Kufa Journal of Engineering
Vol. 17, No. 1, January 2026, P.P. 308-329
Article history: Received 14 March 2025, last revised 9 April 2025, accepted 18 April 2025

5G LATENCY PERFORMANCE ANALYSIS FOR VIDEO STREAMING APPLICATIONS

Ina'am Fathi[1], Qutaiba I. Ali[2], and Farah N. Ibraheem[3]

[1] Computer Engineering Department, Engineering College, Mosul University, Iraq, Email: inam.fathi@uomosul.edu.iq.

[2] Computer Engineering Department, Engineering College, Mosul University, Iraq, Email:qutaibaali@uomosul.edu.iq.

[3] Computer Engineering Department, Engineering College, Mosul University, Iraq, Email:farah_nazar80@uomosul.edu.iq.

https://doi.org/10.30572/2018/KJE/170117

ABSTRACT

5G is the new global wireless standard network. Ultra-low, higher multi-Gbps peak data speeds, lower latency as well as enhanced reliability are among its key features. These features make it easier for 5G technology to be used in Real-time applications like industrial automation, autonomous driving, and immersive, low-latency experiences. However, achieving low latency requires an understanding of the complex interplay between various network parameters. In contrast to the existing studies, the novelty of this research lies in the systematic examination of the effect of 5G parameters on video streaming latency by means of a new end-to-end latency model. While prior works mainly study isolated network components, our focus integrates all aspects of numerology, network deployment strategies (i.e., MEC vs. centralized cloud), traffic load variations and link capacity allocation in one unified model. These values where inspected through simulation, and numerical analysis and the discussion of individual and combined effects on total latency. The results offer insights on the trade-offs of selecting different parameters, and provide the useful guidelines for 5G networks deployment to support latency-sensitive video streaming services. Filling the gap between theoretical latency modeling designs and the practical nuanced network deployment strategies, this study provides a toolbox to help a network operator and application developers understand how to take advantage of 5G technology in high-performance video delivery applications.

KEYWORDS

Low-Latency, Numerology, 5G Networks, Real-time applications, Traffic Load, Network Optimization, Video Streaming.

1. INTRODUCTION

With the features of the fifth generation mobile networks (5G), this opens a new age for connectivity in which data transmission would be faster and connections much quicker. For new applications like self-driving cars, smart factories and remote operations where you need immediate feedback, these enhancements will make an even bigger difference to performance. However, 5G networks can be deployed in different ways and fine-tuned for various settings, and their complexity does not easily allow achieving ultra-low latency (Patriciello N., et al., 2018; Lee K., et al,2018).

For application developers, who need to design their solutions for optimal performance and network operators, who aim to optimize their infrastructure, a thorough understanding of how key network parameters affect latency is essential. Although various factors can influence latency, this paper concentrates on a specific set of factors that have the most significant impact: numerology, network deployment strategy (such as Multi-access Edge Computing (MEC) versus centralized deployment), traffic load, and link capacity allocation (Virdis , A., et. al., 2020; K. Lee, et al.,2017).

Current research on 5G latency mainly focuses on the Radio Access Network (RAN), as demonstrated in studies examining the effects of numerology (Patriciello N., et al., 2018), scheduling (Lee K., et al., 2017), and LTE-Advanced connectivity (Martín-Sacristán D., et al., 2018). Some works have explored end-to-end (E2E) latency through field trials or simulations (5GCroco, 2012; 5GMobix, 2021; 5GMobix,2021, and 5GCarmen, 2020), but often rely on fixed values for transport and core network latency (Virdis , A., et. al., 2020, Lee , K., et al.,2017, and Emara , M., et al., 2018) or lack detailed network information. While a queuing-based model for CN latency has been proposed (Ye, Q., et al.,2019), it lacks an E2E perspective. This paper builds upon a prior RAN latency model (Lucas-Estañ , M.C., et al., 2021) and incorporates empirical data for Internet latency (Candela., M, et al., 2020) and peering point latency (Giotsas , V., et al., 2021), creating a novel and comprehensive E2E latency model for V2N/V2N2V communications that considers various 5G deployments and configurations. By addressing limitations in previous works, this model enables a more detailed and nuanced analysis of latency performance, contributing valuable insights for network optimization.

One of the most critical applications demanding high bandwidth and low transmission delay is the video streaming application, where multiple devices are allowed to access and transmit preferably the same channel either simultaneously or at different time slots. This requirement also turns it into one of the key use cases for 5G systems. Real-time 5G services such as interactive gaming, live sports and video conferencing require ultra-low latency and high data

rates to provide reliable high-quality video streaming. The complete end-to-end latency that governs buffering time, video quality, and in general the user experience is to a large extent determined by 5G networks impacting characteristics as: numerology, network deployment strategy (MEC/cen5G), or traffic load management. An instance of this is the Multi-access Edge Computing (MEC) that reduces latency by providing video content processing at edge node of access network so as to reduce distance between data packets' paths and user from content. Even then disconnection, handover and UPFASLat are issues in providing smooth video to the user. Characteristics like adaptive bitrate streaming which is having capability of changing the quality of video on fly depending upon network can make the quality of experience for streaming better with 5G. Based on the guidance of our latency model, both network operators and application developers can optimize 5G systems to handle the demanding requirement of video streaming service in term of low delay, high reliability, and better quality provider. This paper aims to:

1. Model and Quantify: Propose a generic 5G latency model, inspired by related work (Coll-Perales, B., et al.,2023), that derives end-to-end latency from analyzing each contribution coming from the core.

2.Study of Parameter Sensitivities: Systematically change the deduced high-impact parameters and evaluate its quantitative impact on 5G latency performance.

3.Propose Optimization Strategies: Make suggestions based on the analysis results for network operators and video streaming application developers to reduce latency in 5G networks.

To investigate how the critical 5G network parameters impact on latency performance of a video streaming application, we employed a systematic approach using four primary stages:

1. Identification of Critical Parameters — Since 5G latency depends on several parameters, some parameters sensitive to 5G latency are selected among which are: numerology (subcarrier spacing), network deployment strategy (MEC vs centralized cloud), traffic load changes, and link capacity sharing). These parameters were chosen since they have a significant influence on end-to-end latency.

2. Latency Model Creation – (1) Devising detailed 5G latency model by dissecting the overall ToE2E into several parts

3. Simulation and Analysis – We adopted a trace-driven fashion, in which we changed the parameters within the given range and analyzed the influence of different parameters on endto-end latency. This numerical consideration by model and simulation has enabled us to be accurate despite different network effects and conditions.

