in the cloud, while an edge-based control plane operates at the network edge. The cloud control plane oversees high-level tasks and directs the edge control plane to manage low-level behaviors of the data plane. This split architecture significantly reduces latency and enhances system performance for URLL devices.

## 2 BACKGROUND

The 5G network consists of three main components: a cellular device (UE), a base station (gNB), and the 5G core [**?** ] . Figure 1 illustrates a simplified 5G infrastructure. The mobile or end-user device connects to the base station (gNB) over the radio channel. The base station then connects to the 5G core, which is typically deployed in the cloud. The 5G core forwards packets from users to the internet, and similarly, for downlink traffic, the core forwards traffic from the internet to the end-user device.

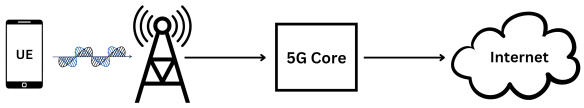


**Figure 1: Simplified Cellular 5G Infrastructure; (from left) UE, gNB, 5G core, and internet**

The 5G core acts as the brain of the network, managing data packets, determining policies, and forwarding them to their requested destinations. The 3rd Generation Partnership Project (3GPP) standardizes 5G and its components, introducing a service-based architecture (SBA) [1] for the 5G core. In this architecture, core network functionalities are modularized into services and deployed as Network Functions (NFs). The 5G SBA comprises various NFs, such as the Access and Mobility Function (AMF), Session Management Function (SMF), User Plane Function (UPF), and Policy Control Function (PCF), among others. This architecture implements Control and User Plane Separation (CUPS) [7] to enable centralized network management, allowing for dynamic traffic flows and on-demand scalability. The concept draws inspiration from the success of Software-Defined Networking (SDN). As depicted in Figure 2, the SMF serves as the controller, while the UPF acts as the user plane (or forwarding plane in SDN terms), forwarding data between user equipment (UE) and the internet.

The concept of cloud-native 5G emerged from SDN-based service architecture, revolutionizing how 5G network functions are deployed. Now, 5G network functions (services) can be hosted in the cloud, allowing the user plane to be positioned closer to the user while being controlled by the

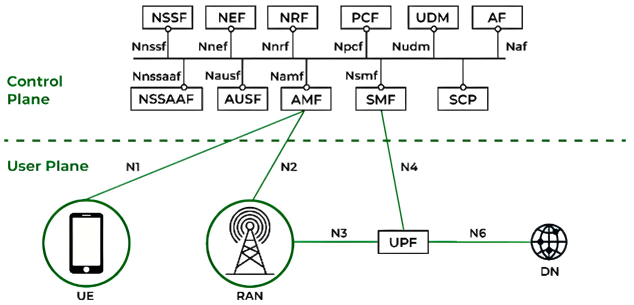


**Figure 2: Service-based architecture of 5G Infrastructure: Control plane functions (i.e., AMF, SMF) are deployed in the cloud; User Plane Function (i.e., UPF) resides at the edge**

control plane from a different location. This approach enhances flexibility, scalability, and performance, ensuring that the network can efficiently meet the demands of modern applications requiring low latency and high throughput.

Although cellular networks were initially designed for mobile phones, 5G now supports a variety of devices, each with unique requirements. This diversity demands specialized 5G solutions tailored to different sets of devices. By customizing 5G technology to meet the specific needs of various devices, we can optimize performance, enhance user experience, and fully realize the potential of 5G in transforming industries and daily life.

In recent efforts, the 3GPP introduced NB-IoT (Narrowband Internet of Things), demonstrating that standard bodies are taking the concept of specialized cellular systems seriously. NB-IoT is specifically designed for low-power, wide-area applications, such as smart meters, environmental monitoring, and asset tracking. This highlights the importance of tailoring cellular technologies to meet the distinct requirements of various devices, ensuring optimal performance and efficient use of network resources.

## 3 WHY REDESIGN 5G CORE?

In recent years, we have witnessed the rapid adoption of cellular IoT devices and edge applications. With the advent of 5G, cellular networks can now support high throughput and low latency applications. However, the core of these networks remains primarily focused on mobile phones. While mobile phones constitute the majority of devices connected to 5G, they are not the only ones. Figure: 3 illustrates different devices using 5G network. Different devices have varying requirements, therefore it is time to develop specialized 5G cores for different applications to enhance system efficiency.

The release of the NB-IoT 5G specifications by 3GPP highlights the importance of specialized network cores for diverse applications.