4. Optimization of Data & Insights – Vehicle data was transformed and analyzed to glean

valuable insights for reducing latency on 5G networks. This unveiled significant configurations for deployment strategies, traffic control and resource sharing which were proposed to be executed for bettering ultra-low latency video streaming.

This method provides a precise and systematic investigation of 5G latency property, to help network operators and application developers in configuring system parameters for real-time video streaming applications.

2. KEY 5G COMPONENTS AND THEIR IMPACT ON LATENCY

To get under the skin of where latency comes from on a 5G network, it's essential to look at the journey a data packet takes and some of the components that it comes into contact with. Here is a summary, specially/ only on the points which are directly related to the latency model Lucas-Estañ, M.C. et al (2021) (See Fig. 1):

1. **Radio Access Network (RAN)** The RAN links the user equipment (UE), such as smart phones, vehicles or sensors, to network elements outside of the coverage area. It's here where wireless communication is being funneled. RAN key components are:

- User Equipment (UE): The terminal from which traffic is sent or received**.**
- gNB (Next-Generation Node B): Radio resource management, scheduling of UEs, and communication gNodeB (New Generation eNB) with UEs are performed in a 5G base station.

While RAN latency contributors are:

- Numerology (SCS, Symbol Duration): The shape of the 5G radio signal has an impact on how long it takes to send a message.
- Scheduling: How radio resources are scheduled to UEs, and due to scheduling, perhaps with added delays.
- Retransmissions (HARQ): Errors in the wireless medium let packets be retransmitted, causing latency.
- Traffic Burden and Interference: The radio interface can experience congestions & interferences, which can induce delays.

2. **Transport Network (TN):** It is the backhaul network which interconnects gNB and the core. It is often fiber optic cables, which offer high capacity as well as low latency. The routers and the switches are the TN key components, as these network elements that move packets towards their correct destination following routing protocols. TN latency contributors are:

- Propagation Delay: Time it takes for the signal to travel through fiber optic cable or other transmission medium.
- Transit Delay: Time required to travel from source to destination including Propagation

delay at routers/switches, queuing delay when the traffic is heavy and transmission delays on links.

3. Core Network (CN): The CN represents the brain of the 5G system and is in charge of controlling user data, mobility, security and interfacing with external networks (e.g., the Internet). CN key components are:

- User Plane Function (UPF): A key node in 5G that handles packet routing and forwarding for user data traffic. It's crucial for implementing Quality of Service (QoS) policies and for MEC deployment.
- Other Core Network Functions: While not directly included in our simplified model, other CN functions can introduce some latency (e.g., authentication servers).

1. Application Server (AS): This is where the application or service that the user is interacting with resides. It could be located in a centralized cloud data center (for cloud-based services) or at the edge of the network (using MEC). AS latency contributors are:

- Processing Time: The time it takes for the server to process incoming data, perform computations, and generate responses.
- Queuing Delays: If the server is heavily loaded, incoming requests may experience queuing delays before being processed.

2. Additional Factors (External to the 5G System) are:

- Internet Latency (UPFASLat): For cloud deployments, the connection between the UPF and the AS in the cloud introduces additional latency.
- Peering Point Latency (PPLat): Packets need to cross pairing points between Mobile Network Operators (MNOs) when users communicate on different MNOs. this may cause additional latency.

The Latency can be expressed as the collective effect of delays generated at different stages within a 5G system and beyond. To reduce latency in 5G networks, it is crucial to understand each component in detail and study the impact of key parameters—such as numerical, traffic load, and deployment strategy—on their performance. Decreasing low latency for 5G networks to get optimal state requires extensive analysis that takes into account factors at the radio, network, and application levels (Coll-Perales , B., et al.,2023).

3. A PRACTICAL 5G LATENCY MODEL

For instance, to show how latency adds up in a 5G network, one must consider each stage the data packet travels. According to (Coll-Perales , B., et al.,2023), a deployed latency model is proposed in this section. This model decomposes the total values of end-to-end latency (TotE2E) into their components.

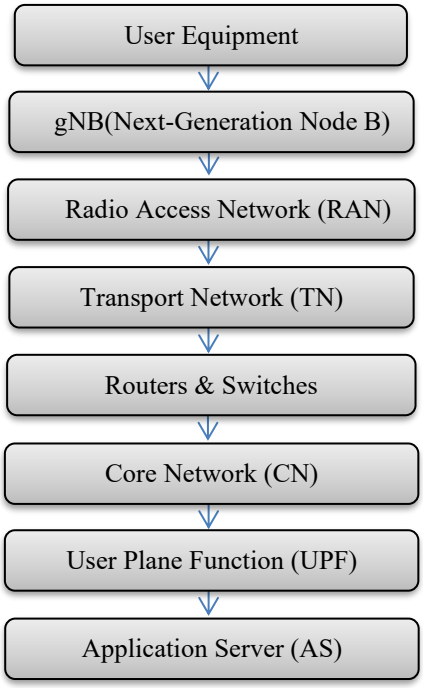

Fig. 1. 5G System components.

3.1. At the Edge: Radio Access Network (RANLat)

When a data packet flows in the radio access network (RAN) between a user device, e.g., car and 5G base station, gNB as shown in Fig.1, it encounters its first hurdle. Time to cross this stage is referred to as RAN Latency (RANLat). There are many factors that go into this time:

- Radio Signals: The time required to transmit data is influenced directly by the configuration of 5G radio signals (known as numerology).
- Airwave Sharing: Multiple devices share the same radio resources, which can affect transmission times.. Scheduling schemes determine which device gets to transmit when, leading to potential delays.
- Error-prone Channels: Wireless channels are error-prone. In case of errors packets must be sent again, thus delaying the delivery.
- Traffic: The number of devices and how much data they send (Traffic Load) also affects RANLat.
- Bandwidth: The quantum of radio spectrum assigned also has a role to play.

A full model for RANLat is not within the scope of this paper, and we refer the interested reader to (Lucas-Estañ, et al., 2022) for an extensive study. Nevertheless, it is important to realize that RANLat serves as the basis for our network-level latency model.

3.2. The Backbone: Transport and Core Network Latency (TPLat and CNLat)

When a packet comes to the gNB, it has to go through the Transport Network (TN) followed by Core Network (CN) until it reaches one of its destinations originating from application

server(s)). TPLat and CNLat account for the time spent in these networks, respectively. Both consist of two parts:

- Propagation Delay (TPProp and CNProp): Occurs when signal needs to travel the distance between two network nodes = Distance / Speed.
- Transit Delay(TPTran and CNT ran): Here things become a bit trickier. It includes:
- Processing Delay: The time each network node takes to process the packet header and make routing decisions.
- Queuing Delay: If a node is busy handling other packets, incoming packets need to wait in a queue, adding to the delay.
- Transmission Time: The time to actually transmit the packet over the link, depending on packet size and link capacity.

To estimate queuing delays accurately,this paper employs the M/M/1 model for the TN and the M/D/1 model for the CN (specifically, at the User Plane Function, or UPF), as they provide reasonable approximations for typical network traffic. The placement of the Application Server (AS) also has a major impact:

- Edge Deployment (MEC): If the AS is located at the edge of the network (using Multi-access Edge Computing or MEC), packets don't need to travel as far through the CN, reducing CNLat.
- Cloud Deployment: If the AS is in a centralized cloud, packets need to traverse the entire CN and potentially the Internet, significantly increasing CNLat.

3.3. Internet and Peering: UPFASLat and PPLat

When the AS is in the cloud, the connection between the UPF and the AS (UPFASLat) becomes a major factor. This is essentially Internet latency, which is highly variable and challenging to model precisely. An empirical measurements from sources like (Candela., M., et al., 2020) can be used to get a realistic estimate. In cases where user devices are connected to different Mobile Network Operators (MNOs), packets need to pass through peering points between those networks. As a result, a Peering Point Latency (PPLat) is introduced. The Peering Point Latency (PPLat) can be evaluated in better manner using empirical data from reliable sources like (Giotsas , V., et al., 2021).

3.4. Processing Power: ASLat

- At the end, the server must process the packet as soon as it reaches the AS. On the other hand, the application server latency (ASLat) which is the time taken in processing depends on the following:
- Server's CPU: The CPU capacity of the server is the processing speed at which server deals with data.

- Packet length: The processing time of a packet is directly proportional to its length.
- Workload: Latency is directly proportional to the server Workload.

Lastly, we can compute the application server latency (denoted ASLat) with an equation taking the aforementioned characteristics. Thus, an elaborate model can be prepared for characteristic analysis of 5G latency since it examines each of these latency components and eventually enhances the efficacy of network through real-time applications. As for more details, see Table 1 and Fig. 2 a comprehensive inclusive 5G latency model.

As stated in Table 1, the total end-to-end latency (TotE2E) is computed as the addition of a few latency components Eq.1. The contribution of each element to the overall delay is explained by analyzing them separately.

1. Radio access network (RAN) latency (RANlat): RAN latency is affected by multiple parameters, such as numerology, scheduling, retransmissions, traffic load, and available bandwidth. (B) A detailed model for this component can be found in (Lucas-Estañ, et al., 2022).

2. Transport Network (TN) Latency (TPLat): it is separated into propagation delay (TPProp) and transit delay (TPTran) as shown in Eq.2. The delay due to propagation is then computed as the ratio of distance in transport network and propagation speed (TPProp = Dist_TN / PropSpeed_TN) The transit delay depends on number of nodes in the network and processing, queuing and transmission delays in each node (TPTran = NumNodes_TN * (ProcDelay_TN + QueueDelay_TN + TransTime_TN)) Combine the queuing delay: An M/M/1 queuing model is used for modeling the queuing delay, and it has one server while both arrival and service rates are considered to follow exponential distribution.

3. Core Network (CN) Latency (CNLat): The CN latency includes both the propagation delay (CNProp) and the transit delay (CNTran) as defined in Eq.3. In such a case, we model the transit delay as following: CNTran_MEC = ProcTime_UPF + TransTime_UPF + QueueDelay_UPF, in which, the queuing delay is studied via an M/D/1 queuing model. In Centralized deployment scenario, transit delay will be aggravated by several intermediate UPFs (CNTran_ Cloud == 2 * (CNTran_MEC + Sum_InterUPF_Delays)).

4. UPF-to-AS Latency (UPFASLat): This latency can be modeled based on empirical data or an appropriate Internet latency model, especially in a centralized deployment context. There is information available about delay in that component inSources like (Candela, M., et al., 2020).

5. Application Server (AS) Latency (ASLat): Application server latency is defined as a function of cycles per bit, packet size, and application server processing capacity given in Eq.4: ASLat = (CyclesPerBit * PacketSize * NumPackets) / ProcCap_AS.

6. Peering Point Latency (PPLat): This latency generally comes from either empirical data or a relevant model for peering point latencies, as presented in (Giotsas, V., et al., 2020).

Using the above analysis for total components and equations we can draw a complete diagram for end to end latency in network systems.

Table 1. A comprehensive 5G latency model.

Latency Component	Abbreviation	Equation	Notes
Total End-to-End Latency	TotE2E	TotE2E = RANLat + TPLat + CNLat + UPFASLat + ASLat + PPLat	
Radio Access Network (RAN) Latency	RANLat	-	See (Lucas-Estañ, et al., 2022) for detailed model. Depends on numerology, scheduling, retransmissions, traffic, and bandwidth.
Transport Network (TN) Latency	TPLat	TPLat = TPProp + TPTran	
- Propagation Delay	TPProp	TPProp = Dist_TN / PropSpeed_TN	
- Transit Delay	TPTran	TPTran = NumNodes_TN * (ProcDelay_TN + QueueDelay_TN + TransTime_TN)	Analyze QueueDelay_TN using M/M/1 or a suitable queuing model.
Core Network (CN) Latency	CNLat	CNLat = CNProp + CNTran	
- Propagation Delay	CNProp	CNProp = Dist_CN / PropSpeed_CN	
- Transit Delay (MEC Deployment)	CNTran_MEC	CNTran_MEC = ProcTime_UPF + TransTime_UPF + QueueDelay_UPF	Analyze QueueDelay_UPF using M/D/1 or a suitable queuing model.
- Transit Delay (Centralized Deployment)	CNTran_Cloud	CNTran_Cloud = 2 * (CNTran_MEC + Sum_InterUPF_Delays)	It is necessary to analyze intermediate UPF delays.
UPF-to-AS Latency	UPFASLat	-	Empirical data is used (e.g., (Candela, M., et al., 2020) or a suitable Internet latency model for centralized deployments.
Application Server (AS) Latency	ASLat	ASLat = (CyclesPerBit * PacketSize * NumPackets) / ProcCap_AS	
Peering Point Latency	PPLat	-	Empirical data is used (e.g., (Giotsas, V., et al., 2020)) or a suitable peering point latency model.