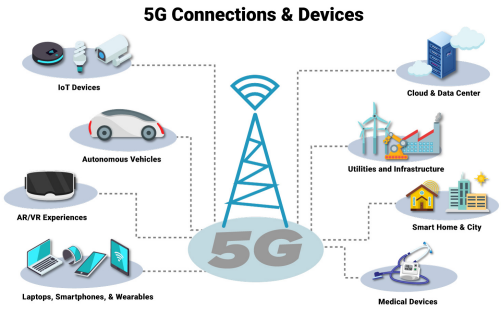


**Figure 3: Multitude of devices connected to 5G network**

In this paper, we propose a specialized core for ultra-reliable and low latency (URLL) devices, such as AR/MR applications, online gaming, and medical devices. We assume that most URLL devices are stationary and do not require updates when idle. Based on these assumptions, we have identified inefficiencies in the current 5G core for supporting URLL devices. In the following subsections, we will discuss these issues in detail:

## 3.1 Multiple Controllers

3GPP adopted an SDN-inspired approach to the 5G core, as shown in figure: 4 by separating the data and control plane. In this architecture, the Session Management Function (SMF) acts as the controller, while the User Plane Function (UPF) represents the data plane. The SMF defines rules for the UPF, and the UPF forwards data packets from the User Equipment (UE) to the Internet. A key aspect of SDN is its centralized control architecture, where a single controller handles the data plane [9]. This centralized controller defines the policies for the data plane, streamlining overall performance and ensuring consistent policy enforcement. In contrast, the 5G core architecture does not adhere to this singular control model. Although the User Plane Function (UPF) can be managed exclusively by the Session Management Function (SMF), a closer examination reveals that multiple other components

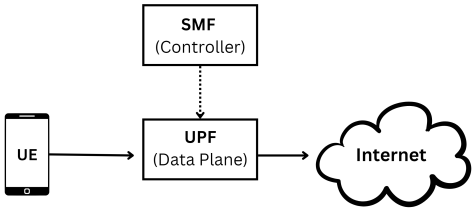


**Figure 4: 3GPP view of 5G Core**

also influence the data plane via the SMF. This multiplicity of controllers complicates data plane management, leading to inefficiencies and potential performance bottlenecks.

For instance, functions such as the Policy Control Function (PCF), Access and Mobility Management Function (AMF), and the Network Slice Selection Function (NSSF) all interact with the SMF to impact the data plane, as illustrated in Figure 5. This collaborative but fragmented approach contrasts sharply with the streamlined, centralized model of SDNs, where a single point of control facilitates more straightforward and efficient management.

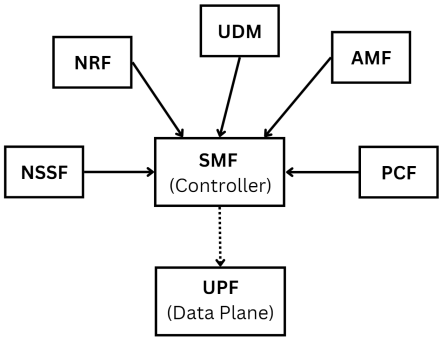


**Figure 5: How 5G works**

As a consequence, 5G core, unlike SDNs, suffers from increased complexity and reduced efficiency. This fragmentation can impact performance, especially in scenarios demanding low latency and high throughput, such as online gaming, AR/VR applications, and real-time video streaming [5]. Therefore, simplifying the control plane by reducing the number of influencing components and aligning more closely with the SDN model could significantly enhance the performance of the 5G core.

## 3.2 Unnecessary Control and User Plane Interaction

In the 5G core, unnecessary interactions between the control and user planes occur in procedures such as reflective QoS and charging functions. These interactions introduce additional signaling overhead, which can negatively impact performance, especially for low-latency, high-throughput devices. For instance, reflective QoS requires coordination between the control plane, which manages policies, and the user plane, which handles data traffic. Similarly, charging functions involve frequent updates and synchronization between the planes to accurately track data usage.

These continuous interactions not only add complexity but also increase latency and reduce the overall efficiency of the network. For applications that demand ultra-low latency

and high throughput, such as augmented reality (AR) and online gaming, this added latency can degrade the user experience by causing delays and interruptions. By simplifying the control plane and minimizing these interactions, we can significantly enhance the performance and responsiveness of 5G networks, ensuring they meet the stringent requirements of modern, high-demand applications.

## 3.3 Idle Mode

Idle mode in 5G [4] refers to a state where a device is not actively engaged in data transmission but remains connected to the network to quickly resume activity when needed. This mode helps conserve battery life and reduce network resource usage while ensuring that the device can swiftly transition to an active state when required.