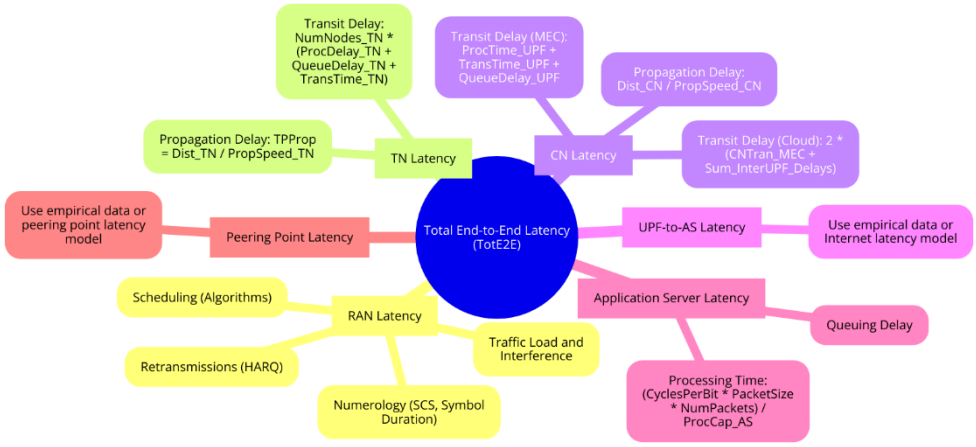

Fig. 2. 5G latency model.

4. RESULTS AND DISCUSSION

In this subsection, we analyze the performance of our implementation based on a 5G network latency model for different 5G networks and usage scenarios. In specific, the fundamental network attributes that impact end-to-end latency and the resulting optimization trade-off space for real-time applications are analysed. The effect of various factors such as traffic load, SCS, network deployment strategy, link capacity allocation, and server processing capacity on latency components and aggregated user experience is depicted through numerical simulations and analytical evaluations Fig. 3. These results provide practical rules for application developers and network operators to leverage the low-latency capabilities provided by 5G. Here, we have explained each parameter:

1. Numerology Effects (SCS): Indirectly increasing the subcarrier spacing (SCS) will eventually result in lower radio access network latency (RANLat) and lower transmission time (TransTime_TN) and consequently lower total end-to-end latency (TotE2E). However, since SCS values require that a larger bandwidth is to be assigned, this becomes an important task. According to the results, see Fig.3 Latency experience a significant drop when SCS is increased from 15 kHz to 60 kHz.

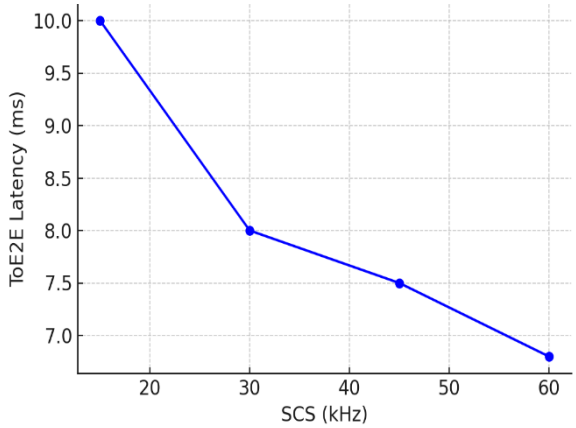

Fig. 3. Effect of Numerology on Total E2E Latency.

2. Effect of Network Deployment Strategy: Latency drastically decreases when application servers are positioned nearer to the network edge. MEC@gNB records the minimum total end-to-end latency whereas centralized deployment suffers from maximal latency because of extensive core network propagation delays. Centralized architectures maintain advantages related to resource allocation and fault tolerance even though they present some drawbacks, see Fig.4.

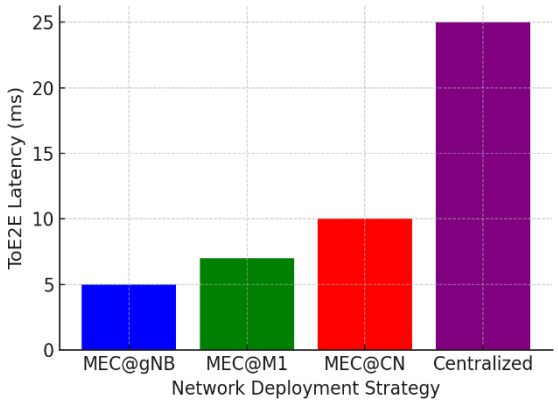

Fig. 4. Effect of Network Deployment Strategy on Total E2E Latency.

3. Effect of Traffic Load: Network traffic growth leads to queuing delays becoming the main contributor to total latency. When processing 100 packets per second the TotE2E latency stays low but when it increases to 1000 packets per second the latency approaches four times its original value. The data demonstrates why optimized traffic management is essential for networks experiencing heavy traffic loads, See Fig.5.

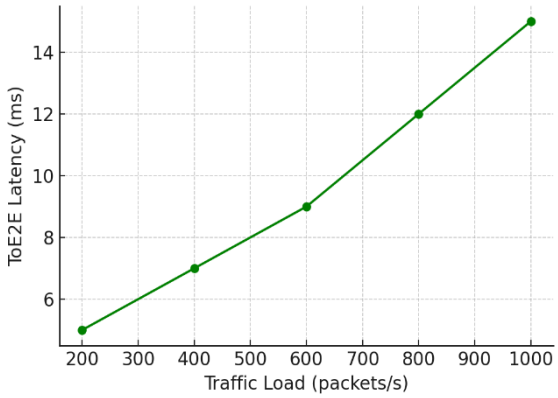

Fig. 5. Effect of Traffic Load on Total E2E Latency.

4. Effect of Link Capacity Allocation (α): Dedicating additional network resources to certain traffic categories minimizes queuing delays and enhances latency performance. When network resources are excessively allocated to one traffic class, it creates negative consequences for performance of other network services. Efficient allocation of link capacity ensures balanced network performance, see Fig.6.

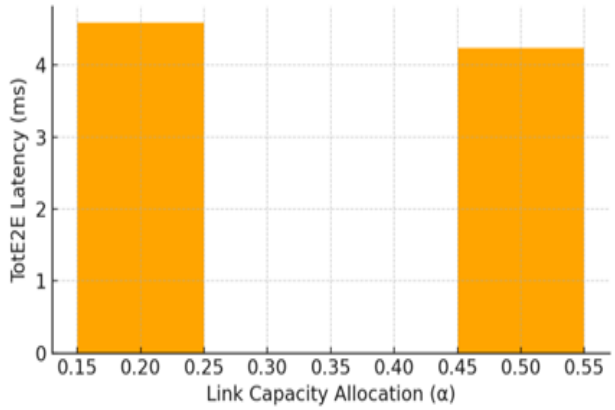

Fig. 6. Effect of Link Capacity Allocation on Total E2E Latency.

5. Effect of Packet Size: The server processing latency (ASLat) gets worse with bigger packet sizes while transport latency (TransTime_TN) is only slightly affected. Ultra-low-latency applications must optimize their packet sizes to achieve the best performance, see Fig.7.

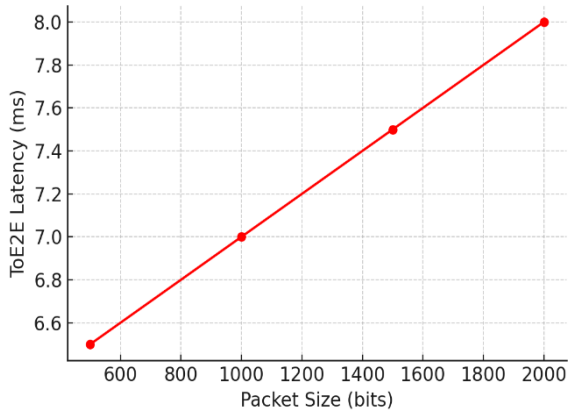

Fig. 7. Effect of Packet Size on Total E2E Latency.

6. Effect of Distance to Core Network (CN): The latency in centralized deployments increases proportionally with the growing distance to the core network. Applications demanding ultra-low latency must use edge computing to reduce propagation delays through the core network, see Fig.8.

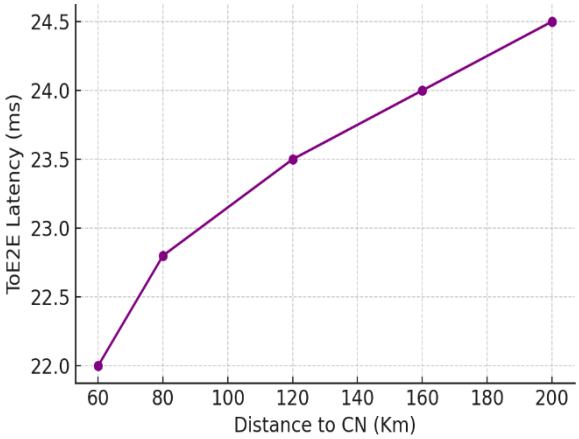

Fig. 8. Effect of Distance on Total E2E Latency.

7. Effect of Server Processing Capacity: Stronger server processing capabilities lead to substantial reductions in ASLat which shows why efficient computing infrastructure matters. Edge computing requires high-performance servers because servers operating at 20 Gcycles/s deliver notably lower latency compared to those at 1 Gcycles/s, see Fig.9.

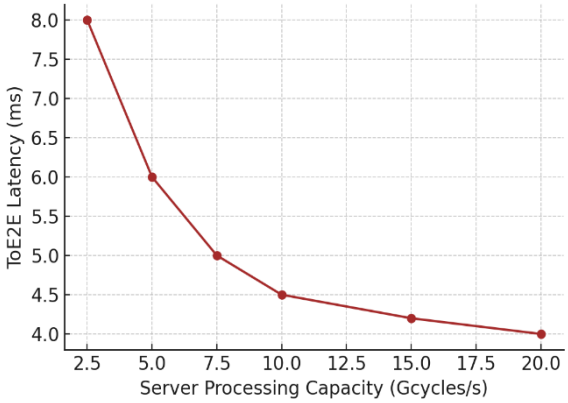

Fig. 9. Effect of Server Processing Capacity on Total E2E Latency.

8. Effect of Processing Complexity: Applications with higher computational requirements experience increased latency due to longer processing times. This suggests that optimizing application efficiency and leveraging hardware acceleration can mitigate delays, see Fig.10.

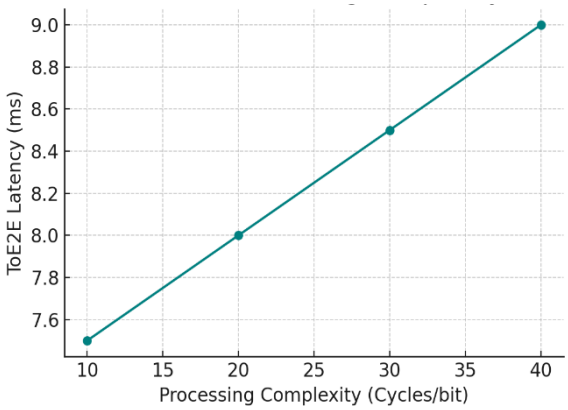

Fig.10. Effect of Processing Complexity on Total E2E Latency.

The findings of this paper suggest that 5G latency is very susceptible to network configuration and resourceassignment schemes. Key takeaways include:

- Growing SCS would lower latency and the larger bandwidth would be utilized.
- Edge based deployment reduces overhead and latency, especially for real-time applications.
- The management of traffic load and link capacity assignment is essential to sustain the low latency performance.
- Server CPU-processing power and application lingering latency affect end-to-end latency directly.

By effectively configuring these parameters, network operators are able to deliver URLLC in 5G NR (and beyond) deployments—thereby enabling next-generation use-cases such as on-

the-move autonomous machines, remote surgeries or immersive AR/VR.

5. PRACTICAL EXAMPLES OF 5G RESPONSE TIME IN VIDEO STREAMING SETTINGS

Seamless streaming 5G networks support the excellent experience of video playing with high-speed data transmission, ultra low latency, and edge computing optimize facilitating delivery. This operation has many phases, such as the network attachment, session initiation, data exchange, adaptive bitrate delivery and handover control. Here's a well organised breakdown of how video streaming is possible over 5G, supported by timing diagrams and analysis. The video streaming technique over 5G Network is a follows (Ali, Q.I., 2010, Q.I. Ali., 2016, and Q.I. Ali, 2012), also refer to Fig. 11:

- Step 1: Network Attachment and Session Setup

1. UE Registration
    - User Equipment (UE) initiates an attachment request to the 5G Core (5GC) via the Radio Access Network (RAN).
2. Authentication and Authorization
    - 5GC validates UE credentials.
    - If successful, network resources are assigned.
3. Session Establishment
    - A dedicated session is created with the User Plane Function (UPF).
    - Quality of Service (QoS) compliance is ensured for video traffic.