5G idle mode adds initial latency due to the need to reestablish the connection. For example, when an AR device transitions from idle mode to connected mode, it must reestablish the connection, resulting in initial latency. This delay can be problematic for applications requiring immediate responsiveness. Although idle mode is intended to save power, the re-connection process can compromise the performance of real-time applications like AR, where seamless and instantaneous connectivity is crucial. Consequently, optimizing idle mode to reduce or eliminate this initial latency is essential for enhancing the user experience and efficiency in low-latency, high-throughput applications. Figure: 6 The initial latency after the device is started shows the transition from idle to connected mode.

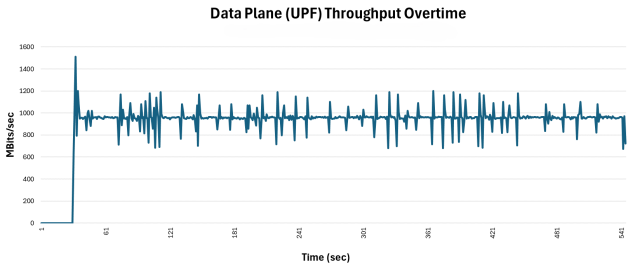


**Figure 6: Variation in data plane throughput over time after device is started**

In idle mode, a device periodically wakes up at regular intervals. For ultra-reliable low-latency (URLL) applications, this behavior is undesirable. For instance, an AR device only needs to be active when worn; if it is lying down, regular connection updates are unnecessary. Although idle mode is designed to save power, it still consumes a significant amount because the device spends most of its time in this state. This continuous power draw undermines the energy efficiency of the device, particularly in applications where maintaining a constant connection isn't crucial when the device is not actively in use. Therefore, optimizing power management for such scenarios is essential to enhance overall efficiency and performance. The concern of energy utilization is also discussed in the subsequent section.

## 3.4 Energy Concerns

The current 5G core raises energy concerns when it comes to ultra-reliable low-latency (URLL) devices. For example, when an AR device is lying idle, it does not need to be regularly updated. However, the device still performs periodic registrations to stay connected to the internet and check for updates. It also performs the Registration Area Update (RAU) procedure to inform the network about its location/registration area. Additionally, the core broadcasts messages to the device, known as paging. All these procedures are undesirable and contribute to increased energy consumption.

The following figure shows the energy consumption of a 5G device during idle and connected modes. It can be seen that during idle mode, the device periodically wakes up, indicated by the spikes in energy levels.

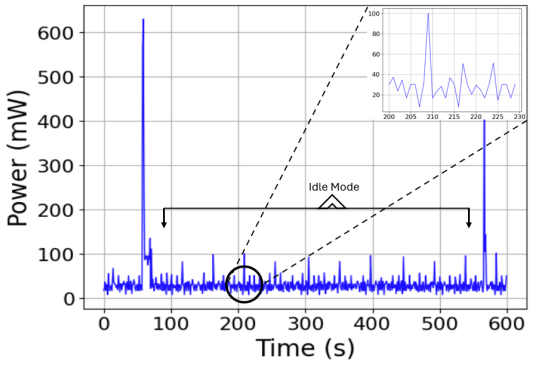


**Figure 7: Energy consumption of a 5G device during idle mode. Periodic wake-ups in idle mode are shown by spikes in energy levels. (Bigger spikes show when device move from idle mode to connected mode)**

This continuous power draw, even when the device is not in active use, undermines energy efficiency. For URLL applications, where maintaining a constant connection isn't crucial when the device is inactive, these periodic updates and paging messages are particularly wasteful. Optimizing power management to reduce or eliminate unnecessary procedures in idle mode is essential for enhancing overall efficiency and performance, especially for devices that demand both reliability and low latency.

# 4 CYGNUS DESIGN

In this section, we introduce CYGNUS, our lightweight core specifically designed for URLL applications. CYGNUS addresses the demands of these applications by minimizing initial latency and optimizing power efficiency, ensuring seamless and instantaneous connectivity. Our design prioritizes the performance and responsiveness critical for applications like AR, providing an energy-efficient solution specific to the needs of modern high-performance devices.

## 4.1 Design Overview

Although the design of CYGNUS is inspired by software-defined networks (SDNs), we propose a split control architecture that enhances latency and responsiveness. In this architecture (as shown in figure: 8), a minimalist control plane is positioned in the cloud to handle essential tasks, and an edge-based control plane is for immediate, localized operations. This dual-layered approach ensures rapid decision-making and efficient processing, significantly improving the performance of the control plane for ultra-reliable low-latency applications [10].