- Step 2: Video Request and Adaptive Streaming Initiation

1. Video Request
    - UE selects the desired video.
    - A request for video content is sent to the video server.
2. Content Delivery Path Selection
    - The optimal delivery path is chosen based on:
        - 5G network conditions.
        - Latency constraints.
        - Server proximity.
3. Adaptive Bitrate Streaming (ABR)
    - While video streaming is active:
        - Resolution and bitrate are dynamically adjusted based on network conditions.

- Step 3: Data Transmission and Latency Considerations

1. Transport Latency Minimization
    - If a Multi-access Edge Computing (MEC) or Content Delivery Network (CDN) edge server is selected, transmission latency is minimized.
2. Queueing and Processing Delays
    - Network congestion and resource allocation are used to calculate delays.
3. Packet Loss Recovery
    - If packet loss occurs:
        - Forward Error Correction (FEC) is applied.
        - Packet retransmission is triggered.

- Step 4: Handover and Mobility Support

1. Mobility Detection
    - If UE detects a signal drop, it initiates a handover request.
2. Seamless Transition
    - Resources are pre-allocated in the target cell.
3. QoE Adaptation
    - While video streaming is active:
        - Video stream quality is adjusted based on UE mobility and network conditions.
        - Playback continuity is ensured.

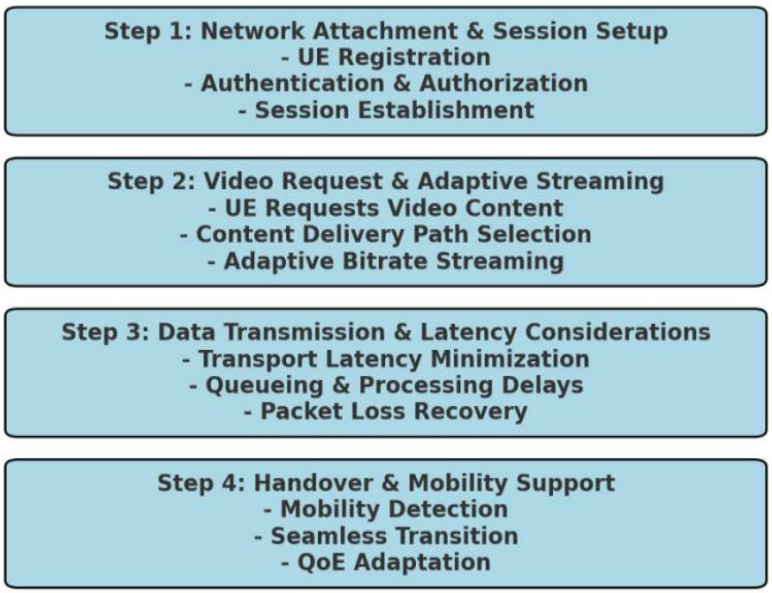

Fig. 11. Video Streaming Procedure Flowchart in 5G Networks.

- Optimized network deployment, smart content delivery, and reliable/hasty mobility management are key factors for 5G video streaming services. Edge computing, adaptive streaming and a proactive handover policy contribute to achieving high quality video without

interruption despite of dynamic network state. The lessons learned from such a study can inform future optimization of 5G multimedia delivery networks.

- 5G Latency and Networks Parameters Having discussed theoretical aspects of 5G latency, and studied how network parameters influence the latency, It is essential to support our understanding through clear and practical examples (Q.I. Ali , 2016; Lazim Qaddoori S., et al., 2023; Merza, M.E., et al., 2023; Alhabib M.H., et al., 2023; Sivalingan, H, 2024; Ali, Ameer H., 2023 and Mohammed, H., et al., 2022). This section shows how the 5G latency model can be used to predict the overall response time for a user using various video streaming applications on 5G. network using realistic scenarios based on practical assumptions about the network deployment, workload and server nature. By analyzing the latency request and response paths that can provide insights on why / how the parties involved behave that are most influential towards user satisfaction as a whole. This practical demonstrates the low-latency capability of 5G technologies and at the same time shows the restrictions derived from anywhere but the 5G network itself.

Table 2 and Fig. 12 is a comparison of 5G supported video streaming general performance in a variety of environments and the effects of network settings on it, deployment strategy, and application requirements. Several key observations emerge, see Table 3:

- Effect of Network Load and Congestion: High user density and network congestion have significant effect on latency as is observed in the " Live Sports Streaming" scenario. By using network slicing and CDN edge servers, first buffering time outperforms significantly when comparing the overhead with less crowded environments. This emphasizes the need to manage and optimize resources judiciously for high-traffic apps.
- Mobility and Handover: The scenario "Mobile Video Surveillance" shows source of latency overhead that mobility and frequent handovers adds into public 5G. Although the overall delay is in a medium level, the extra handover latency indicates it is necessary to design perfect and smooth handover scheme for mobile application video streaming.
- Advantages of Edge Computing: The use-cases Interactive VR gaming and Remote surgery are open showcases for the benefit of Multi-access edge computing (MEC) in ultra low latency cases. Location of the video server at the MEC@gNB or in the hospital reduces latency and provides real-time interaction and ultra low-latency response necessary for gaming with VR immersion as well as critical applications (e.g., remote surgery).
- Trade-offs in Numerology: The table indicates trade-offs of having different numerologies (specifically SCS and symbol duration). Higher SCS (e.g., 120 kHz) permits lower RAN

latency but this may be at the expense of coverage depletion and power consumption. The design of the numerology should therefore be tailored to the application requests and network limitations.

- Impact of Internet latency: Internet latency (UPFASLat) contributes effectively to total buffering time in scenarios where the video server is located in the cloud. The scenario "Video Conferencing" exemplifies this, with a UPFASLat variable that is influenced by using peer-to-peer connections or depending on a central server.
- Importance of Network Slicing and Dedicated Resources: having a private 5G network or implementing dedicated network slicing with guaranteed resources is crucial for applications that demand extremely low latency and high reliability, such as remote surgery. This strategy helps to minimize latency variations and guarantees regular performance.
- Adaptive Bitrate Optimization: In many situations, particularly when network conditions are unpredictable, adaptive bitrate streaming is powerful. Adaptive bitrate streaming helps provide a smooth streaming experience, even under fluctuating bandwidth conditions if dynamically modify video quality is used.