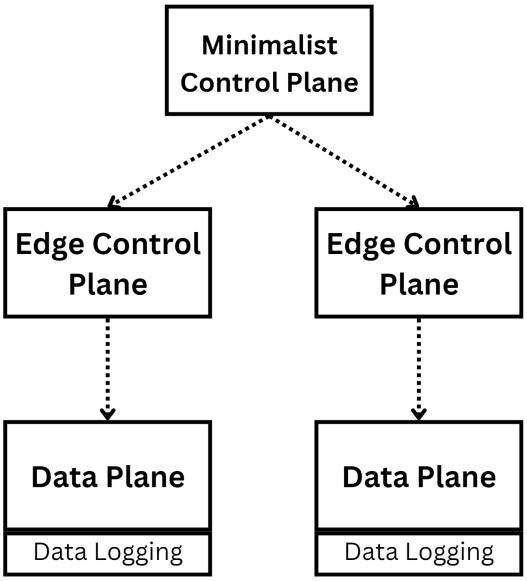


**Figure 8: Proposed Cygnus design; A centralized lightweight minimalist control (For the tasks that don't require immediate response) connected to an edge-based control plane (for tasks that require immediate response, thus reducing latency). Which is then connected to edge-based data plane. The data plane collects the logs of sessions through data logging which is designed for charging functions etc**

The data plane is deployed at the edge and is directly managed by the edge control plane. To minimize frequent interactions between the control and user planes, we have incorporated a data logging unit. This unit logs session data and calculates data packet usage for charging functions and other purposes. It reduce control overhead, enhances efficiency, and ensures that essential data is readily available for billing and analytics without compromising on performance.

## 4.2 Minimalist Control Plane

Sitting at the cloud minimalist control plane performs high-level and important tasks which don't require an immediate response from core. Figure: 9 illustrates the design overview of minimalist control. As shown in the figure: 9 Its primary responsibilities include device authentication and policy management, which are critical for maintaining network security and ensuring that devices comply with predefined rules and protocols. By concentrating on these core tasks, the minimalist control plane minimizes unnecessary control signaling and overhead, which can otherwise introduce latency and consume additional power

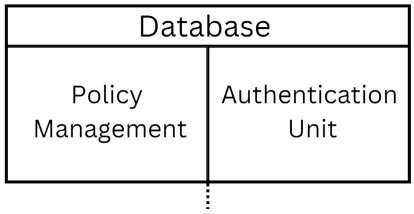


**Figure 9: The design of Minimalist control plane**

Moreover, this minimalist control plane functions as the brain of the network. It is similar to an SDN controller and manages the network at a high level and maintains a comprehensive view of the entire network. This control plane dictates network policy and ensures that packet forwarding occurs according to preset rules. One of the critical functionalities of this control plane is device authentication. It maintains a user database, which it uses to authenticate devices and perform network registration. By centralizing these functions, the control plane can efficiently manage the network, streamline operations, and maintain high-performance standards

## 4.3 Edge Control plane

At the edge, we deploy an edge-based control plane that translates high-level commands from the minimalist cloud-based controller into low-level controls on the data plane. This edge-based control plane enforces the policies dictated by the centralized control plane. In the context of 5G and

Ultra-Reliable Low-Latency Communications (URLLC) applications, this architecture can improve performance. URLLC applications, such as online gaming, Mixed Reality (MR), Augmented Reality (AR), and medical devices, demand minimal latency and high reliability. By offloading certain 5G functions to the edge-based control plane, we can significantly reduce latency and improve the performance of these applications.

The edge-based control plane manages several critical control functions to enhance the performance and efficiency of the 5G network. Firstly, it handles packet forwarding and routing, ensuring data packets are directed accurately, and minimizing latency. Secondly, it oversees Quality of Service (QoS) management and monitoring, maintaining optimal performance levels for applications requiring high reliability and low latency. Thirdly, the edge-based control plane manages device re-registration, streamlining the process of re-authenticating devices and reducing the signaling burden on the core network. Additionally, it plays a crucial role in network security by implementing localized security measures, thereby enhancing overall network protection. Lastly, it manages network slices, ensuring that specific network resources are allocated efficiently to different applications and services based on their unique requirements. By performing these functions locally, the edge-based control plane significantly improves responsiveness and reliability, particularly for ultra-low latency and high throughput applications.

In this design, we have eliminated the functionality of mobility management because most URLL devices, except vehicles, are stationary. Consequently, mobility management becomes a redundant procedure that unnecessarily adds to the control signaling overhead.

Similarly, idle mode procedures have also been removed. As discussed in section 3.3, idle mode can introduce latency and increase power consumption. By eliminating these procedures, we can further streamline the network, reduce unnecessary signaling, and improve overall efficiency.

## 4.4 Data Plane Logging

In this design, we introduce the concept of data-plane logging, where data session activities are logged directly at the user plane. This logging provides a detailed account of data transactions and interactions in real time. As a result, the control plane doesn't need to constantly query the user plane for updates, which helps reduce signaling overhead. For example, activities like session establishment, data transfer, and termination are logged as they happen. This allows the control plane to access these logs when needed, without starting new signaling procedures.

Data-plane logging is also valuable for the charging function. Traditionally, charging systems depend heavily on control plane interactions to track data usage, increasing the signaling load. With data-plane logs, the network can track and record data usage metrics directly at the edge. These logs can be processed periodically or on-demand by the charging system, ensuring accurate billing without continuous control plane involvement.

## 4.5 Performance

Cygnus is expected to perform better than current 5G architecture for the URLL devices. With critical control functions positioned at the network edge, the architecture minimizes latency and enhances responsiveness, which is crucial for ultra-reliable low-latency (URLL) applications like augmented reality (AR) and virtual reality (VR). Similarly decentralizing control tasks, immediate actions can be executed locally without relying on a centralized core, ensuring rapid decision-making and adaptability to changing network conditions. Moreover, this edge-centric approach optimizes resource utilization, reduces power consumption, and enhances scalability, making the architecture well-equipped to handle evolving traffic demands and accommodate future growth.

# 5 CONCLUSION

This paper has examined the inefficiencies of the current 5G core architecture, particularly for ultra-low latency and high throughput applications such as online gaming and MR/AR devices. Our analysis highlighted that traditional 5G procedures, including Registration Area Update (RAU) and paging, contribute to significant control signaling overhead and energy inefficiency. Recognizing the restricted mobility of most URLL devices, we proposed a redesign of the 5G core by eliminating mobility management and associated procedures. We proposed CYGNUS, a lightweight and distributed control plane for the 5G data plane. We showed that this control plane is SDN-like and helps improve the URLL devices' latency constraints.

# ACKNOWLEDGMENT

We appreciate and acknowledge the efforts of ECE-690 course instructor Dr. Taqi Raza for proposing the idea and providing continuous support and guidance over the semester.

# REFERENCES

[1] Gabrial Brown. 2017. Service-based architecture for 5g core networks. *Huawei White Paper* 1 (2017).

[2] Hossam Fattah. 2018. *5G LTE Narrowband Internet of Things (NB-IoT)* (1st ed.). CRC Press, Inc., USA.

[3] R. Guerzoni, R. Trivisonno, and D. Soldani. 2014. SDN-based architecture and procedures for 5G networks. In *1st International Conference*

*on 5G for Ubiquitous Connectivity*. 209–214. https://doi.org/10.4108/icst.5gu.2014.258052

[4] Sofonias Hailu, Mikko Saily, and Olav Tirkkonen. 2018. RRC state handling for 5G. *IEEE Communications Magazine* 57, 1 (2018), 106–113.

[5] Omar Hayat, Zeeshan Kaleem, Muhammad Zafarullah, Razali Ngah, and Siti Zaiton Mohd Hashim. 2021. Signaling overhead reduction techniques in device-to-device communications: Paradigm for 5G and beyond. *IEEE Access* 9 (2021), 11037–11050.

[6] Wolfgang Kiess, Ekkehard Lang, Klaus Hoffmann, Hans-Jochen Morper, and József Varga. 2019. Ultra-reliable low latency services: 5G architecture and operational alternatives with cost analysis. In *2019 IEEE Wireless Communications and Networking Conference (WCNC)*. 1–8. https://doi.org/10.1109/WCNC.2019.8885604

[7] Milind Nadkarni and Sanjeev Panem Jaya. 2022. INNOVATIVE POLICY ENFORCEMENT IN A 5G/CUPS NETWORK USING HEADER ENRICHMENT. (2022).

[8] Amir Roozbeh. 2015. Distributed cloud and de-centralized control plane: A proposal for scalable control plane for 5G. In *2015 IEEE/ACM 8th International Conference on Utility and Cloud Computing (UCC)*. IEEE, 348–353.

[9] V Thirupathi, CH Sandeep, Naresh Kumar, and P Pramod Kumar. 2019. A comprehensive review on sdn architecture, applications and major benifits of SDN. *International Journal of Advanced Science and Technology* 28, 20 (2019), 607–614.

[10] Shahin Vakilinia and Halima Elbiaze. 2020. Latency control of icn enabled 5g networks. *Journal of Network and Systems Management* 28 (2020), 81–107.