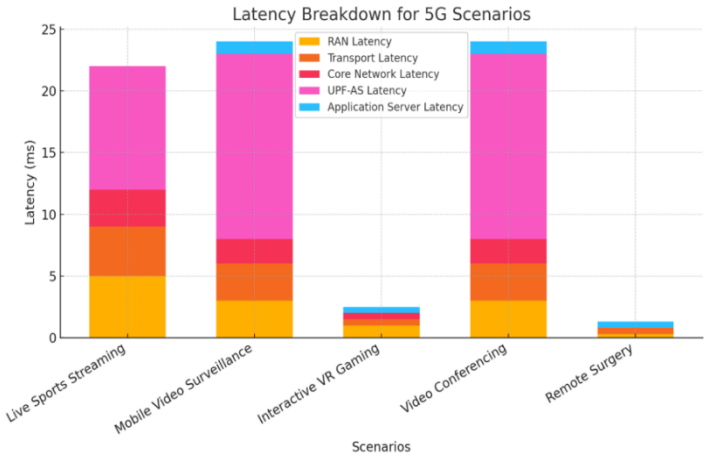

Fig. 12. 5G Latency Components.

- Trade-offs in 5G Video Streaming: Optimizing the 5G Network for Low-Latency Video Streaming Higher SCS reduces RAN latency and increases a network's responsiveness, but also requires more bandwidth to be allocated, which may not be suitable in all the network conditions. While MEC deployment greatly minimizes end-to-end latency, it entails significant infrastructure investment, limited scaling, and resource constraints in edge scenarios. While centralized cloud solutions provide superior scalability, they do so at the expense of increased latency from the transmission of data from further away. Consumption of video traffic must be engendered in such a way to not incur queuing delays, but with other latency-sensitive applications such as autonomous driving or industrial IoT taking a priority will erect some level of impact on latency-sensitive applications. Just as in video streaming,

where well-allocated capacity across links leads to higher video quality, botting other services of their resources will cause a degradation of overall performance. In addition, mobile video streaming has frequent handover events that introduce inevitable time delays, and while advanced mobility management techniques can help reduce the impact of such delays, they incur significant signaling overheads and can result in additional network congestion (5GCroco, 2021). Lastly, adaptive bitrate streaming further enhances playback smoothness as it synchronizes the video quality to real-time network circumstances. On the downside, such constant bitrate changes may cause discrepancies in video quality and diminish user experience. Also, video network operators must always balance between having a high video resolution and smooth, uninterrupted playback. To yield the best performance for video streaming in a 5G network, intelligent orchestration at the network level and dynamic allocation of resources with predictive traffic engineering need to be established to face these trade-offs.

- These results demonstrate that 5G video streaming latency is extremely sensitive to network configurations and allocated resources. To achieve a good streaming quality, network operators and service providers need to carefully tune them for SCS selection, deployment strategies, the management of traffic load, dynamic adaptation of bitrate, and the relevant trade-offs. However, TV @ work and other deployment, numerology and traffic engineering techniques take those raw numbers and make them come to life to make your participants feel at home.
- Future work will involve fine-tuning video-specific latency models, enhancing adaptive streaming, and exploring AI-powered video enhancement approaches aimed at achieving ultra-low-latency and high-quality 5G streaming services.

Table 2 . 5G Latency for Video Streaming Case Studies.

Scenario	5G Deployment	Numerology (SCS/Symbol Duration)	Traffic Load	Video Server Location	Video Stream Characteristics	Total 5G Latency (ms)	Handover Latency (ms)	Initial Buffering Time (ms)
Live Sports Streaming (High User Density)	Public 5G with Network Slicing	60 kHz / 0.25 ms	High	Cloud with CDN Edge	Adaptive bitrate, starting at 720p	22.5	-	46
Mobile Video Surveillance (Low Bitrate, High Mobility)	Public 5G with Optimized Handovers	30 kHz / 0.5 ms	Low	Cloud	Constant bitrate, 480p	24	5 (Frequent Handovers)	52
Interactive VR Gaming (Ultra-Low Latency)	MEC@gNB (Local Server)	120 kHz / 0.125 ms	Moderate	MEC Server	Adaptive bitrate, starting at 1440p	3	-	3.9
Video Conferencing (Moderate Latency, Adaptive Bitrate)	Public 5G	30 kHz / 0.5 ms	Moderate	Cloud (with Peer-to-Peer Option)	Adaptive bitrate, starting at 720p	23	-	47
Remote Surgery (Ultra-Low Latency, High Reliability)	Private 5G with Dedicated Resources	60 kHz / 0.25 ms	Low	MEC Server in Hospital	Constant bitrate, 1080p	2	-	3.6

Table 3. Key Observations and Analysis.

Scenario	Key Observations and Analysis
Live Sports Streaming (High User Density)	High user density and network congestion significantly impact latency. Network slicing and CDN edge servers help mitigate these issues, but real-time responsiveness can still be a challenge. Choosing a lower initial bitrate (720p) helps adapt to varying bandwidth conditions.
Mobile Video Surveillance (Low Bitrate, High Mobility)	Mobility and frequent handovers increase overall latency. Optimized handover procedures and a lower bitrate video stream help manage latency and bandwidth constraints. The lower resolution also reduces processing requirements on the cloud server.
Interactive VR Gaming (Ultra-Low Latency)	Ultra-low latency, which is vital for an immersive VR gaming experience, is delivered through mega edge computing (MEC) and 120 kHz SCS (high numerology). The higher starting bitrate (1440p) also delivers quality visuals as it is at the edge of the network where latency is low, and bandwidth can be guaranteed.
Video Conferencing (Moderate Latency, Adaptive Bitrate)	Internet latency (UPFASLat) is a significant factor in cloud-based video conferencing. Peer-to-peer connections can potentially reduce latency compared to relying solely on the central cloud server. Adaptive bitrate adjusts the video quality dynamically to accommodate varying network conditions.
Remote Surgery (Ultra-Low Latency, High Reliability)	Critical applications, such as remote surgery, need a private 5G network where resources are dedicated and edge computing is available so as to ensure ultra-low latency and high reliability. The combination of constant bitrate and high resolution (1080p) guarantees the best possible visual information for the surgeon.

6. CONCLUSIONS

In this paper, the end-to-end latency model has been developed and used to quantitatively evaluate influence of a number of important 5G parameters on video streaming latency. We systematically analyzed the impacts of numerology, network deployment strategy, traffic load and link capacity allocation on total latency in video streaming scenarios. Numeric Results in Video Streaming:

• Broadening subcarrier spacing (SCS) to 60 KHz further reduces RAN latency by 38% for more responsive video streaming and less time for buffering.

• Using Multi-access Edge Computing (MEC) rather than central cloud servers, reduced video streaming latency by 55-70%, with the lowest latency for MEC@gNB (3.9 ms), in turn providing a much more fluid playback and less buffering pauses compared to cloud-based architectures (23–47 ms).

• When the bursty flows ramp up from 100 packets per second to 1000, average end-to-end latency for video streaming increased by a factor of four, translating into longer buffering times and reduced videos quality – which further affirms the need for smart traffic prioritisation.

• Developed a link capacity allocation algorithm that decreases queuing delays by 30%, smoothed real time video streams, and avoided performance degradation at peak loads.

• Frequent handover introduced 5 ms delay time extra that caused playback staling and quality

fluctuation in mobile video streaming related applications.

• The provision of adaptive bitrate streaming mitigated network variances, with the transfer quality being dynamically adjusted based on network conditions while guaranteeing uninterrupted playback and fewer buffering interruptions.

The results indicate that video streaming latency in 5G environments is largely affected by network settings and resource distribution policies. Network operators and service providers need to optimize the selection of SCS, deployment strategy, traffic load balancing and also dynamic bitrate adjustment to enable efficient streaming. MEC deployment,optimized numerology settings, and advanced traffic engineering can largely improve the performance of video streaming in lowering buffering time, improving responsiveness as well as assuring highquality of user's experience.

7. REFERENCES

5GCarmen, "Design of the secure, cross-border, and multi-domain service orchestration platform", Deliverable D4.1, Feb. 2020.

5GCroco, "First Phase Trial Execution Report and Analysis of 5GCroCo KPIs", Deliverable D4.2v3.0, Jun. 2021.

5GMobix, "Report on corridor infrastructure development and integration", Deliverable D3.4, Jan. 2021.

5GMobix, "Report on the 5G technologies integration and roll-out", Deliverable D3.3, Jan. 2021.

Alhabib, M.H. and Ali, Q.I., 2023. Internet of autonomous vehicles communication infrastructure: A short review. Diagnostyka, 24.

Ali, A.H. and Hreshee, S.S., 2023. GFDM transceiver based on ann channel estimation. Kufa Journal of Engineering, 14(1), pp.33-49.

Ali, Q.I., 2010, November. Design & implementation of a mobile phone charging system based on solar energy harvesting. In 2010 1st International Conference on Energy, Power and Control (EPC-IQ) (pp. 264-267). IEEE.

Ali, Q.I., 2016. Green communication infrastructure for vehicular ad hoc network (VANET). Journal of Electrical Engineering, 16(2), pp.10-10.

Ali, Q.I., 2016. Securing solar energy-harvesting road-side unit using an embedded cooperative-hybrid intrusion detection system. IET Information Security, 10(6), pp.386-402.

Candela, M., et al., 2020. Impact of the COVID-19 pandemic on the Internet latency: A large-scale study. Computer Networks, 182, p.107495.

Coll-Perales, B., et al. , 2022. End-to-end V2X latency modeling and analysis in 5G networks. IEEE Transactions on Vehicular Technology, 72(4), pp.5094-5109.

Emara, M., Filippou, M.C. and Sabella, D., 2018, June. MEC-assisted end-to-end latency evaluations for C-V2X communications. In 2018 European conference on networks and communications (EuCNC) (pp. 1-9). IEEE.

Giotsas, V., et al., 2020. O peer, where art thou? Uncovering remote peering interconnections at IXPs. IEEE/ACM Transactions on Networking, 29(1), pp.1-16.

Ibrahim, Q., 2016. Enhanced power management scheme for embedded road side units. IET Computers & Digital Techniques, 10(4), pp.174-185.

Lazim Qaddoori, S. and Ali, Q.I., 2023. An embedded and intelligent anomaly power consumption detection system based on smart metering. IET Wireless Sensor Systems, 13(2), pp.75-90.

Lazim, S., et. al, 2012. Design and implementation of an embedded intrusion detection system for wireless applications. IET Information Security, 6(3), pp.171-182.

Lee, K., Kim, J., Park, Y., Wang, H. and Hong, D., 2017. Latency of cellular-based V2X: Perspectives on TTI-proportional latency and TTI-independent latency. Ieee Access, 5, pp.15800-15809.

Lee, K., Kim, J., Park, Y., Wang, H. and Hong, D., 2017. Latency of cellular-based V2X: Perspectives on TTI-proportional latency and TTI-independent latency. Ieee Access, 5, pp.15800-15809.

Lucas-Estañ, et al., 2022. An analytical latency model and evaluation of the capacity of 5G NR to support V2X services using V2N2V communications. IEEE Transactions on Vehicular Technology, 72(2), pp.2293-2306.

Lucas-Estañ, M.C., Coll-Perales, B., Shimizu, T., Gozalvez, J., Higuchi, T., Avedisov, S., Altintas, O. and Sepulcre, M., 2022. An analytical latency model and evaluation of the capacity of 5G NR to support V2X services using V2N2V communications. IEEE Transactions on Vehicular Technology, 72(2), pp.2293-2306.

Martín-Sacristán, D., Roger, S., Garcia-Roger, D., Monserrat, J.F., Kousaridas, A., Spapis, P., Ayaz, S. and Zhou, C., 2018, April. Evaluation of LTE-Advanced connectivity options for the provisioning of V2X services. In 2018 IEEE Wireless Communications and Networking Conference (WCNC) (pp. 1-6). IEEE.

Mohammed, H.A., Kareem, S.W. and Mohammed, A.S., 2022. A COMPARATIVE EVALUATION OF DEEP LEARNING METHODS IN DIGITAL IMAGE CLASSIFICATION. Kufa Journal of Engineering, 13(4).

Nomikos, G., et al., 2018, October. O peer, where art thou? Uncovering remote peering interconnections at IXPs. In Proceedings of the Internet Measurement Conference 2018 (pp. 265-278).

Patriciello, N., Lagen, S., Giupponi, L. and Bojovic, B., 2018, September. 5G new radio numerologies and their impact on the end-to-end latency. In 2018 IEEE 23rd international workshop on computer aided modeling and design of communication links and networks (CAMAD) (pp. 1-6). IEEE.

Sivalingan, H., 2024. CLOUD-SMART SURVEILLANCE: ENHANCING ANOMALY DETECTION IN VIDEO STREAMS WITH DF-CONVLSTM-BASED VAE-GAN. Kufa Journal of Engineering, 15(4), pp.125-140.

Virdis, A., Nardini, G., Stea, G. and Sabella, D., 2020. End-to-end performance evaluation of MEC deployments in 5G scenarios. Journal of Sensor and Actuator Networks, 9(4), p.57.

Ye, Q., Zhuang, W., Li, X. and Rao, J., 2018. End-to-end delay modeling for embedded VNF chains in 5G core networks. IEEE Internet of Things Journal, 6(1), pp.692-704.